

SYLED - CLA2T

Université de la Sorbonne Nouvelle - Paris 3

Explorations textométriques



Volume 2 : séries textuelles chronologiques

*Sous la direction de
André Salem et Serge Fleury*

A. Salem, J-M. Leblanc,
Li-Chi Wu, S. A. de Sousa

2009

Nous avons rassemblé plusieurs compte-rendus d'expériences réalisées avec les logiciels de la famille Lexico au cours de nombreuses recherches et dans le cadre de collaborations diverses. Les navigations rassemblées ici ont été choisies pour mettre en évidence la très vaste gamme des domaines d'application des méthodes textométriques ainsi que les fonctionnalités des logiciels **Lexico3** et **mkAlign**. Elles sont publiées sous la forme de trois volumes (**volume 1** : *corpus et problèmes*, **volume 2** : *séries textuelles chronologiques*, **volume 3** : *corpus multilingues*).

Lexico3

<http://www.cavi.univ-paris3.fr/ilpga/ilpga/tal/lexicoWWW/>

mkAlign

<http://tal.univ-paris3.fr/mkAlign/>

Lexicometrica

<http://www.cavi.univ-paris3.fr/lexicometrica/>

Fonctionnalités de Lexico3

Tableau des fonctionnalités

Pour présenter les fonctionnalités mises en œuvre dans les sections suivantes, nous avons réunis ci-dessous les différentes icônes associées aux fonctionnalités visées dans leur contexte d'utilisation :

Fenêtre/barre principale	
Fenêtre « carte des sections »	Fenêtre PCLC
Fenêtre « groupe de formes »	

Le tableau ci-contre rassemble, nomme et présente brièvement l'ensemble des fonctionnalités du logiciel **Lexico3** utilisées pour mener à bien l'exploration textométrique proposée dans les sections suivantes. On pourra aussi se reporter aux différents manuels du logiciel disponibles en ligne.

N°	Nom	Paramètres	Localisation	icône
1	SEGMENTATION	Liste de délimiteurs <i>Par défaut :</i> .,:;! ?/_-\'' () [] {} \$\$	Barre principale	
3	CONCORDANCE	Forme (ou Type Généralisé)	Barre principale	
4	SEGMENTS REPETES		Barre principale	
5	PCLC	Une fois la partition construite, on peut accéder au tableau présentant les Principales Caractéristiques lexicométriques de la partition.	Barre principale	
6	PARTITION	Une clé définissant une partition dans le corpus original est du type : <CLE= « valeur »> C'est le nom de la clé qui est donné ici pour construire la partition visée	Barre principale	
6	VENTILATION	Forme ou groupe de formes		
7	CARTE DES SECTIONS	délimiteur de section	Barre principale	
8	GROUPE DE FORMES	Cette fonctionnalité produit des listes de formes qu'il est possible de mémoriser, d'exporter ou de « projeter » sur les graphiques construits par Lexico3 . Elle permet surtout de faire des recherches de formes ou de groupes de formes en utilisant la notion d'expression régulière.	Barre principale	
5.3	AFC		Fenêtre des PCLC	
5.1, 7.2	SPECIFICITES (POSITIVES NEGATIVES)	Partie ou section du corpus	Fenêtre des PCLC Carte des sections	 

Lexico3, Tableau des Fonctionnalités

Glossaire

segmentation - opération qui consiste à délimiter des unités minimales dans un texte. Les **unités minimales** (pour un type de segmentation) - unités que l'on ne décompose pas en unités plus petites pouvant entrer dans leur composition (ex : dans la segmentation en formes graphiques les formes ne sont pas décomposées en fonction des caractères qui les composent)

caractères délimiteurs / non-délimiteurs : distinction opérée sur l'ensemble des caractères qui entrent dans la composition du texte, permettant aux procédures informatisées de segmenter le texte en occurrences (suite de caractères non-délimiteurs bornée à ses extrémités par des caractères délimiteurs).

On distingue parmi les caractères délimiteurs:

- les caractères délimiteurs d'occurrence (encore appelés "**délimiteurs de forme**") qui sont en général : le blanc, les signes de ponctuation usuels, les signes de préanalyse éventuellement contenus dans le texte.
- les caractères **délimiteurs de séquences** : sous-ensemble des délimiteurs d'occurrence correspondant, en général, aux ponctuations faibles et fortes contenues dans la police des caractères.
- les caractères **séparateurs de phrase** : (sous-ensemble des délimiteurs de séquence) qui correspondent, en général, aux seules ponctuations fortes.

forme ou "**forme graphique**" : archétype correspondant aux occurrences identiques dans un corpus de textes, c'est-à-dire aux occurrences composées strictement des mêmes caractères non-délimiteurs d'occurrence

partition (d'un corpus de textes) : division d'un corpus en *parties* constituées par des fragments de texte consécutifs, n'ayant pas d'intersection commune et dont la réunion est égale au corpus.

ventilation (des occurrences d'une unité dans les parties du corpus) : La suite des n nombres (n = nombre de parties du corpus) constituée par la succession des sous-fréquences de cette unité dans chacune des parties, prises dans l'ordre des parties

motif : un ensemble d'objets possédant une propriété reconnaissable.

analyse factorielle : famille de méthodes statistiques d'analyse multidimensionnelle, s'appliquant à des tableaux de nombres, qui visent à extraire des "facteurs" résumant approximativement par quelques séries de nombres l'ensemble des informations contenues dans le tableau de départ.

spécificité positive : pour un seuil de spécificité fixé, une forme i et une partie j données, la forme i est dite spécifique positive de la partie j (ou forme caractéristique* de cette partie) si sa sous-fréquence est "anormalement élevée" dans cette partie. De façon plus précise, si la somme des probabilités calculées à partir du modèle hypergéométrique pour les valeurs égales ou supérieures à la sous-fréquence constatée est inférieure au seuil fixé au départ

spécificité négative : pour un seuil de spécificité fixé, une forme i et une partie j données, la forme i est dite spécifique négative de la partie j si sa sous-fréquence est anormalement faible dans cette partie. De façon plus précise, si la somme des probabilités calculées à partir du modèle hypergéométrique pour les valeurs égales ou inférieures à la sous-fréquence constatée est inférieure au seuil fixé au départ

partie (d'un corpus de textes) : fragment de texte correspondant aux divisions naturelles de ce corpus ou à un regroupement de ces dernières.

section : portion de texte comprise entre deux délimiteurs de section (exemple : le paragraphe, etc.).

segment répété (ou polyforme répétée) : suite de forme dont la fréquence est supérieure ou égale à 2 dans le corpus.

Les expressions régulières avec *Lexico3*

Dans les sections qui suivent on utilisera à plusieurs reprises la notion d'expression régulière en particulier à travers la fonction «**GROUPE DE FORMES**». Nous rappelons ci-dessous brièvement cette notion et les différents opérateurs disponibles avec *Lexico3* pour écrire de telles expressions. Les expressions régulières permettent de représenter de manière générique des motifs textuels : un *motif* est un ensemble d'objets possédant une propriété reconnaissable, par exemple tous les mots terminés par le suffixe «*able*» ou commençant par le préfixe «*pré*». Les expressions régulières permettent ainsi de décrire des portions de texte à l'aide d'opérateurs particuliers. Le tableau suivant rassemble l'ensemble des opérateurs disponibles avec *Lexico3* pour écrire des motifs sous la forme d'expression régulière :

Opérateur	Fonction	Application
.	(le point) Représente n'importe quel caractère	L'expression "m.l" représente des séquences comme : mal, mol...
*	0 ou n occurrences du caractère qui précède	L'expression "com*e" représente des séquences comme : coe, come, comme, comme...
+	1 ou n occurrences du caractère qui précède	L'expression "com+e" représente des séquences comme : comme, comme, ...
\b	Représente un début de mot	L'expression "\bcapital" représente des séquences comme : capital, capitale, capitalisme...
\b	Représente une fin de mot	L'expression ".*isme\b" représente des séquences comme : syndicalisme, capitalisme...
[]	Représente un ensemble de caractères	L'expression "[aeiou]" représente des séquences comme : un des caractères de l'ensemble des voyelles minuscules. L'expression "[a-z]" représente un des caractères minuscules compris entre a et z.
[^]	Représente la négation du contenu de l'ensemble de caractères	L'expression "[^aeiou]" représente un des caractères parmi ceux qui ne sont pas ceux de l'ensemble des voyelles minuscules

Sommaire

Fonctionnalités de Lexico3	4
Tableau des fonctionnalités	4
Glossaire.....	6
Les expressions régulières avec <i>Lexico3</i>	7
Tutoriel n°2 : Séries textuelles chronologiques.....	9
1 L'analyse des séries textuelles chronologiques : analyses de la série textuelle Duch8M... 9	9
2 Rechercher les formes prototypiques de l'évolution.....	13
3 Faire varier l'échelle d'analyse	22
4 Organiser le retour au contexte	23
5 Proposer une interprétation	25
6 Conclusion.....	26
7 Références	26
8 Principales fonctionnalités <i>Lexico3</i> utilisées.....	27
Vœux présidentiels	29
1 Contexte de la recherche	29
2 Affinités et contradictions lexicales	29
3 L'originalité Gaullienne ?	30
4 Analyse factorielle et temps lexical	32
5 Des ethos fortement contrastés - Spécificités des marques énonciatives.....	32
6 Registres personnels ou chronologie ?	33
7 Références	37
8 Fonctionnalités <i>Lexico3</i> utilisées dans cette navigation.....	37
Essai de textométrie politique chinoise.....	38
1 Contexte de la recherche	38
2 Recueil des données, préparation du corpus	40
3 La série chronologique 1921-2007.....	44
4 Explorer les dimensions chronologiques du corpus.....	52
5 Faire varier les échelles de l'analyse : un corpus restreint (1982-2007).....	60
6 Conclusion.....	67
7 Références	67
Le discours de Fidel Castro. Essai de lexicométrie politique.	68
1 Contexte de la recherche	68
2 Le corpus	70
3 Exploration lexicométrique	72
4 Conclusion.....	94
5 Références	94

Tutoriel n°2 : Séries textuelles chronologiques

[STC]

André Salem

salem@msh-paris.fr

Apprendre à :

1. Constituer une série textuelle chronologique
2. Repérer la structure chronologique
3. Repérer les formes prototypiques de l'évolution
4. Utiliser les outils d'analyse chronologique
5. Organiser le retour au texte

Les *séries textuelles chronologiques* sont des corpus constitués par la réunion de textes similaires produits par une même source textuelle au cours d'une période de temps. Comme nous allons le voir, la prise en compte de la dimension chronologique de tels corpus permet de mettre en évidence des variations qui surviennent au cours du temps dans l'emploi du vocabulaire, de mettre en évidence des moments importants dans l'évolution de celui-ci¹. L'analyse des séries textuelles chronologiques s'appuie sur une série de méthodes tout particulièrement adaptées à ces objectifs.

Les 96 livraisons qui constituent le corpus **Duchn** constituent un découpage en unités trop ténues pour permettre aux analyses quantitatives de mettre en évidence des variations chronologiques. Comme nous l'avons vu plus haut, les analyses opérées à partir de ce type de découpage font ressortir des oppositions qui concernent avant tout des différences stylistiques dans la confection des différents numéros. Nous allons voir dans ce qui suit que dès que l'on regroupe les livraisons publiées au cours de la même semaine, de la même quinzaine ou du même mois les analyses textométriques mettent en évidence des évolutions chronologiques liées à l'évolution du vocabulaire au cours du temps.

1 L'analyse des séries textuelles chronologiques : analyses de la série textuelle Duch8M

La partition du corpus **Duchn** en huit parties regroupant chacune les livraisons parues au cours d'un même mois permet de centrer plus précisément nos observations sur l'évolution du vocabulaire au cours de la période couverte par le corpus.

¹ Sur l'analyse des séries textuelles chronologiques on consultera, par exemple [Salem 1993].

1.1 Typologie sur les 8 périodes

L'AFC réalisée à partir de ce regroupement en 8 parties met en évidence met clairement en évidence la proximité relative des périodes consécutives et leur échelonnement dans le temps.²

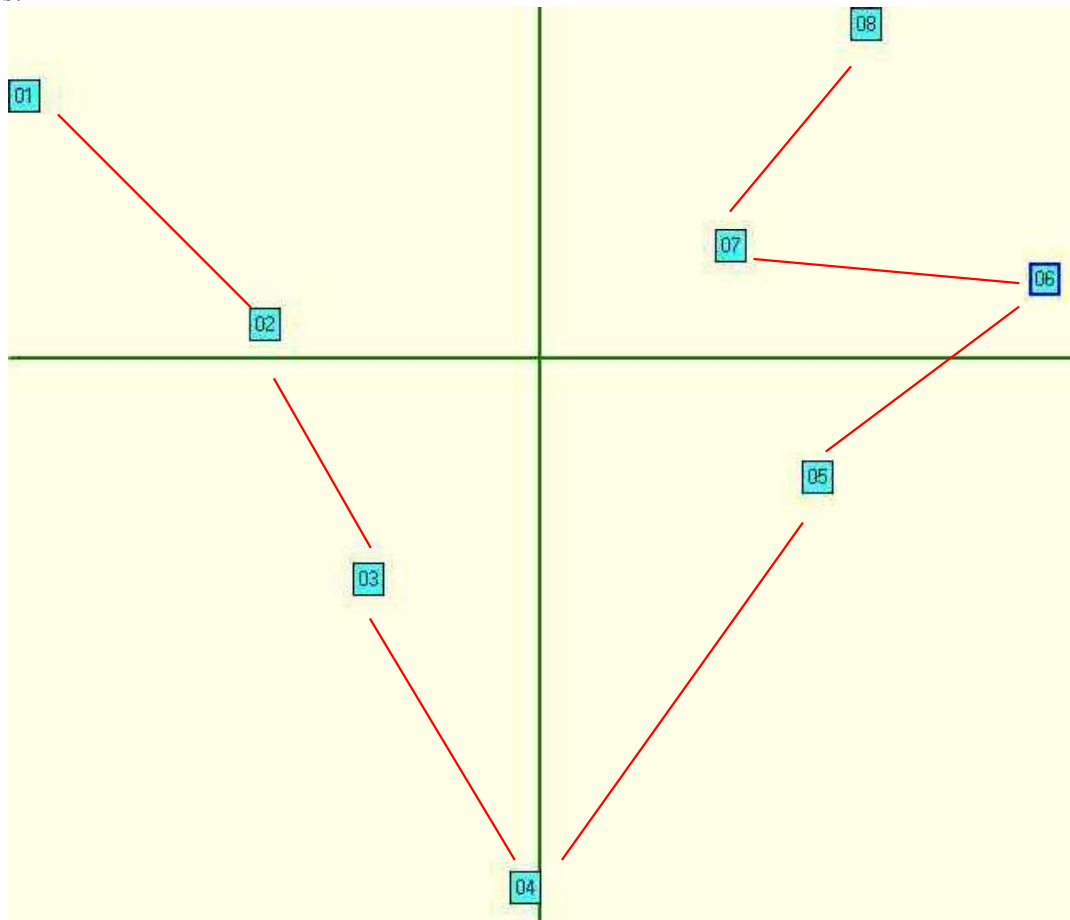


Figure 1

AFC sur le corpus *Duchn*

8 périodes d'un mois x 1420 formes de fréquence ≥ 10

1.2 Interprétation des analyses chronologiques

==== Repère méthodologique ==== l'effet Guttman

L'AFC réalisée à partir des séries textuelles chronologiques produit souvent des figures particulières, dont l'interprétation doit être faite selon des règles spécifiques.

L'existence d'une évolution progressive du vocabulaire au fil des parties d'un corpus chronologique induit un système de distances particulier sur l'ensemble des parties soumises à l'analyse :

- les parties consécutives dans les temps se révèlent plus proches entre elles que les parties séparées par une période de temps plus important.

² Les traits rouges qui relient les parties consécutives dans le temps ont été rajoutés par nous sur le résultat de fournis par l'analyse. Signalons que l'on rencontre très régulièrement ce genre de résultats lorsqu'on analyse des corpus chronologiques (cf. [Salem 19xx]).

La décomposition de ces distances selon les différents axes de la représentation factorielle constitue alors une suite d'approximations progressive plutôt plus complexes que le phénomène relativement simple dont elles à rendent compte.

Le schéma théorique correspondant à une telle évolution est le suivant :

- le premier facteur rend compte d'une évolution linéaire.
- les facteurs suivants sont des fonctions de degré plus élevé (2, 3, 4, etc.) de ce premier facteur.

L'analyse des corpus chronologiques réels relève pour partie de ce schéma même si d'autres oppositions peuvent venir perturber le schéma théorique décrit plus haut. :

Pour en savoir plus :

P on consultera par exemple :

Salem, A. : « Approches du temps lexical », in *Histoire et mesures*, **XXXXX**, Klincksieck, 19xx

Guttman 1941]. On trouvera un exposé plus accessible dans [Benzécri 1973] p 192-196, et des applications dans [Van Rijckevorsel 1987]..

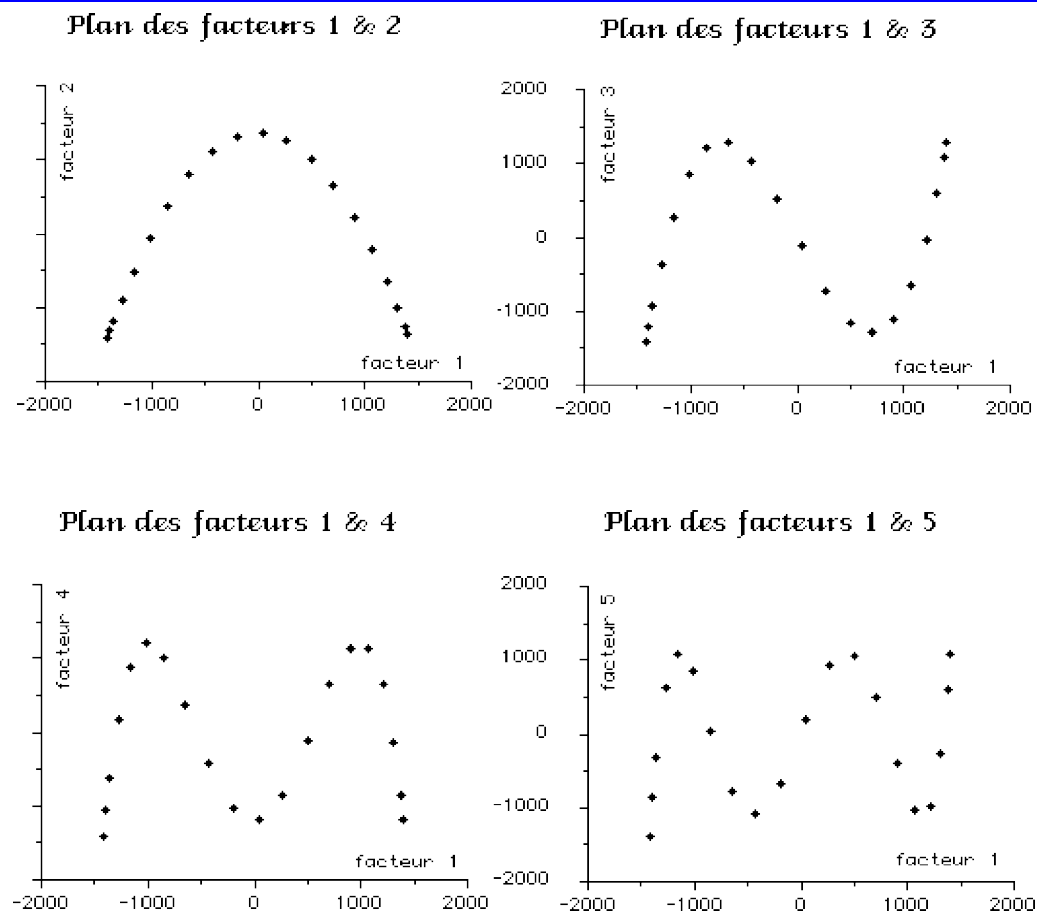


Figure 2

Les facteurs issus de l'analyse du scalogramme 20 x 20

Guide de lecture pour la figure 2

On voit sur cette figure les valeurs prises par les cinq premiers facteurs issus de l'analyse du scalogramme 20x20.

Les points dessinent une parabole sur le plan des deux premiers facteurs si on les relie dans cet ordre. L'étude des facteurs suivants montre que ces derniers sont également des fonctions, de degré supérieur, du premier facteur. La prise en compte de facteurs supplémentaires apporte, dans ce cas, une information d'ordre systématique qui est peu intéressante.

Tableau 1
Scalogramme de dimension 20 x20

	C 1	C 2	C 3	C 4	– –	C 18	C 19	C 20
L1	1	1	0	0	–	0	0	0
L2	1	1	1	0	–	0	0	0
L3	0	1	1	1	–	0	0	0
L4	0	0	1	1	–	0	0	0
.....								
L18	0	0	0	0	–	1	1	0
L19	0	0	0	0	–	1	1	1
L20	0	0	0	0	–	0	1	1

Le modèle de l'effet Guttman nous fournit un éclairage important pour comprendre la disposition des parties sur la courbe présentée à la figure 1. Devant des résultats de ce type on évitera donc de commenter séparément les oppositions constatées sur chacun des axes factoriels pris isolément. Le schéma classique de l'interprétation d'une typologie que l'on affine au fur et à mesure par la prise en compte de nouveaux axes factoriels doit faire place à la reconnaissance d'une situation caractéristique globale liée à l'existence et à la dominance d'une gradation progressive.

Au delà de cette mise en évidence, le modèle d'analyse attire notre attention sur les moments particuliers de l'évolution lexicale. La particularité de l'évolution constatée ici à propos de l'évolution de la série **Duch8M** tient au fait que la période M7 ne s'inscrit pas totalement dans l'évolution d'ensemble.

1.3 Retour à l'analyse des tableaux lexicaux

A la différence du scalogramme soumis à l'AFC dans la section qui précède, les tableaux lexicaux tels que celui que nous avons soumis à l'AFC à la section §1 comptent en général plusieurs dizaines de milliers de cellules. De plus, ces tableaux résultent de décomptes réels effectués sur des textes et non d'une construction formelle comme c'était le cas pour le scalogramme. Si les résultats issus de l'analyse factorielle appliquée à ces deux tableaux présentent des similitudes, ils présentent aussi des différences profondes. Dans les tableaux lexicaux, la propriété d'autocorrélation des facteurs ne concerne pas tous les facteurs issus de l'analyse bien que les premiers facteurs possèdent en général cette propriété. De plus, comme on a pu le vérifier sur la figure 1, le schéma des distances que l'on construit sur la base d'un tableau lexical ne présente pas la régularité de celui que l'on obtient à partir d'un scalogramme.

Comme dans le cas du scalogramme, les premiers facteurs issus de l'AFC du corpus **Duch8M** possèdent la propriété de rapprocher les périodes consécutives dans le temps. Cette propriété d'une variable s'appelle l'**autocorrélation**. On vérifie sur la figure XX que tel est bien le cas pour le premier facteur issu de l'analyse du tableau (1420 formes x 8 périodes).

L'expérience prouve que dans le cas de l'analyse des tableaux réalisés à partir de séries chronologiques

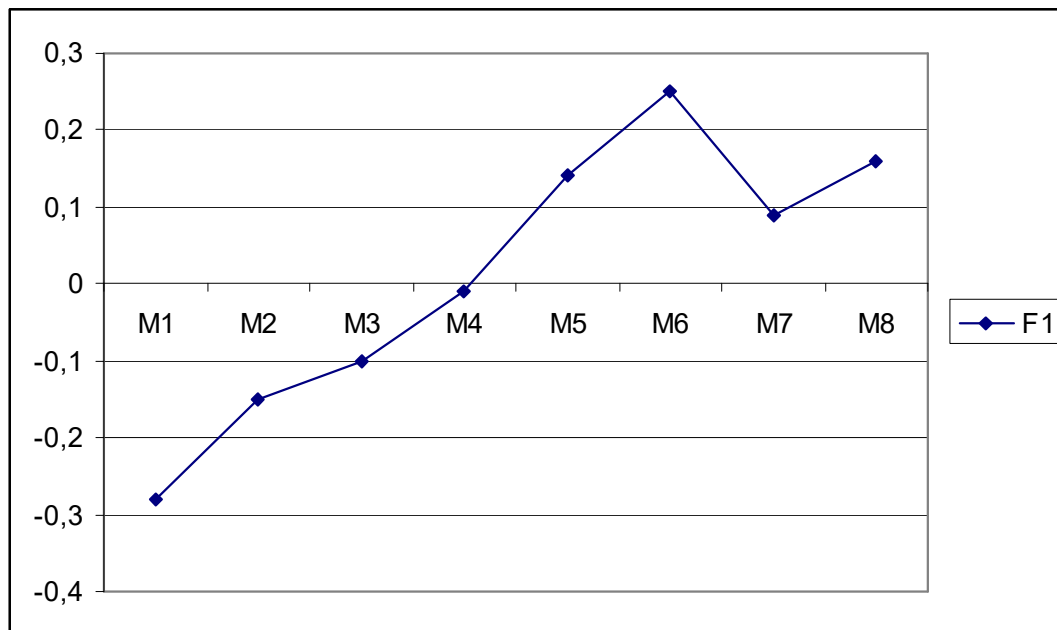


Figure 3

Le premier facteur issu de l'analyse du tableau
8 périodes d'un mois x 1420 formes de fréquence ≥ 10

1.4 Mesures de contiguïté

==== Repère méthodologique ==== le coefficient de Geary

Parmi d'autres coefficients du même type, le *coefficient de Geary* permet de mesurer l'autocorrélation d'une série de nombres : $x_1, x_2, \dots, x_n$.

Une valeur faible du coefficient indique que les valeurs successives sont « en moyenne » plus proches

Pour en savoir plus :

Lebart, L., Salem, A. : *Statistiques textuelles*, Paris, Dunod, 1994.

2 Rechercher les formes prototypiques de l'évolution

La disposition des 8 parties que l'on voit sur la figure 1, nous laisse penser que l'évolution dans le temps de la série Duch8M n'est pas complètement régulière. Des formes voient leur importance croître au cours du temps mais on peut penser que cette croissance connaît un ralentissement en M7 pour reprendre dans la huitième période.

Une des premières questions qui se posent est celle de mettre en évidence tout ou partie de l'ensemble des formes lexicales dont le profil ressemble à cette évolution. Plusieurs méthodes statistiques peuvent nous aider à mettre ces formes en évidence³. Dans ce qui suit nous avons

³ A partir des résultats de l'AFC, les formes possédant une coordonnée factorielle positive et également pourvues d'une forte contribution relative au facteur constituent de bons candidats pour ce type de sélection.

sélectionné une classe de formes construite par une méthode de classification hiérarchique à partir des profils des formes.

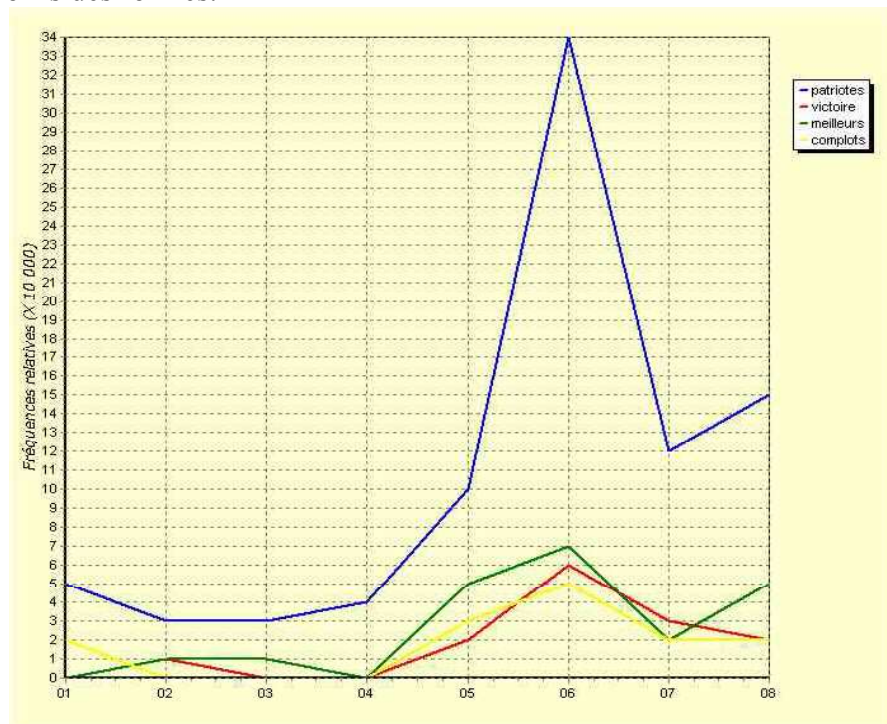


Figure 4

Quelques formes prototypiques de l'évolution de la série Duchn8M

Dans une classe plus large qui contient : *patriotes*, *meilleurs*, *complots*, *surplus*, *accusent*, *victoire*, *réjouissance*, *modérés*, *revient*, *tricherie*.

2.1 Calcul des spécificités chronologiques

Le calcul des *spécificités chronologiques* permet de mettre en évidence les formes qui ont une ventilation particulièrement centrée sur une période ou un groupe de période consécutives du corpus.⁴ Ce calcul signale comme le fait le plus saillant, sur l'ensemble du corpus, la répartition de la forme *patriotes* centrée sur la période M6.

Tableau 2 :

Spécificités chronologiques majeures du corpus *Duchn*

⁴ Sur le calcul des spécificités chronologiques, on consultera [Salem 1993] et [Lebart & Salem 1994]

Formes/SR	F	f	coeff	Partie(s)
patriotes	152	60	18	06
constitution	72	53	16	01 - 02
*paris	171	78	14	01 - 02
vous	1097	705	13	01 - 04
*phélipotin	13	13	13	06
avez	171	55	12	04
*custine	70	23	12	01 - 02
foutre	1100	784	10	07 - 08
phélippotins	14	14	10	07 - 08
nobles	42	21	9	01
elle	368	90	9	01
*courtille	13	3	9	08
*vincent	15	14	9	06 - 07
républicains	129	91	9	05 - 08
tu	296	200	9	03 - 06
on	856	288	9	05 - 06
*marat	65	27	9	01
nouvelle	46	21	9	06
ça	8	8	8	02
*ami	8	8	8	01
je	979	561	8	04 - 07
*cordeliers	17	6	8	08
sociétés	15	10	8	08
ai	202	114	8	04 - 06

Guide de lecture pour le tableau 2

Le tableau fournit les spécificités chronologiques majeures calculées pour le corpus **Duchn** muni de la partition en 8 périodes d'un mois.

- La première colonne du tableau indique le terme pour lequel le diagnostic de spécificité chronologique a été calculé ;
- la seconde *F*, donne la fréquence du terme dans l'ensemble du corpus ;
- la troisième *f* la fréquence de ce même terme dans la période considérée ;
- la quatrième *Coeff.* fournit le coefficient de spécificité chronologique calculé pour le terme et la période ;
- on trouve dans la dernière colonne la période pour laquelle le diagnostic a été calculé.

==== Repères méthodologiques ====

Spécificités chronologiques - Accroissements spécifiques

Pour chaque unité textuelle (forme, segment répété, etc.) d'un corpus divisé en *périodes*, le calcul des **spécificités chronologiques** permet de déterminer l'intervalle temporel (période ou groupe de périodes consécutives) au cours duquel cette unité connaît un emploi spécifique maximal (positif ou négatif). Les spécificités ainsi calculées sont ensuite triées par coefficient de spécificité décroissant (i.e. les plus remarquables en tête) afin de mettre en évidence les distorsions les plus remarquables dans l'emploi des termes par les différentes périodes (ou groupes de périodes consécutives).

Pour chaque période du corpus (à partir de la seconde), le calcul des **accroissements spécifiques** permet de mettre en évidence le vocabulaire et les segments dont l'utilisation connaissent un accroissement particulièrement

important au cours de la période (*accroissements spécifiques positifs*) ou au contraire ceux dont l'utilisation se raréfie brutalement après une utilisation plus intensive *accroissements spécifiques négatifs*).

Pour en savoir plus :

Pour un exposé et des exemples d'application de l'analyse des spécificités à l'étude des corpus de textes, on consultera par exemple :

Salem, A. : « Approches du temps lexical », in *Histoire et mesures*, **XXXXX**, Klincksieck, 19xx

Lebart, L., Salem, A. : *Statistiques textuelles*, Paris, Dunod, 1994.

Les diagnostics suivants concernent, dans l'ordre décroissant d'importance : l'abondance des formes *constitution* et *paris* dans les deux premières parties du corpus puis, plus loin dans la liste, d'autres diagnostics dont plusieurs concernent à nouveau la période M6 (phélipotin, *nouvelle*, etc.)

==== **Lexico3** ==== *Liste des spécificités chronologiques*

- ✓ Sélectionner l'icône **PCLC** (5^{ème} icône à partir de la gauche)
- ✓ Sélectionner une partie ou un groupe de parties
- ✓ Appuyer sur le bouton **SpecDia** (à droite de la fenêtre)
- ✓ Les résultats apparaissent dans un onglet spécifique dans la fenêtre centrale. On peut trier ces diagnostics en cliquant sur le titre de chaque colonne.

2.2 Calcul des accroissements spécifiques

Tableau 3

Spécificités majeures pour les formes et les segments répétés dans la période M6 du corpus **Duch8M**

Terme	Frq. Tot.	Fréquence	Coeff.
<i>patriotes</i>	152	60	18
<i>*phélipotin</i>	13	13	13
<i>les patriotes</i>	91	36	11
<i>nouvelle</i>	46	21	9
<i>c est là que</i>	10	8	7
<i>*vincent</i>	15	10	7
<i>est là</i>	36	15	6
<i>là que</i>	12	8	6
<i>c est là</i>	28	13	6
<i>conspirateurs</i>	48	18	6
<i>ai</i>	202	48	6
<i>les</i>	4748	687	6
<i>accuser</i>	10	7	5
<i>diviser</i>	16	9	5
<i>mes</i>	120	30	5
<i>contre</i>	298	60	5
<i>les meilleurs</i>	31	13	5
<i>généraux</i> <i>*sans</i>	10	7	5
<i>culottes</i>			
<i>aristocrates</i>	133	32	5

Forme	Frq. Tot.	Fréquence	Coeff.
patriotes	111	60	21
les patriotes	68	36	13
*phélipotin	13	13	11
*vincent	10	10	9
nouvelle	42	21	8
les	3667	687	8
conspirateurs	35	18	7
*camille	8	8	7
généraux *sans culottes	7	7	7
*bouchotte	9	8	7
c est là	24	13	6
diviser	13	9	6
aristocrates	94	32	6
*phélipotins	6	6	6
est là	30	15	6
c est là que	10	8	6
les meilleurs	22	13	6
ai	179	48	5
victoires	7	6	5
victoire	20	11	5
mes	101	30	5
contre	243	60	5
là que	12	8	5
accuser	9	7	5
en réjouissance de	8	6	5
meilleurs	29	13	5
réjouissance	8	6	5
les conspirateurs	19	10	5
en réjouissance	8	6	5
des patriotes	10	7	5

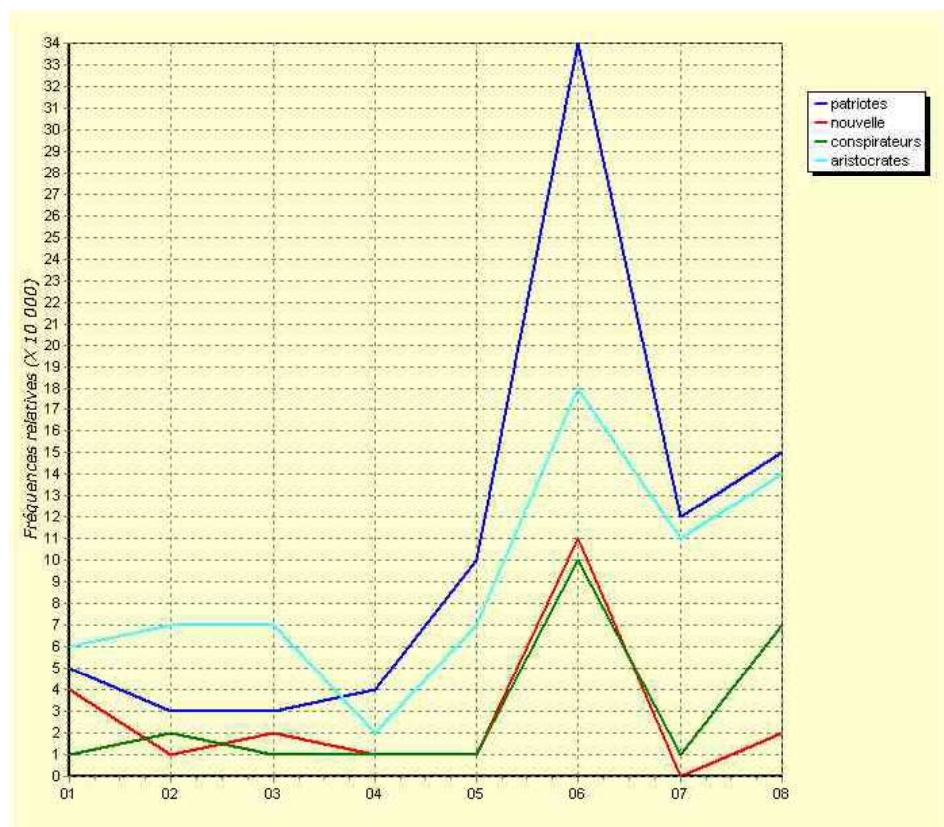


Figure 5

2.3 Un vocabulaire « offensif »

L'analyse des *spécificités positives* (unités textuelles particulièrement abondantes dans chacune des parties du corpus) montre que le caractère le plus marquant de cette évolution lexicale réside dans l'augmentation progressive, particulièrement remarquable dans la sixième partie M6, d'un vocabulaire dont on peut remarquer qu'il est lié au conflit politique et que nous appellerons, dans le cadre de cette étude, *vocabulaire offensif*. Ce vocabulaire, relativement sous-utilisé dans les premières parties du corpus (M1, M2, M3) connaît par la suite une utilisation croissante qui trouve un paroxysme en M6. Du point de vue de cette évolution d'ensemble, la période M7 marque un net recul dans l'utilisation de ce vocabulaire qui sera à nouveau utilisé en M8.

On trouve à la figure 6 la ventilation dans les 8 parties du corpus de quelques-unes des formes les plus caractéristiques de cet ensemble : les formes *patriotes*, *nouvelle*, *conspirateurs*⁵. Comme on le voit sur cette figure, le schéma de l'évolution fréquentielle de ces formes, exprimé à la fois en terme de fréquence relative (à gauche) et en termes de spécificité (à droite), correspond bien au type d'évolution chronologique que nous avons décrit plus haut.

⁵ J.C Deroubaix a proposé d'appeler ces formes, particulièrement représentatives d'un type de distribution remarquable des *parangons* de leurs classes respectives.

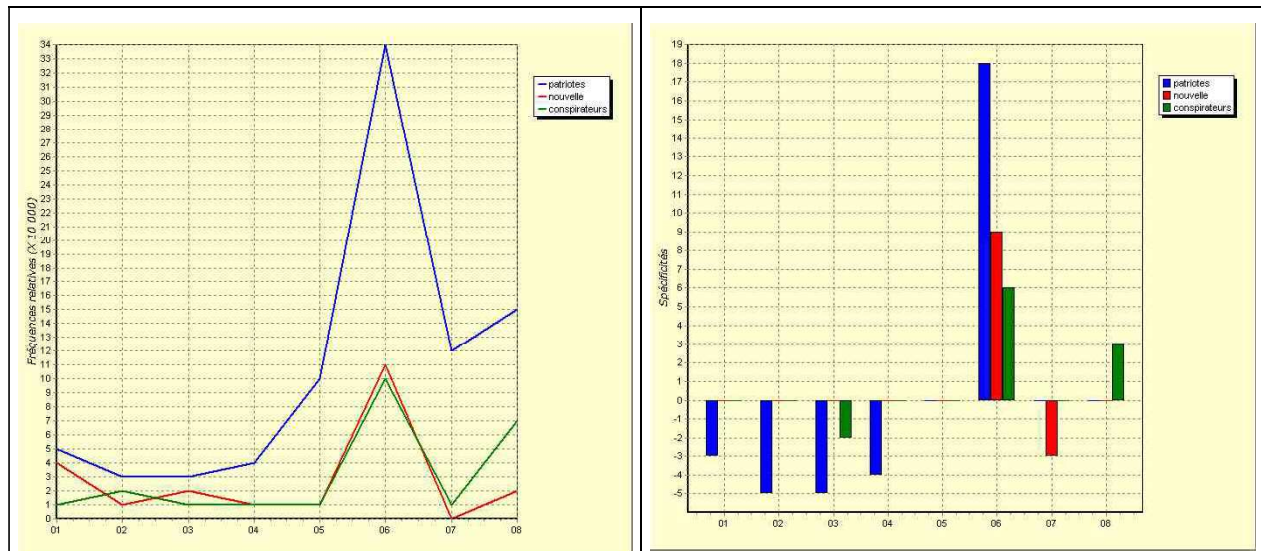


Figure 6

Ventilation des formes *patriotes*, *nouvelle*, *conspireurs*
dans les 8 périodes du corpus **Duch8M**

==== Lexico3 ==== Statistiques par parties

- ✓ Sélectionner l'icône **Statistique par parties** (6 ème icône à partir de la gauche)
- ✓ Sélectionner une clé pour créer une partition du corpus (ici : mois)
- ✓ Traînez sur la fenêtre ainsi créée une ou plusieurs forme ou un groupe de formes
- ✓ Choisir le mode de représentation (*fréquence relative*, *absolue* ou *spécificité*) en sélectionnant sous le graphique le bouton correspondant.

2.4 Vers une sémantique de corpus

Examinons maintenant la liste des formes les plus spécifiques de la partie M6 (tableau 3).

Pour plusieurs de ces formes : *conspireurs*, *accuser*, *diviser*, *fripes*, etc., l'appellation *vocabulaire offensif* paraît tout à fait légitime dans la mesure où ces termes véhiculent dans leur définition et dans tous leurs emplois discursifs une charge polémique évidente.

Pour des termes comme *aristocrates* ou *phélipotins* (partisans d'un certain *Phélipot*), une connaissance, même superficielle, de la période historique permet de reconstituer la charge polémique que ces termes ont pu revêtir dans le contexte de la lutte révolutionnaire.

La situation est plus délicate lorsqu'il s'agit d'expliquer l'appartenance à cet ensemble de formes comme *nouvelle*, *meilleurs*, etc. auxquelles on a pas coutume d'attacher de valeur polémique dans la langue courante. Pour comprendre la raison de la présence de ces formes dans des situations de conflit, il nous faut retourner à l'ensemble des contextes de leur utilisation.

Sur le tableau 4, les contextes de la forme *nouvelle* à l'intérieur de la partie M6 du corpus ont été triés en fonction de la forme qui suit le pôle. Un rapide examen de ces contextes permet de vérifier que, si l'on excepte les quelques occurrences de la forme qui renvoie à la forme substantivale *une (bonne/mauvaise) nouvelle*, la majorité des contextes renvoie à la forme adjectivale *une nouvelle X*. Dans un second temps, on peut vérifier que les contextes droits de cette forme adjectivale renvoient le plus souvent à des entités auxquelles le contexte

historique permet d'attribuer une valeur négative, ex : *une nouvelle *vendée*⁶, soit à des entités présentées comme négatives par l'emploi de formes explicitement dépréciatives (*une nouvelle clique d'aristocrates, une nouvelle clique de modérés*).⁷

Dans ces contextes, les entités auxquelles on fait allusion renvoient à des ennemis X dont on pensait s'être débarrassé dans un passé récent. L'adjectif *nouvelle X* introduit l'idée que la lutte n'est pas terminée et qu'il convient de la poursuivre.

Tableau 4
Concordances pour les formes *nouvelle* et *meilleurs*
dans la période M6 du corpus *Duch8M*

Partie : 06, Nombre de contextes : 21

lipoterie ; s ' il propose de former une	nouvelle	*vendée en mettant en liberté ses bons
recruter tous les brigands et former une	nouvelle	*vendée en proposant d ' ouvrir les pri
e est venu à son secours . " formons une	nouvelle	*vendée , lui a dit l ' animal cornu ,
andelle ; je savais qu ' il existait une	nouvelle	clique d ' aristocrates , de modérés ,
colère du *père *duchesne . de voir une	nouvelle	clique de modérés , de feuillants , d '
gtemps que je furetais pour découvrir la	nouvelle	conspiration qui est maintenant sur le
courriers extraordinaires ont apporté la	nouvelle	d ' une nouvelle victoire , on a annon
lus . la convention vient de recevoir la	nouvelle	de la destruction totale des brigands ;
bonnets rouges . ces *sans - culottes de	nouvelle	fabrique se sont répandus dans les sect
l débite gratis . au dénouement de cette	nouvelle	farce contre - révolutionnaire , nous d
ont unis , ne sera pas le jouet de cette	nouvelle	intrigue . après avoir sauvé la *france
' est pas assez , il faut savoir quelle	nouvelle	liste civile paye les rhapsodies qu ' i
a *vendée , ou plutôt pour en former une	nouvelle	, mais suivons pied à boule notre endor
rahisons , serait le connétable de cette	nouvelle	majesté égorgeante ; *brissot serait so
urs desseins . quand je découvre quelque	nouvelle	manigance contre les *sans - culottes ,
: victoire , vive la république . grande	nouvelle	, mes amis , leur ai je dit ; femmes ,
ommères , en réjouissance de cette bonne	nouvelle	, qui donne le coup de grâce aux aristo
parole . voilà , mon vieux , une grande	nouvelle	qui va faire gueuler , dans tous les co
uronnés , quand ils vont apprendre cette	nouvelle	. sa grande colère contre certains jean
dinaires ont apporté la nouvelle d ' une	nouvelle	victoire , on a annoncé que les lignes
détruire , lui donnent au contraire une	nouvelle	vie . ils tiennent les patriotes en hal

Partie : 06, Nombre de contextes : 13

vention même ; cries , gueule contre les	meilleurs	*montagnards , accuse le comité de sal
s manigances des *phélipotins contre les	meilleurs	citoyens , et surtout contre les génér
u plus d ' audace ? celui qui traite les	meilleurs	patriotes de bourreaux , d ' assassins
s osent lever la crête , et accusent les	meilleurs	patriotes de tous les coups de chien q
nt de sauver la république ; dénonce les	meilleurs	patriotes . " § le monstre *phélipotin
à la porte de ce comité pour accuser les	meilleurs	patriotes . les amis de la royauté , l
artes à la convention , en dénonçant les	meilleurs	patriotes pour faire revenir sur l ' e
elle branches se raccrocher accusent les	meilleurs	patriotes pour qu ' on les oublie , ou
, et tous les nains qui asticotaient les	meilleurs	patriotes sont rentrés cent pieds sous
ents de *pitt et *cobourg , accusent les	meilleurs	républicains , et les *sans - culottes
vu les plus vils intrigants accuser les	meilleurs	républicains ; je me doutais bien que
s sa fureur de mordre et de déchirer les	meilleurs	républicains on l ' étouffera comme un
ssassiner la patrie , qui ont égorgé les	meilleurs	républicains subirent tous le supplice

L'examen des contextes de la forme *meilleurs* dans cette même partie M6, montre que cette forme introduit une distinction, particulièrement insistante dans la période M6, parmi des catégories d'actants présentés jusqu'alors comme positives dans leur totalité. L'apparition de l'adjectif *meilleurs* devant chacun de ces groupes induit la question de l'hétérogénéité de la catégorie et la possibilité de l'existence d'intrus parmi les gens susceptibles de s'en réclamer. Pour plus de clarté, *les meilleurs X* sont d'ailleurs présentés comme faisant l'objet d'intrigues,

⁶ La Vendée est une province française qui a connu un soulèvement armé contre le pouvoir révolutionnaire. Ce soulèvement a fait l'objet d'une répression très dure, qui apparaît comme terminée à l'époque où sont produits ces textes.

⁷ Un examen similaire sur les contexte de la forme *nouveau* montre un fonctionnement identique de la forme (*nouveau complot, nouveau coup de chien*), cette fois dans la partie M5 du corpus.

de manigances, d'accusations diverses de la part d'ennemis plus difficiles à identifier. Dans ce sens, la spécificité de la forme *meilleurs* dans la partie M6 peut tout à fait être rattachée au vocabulaire « offensif » signalé plus haut.

2.5 Le cas de la forme *les*

La forme *les*, que la méthode des spécificités signale comme particulièrement représentée dans la période M6 du corpus *Duchn8M*, mélange, compte tenu de la méthode de segmentation en *formes graphiques* utilisée depuis le début de cette expérience, les occurrences de l'article défini pluriel *les* et celles du pronom personnel homographe. De plus, comme dans la plupart des textes rédigés en français que nous avons étudiés, cette unité est très fréquente dans le corpus (4 748 occurrences). Quelle signification peut-on accorder à la présence abondante des occurrences de cette forme dans les parties qui recourent fortement au vocabulaire que nous avons qualifié d'*offensif* ?

L'analyse des segments les plus fréquents commençant par la forme *les* nous donne quelques indications. Comme on pourra le vérifier en annexe B, en se reportant à une liste plus fournie de ces segments, l'article défini *les* introduit souvent les actants, en majorité hostiles, que nous avons évoqués plus haut en les constituant comme des groupes dont l'unité ne peut pas être remise en cause au moment de l'énonciation : *les patriotes* (91 occ.), mais aussi : *les traîtres* (86 occ.), *les aristocrates* (75 occ.), *les fripons* (64 occ.), *les brigands* (63 occ.), etc..

Par conséquent, la condensation des occurrences de , la ventilation de cette forme graphique constitue également un indicateur de l'accroissement du discours offensif que nous avons évoqué plus haut. Ce que l'on peut vérifier dans le paragraphe présenté ci-dessous à titre d'exemple.

[§ 487] /.../ **les** *brissotins, comme **les** jean-foutres de conseillers de *capet, ont persécuté **les** patriotes et **les** ont accusé également d'être des anarchistes. on se sert aujourd'hui d'un autre mot pour avilir **les** ardents républicains. on **les** appelle des ultra-révolutionnaires, parce que **les** mots de factieux et de désorganisateurs sont usés; mais foutre, le peuple ne prendra pas le change, ce sont encore **les** apôtres du modérantisme, **les** amis de la royauté et **les** aristocrates déguisés qui **les** accusent. /.../

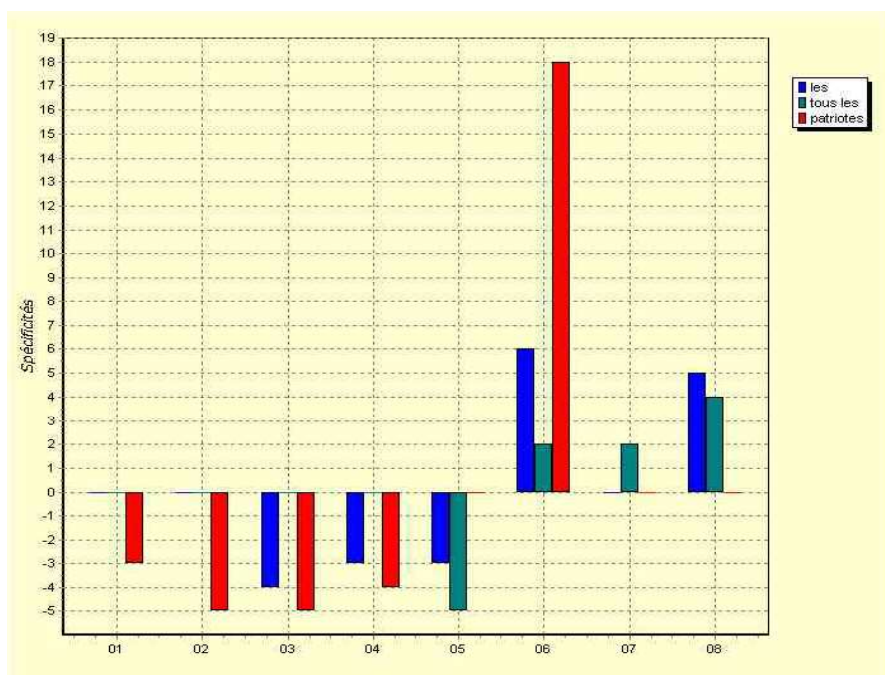


Figure 7

Ventilations des termes *patriotes*, *les* et *tous les*
dans les 8 périodes du corpus **Duch8M**

3 Faire varier l'échelle d'analyse

Comme on l'a vu plus haut, la division du corpus **Duchn** en 8 périodes d'un mois permet de mettre en évidence une évolution du vocabulaire employé au fil du temps qui n'était pas repérable sur le corpus découpé en unités trop fines (les 96 numéros). On peut voir sur la figure XX, le résultat d'une analyse portant sur le même corpus découpé cette fois en 16 parties correspondant chacune à un intervalle de temps de deux semaines. Les quinzaines ont été numérotées à l'aide de deux chiffres. Le premier de ces chiffres correspond au mois de parution également mentionné dans l'analyse en 8 parties. Le second indique s'il s'agit de la première ou de la seconde quinzaine.

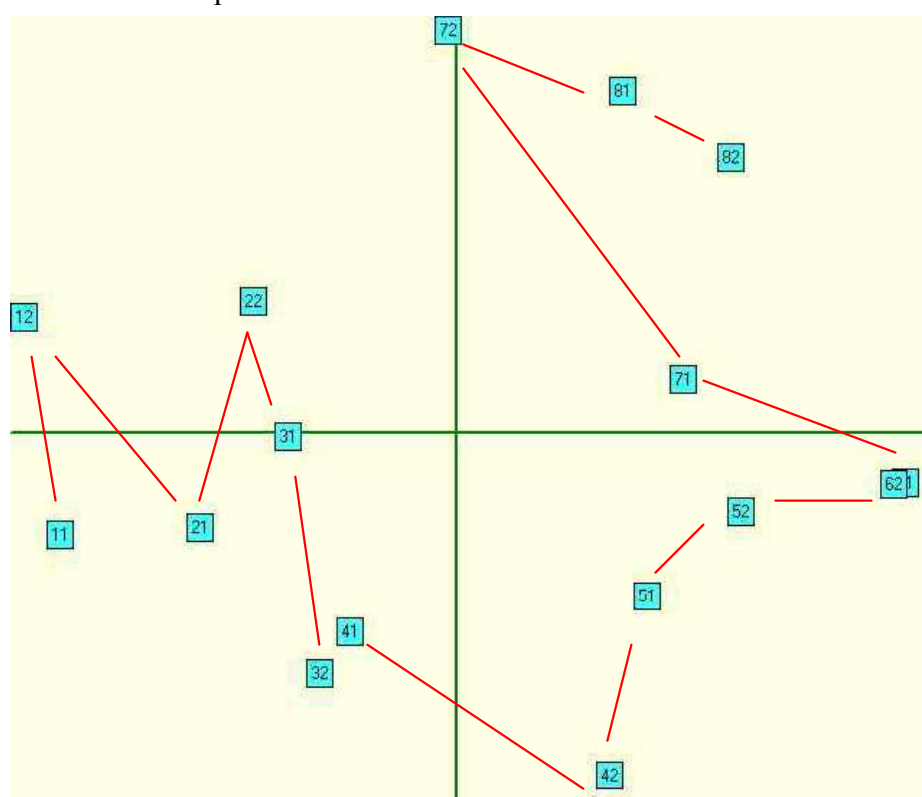


Figure 8

AFC sur le corpus **Duchn**

16 périodes d'une semaine x 1420 formes de fréquence ≥ 10

Comme on le voit sur cette seconde analyse, la partition du corpus en unités plus fines amène des résultats un peu moins lisibles, du fait de la plus grande variabilité de ces parties. En revanche, l'analyse permet cette fois de porter des jugements sur des périodes de temps plus courtes et par la même de tenter de produire un commentaire qui colle de plus près à la chronologie des événements.

Tout le problème devient alors de faire le départ entre les variations qui traduisent des modifications chronologiques survenues sur des périodes courtes, hautement pertinentes pour le chercheur, et des variations qui sont à mettre sur la réduction de la taille des unités

statistiques dans lesquelles sont effectuées les décomptes.⁸ On note que pour cette seconde analyse opérée sur le découpage en quinzaines, les parties proches dans le temps occupent toujours des positions voisines sur le premier plan factoriel. Si l'on conserve le modèle d'interprétation chronologique proposée à partir de l'analyse en 8 parties, on peut préciser le schéma d'évolution proposé plus haut en situant désormais dès la première quinzaine de la partie 7 le repli opéré par le Père Duchesne par rapport à l'évolution chronologique d'ensemble du vocabulaire.

4 Organiser le retour au contexte

Il nous reste à vérifier en retournant au contexte que l'hypothèse de la montée d'un vocabulaire offensif rend bien compte de l'évolution lexicale au sein du corpus étudié.

Sur le tableau XX nous avons projeté le type constitué par les formes les plus spécifiques de la partie M6. Le logiciel a ensuite calculé, pour chaque section, la spécificité attachée à ce type dans chacun des paragraphes du corpus. Un seuillage de ce diagnostic a permis de colorier de manière différenciée les paragraphes. Les paragraphes en couleur sombre contiennent en principe un nombre important de formes et de segments appartenant au type étudié (ici, le type : formes les plus fréquentes du vocabulaire spécifique de la partie M6). Remarquons qu'un certain nombre de ces parangons n'appartiennent pas à la partie M6. on en conclura que ces parangons extérieurs anticipent (pour ceux qui sont situés dans la période M5, par exemple) l'expression particulière qui va avoir lieu dans la période M6. Le parangon situé dans la période M8 constitue au contraire un écho de cette expression

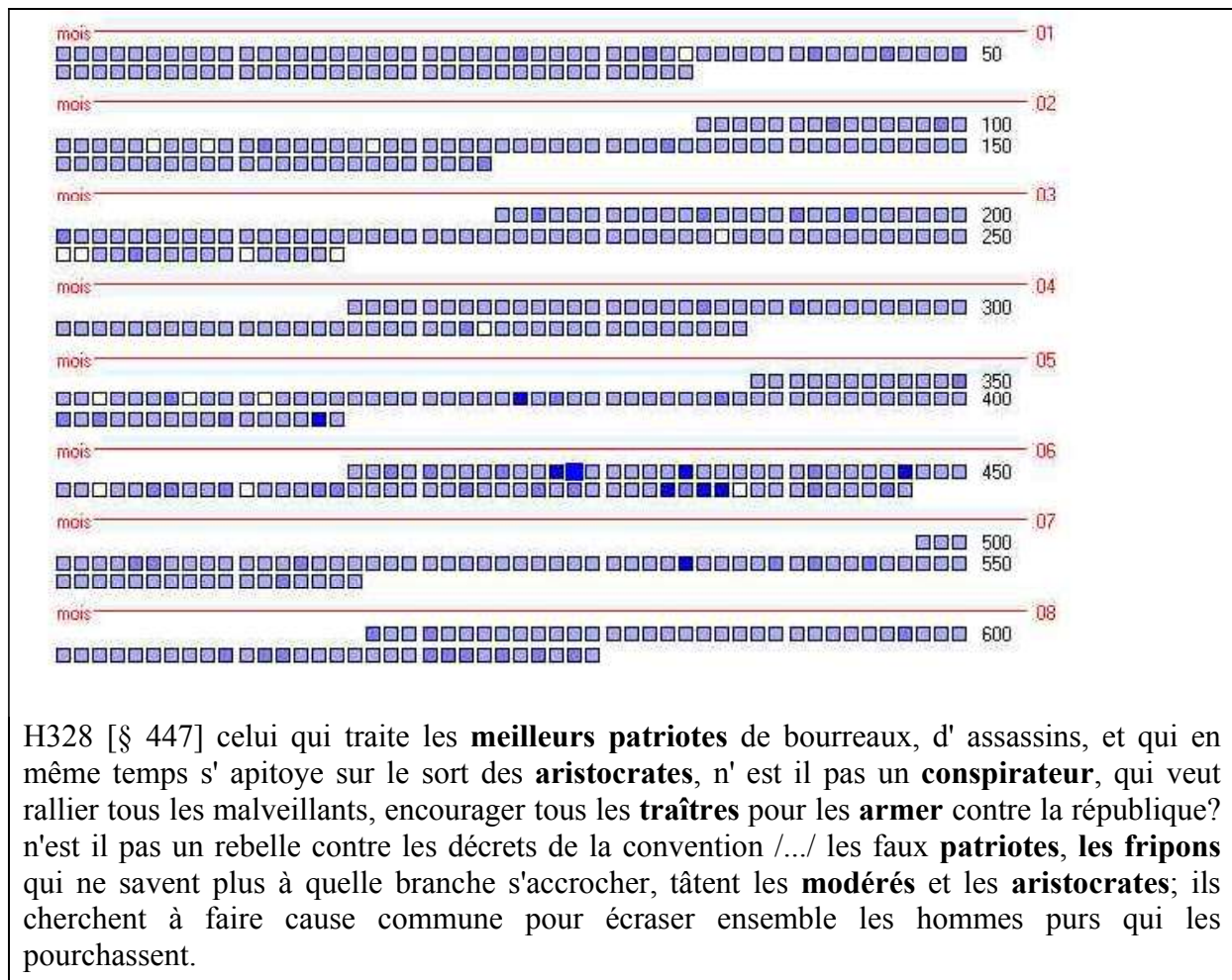
Parmi d'autres parangons signalés en couleur sombre, le paragraphe sélectionné au tableau XX révèle une tonalité particulièrement offensive. Aux formes qui ont servi à construire le type viennent s'ajouter plusieurs formes comme : *rebelles*, *malveillants*, *bourreaux*, *assassins*, *écraser* que leur faible fréquence a écarté de notre sélection mais qui auraient tout à fait leur place dans liste des termes polémiques.

==== Lexico3 ==== Repérage des parangons pour un type donné

- ✓ Construire le type que l'on désire spatialiser (forme, groupe de formes, liste ...)
- ✓ Construire une carte des sections en fixant un délimiteur de section (§ 2.1)
- ✓ Traîner le type sur la carte des sections
- ✓ Cocher la case « seuillage » (en haut sur la carte des sections) après avoir éventuellement modifié le niveau des seuils (bouton à gauche de cette case)

⁸ Plusieurs modèles statistiques ont été proposés par les statisticiens pour aider les chercheurs face à ce type d'interrogation.

Tableau 5 :
Repérage des parangons pour un type donné



4.1 Les types caractéristiques

Une *sélection* ou *zone textuelle* est un ensemble d'occurrences sélectionnées dans le corpus. Une zone peut être constituée par un ensemble de partie, de paragraphes ou de phrases consécutives ou non ou encore par une sélection arbitraire d'occurrences éparpillées dans le texte. Pour une zone donnée, on appellera *type caractéristique*, ou *type spécifique* toute liste des formes constituées par les formes les plus caractéristiques de la zone.

Dans l'exemple ci-dessus, les formes les plus caractéristiques prélevées parmi les spécificités majeures de la partie M6, dont nous avons analysé la ventilation dans les paragraphes du corpus constituent donc un type spécifique pour la zone M6.

Le type caractéristique présenté plus haut nous a permis de localiser les parangons de la partie M6 (paragraphes particulièrement saturés en formes spécifiques de la partie M6). Le graphique présenté à la figure 9 permet d'apprécier l'évolution de ce que nous avons appelé le vocabulaire offensif dans un découpage plus fin du corpus.

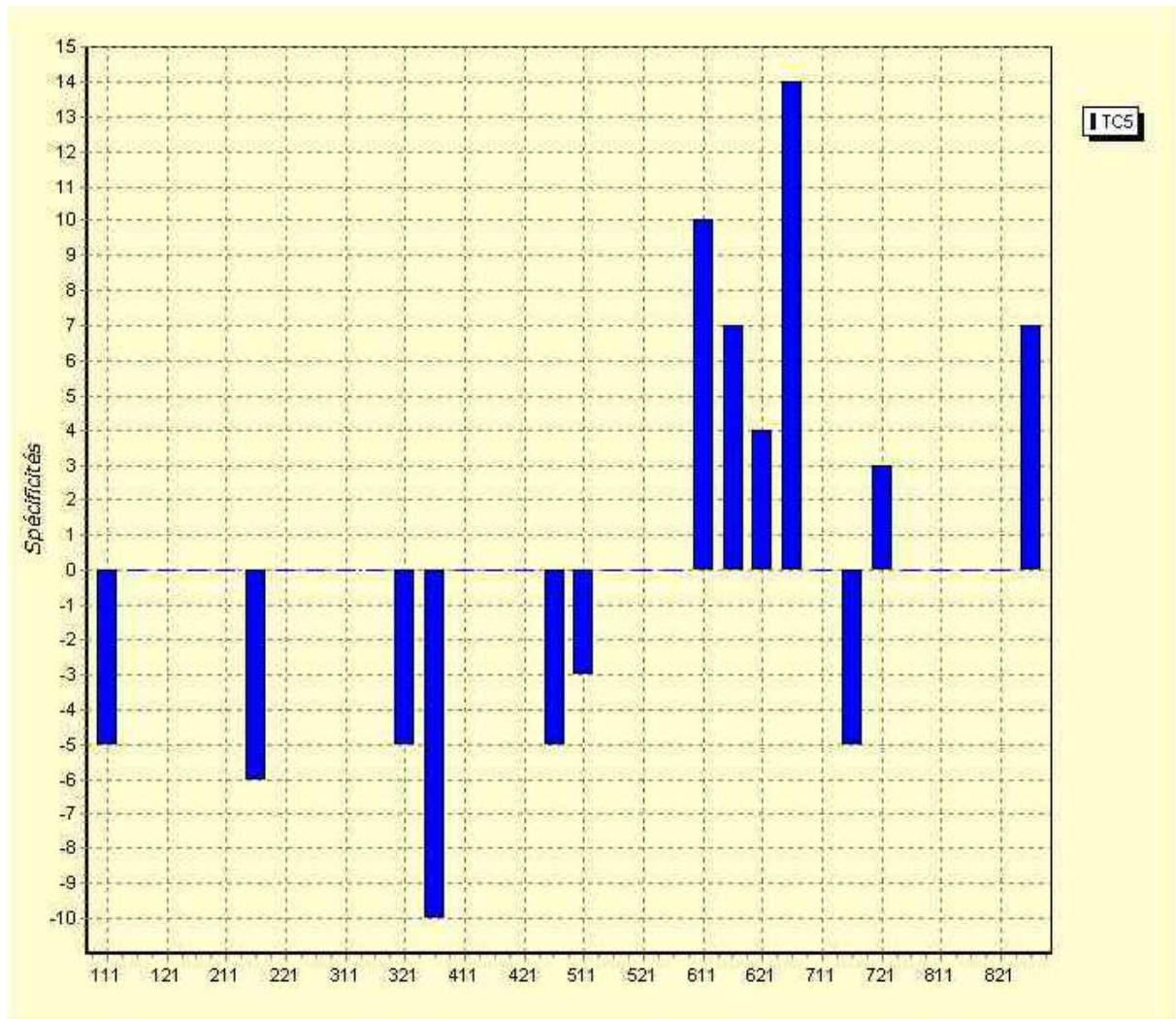


Figure 9 :

Ventilation dans les 32 semaines du corpus *Duchn* du type spécifique T5 (M6)
(formes de spécificité $< 1/10^5$ dans la partie M6)

L'histogramme présenté à la figure XX permet avec plus de précision l'apparition et l'augmentation du vocabulaire *offensif* introduit plus haut.

==== *Lexico3* ==== Etude de la variation des types caractéristiques

- ✓ Construire le type que l'on désire étudier (forme, groupe de formes, liste ...)
- ✓ Construire une partition en choisissant une clé
- ✓ Traîner le type sur le graphique
- ✓ Choisir le mode de représentation (fréquence relative, absolue ou spécificité)

5 Proposer une interprétation

Dans le cadre de ce seul paragraphe, nous tenterons d'opérer un rapprochement entre les particularités observées dans la répartition des formes lexicales mises en évidence par les analyse textométriques et les principaux événements survenus au plan politique dans la

période couverte par le corpus. En d'autres termes, d'esquisser une tentative d'interprétation au plan historique des phénomènes constatés⁹.

Les premières périodes du corpus (M1-Mx) voient un renforcement des positions défendues par l'auteur du *Père Duchesne* au détriment des ennemis politiques qu'il ne cesse de dénoncer (*Brissotins*, i.e. partisans de Brissot, etc.). Les 31 députés girondins qui constituent le parti des *modérés*, parmi lesquels Brissot, seront exécutés le xxx qui correspond à la période Mxx du corpus. Après cette victoire politique, Hébert entre en opposition avec le mouvement jacobin et plus particulièrement avec les montagnards du Comité de salut public (Robespierre, St Just, etc.). Dans le but d'influencer l'action du Comité de salut public, il redouble d'ardeur dans la dénonciation des *conspirateurs*, des *faux patriotes*, des *nouvelle(s) clique(s) de modérés*, de ceux qui veulent allumer une *nouvelle Vendée* et appelle à l'union des *meilleurs patriotes*.

Les jacobins réagiront d'abord par la voix de Camille Desmoulins qui attaquera Hébert dans son journal *Le vieux cordelier* (période M7)¹⁰. Le *Père Duchesne* tentera dans un premier temps de reprendre l'offensive en faisant à nouveau appel aux thèmes qui ont assuré sa popularité des premières périodes : lutte pour la déchristianisation, lutte contre la présence de nobles à la tête des armées. Devant l'échec de cette tentative, il reprendra, sans succès au cours de la dernière période M8 qui verra son arrestation et son exécution, l'offensive amorcée en M6.

6 Conclusion

L'exploration du corpus *Duchn*, à l'aide des méthodes textométriques met en évidence une importante évolution du vocabulaire au cours des huit mois sur lesquels s'étend le corpus.

Les analyses quantitatives sur la partition en 96 livraisons, mettent en évidence des différences stylistiques liées à une alternance de genre entretenue par l'auteur du corpus. De ce fait, elles ne permettent pas d'apprécier l'évolution lexicale du corpus.

Un regroupement des livraisons en périodes de 30 jours consécutifs permet par sa part de cerner l'évolution lexicale de manière nettement plus satisfaisante. Les méthodes quantitatives permettent alors tout à la fois : de mettre en évidence un vocabulaire offensif qui trouvera un emploi particulièrement remarquable dans la période M6. Le retour au contexte permet de préciser ces analyses.

7 Références

Lamalle C., Salem A., « Types généralisés et topographie textuelle dans l'analyse quantitative des corpus textuels », in *actes des 6emes journées d'analyse statistique des données textuelles*, Inria, St Malo, 2002
<http://www.cavi.univ-paris3.fr/lexicometrica>, 1997

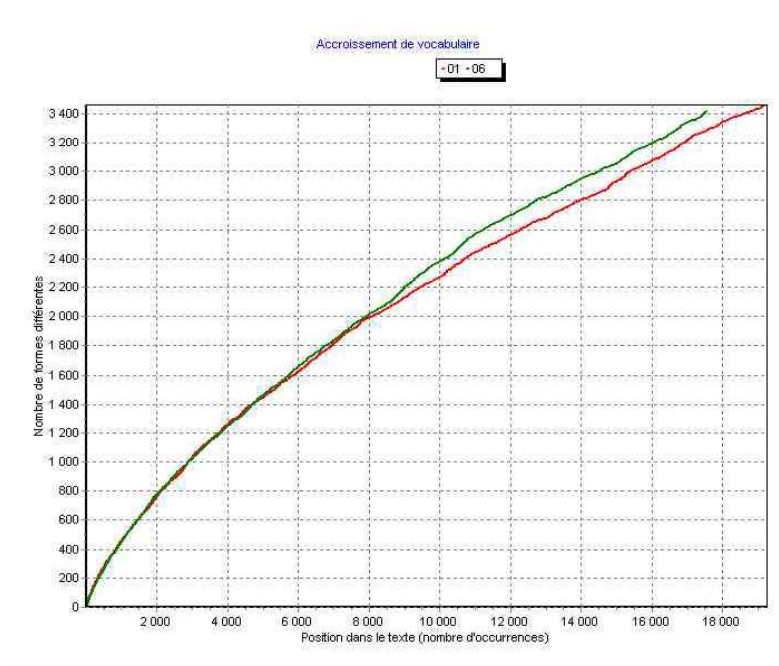
⁹ Cette tentative d'interprétation est née dans le cadre d'une collaboration étroite et ancienne entre l'auteur de ces lignes, nullement spécialiste de ces problèmes et des historiens spécialistes de la période concernée, tout particulièrement J. Guilhaumou, connaisseur des langages de la Révolution française.

¹⁰ On trouve d'ailleurs, dans le journal de Hébert, des traces de cette polémique en M6 et M7 qui concernent, le /gentil /renégat/ benêt/ niais/ Camille.

8 Principales fonctionnalités Lexico3 utilisées

<i>N°</i>	<i>Fonctionnalité</i>	<i>Résultat</i>
<i>2</i>	Partition (clé a, pour année)	
<i>5</i>	Principales car lexicom (PCLC)	<i>Tableau 2</i>
<i>5.6</i>	Accroissement du vocabulaire (corpus)	<i>Figure 1</i>
<i>5.6</i>	Accroissement du vocabulaire (P92, P93)	<i>Figure 2</i>
<i>4</i>	Segments Répétés (seuil minimal =2)	
<i>8</i>	Sélection d'un Type (occurrence de SR long>10)	
<i>7</i>	Carte des sections (paragraphe, présence SR de long>10)	<i>Figure 3</i>

Annexe



Vœux présidentiels

[Vœux]

Jean-Marc Leblanc

leblanc.jeanmarc@free.fr

Résumé : Un corpus de vœux présentés français à l'occasion de la nouvelle année par les différents présidents de la cinquième république permet d'étudier les évolutions lexicales survenues dans ce type d'intervention au cours des quarante dernières années. On esquisse ensuite une étude de l'énonciation de chacun des locuteurs notamment du point de vue de l'emploi des pronoms personnels.

Abstract : A corpus of New Year addresses by the French President of the Republic offers an opportunity to study the lexical evolutions in this type of speech during the last forty years. An analysis can then be carried out on each speaker's enunciation particularly regarding the use of personal pronouns.

1 Contexte de la recherche

Le corpus rassemble les quarante-trois messages de vœux adressés l'occasion de la nouvelle année par les cinq présidents qui se sont succédés à la tête de la cinquième République de 1959 à 2001. Ce corpus lexicométrique compte 41 125 occurrences pour 5 200 formes qui se répartissent de la manière suivante entre les différents chefs d'état.

Partie	occurrences	formes	hapax	Fréq. Max	Forme
De Gaulle	11498	2407	1440	531	de
Pompidou	2850	890	577	118	de
Giscard	6066	1360	833	262	de
Mitterrand	11991	2521	1490	546	de
Chirac	8720	1798	1048	421	de

Tableau 1

Principales caractéristiques de la partition locuteur

2 Affinités et contradictions lexicales

L'analyse factorielle réalisée selon la partition locuteur caractérise les grandes oppositions du corpus. Sur le deuxième axe, un clivage essentiellement diachronique oppose le lexique des trois premiers présidents aux deux derniers. Sur le premier axe, De Gaulle s'oppose à l'ensemble de ses successeurs. On observe par ailleurs des proximités entre Giscard et Pompidou d'une part et Mitterrand Chirac par ailleurs, dont on peut penser qu'elles sont dues en partie à la chronologie et à l'événementiel, d'autant que l'opposition la plus forte, matérialisée par la diagonale de l'analyse factorielle réside entre De Gaulle et les deux derniers présidents de la République.

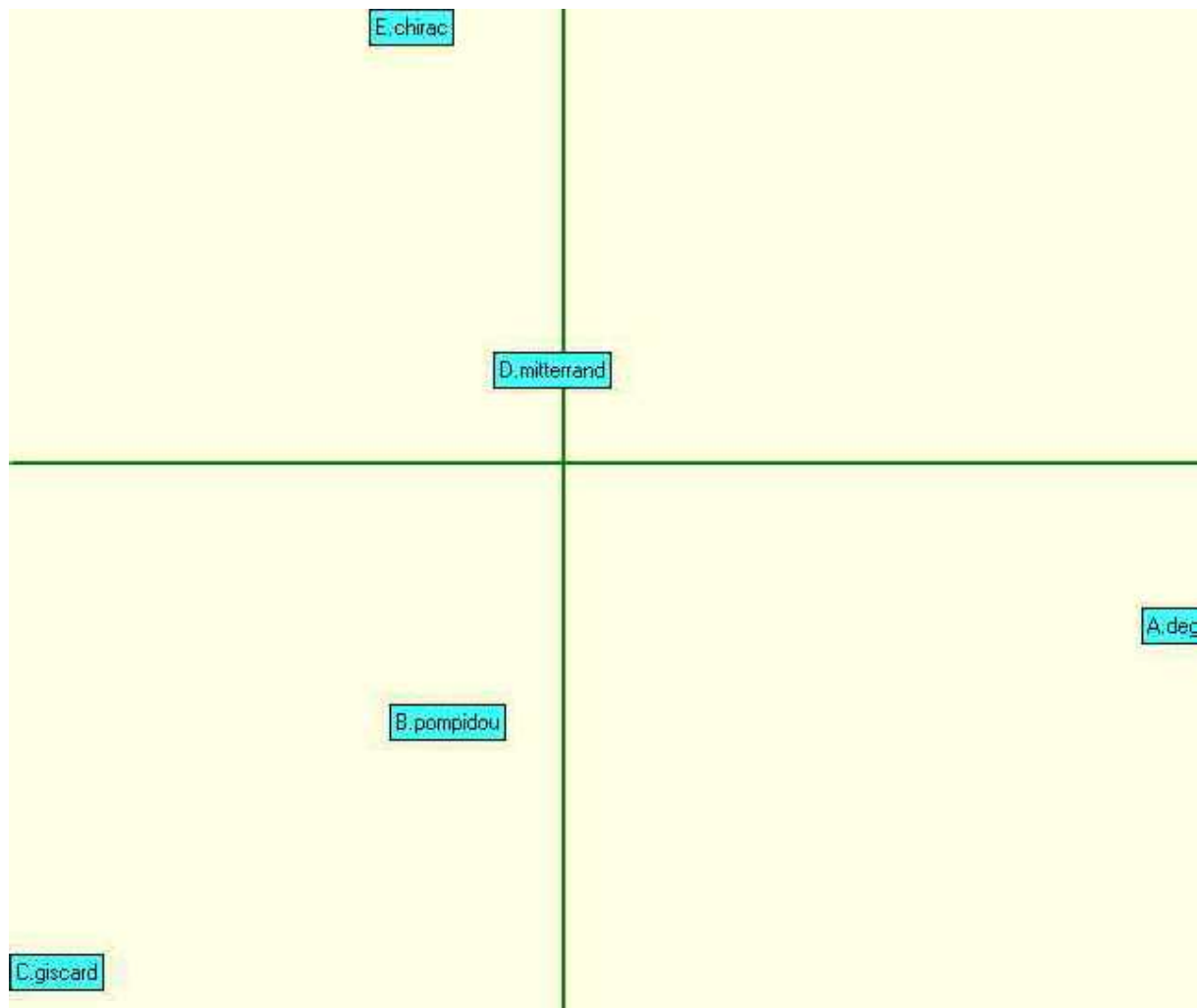


Figure 1 :
AFC du corpus vœux selon la partition locuteur.

La représentation factorielle ne laisse en rien présager d'une quelconque affinité entre De Gaulle et Mitterrand, à qui l'on prête souvent volontiers quelques analogies dans leur façon de concevoir la fonction et quelques traits communs dans leur vision de la France. A cela plusieurs explications peuvent être avancées : vingt trois années séparent la première intervention du général De Gaulle de celle de François Mitterrand qui clôt le corpus. Il est donc probable que les usages lexicaux aient subi une importante mutation au cours de cette période. D'autre part l'analyse factorielle met en évidence les seules oppositions relevant de la distribution du stock lexical. C'est pourquoi il convient de pousser plus avant l'analyse avant de tirer des conclusions que pourrait inspirer une interprétation hâtive de la configuration de la représentation factorielle. Celle-ci n'en a pas moins le mérite de susciter des interrogations, d'orienter l'analyse.

3 L'originalité Gaullienne ?

La position particulière qu'occupent les textes produits par le général De Gaulle sur la représentation factorielle soulève quelques questions. On peut se demander s'il n'est pas plus proche discursivement et lexicalement des pratiques de la quatrième République. Pour tenter de répondre à cette interrogation, nous avons, à titre expérimental, intégré les seuls messages de vœux existants des présidents de la Quatrième République. (Vincent Auriol, décembre

1946, 1948, 1953). Étonnamment ces discours se placeraient au voisinage de Pompidou et Giscard, de Gaulle conservant la singularité qu'on a pu observer. Ainsi, la chronologie et l'événementiel ne seraient pas les seuls facteurs explicatifs de ce phénomène. S'agit-il alors d'une opposition relevant de registres personnels ? Il est probable que la personnalité des locuteurs soit pour beaucoup dans la configuration de l'analyse factorielle. Une seconde A.F.C., réalisée selon la partition année conforte cette hypothèse.

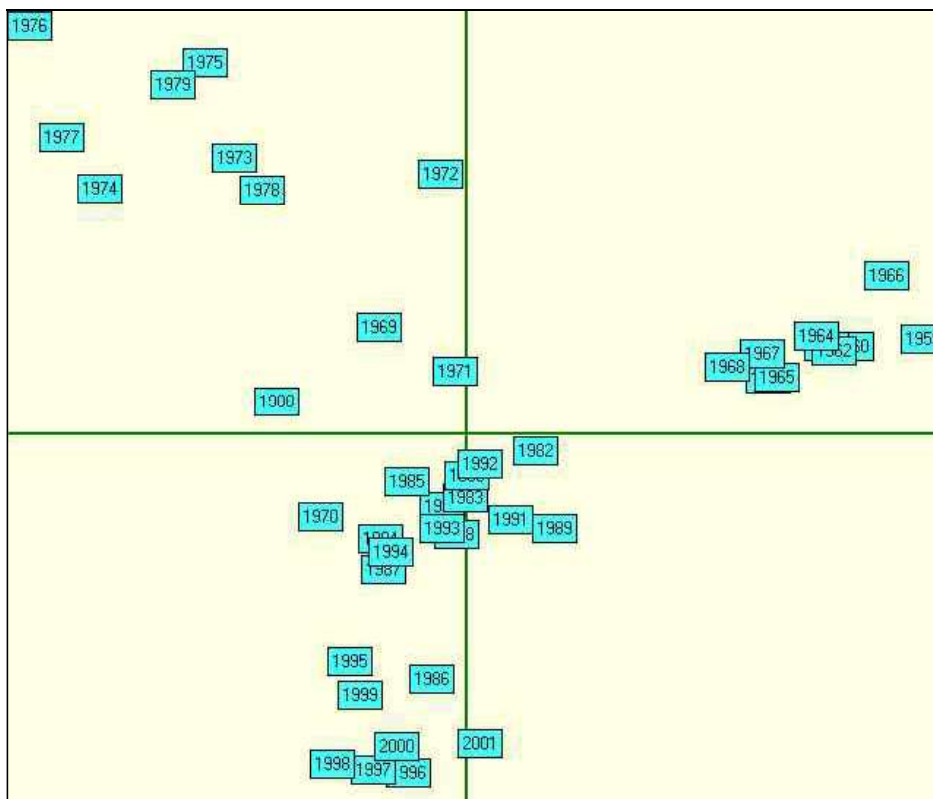


Figure 2 :

AFC du corpus vœux selon la partition année.
Points superposés : 93(88), 90(83), 62(60).

De cette deuxième représentation on retiendra la configuration remarquable de l'AFC où la disposition des années permet d'identifier aisément les locuteurs. L'ensemble des interventions du général De Gaulle se trouve regroupé dans une zone restreinte du plan, en marge de toutes les autres années du corpus. Il en ressort que son discours est à la fois très différent des autres dans le lexique et très stable d'une année sur l'autre. La faible dispersion des années 1996-2001 et 1981-1995 nous conduit à établir une conclusion identique : Bien que proches l'un de l'autre, les deux derniers présidents affichent une grande stabilité lexicale. Le constat est quelque peu différent pour Giscard et Pompidou. L'amplitude est beaucoup plus importante. On y verra peut-être la volonté novatrice de Giscard dans sa stratégie de communication et la recherche de compromis de Pompidou, entre rupture et continuité, fidélité au général De Gaulle et volonté de renouvellement.

Si la première représentation illustre les affinités et divergences des locuteurs, la seconde souligne la grande stabilité des présidents, plus particulièrement de Gaulle, Mitterrand et Chirac. Au-delà de l'aspect chronologique, les A.F.C semblent mettre en évidence des oppositions de registres personnels.

4 Analyse factorielle et temps lexical.

La diachronie semble cependant constituer un facteur déterminant de ces représentations factorielles. L'AFC sur les locuteurs en particulier représente leur succession selon l'ordre chronologique. Pour autant, ces représentations ne reproduisent pas précisément le modèle de l'effet *Guttman* relatif au phénomène de *temps lexical*¹¹. En effet, d'après ce phénomène de temps lexical, si le seul facteur de modification du lexique était le temps, les points portés sur les deux axes factoriels seraient ordonnés selon une courbe idéale, incurvée en son centre. Ainsi, les années ou périodes contiguës seraient proches sur le graphique. Le facteur chronologique est pourtant nettement visible. A cela plusieurs explications.

Les messages sont relativement courts, ce qui rend difficile l'émergence de faits liés à la chronologie, en particulier sur la partition par année. De plus, bien que les conditions d'énonciation soient quasiment identiques, on ne peut considérer qu'on est en présence d'un locuteur collectif mais bien d'une succession de locuteurs. Il est donc probable que les ethos différents viennent brouiller le phénomène chronologique. Notons par ailleurs que bien que la périodicité soit régulière, le corpus se place plutôt dans la ponctualité que dans la continuité.

5 Des ethos fortement contrastés - Spécificités des marques énonciatives

La personnalité des locuteurs transparait au travers de profils énonciatifs contrastés. Le tableau qui suit synthétise la répartition des emplois des pronoms personnels et adjectifs possessifs en termes de spécificités.

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

Figure 5 :
Spécificités des pronoms personnels et adjectifs possessifs.

¹¹ Cf., sur ce problème du temps lexical le tutoriel n°2 du présent volume.

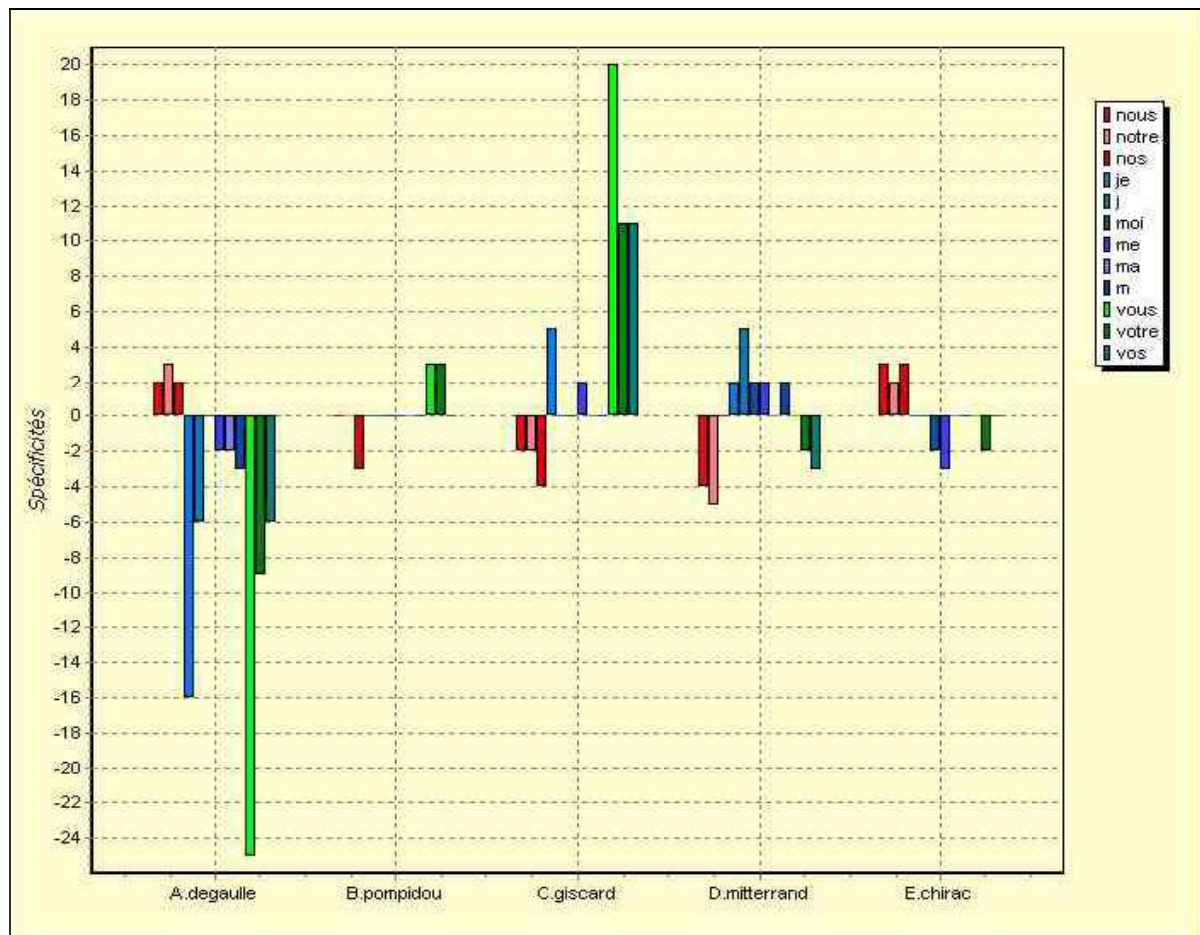


Figure 6 :
Pronom personnels et adjectifs possessifs. Histogramme des spécificités

Ces oppositions relèvent-elles uniquement de la personnalité du locuteur où trouvent-elles également leur origine dans une modification plus profonde du genre vœux présidentiels au cours des années que nous étudions ?

6 Registres personnels ou chronologie ?

Les spécificités chronologiques font apparaître un phénomène général qui n'est pas simplement le fait de personnalités différentes mais témoigne d'une évolution du discours politique, et de la conception de la fonction présidentielle. Il semblerait qu'au fil du temps le président de la République assume son énoncé, s'implique d'avantage dans son propre discours. Le tableau des spécificités évolutives de la première personne du singulier montre une tendance à la personnalisation du discours centrée sur le *je* qui ne semble pas cependant se poursuivre chez Chirac avec la même intensité.

L'observation des fréquences relatives de la première personne du singulier chez le dernier président de la République affine ce constat.

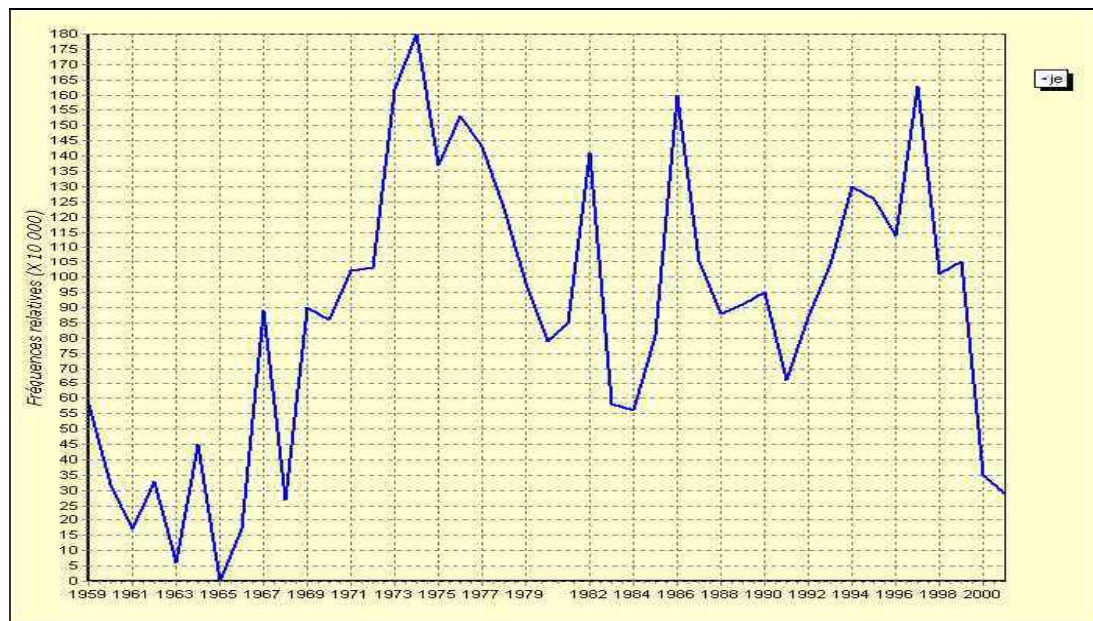


Figure 7 :

Fréquences relatives des marques de la première personne du singulier.

On note une diminution sensible des marques de la première personne du singulier (le pronom personnel suivant globalement la même évolution que le groupe *je, j' mon, ma, mes, me, m', mien(s), mienne(s)*). La personnalisation du discours n'a cessé de décroître sur l'ensemble de la période considérée, malgré un pic important en 1997. Cette période coïncidant avec le début de la cohabitation, doit-on en déduire que le chef de l'État s'est alors personnellement impliqué dans le discours pour revenir sur les raisons de la dissolution, pour redéfinir son rôle, se présenter comme le garant des institutions, légitimer sa présence dans la nouvelle configuration politique qui se présente à lui ?

Autour du *je* présidentiel...

Examinons maintenant le statut du *je* présidentiel réduit à sa forme graphique en explorant ses espaces cooccurentiels dans les messages de vœux aux Français.

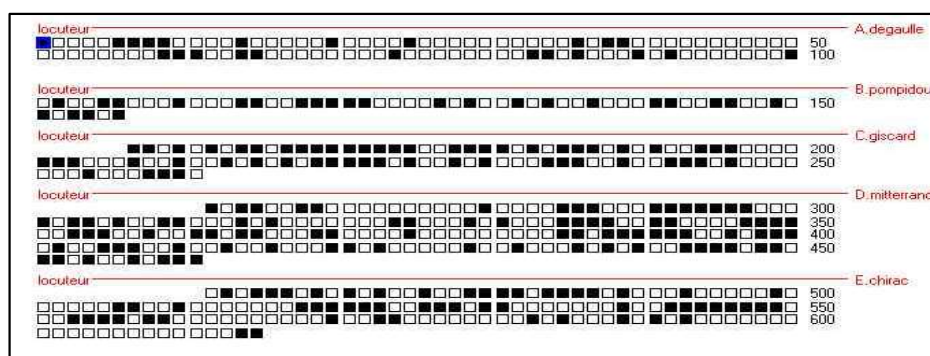


Figure 8 :

Carte des sections (paragraphe) de la forme *je* sur la partition locuteur.

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

Dans cette « topographie textuelle », [Lamalle, Salem, 2002] chaque carré du tableau 5 représente un paragraphe du texte. Les unités colorées indiquent les paragraphes qui contiennent 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 rares îlots chez De Gaulle.

Le calcul des cooccurents 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 si ce n'est que les parties sur lesquelles porte la comparaison ne sont plus constituées sur la base d'une partition en locuteurs mais à partir de la présence ou de l'absence du pronom personnel *je*. 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.

Tableau 9 :
Cooccurents spécifiques majeurs de la forme *je*

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

Les spécificités positives montrent la forte proportion des verbes qui gravitent autour du référent-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 cooccurents spécifiques. On note aussi de façon plus inattendue la présence d'un verbe à la troisième personne du pluriel : *vivent* (+3), dont on trouve les occurrences 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).

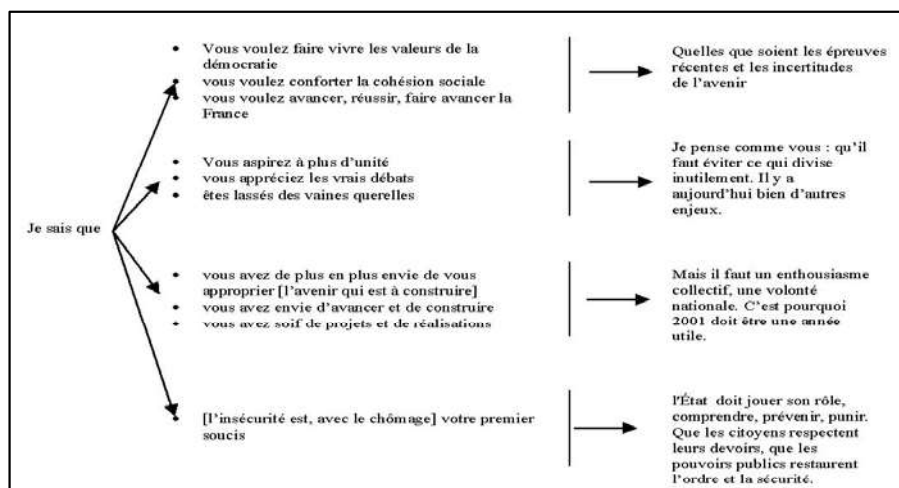
Ces messages sont donc particulièrement marqués par des verbes de « circonstance », (*souhaiter, adresser former*), par des volitifs et des verbes exprimant la connaissance. Cependant, cette interprétation sémantique a priori doit être corrigée par l'examen des contextes.

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 vœux 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'événementiel (*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 10.

Tableau 10 :
Marques de l'empathie chez J. Chirac et relances incitatives.
L'empathie dans les messages de vœux, les cognitifs...



L'examen des verbes associés à la première du singulier a révélé une caractéristique importante des messages de vœux : l'empathie et la compassion, au moins à partir des années Pompidou. Ces emplois s'accompagnent souvent de procédés argumentatifs articulés autour de deux schémas essentiels : *Je sais donc, je sais mais...* Ceci nous engage à explorer deux pistes nouvelles : quelles autres traces de l'empathie peut-on relever au sein des vœux présidentiels ? Les cognitifs y tiennent-ils une place particulière comme on l'a constaté avec le verbe savoir ?

Groupe de formes sais, vois, constate,distribution par locuteur et par année.

7 Références

Lamalle C., Salem A., « Types généralisés et topographie textuelle dans l'analyse quantitative des corpus textuels », in *actes des 6èmes journées d'analyse statistique des données textuelles*, Inria, St Malo, 2002

8 Fonctionnalités Lexico3 utilisées dans cette navigation

N°	Fonctionnalité	Résultat
2	Partition (clé a, pour année)	
5	Principales car lexicom (PCLC)	Tableau 2
5.6	Accroissement du vocabulaire (corpus)	Figure 1
5.6	Accroissement du vocabulaire (P92, P93)	Figure 2
5.6	Accroissement du vocabulaire ([P92,P93] et [P89,P90])	Figure 4
4	Segments Répétés (seuil minimal =2)	
8	Sélection d'un Type (occurrence de SR long>10)	
7	Carte des sections (paragraphe, présence SR de long>10)	Figure 3

Essai de textométrie politique chinoise

Les congrès du Parti Communiste Chinois (1921-2007)

[Congrès PCC]

André Salem, Li-Chi Wu¹²

salem@msh-paris.fr, wulucie@yahoo.com.tw

Résumé : L'analyse textométrique d'une série de discours prononcés par les principaux dirigeants du Parti communiste chinois, à l'occasion des congrès survenus depuis sa création, permet d'étudier l'évolution du vocabulaire employé au cours de son histoire. Des typologies d'ensemble permettent de vérifier le caractère progressif de cette évolution. L'analyse des spécificités majeures pour chacune des périodes permet d'en dégager le vocabulaire caractéristique. L'analyse des *chroniques spécifiques* permet de localiser des moments importants de l'évolution autour desquels on peut ensuite organiser avec précision le retour au contexte. Les mêmes méthodes appliquées à l'étude d'une série plus restreinte (1982-2007) permettent de décrire plus précisément les évolutions survenues dans la dernière période.

Mots-clés : Textométrie, séries textuelles chronologiques, discours politique chinois

Abstract : The textometric analysis of a series of speeches made by the main leaders of the Chinese Communist Party during its Congress provides insight into the evolution of vocabulary used since the party's creation. Global typologies of the text show the progressive character of this evolution while the study of major specificities for each one of these periods yields its characteristic vocabulary. The analysis of specific chronicles helps locate important moments in the evolution of these speeches and points out precise areas of contextual interest. The same methods when applied to a narrower period (1982-2007) help describe more precisely the evolutions which occurred in the last chronological period.

Keywords : Textometrics, textual time series, Chinese political discourse

摘要

本文运用词量法来分析自中国共产党成立以来, 历届重要领导人在共产党代表大会上所作的报告, 从而对报告中词汇使用的历史演变进行研究。可以通过整体类型学来验证这一演变的特性。通过历时性特征分析, 可以勘测出词汇使用变化的几个关键时期, 并将其回归到上下文从而进行详细考察。文章最后, 特别针对党代表大会的后半阶段(1982-2007), 采用相同的研究方法进行分析, 更详尽地推敲此时期词汇的演变。

关键词 : 词量法 ; 年代文本系列 ; 党代表大会报告

1 Contexte de la recherche

Il y a peu de temps, l'étude que nous présentons aujourd'hui, avec des défauts et des lacunes que nous espérons provisoires, n'aurait pu être envisagée. En effet, si la textométrie des séries chronologiques connaît depuis une dizaine d'années de nombreuses applications portant sur des corpus rassemblés dans des sphères d'activités diverses, le traitement textométrique de larges séries chronologiques de textes chinois restait, pour plusieurs raisons, totalement inaccessible tout au long de ces dernières années.

D'abord les difficultés inhérentes à la prise en charge informatique des textes rédigés dans des langues qui, à l'instar du chinois et à la différence des langues occidentales, nécessitent pour leur transcription un très grand nombre de signes graphiques. En second lieu, l'impossibilité

¹² Les auteurs remercient Miao Jun, Serge Fleury et Cédric Lamalle pour leur aide et leurs conseils précieux dans la réalisation de ce travail.

qu'il y avait jusqu'à ces derniers temps à se procurer, à des fins d'analyse textométrique, des matériaux textuels concernant une longue période chronologique, librement utilisables et présentés sur support informatisé avec une homogénéité suffisante.

Toutes ces limites sont brusquement tombées dans la dernière période du fait d'un développement rapide et parallèle de deux séries d'activités. D'une part, des systèmes d'encodage et de restitution des textes adaptés à la grande variété des systèmes d'écriture existant dans le monde ont été adoptés par l'ensemble de la communauté des informaticiens. Parallèlement, des standards de présentation se sont progressivement imposés à tous les internautes désireux de manifester une présence sur le web (organisations internationales, partis politiques, institutions gouvernementales et indépendantes, entreprises commerciales, etc.). La nécessité de mettre à disposition du public un certain nombre d'archives textuelles présentant le périmètre de l'activité de l'intervenant, permettant également d'appréhender son histoire, a progressivement été intégrée au nombre de ces standards.

Ces circonstances nous ont permis de recueillir sans trop de difficultés un vaste corpus de textes politiques chinois qui s'étend sur près d'un siècle afin de le soumettre à des analyses textométriques dont les résultats font l'objet du présent article.

Avant de passer à l'exposé des résultats obtenus, il convient de préciser notre propre vision critique du type d'activité auquel nous sommes livrés à propos de ces textes. Notre sentiment est que nous avons eu la chance d'être parmi les premiers à projeter, sur une masse de textes produits dans des conditions historiques que nous connaissons de manière insuffisante, un éclairage textométrique que nous avons appris à maîtriser sur d'autres séries textuelles chronologiques rassemblées selon les mêmes principes dans des langues et à des époques différentes.

Nous sommes conscients que cet avantage méthodologique ne nous a pas transformés, pour autant, en spécialistes de la période historique considérée. Tout au long de cette étude, nous avons tenté de résister aux tentations de réintroduire par un biais ou un autre, afin d'expliquer les variations les plus importantes constatées au plan textométrique, les quelques éléments d'histoire de la période connus de tous. Dans certains cas cependant, il nous a semblé indispensable, pour la compréhension du lecteur, d'établir un lien entre des résultats obtenus de manière formelle et certains éléments du contexte historique dans lequel les textes avaient été produits. Pour souligner le caractère hétérogène des remarques de ce dernier type par rapport aux résultats que nous avons obtenus en utilisant des méthodes de traitement formalisées et automatisées, nous avons fait précéder ces tentatives d'interprétation de la mention *commentaire*. Ces *commentaires* constituent dans notre esprit autant d'incitations adressées aux spécialistes de l'étude des textes politiques chinois à prolonger, en se servant de leur connaissance du contexte sociopolitique, l'indispensable processus d'interprétation des constats obtenus à l'aide des procédures textométriques formalisées.

Ainsi, nous nous sommes efforcés de présenter nos analyses de la manière la plus vérifiable possible et de proposer des modes de description compréhensibles par d'autres chercheurs. Notre but serait que les constats que nous avons opérés puissent leur être utiles, qu'ils confirment certaines de leurs connaissances, en infirment d'autres, qu'ils suggèrent surtout de nouvelles pistes de recherche multidisciplinaires dans un cadre de description stabilisé. Nous espérons que l'exposé qui suit pourra convaincre le lecteur que nous ne nous sommes pas trop écartés de ces principes posés au départ.

2 Recueil des données, préparation du corpus

Pour cette première étude sur l'évolution du discours politique chinois, nous avons choisi de rassembler des textes produits dans les congrès du Parti communiste chinois (PCC), organisation politique créée à Shanghai en 1921 et au pouvoir en Chine populaire depuis 1949. Pour chacun des dix-sept congrès, qui se sont tenus durant cette période, nous avons sélectionné le discours prononcé par le principal responsable du parti. Nous avons appelé **CongPCC** ce corpus chronologique qui s'étend sur près de 90 ans.

L'ensemble du corpus compte 331 239 occurrences de 2 336 caractères (*hanzi*) différents. Le corpus peut être divisé en dix-sept parties qui correspondent chacune à un des congrès dont on trouvera la liste au tableau 3.

2.1 Rappels sur le chinois et son système d'écriture¹³

Le chinois possède un système d'écriture dans lequel chaque caractère correspond plus ou moins à un morphème et à une syllabe de l'oral. Les caractères (appelés aussi *sinogrammes* ou *hanzi*) utilisés par ce système d'écriture se comptent en milliers (parfois en dizaines de milliers, si l'on tente d'inclure des graphies rares ou tombées en désuétude)¹⁴. Cette profusion de signes graphiques a longtemps constitué un obstacle majeur au traitement des textes chinois sur ordinateur. L'avènement du système Unicode permet désormais de coder l'ensemble des caractères chinois, aussi bien sous leur version traditionnelle que sous la forme simplifiée adoptée en 1950¹⁵.

Bien que l'écriture chinoise n'intègre pas d'espace entre les unités lexicales¹⁶, la notion de *mot* est présente dans la grammaire chinoise. Pour appréhender un texte, le lecteur chinois découpe la chaîne textuelle en unités distinctes, à partir de sa connaissance du lexique. Les mots chinois sont composés d'un ou de plusieurs sinogrammes. Les textes s'écrivent sans que des espaces ne viennent séparer les mots (cf. tableau 1, infra). Un même mot peut avoir différentes fonctions grammaticales en fonction du contexte dans lequel il entre (ex : 来回 *lai hui* verbe : aller et venir¹⁷ ; substantif : un aller-retour ; proposition adverbiale : en allant et revenant). La plupart du temps, le genre et le nombre des noms ne sont pas marqués explicitement (ex : 工人 *gong ren* ouvrier / ouvriers / ouvrière/ ouvrières).

2.2 Segmentation du texte

Les comparaisons textométriques supposent que l'on définisse des unités de décompte dont on étudiera les variations de fréquence au sein des différentes parties du corpus.

¹³ On trouvera une description rapide du système d'écriture chinois et des problèmes du dépouillement automatique des corpus de textes chinois dans [Miao et Salem 2008] pour une description plus complète du système d'écriture chinois, on se reportera par exemple à [Alleton 1997].

¹⁴ Le *Kangxi zidian* (Dictionnaire de Kangxi), élaboré sur commande impériale en 1716, compte au total 47 035 caractères, intégrant des raretés, des variantes et des sinogrammes obsolètes. Le *Xinhua zidian* (Dictionnaire Xinhua), la référence actuelle pour ce qui est des dictionnaires de caractères, en compte environ 10000.

¹⁵ A partir de 1952, le gouvernement de la République Populaire de Chine commence à promouvoir un système d'écriture en *caractères simplifiés*, appelés à remplacer les caractères traditionnels. Un comité pour la réforme de l'écriture est créé pour superviser cette tâche. En 1964, une liste de caractères simplifiés, qui sera plusieurs fois mise à jour par la suite, est publiée dans un document intitulé « Table intégrale des caractères chinois simplifiés » (简化字表 *jianhua zi congbiao*).

¹⁶ A l'instar de nombreux systèmes d'écritures de l'antiquité gréco-latine.

¹⁷ Dans ce qui suit, nous nous efforcerons de respecter les conventions typographiques suivantes : les séquences textuelles citées en chinois sont suivies de leur transcription *pinyin* (en italique) ainsi que de leur traduction française, ex : 工人 *gong ren* ouvrier.

Tableau 1

Extrait du discours du secrétaire général du PCC au 17^e congrès (2007)
reproduit sur le site web de l'agence *Xin Hua* 新华网¹⁸

高举中国特色社会主义伟大旗帜 为夺取全面建设小康社会新胜利而奋斗
——在中国共产党第十七次全国代表大会上的报告

(2007年10月15日)

胡锦涛

同志们：

现在，我代表第十六届中央委员会向大会作报告。

中国共产党第十七次全国代表大会，是在我国改革发展关键阶段召开的一次十分重要的大会。大会的主题是：高举中国特色社会主义伟大旗帜，以邓小平理论和“三个代表”重要思想为指导，深入贯彻落实科学发展观，继续解放思想，坚持改革开放，推动科学发展，促进社会和谐，为夺取全面建设小康社会新胜利而奋斗。

中国特色社会主义伟大旗帜，是当代中国发展进步的旗帜，是全党全国各族人民团结奋斗的旗帜。解放思想是发展中国特色社会主义的一大法宝，改革开放是发展中国特色社会主义的强大动力，科学发展、社会和谐是发展中国特色社会主义的基本要求，全面建设小康社会是党和国家到二〇二〇年的奋斗目标，是全国各族人民的根本利益所在。

Pour pouvoir mettre en œuvre des comparaisons textométriques, on utilise dans un premier temps des outils informatiques qui permettent de découper automatiquement au fil du texte les unités dont on pourra ensuite décompter les occurrences. Plusieurs logiciels réalisent désormais des segmentations automatiques des textes chinois en mots. Pour dépouiller notre corpus, nous avons utilisé un seul de ces logiciels dans le but d'obtenir une cohérence maximale pour l'ensemble du dépouillement¹⁹.

On peut voir, dans le haut du tableau 2, un extrait de ce même texte segmenté automatiquement. On trouve une traduction de ce dernier extrait dans le bas du même tableau. Des caractères ‘#’ ont été insérés pour délimiter les divisions qui correspondent aux paragraphes que l'on peut repérer dans le texte original²⁰.

¹⁸ Téléchargeable sur : http://news.xinhuanet.com/ziliao/2004-11/24/content_2255749.htm.

¹⁹ Dans le présent travail, nous avons utilisé exclusivement le logiciel : ICTCLAS (Institute of Computing Technology, Chinese Lexical Analysis System), mis au point par Kevin Zhang à l'Institute of Computing Technology, Chinese Academy of Sciences. Signalons que l'observation de la ventilation des unités ainsi découpées dans le corpus nous a quelquefois conduits à corriger des erreurs manifestes produites par la segmentation automatique du texte (segmentations différentes pour une même séquence de caractères dans différentes portions du texte).

²⁰ L'écriture chinoise possède ses propres signes de ponctuation. Pour faciliter le traitement informatique par des logiciels qui ne prennent pas encore en compte la ponctuation chinoise, nous avons remplacés les principales ponctuations du chinois par leurs équivalents occidentaux. Le caractère de ponctuation *demi-virgule* qui sert en chinois à séparer les

Tableau 2
Corpus textométrique *CongPCC*
 Le début du discours de 2007 balisé et segmenté en mots suivi de sa traduction française

```
#<cong=2007>
#胡锦涛 在 中国 共产党 第 十七 次 全国 代表 大会 上 的 报告
# 高举 中国 特色 社会主义 伟大 旗帜 为 夺取 全面 建设 小康 社会 新 胜利 而 奋斗
# -- 在 中国 共产党 第 十七 次 全国 代表 大会 上 的 报告
# ( 2007 年 10 月 15 日 )
# 胡锦涛
# 同志 们 :
# 现在 , 我 代表 第 十六 届 中央 委员会 向 大会 作 报告 .
# 中国 共产党 第 十七 次 全国 代表 大会 , 是 在 我国 改革 发展 关键 阶段 召开 的 一 次 十 分
重要 的 大会 . 大会 的 主题 是 : 高举 中国 特色 社会主义 伟大 旗帜 , 以 邓小平 理论 和 “
三个 代表 ” 重要 思想 为 指导 , 深入 贯彻 落实 科学 发展 观 , 继续 解放 思想 , 坚持 改革 开放 ,
推动 科学 发展 , 促进 社会 和谐 , 为 夺取 全面 建设 小康 社会 新 胜利 而 奋斗 .
# 中国 特色 社会主义 伟大 旗帜 , 是 当代 中国 发展 进步 的 旗帜 , 是 全党 全国 各族
人民 团 结 奋斗 的 旗帜 . 解放 思想 是 发展 中国 特色 社会主义 的 一 大 法宝 , 改革 开放 是
发展 中国 特色 社会主义 的 强大 动力 , 科学 发展 & 社会 和谐 是 发展 中国 特色 社会主义 的
基本 要求 , 全面 建设 小康 社会 是 党 和 国家 到 二 0 二 0 年 的 奋斗 目标 , 是 全国 各族
人民 的 根本 利益 所在 .
```

```
<cong=2007>
# texte intégral du rapport de hu jintao prononcé au xviie congrès du pcc
# lever bien haut l'étendard du socialisme à la chinoise et lutter pour
réaliser de nouveaux succès dans le développement général d'une société de
moyenne aisance
# -- rapport au xviie congrès du parti communiste chinois
# (15 octobre 2007)
# hu jintao
# camarades,
# au nom du xvie comité central, j'ai l'honneur de présenter ici le rapport
d'activité au xviie congrès du parti.
# le xviie congrès du parti communiste chinois est un congrès particulièrement
important dû au fait qu'il s'est tenu dans une phase cruciale de la réforme et
du développement de notre pays. le thème de ce congrès se définit comme suit :
en brandissant bien haut l'étendard du socialisme à la chinoise et en nous
guidant sur la théorie de deng xiaoping et sur la pensée importante de triple
représentativité, appliquons le concept de développement scientifique de façon
approfondie, continuons à manifester une grande liberté d'esprit, poursuivons
la réforme et l'ouverture sur l'extérieur, encourageons le développement
scientifique, favorisons l'harmonie sociale et luttons pour réaliser de
nouveaux succès dans le développement général d'une société de moyenne
aisance.
# l'étendard du socialisme à la chinoise est un drapeau qui guide la chine
contemporaine vers le développement et le progrès, et qui exhorte tout le
parti et tout le peuple multiethnique à se livrer à la même lutte dans une
parfaite union. pour le développement du socialisme à la chinoise,
l'émancipation d'esprit est une arme particulièrement efficace, la réforme et
l'ouverture sur l'extérieur une force motrice puissante, et le progrès
scientifique et l'harmonie sociale une exigence fondamentale. quant à la
construction in extenso d'une société de moyenne aisance, elle constitue un
objectif que notre parti et notre pays devront atteindre d'ici à 2020, et qui
reflète les intérêts fondamentaux de notre peuple multiethnique.
```

différents termes d'une énumération a été remplacé par le caractère '＆'. Ajoutons que dans l'attente de la prise en compte des fichiers unicode par les futures versions du logiciel Lexico3 que nous avons utilisé, le corpus a été stocké dans l'encodage GB2312 (Windows-936).

2.3 Caractéristiques quantitatives du corpus CongPCC

Dans ce qui suit nous avons choisi de travailler sur des textes dépouillés en occurrences de *mots* découpés par le segmenteur automatique *ICTCLAS*. Le corpus ainsi découpé compte 179 447 occurrences de 10 680 formes graphiques différentes. On recense 4 272 formes hapax, la forme la plus fréquente (la particule 的 *de*) compte 15 273 occurrences.

==== Repères chronologiques sur la période (1921-2007) ====

- 1921** Fondation à Shanghai du Parti Communiste Chinois.
Formation d'un gouvernement nationaliste dirigé par Sun Yat-sen.
- 1925** Mort de Sun Yat-sen - Chiang Kai-shek à la tête du Guomindang.
- 1927** Chiang Kai-shek écrase la révolution à Shanghai et crée un gouvernement.
- 1931** Les communistes établissent la *République soviétique chinoise* à Jiangxi.
- 1934** *Longue Marche* de l'Armée rouge, pour échapper aux nationalistes.
- 1937** Début de la guerre contre le Japon.
- 1946** Guerre civile entre nationalistes et communistes.
- 1949** Les communistes fondent de la *République populaire de Chine*.
Le gouvernement nationaliste du Guomindang se réfugie à Taiwan.
- 1950** Traité d'amitié sino-soviétique. Guerre de Corée (1950-1953).
- 1953** Plan quinquennal, priorité à l'industrie lourde. *Collectivisation agricole*.
- 1957** Campagne des *Cent fleurs*, suivie d'une répression en juin.
- 1958** Lancement du *Grand bond en avant* et des communes populaires.
- 1960** Catastrophe alimentaire, crise sino-soviétique.
- 1966** Début de la *Révolution culturelle*.
- 1975** Campagne des *quatre modernisations* (armée, science, agriculture, industrie)
- 1976** Mort de Zhou Enlai et de Mao Zedong. Fin de la révolution culturelle, arrestation de la *bande des quatre*.
- 1978** Les rénovateurs au pouvoir avec Deng Xiaoping.
- 1979** Création de *zones économiques spéciales*.
- 1984** Relance des réformes économiques.
- 1987** Reprise des relations entre la Chine populaire et Taiwan.
- 1989** Répression de la contestation étudiante sur la place Tian'anmen.
- 1992** Deng Xiaoping lance *l'économie socialiste de marché*.
- 1993** Jian Zemin chef de l'Etat et secrétaire général du PCC.
- 1997** Mort de Deng Xiaoping. Rétrocession de Hong Kong à la Chine.
- 2004** Hu Jintao dirige le PCC.

3 La série chronologique 1921-2007

Comme nous l'avons signalé plus haut, la série des dix-sept congrès du PCC survenus entre 1921 et 2007, bien qu'elle ait été produite par une même organisation politique tout au long des 90 années couvertes par le corpus, ne présente pas une homogénéité totale du point de vue des conditions de production. Nous tenterons, dans un premier temps, d'établir une typologie sur l'ensemble des textes afin de mettre en évidence les principaux clivages dans l'évolution du vocabulaire. Nous analyserons, dans un second temps des groupes constitués sur une base plus homogène afin d'affiner les premiers constats.

Tableau 3

Principales caractéristiques lexicométriques du corpus *CongPCC*

Année	Lieu	occurrences	formes	hapax	Fréq. Max	Forme
1921	<i>Shanghai</i>	1441	481	276	158	的
1922	<i>Shanghai</i>	5067	1529	882	467	的
1923	<i>Guangzhou</i>	467	251	172	28	的
1925	<i>Shanghai</i>	1543	562	328	127	的
1927	<i>Wuhan</i>	3786	991	545	416	的
1928	<i>Moscou</i>	4284	1017	534	485	的
1945	<i>Yanan</i>	22130	3038	1403	2046	的
1956	<i>Pékin</i>	23132	3005	1330	2432	的
1969	<i>Pékin</i>	11002	2254	1155	988	的
1973	<i>Pékin</i>	5003	1440	806	418	的
1977	<i>Pékin</i>	15375	2783	1399	1251	的
1982	<i>Pékin</i>	16501	2840	1419	1556	的
1987	<i>Pékin</i>	16114	2785	1300	1375	的
1992	<i>Pékin</i>	12868	2605	1335	984	的
1997	<i>Pékin</i>	13827	2617	1305	1023	的
2002	<i>Pékin</i>	13669	2545	1261	948	的
2007	<i>Pékin</i>	13238	2587	1276	571	的

Les textes ainsi rassemblés correspondent à ce que nous appelons une *série textuelle chronologique*²¹. En effet, ces textes ont été produits au cours du temps, dans une même langue, par une même source institutionnelle, dans un cadre qui s'est toujours appelé *congrès*. Cependant, la longueur de la période considérée (plus de 90 ans), le statut de l'organisation politique responsable de ces textes, laquelle s'est transformée, durant la période que nous considérons, d'un petit groupe d'intellectuels opposants en un parti politique dominant durablement la vie politique du pays, nous permet de questionner l'homogénéité réelle de conditions de productions qui peuvent sembler proches au premier abord.

²¹ Pour de plus amples informations sur les séries textuelles chronologiques, on consultera par exemple [Salem 94] et [Lebart & Salem 94, p.197-240].

En outre, l'examen du volume des textes produits au cours de la période montre que la production n'a pas été régulière. Aux textes très courts, entre 500 occurrences de mots graphiques (1923) et 5000 occurrences (1922), produits dans la première période ont succédé de longues périodes pendant lesquelles aucun congrès n'a été convoqué. Les congrès de 1945 et de 1956 qui se sont tenus à des intervalles de 10 à 13 ans avec le congrès précédent ont donné lieu à des textes très volumineux (respectivement de 22 000 et de 23 000 occurrences). A partir de 1982, après la fin des troubles liés à la *révolution culturelle*, le congrès est convoqué régulièrement cinq ans après le congrès précédent. Le discours du principal responsable se présente désormais sous la forme d'un texte qui compte entre 13 000 (1992) et 16 000 (1982) occurrences. La série des derniers congrès (1982-2007) constitue du point de vue de l'ensemble de ces critères une série textuelle chronologique nettement plus homogène que la série complète des dix-sept congrès rassemblés dans le corpus.

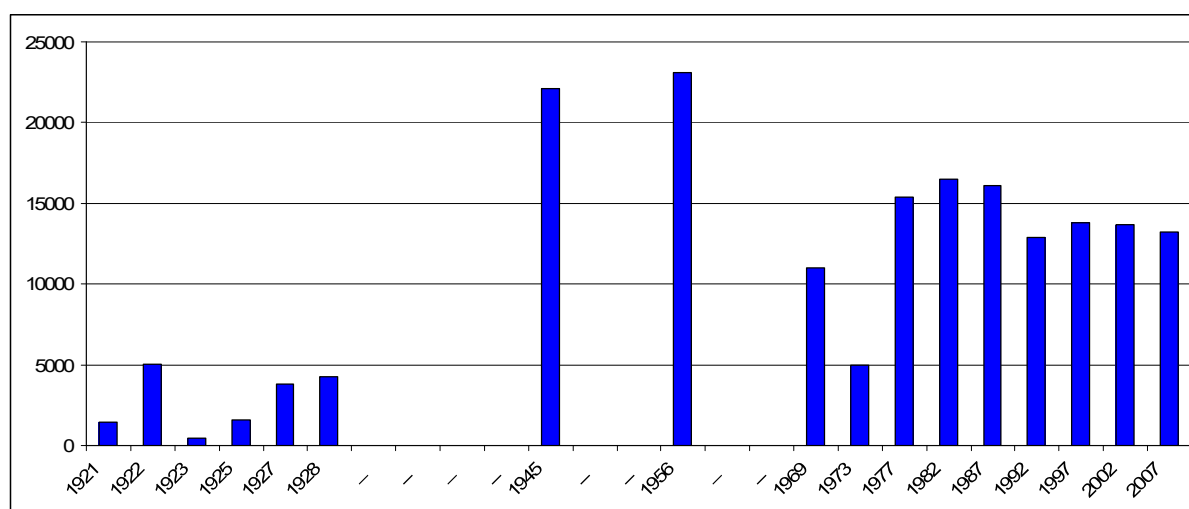


Figure 1

Les 17 congrès du PCC entre 1921 et 2007
et leur taille exprimée en nombre d'occurrences après segmentation en *mots*

3.1 Typologie d'ensemble

Nous commencerons par analyser l'ensemble des textes réunis dans la série complète (1921-2007) afin de mettre en évidence les principaux changements survenus dans l'emploi du vocabulaire entre le début et la fin de la période couverte par notre corpus. Dans un second temps, nous analyserons une série plus restreinte et nettement plus homogène constituée par les textes produits dans les dernières années du corpus (1982-2007) afin de cerner plus précisément les évolutions survenues dans la dernière période.

Comme c'est souvent le cas lorsqu'on analyse une série textuelle chronologique, l'analyse factorielle (AFC, cf. infra) réalisée à partir du corpus constitué par le décompte des formes de fréquence supérieure à dix dans les dix-sept congrès rassemblés dans le corpus **CongPCC** nous amène au constat d'une forte dépendance du vocabulaire par rapport à la période dans laquelle s'est tenue le congrès²². Les congrès séparés par une période de temps courte se

²² Dans les analyses textométriques, on a pris l'habitude d'asseoir ce genre de typologie sur les formes dont la fréquence dépasse un certain seuil afin de restreindre le nombre des unités lexicales. Notons que les typologies du même type appuyées sur les décomptes faisant intervenir la totalité des formes conduisent à des résultats très similaires.

trouvent placés à proximité les uns des autres sur les synthèses factorielles alors que ceux qui sont séparés par un intervalle de temps plus long se trouvent plus distants²³.

On retrouve sur la gauche du graphique, dans un ordre strictement chronologique les congrès survenus entre 1982 et 2007. Les textes, nettement moins longs, qui correspondent aux congrès survenus entre 1921 et 1945 se situent sur la droite du graphique, dans un ordre proche de l'ordre chronologique même si on peut remarquer quelques anomalies, de ce strict point de vue. Un troisième groupe est constitué par les trois congrès survenus entre 1969 et 1977, ce qui témoigne d'une certaine homogénéité de ces groupes pour ce qui concerne l'emploi du vocabulaire.

Les congrès de 1945 et 1956 constituent des cas particuliers à la fois par l'abondance des matériaux textuels auxquels ils ont donné lieu et par les très longues périodes de temps qui se sont écoulées entre les congrès. Un examen attentif des coordonnées factorielles sur axes suivants conduit à ne pas rattacher le congrès de 1945, essentiellement dominé par l'expression de préoccupations liées à la guerre contre le Japon au groupe 1921-1928, même si sa position sur les deux premiers axes témoigne par ailleurs d'une proximité lexicale avec la période initiale.

==== Repères méthodologiques ====

L'analyse statistique des séries textuelles chronologiques

- De nombreuses études ont montré que l'évolution du vocabulaire au fil du temps constitue la caractéristique statistique majeure des corpus constitués par échantillonnage au cours du temps d'une source textuelle homogène.
- L'*Analyse factorielle des correspondances* (AFC) permet de mettre en évidence les principales caractéristiques de cette évolution d'ensemble et de localiser les principaux moments de rupture dans l'emploi du vocabulaire.
- L'*Analyse des spécificités* permet d'illustrer chacune des périodes du corpus par la mise en évidence du vocabulaire qu'elle emploie ou au contraire du vocabulaire qu'elle cesse d'employer par rapport aux autres périodes du corpus.
- L'*Analyse des spécificités chronologiques* permet d'illustrer chacune des périodes du corpus par la mise en évidence du vocabulaire qu'elle emploie ou au contraire du vocabulaire qu'elle cesse d'employer par rapport aux périodes qui ont précédé.

Pour en savoir plus :

Lebart, L., Salem, A., *Statistiques textuelles*, Paris, Dunod, 1994

Salem A., "La lexicométrie chronologique", *Actes du colloque de lexicologie politique "Langages de la Révolution"*, collection "St.Cloud", Paris, Klincksieck, 1994

²³ L'analyse a été réalisée à partir du tableau des 2119 formes dont la fréquence est supérieure ou égale à dix occurrences. Un premier axe se détache très nettement (23% de l'inertie totale). L'ensemble présente les caractéristiques communes aux séries textuelles chronologiques.

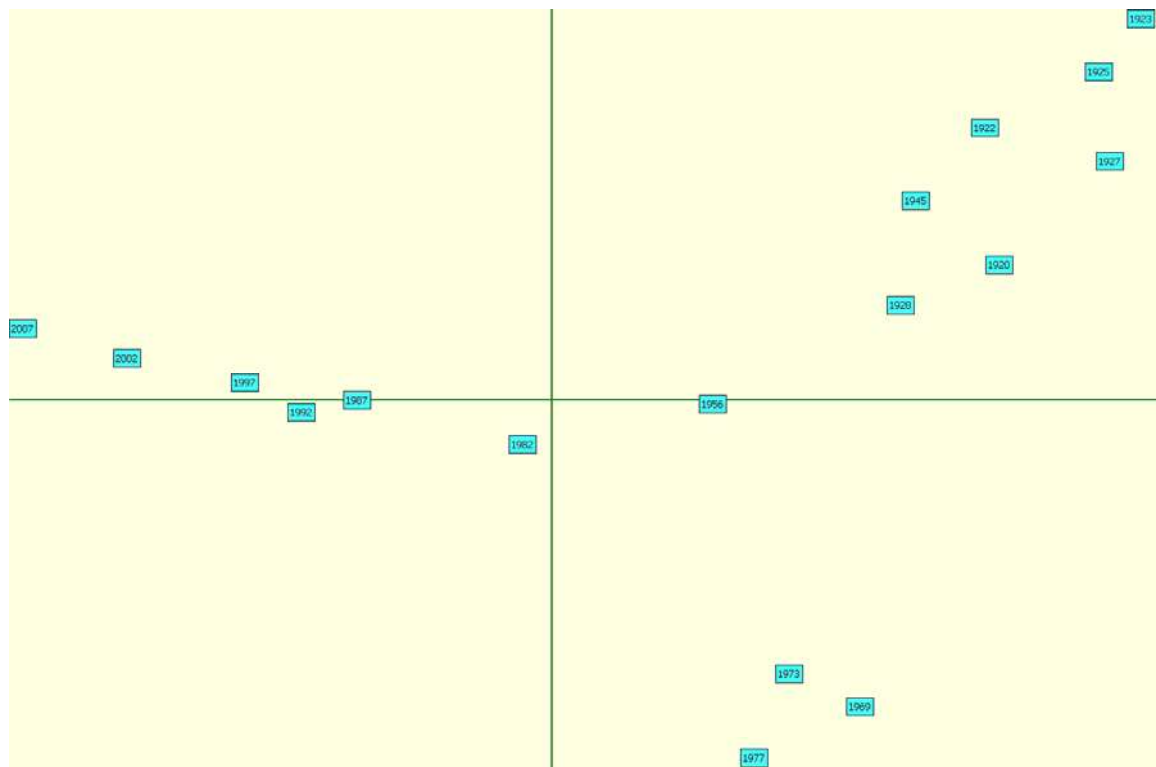


Figure 2
Analyse factorielle des correspondances du corpus

==== Guide de lecture pour la figure 2 ====

- Les congrès sont positionnés sur les deux premiers facteurs issus de l'AFC réalisée à partir du tableau (2 119 formes de fréquence ≥ 10 x 17 congrès).
- Les congrès séparés par un intervalle de temps réduit sont relativement proches sur ces axes.
- Un premier axe (horizontal sur la figure 2) se détache nettement, qui porte 23% de l'inertie totale. Les congrès se disposent sur cet axe selon un ordre qui est proche de l'ordre chronologique (schéma classique pour l'analyse des *séries textuelles chronologiques*). Comme c'est souvent le cas pour les analyses de ce type, pratiquées à partir de séries textuelles chronologiques, ces résultats permettent d'avancer l'hypothèse que c'est transformation du *stock lexical* au cours du temps qui sous-tend l'évolution mise en évidence.
- Enfin, les écarts que l'on remarque par rapport au schéma d'une progression uniforme constituent des points de départ pour une réflexion sur les moments particuliers de l'évolution du vocabulaire.

3.2 Principales périodes chronologiques

Dans ce qui suit, nous proposons de considérer quatre groupes de périodes découpés en tenant compte à la fois de leur proximité sur les axes factoriels et de l'intervalle temporel écoulé entre deux congrès successifs. Nous examinerons tour à tour quatre groupes de périodes : 1921-1928, 1945-1956, 1967-1979, 1982-2007.

Tableau 4
Les spécificités positives majeures de la période 1921-1928

	Forme	Équivalent français	FTot.	FPart	Coeff.
1	革命	révolution	577	213	***
2	军阀	expédition militaire	67	64	***
3	势力	puissance	126	85	***
4	帝国主义	impérialisme	225	132	***
5	运动	mouvement	176	91	47
6	国民	peuple	108	67	42
7	帝国主义者	impérialiste	49	44	40
8	小资产阶级	classe petite bourgeoise	62	46	35
9	所以	donc	78	52	35
10	之	(particule grammaticale)	215	85	33
11	阶级	classe	217	84	32
12	无产阶级	prolétariat	245	89	31
13	中国	Chine	1080	222	30
14	工人	ouvrier	98	53	29
15	民权	droit(s) du peuple	29	28	29
16	暴动	soulèvement	22	22	24
17	压迫	oppression	91	46	24
18	争斗	se battre	20	20	22
19	资本	capital	62	36	22
20	能	pouvoir	338	91	21
21	封建	féodale	57	33	20
22	列强	grande puissance	21	20	20
23	冲突	collision	33	25	20
24	资产阶级	classe bourgeoise	217	67	20
25	他	il	116	46	19
26	农民	paysans	208	65	19
27	意识	conscience	40	27	19
28	工农	ouvrier et paysan	29	22	18
29	群众	masse populaire	583	120	17
30	机会主义	opportunisme	25	19	16

==== Guide de lecture du tableau des spécificités majeures ====

- Les termes (formes et segments répétés) les plus spécifiques sont présentés par ordre d'indice de spécificité décroissant (i.e. en commençant par les termes dont la présence est jugée la plus abondante dans la période) du point de vue du calcul de l'indice.
- La colonne 3 fournit un équivalent français du terme chinois que nous proposons pour faciliter la compréhension du lecteur mais qui ne peut rendre compte de l'ensemble de ses emplois effectifs *en discours*, dans les différentes périodes du corpus.
- On trouve ensuite les fréquences respectives du terme : dans l'ensemble du corpus (col. *FTot*) et dans la partie considérée (col. *FPart*).
- La dernière colonne fournit l'indice calculé par la méthode des spécificités. Un indice de a correspond à une probabilité de l'ordre de 10^{-a} que le terme ait une fréquence supérieure à la valeur constatée. Les valeurs de l'indice supérieures à 50 ont été remplacées par la séquence de symboles ***.
- Un calcul similaire fournit un diagnostic pour chaque segment répété du corpus.

3.2.1 La période 1921-1928

Le tableau 4 rassemble les formes et les segments répétés les plus spécifiques pour les six congrès tenus dans la première période du corpus. On peut vérifier sur la figure 2 que ces congrès, dont la taille est inférieure à celle des autres congrès rassemblés dans le corpus, constituent un ensemble relativement homogène sur les premiers axes factoriels.

Commentaire : Le vocabulaire spécifique majeur des congrès appartenant à ce premier groupe semble avant tout lié à l'analyse économique et politique marxiste de la société chinoise propre aux textes de cette première époque (革命 *geming* révolution, 势力 *shili* puissance, 帝国主义 *diguo zhuyi* impérialisme, 帝国主义者 *diguo zhuyi zhe* impérialistes, 工人 *gongren* ouvrier(s), 小资产阶级 *xiao zichan jieji* classe petite bourgeoise, 无产阶级 *wuchan jieji* prolétariat, etc.). Au fil du temps, une partie de ce vocabulaire sera progressivement remplacée par d'autres formes textuelles en liaison avec l'évolution des catégories d'analyse socio-économique des instances de direction du PCC. Nous tenterons plus loin de dégager les principales tendances de cette évolution.

3.2.2 Les congrès de 1945 et de 1956

Comme nous l'avons signalé plus haut, ces deux congrès plutôt atypiques, surviennent après de longs intervalles sans convocation de l'instance suprême du parti. Ils donnent lieu à la publication de textes relativement abondants. Convoqué après dix-sept années sans congrès, le congrès de 1945 est dominé par les problèmes qui tournent autour de la guerre qui vient de s'achever avec le Japon (抗日 *kangri* combattre le japon, 解放区 *jiefang qu* région libérée, 日本侵略者 *riben qinlue zhe* envahisseur japonais, 侵略者 *qinlue zhe* envahisseurs). La mention de la Chine (中国 *zhong guo*) y est particulièrement fréquente.

Après un intervalle de onze années durant lesquelles le congrès n'est plus réuni à nouveau, le congrès de 1956, introduit un vocabulaire propre qui sera abandonné par la suite (五年计划 *wunian jihua* plan quinquennal, 社会主义改造 *shehui zhuyi gaizao* réforme du socialisme, etc.).

Semblables du point de vue de leurs tailles respectives, ces deux congrès ne constituent cependant pas une classe homogène du point de vue du vocabulaire employé.

3.2.3 La période 1969 -1977

Les congrès survenus entre 1969-1977, forment un groupe particulier sur les deux premiers axes issus de l'analyse factorielle présentée sur la figure 2. On note la présence de toute une série de formes de haute fréquence, quasi-exclusives pour cette période : 毛主席 *mao zhuxi* Président Mao (380 occurrences dans la période sur 380, au total), 四人帮 *sirenbang* la bande des quatre²⁴ (138 occurrences sur 139 au total).

Commentaire : Ces congrès portent la trace lexicale des affrontements internes, largement commentés par la presse internationale, qui ont marqué la période dominée par les luttes politiques autour de la succession politique Mao Zedong mort en 1976. On peut compléter cette liste en remarquant la présence de termes de moindre fréquence, également exclusifs pour ce groupe de périodes comme : 修正主义 *xiuzheng zhuyi* révisionnisme (42/42), 叛徒 *pantu* traître (36/36), etc. qui portent la trace lexicale d'affrontements politiques qui ont marqué la période considérée.

²⁴ Rappelons que sous cette appellation de 四人帮 *bande des quatre*, la direction du PCC désignait à l'époque un groupe de dirigeants ayant exercé une grande influence politique pendant la période de la *révolution culturelle*, démis de leurs fonctions puis arrêtés en 1976, après la mort de Mao Zedong.

Tableau 5
Les spécificités positives majeures de la période 1969-1977

	Forme	Équivalent français	FTot.	FPart	Coeff.
1	毛主席	Président Mao	380	380	***
2	四人帮	bande des quatre	139	138	***
3	主席	président	389	388	***
4	毛	Mao	384	384	***
5	毛主席的	du Président Mao	97	97	***
6	走资派	voie du capitalisme	60	60	46
7	无产阶级专政	dictature du prolétariat	83	68	38
8	苏修	Suxiu (nom propre)	43	43	34
9	无产阶级文化大革命	révolution culturelle du prolétariat	44	44	34
10	刘少奇	Liu Shaoqi	44	44	34
11	革命	révolution	577	221	33
12	修正主义	révisionnisme	42	42	33
13	无产阶级	prolétariat	245	123	32
14	列宁	Lénine	70	57	32
15	资产阶级	classe bourgeoise	217	109	28
16	叛徒	traître	36	36	28
17	斗争	lutte	353	149	28
18	反革命	contre-révolutionnaire	103	67	27
19	批	critiquer	54	44	25
20	搞	faire	89	59	25
21	月	mois	130	74	24
22	伟大	remarquable / important	318	131	24
23	四人帮 的	de la bande des quatre	32	31	23
24	他们	ils	546	191	23
25	指出	montrer	93	58	22
26	复辟	restauration monarchique	28	28	22
27	指示	indiquer	39	34	21
28	了	(particule grammaticale)	1789	470	21
29	阴谋	complot	50	39	21
30	教导	instruire	25	25	20
31	社会主义革命	révolution du socialisme	41	33	19
32	批判	porter un jugement critique	43	34	19
33	路线斗争	lutte de phase	24	24	19
34	毛主席关于	Président Mao à propos	24	24	19
35	夺权	prendre le pouvoir	24	24	19

3.2.4 La période 1982-2007

Nous avons rassemblé dans un dernier groupe les congrès survenus après 1982. Dans cette dernière période, les congrès sont convoqués régulièrement tous les cinq ans, le volume du discours du secrétaire général compte, à chaque occasion, près de 12 000 occurrences. Par ailleurs, l'examen de la figure 2 montre que, sous l'influence d'une transformation progressive du vocabulaire employé, les points représentant chacun de ces congrès s'éloignent vers la gauche de manière régulière au fur et à mesure que l'on avance dans la chronologie interne du corpus. Le tableau 6 présente le vocabulaire spécifique de ce dernier groupe de période.

Tableau 6
Les spécificités positives majeures de la période 1982-2007

	Forme	Équivalent français	FTot.	FPart	Coeff.
1	体制	système	276	274	***
2	建设	édification	1108	920	***
3	社会主义	socialisme	1318	913	***
4	经济	économie	1180	950	***
5	完善	excellent	217	216	***
6	中国特色	à la chinoise	177	177	***
7	坚持	insistance	466	395	***
8	社会	société	709	550	***
9	改革	réforme	650	594	***
10	和	et	4688	2865	***
11	开放	ouvrir au public	258	255	***
12	现代化	modernisation	238	230	***
13	特色	caractéristique	187	185	***
14	改革 开放	réforme et ouverture	161	161	***
15	发展	développement	1480	1138	***
16	文明	civilisation	155	155	50
17	加强	consolidation	449	363	46
18	推进	faire progresser	188	180	46
19	特色 社会主义	socialisme caractéristique	138	138	44
20	中国特色 社会主义	socialisme à la chinoise	136	136	44
21	制度	système	466	368	42
22	提高	améliorer	333	280	42
23	新	nouveau / neuf	621	466	42
24	管理	gestion	228	206	42
25	市场	marché	271	237	42
26	现代化 建设	oeuvre de modernisation	131	130	40
27	精神	élan	211	191	39
28	基本	fondamental	363	293	37
29	结构	structures	107	107	35
30	体制改革	réforme du système	103	103	33
31	加快	activer	116	113	32
32	体系	système	138	129	30
33	经济 体制	système économique	88	88	29
34	创新	innovation	90	90	29
35	有 中国特色	spécifique à la Chine	88	88	29

Commentaire : Les termes sélectionnés portent essentiellement sur les réformes du système économique et sur les problèmes de l'adaptation des principes généraux du socialisme aux spécificités chinoises (改革 开放 *gaige kaifang* réforme et ouverture, 现代化 建设 *xiandai hua jianshe* oeuvre de modernisation, 中国 特色 *zhongguo tese* à la chinoise, 有 中国 特色 *you zhongguo tese* spécifique à la Chine, etc.).

4 Explorer les dimensions chronologiques du corpus

Les typologies présentées dans les sections précédentes permettent de vérifier que le corpus **CongPCC** relève, au plan quantitatif, de la description proposée par le modèle des séries textuelles chronologiques et que l'évolution du vocabulaire constitue bien la principale caractéristique quantitative du corpus. Dans ce qui suit, nous présenterons succinctement quelques méthodes de textométrie chronologique qui permettent de préciser la description d'un corpus chronologique et d'explorer ses dimensions spécifiques.

4.1 Mesurer l'évolution lexicale

A côté des représentations factorielles qui fournissent des descriptions synthétiques calculées à partir de l'ensemble de données la représentation de *chroniques spécifiques* permet de visualiser l'évolution, sur l'ensemble du corpus, de groupes plus restreints constitués de formes caractéristiques (*spécifiques*) pour un sous ensemble particulier du corpus.

==== Repères méthodologiques : les chroniques spécifiques ====

- Pour une période donnée d'un corpus (ici la période 1969-1977), il est possible de dégager un ensemble de formes particulièrement sur-représentées: **les spécificités majeures de la période**.
- Une **chronique spécifique** est constituée par la sélection, sur la base d'un seuil de spécificité, des formes de spécificités maximales (ici : indice de spécificité ≥ 20).
- L'analyse de la ventilation des unités qui constituent une chronique dans l'ensemble du corpus permet de mettre en évidence : l'apparition, la disparition, l'accroissement brusque, la mise en sommeil brutale ou progressive, etc., d'ensembles spécifique d'unités textuelle, de repérer des moments de rupture et de renvoyer le chercheur vers des sites textuels caractéristiques à partir d'observations réalisées à une échelle plus importante.

Pour en savoir plus :

Salem, A. : Les chroniques spécifiques, *Lexicometrica*, décembre 2008.

Pour réaliser le diagramme présenté à la figure 3, nous avons préalablement constitué deux ensemble de formes.

- Le premier que nous appelons C+21-28, en rouge sur le graphique, rassemble les formes dotées d'un indice de spécificité supérieur à 20 dans le premier groupe de congrès (1921-1928).
- Le second que nous appelons C+82-07, en bleu sur le graphique, rassemble les formes dotées d'un indice de spécificité supérieur à 20 dans le dernier groupe de congrès (1982-2007)

On ne s'étonnera pas que les deux chroniques ainsi constituées prennent des valeurs importantes sur les congrès à partir desquels elles ont été constituées (resp. : le premier et le dernier groupe de périodes). L'intérêt de cette représentation vient du fait qu'elle permet de construire des hypothèses nouvelles qui concernent le rythme et les moments forts de la transformation du vocabulaire que l'on observe pratiquement dans toute série textuelle chronologique.

On observe sur ce graphique que le changement du vocabulaire concerné n'est pas tout à fait progressif. Bien que la tendance à l'apparition et à la disparition de chacun des deux groupes de formes soit vérifiée à une grande échelle, des reculs succèdent assez systématiquement aux avancées, particulièrement dans les premières périodes du corpus. A partir des années 70, l'évolution devient plus régulière tant en ce qui concerne les formes dont l'emploi tend à se raréfier qu'en ce qui concerne celles qui deviennent dominantes.

La représentation simultanée des deux chroniques suggère de vérifier l'hypothèse que le congrès de 1969 a été l'occasion d'un recul dans la dynamique de transformation du vocabulaire. Une partie du vocabulaire abandonné lors du congrès précédent par rapport aux débuts de la période considérée a été mobilisée à nouveau. Inversement une partie du vocabulaire dont l'utilisation avait notablement augmenté en 1956 et qui sera mobilisée de manière encore plus importante dans la dernière partie du corpus a été mise en sommeil au cours de ce congrès de 1969.

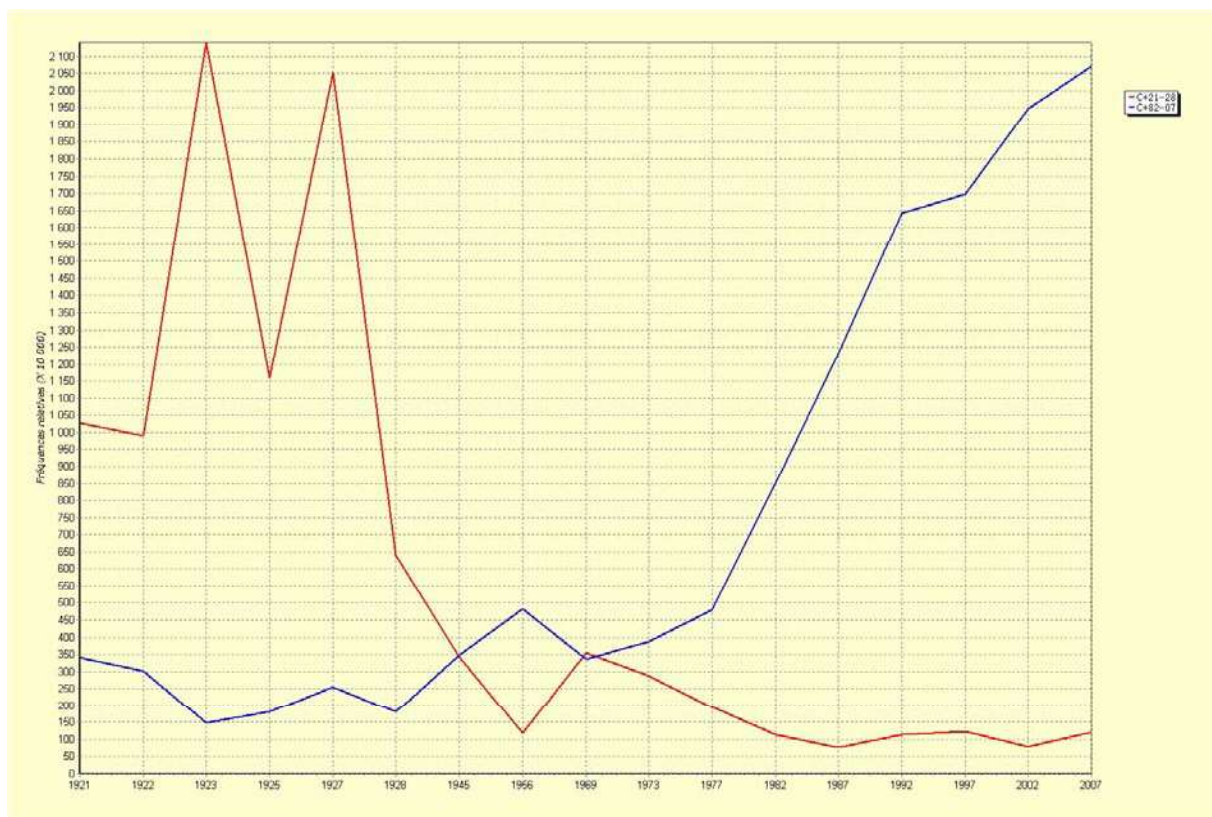


Figure 3
Chronique des spécificités positives majeures
pour les deux groupes de périodes (1921-1928 et 1982-2007)

4.2 Application à l'étude des résonances locales

Comme dans les sections qui précèdent, la chronique présentée sur la figure 4 a été réalisée en sélectionnant les formes dont l'indice de spécificité positive est supérieur à 20 dans le groupe des périodes 1969-1977. On peut voir la liste de ces formes au tableau 5. L'examen de la ventilation de ces formes, considérées, le temps d'une expérience, comme une même entité, nous amène à conclure que :

1. le groupe des formes qui constitue la chronique connaît deux périodes d'utilisation maximale en 1969 et en 1977, interrompues par une période (1973) pendant laquelle cet ensemble de formes est un peu moins utilisé ;
2. cette chronique reprend en partie un vocabulaire mis en place lors du congrès de 1927 ;
3. les termes qui constituent cette chronique tomberont en désuétude lors des congrès de la période 1982-2007.



Figure 4
Chronique des formes les plus spécifiques pour la période 1969-1977
(indice de spécificité ≥ 20 dans le groupe 69-77)

En s'appuyant sur la représentation de cette chronique sur la carte des sections, représentée sur la figure 5, on peut localiser plusieurs paragraphes qui nous permettront d'illustrer des similitudes entre les congrès de 1927, 1969 et 1977 dans l'utilisation des formes lexicales. Nous avons rassemblé, au tableau 7, trois paragraphes particulièrement représentatifs, prélevés dans les textes des congrès pour lesquels la chronique prend des valeurs fortes et localisés à l'aide d'un calcul de spécificités appliqué à chacun des paragraphes du texte.

Tableau 7
Trois sections spécifiques pour la chronique **C20-69-77**

Section 161 (congrès de 1927)

上海工人的经济罢工, 加上**了革命**的政治色彩之后, 那些和帝国主义有密切关系的大**资产阶级**, 立刻起来消灭这个运动. 结果, **他们破坏了革命**的战线, 帮助了帝国主义. 关税会议和法权会议, 并非帝国主义者为报答**资产阶级**背叛**革命**的功劳而开的, 乃是**无产阶级**和民权势力用**革命**的行动在帝国主义者手中所强夺来的. #
所领导的**资产阶级**民族主义的政纲, 是想用军事的胜利统一中国, 然后与帝国主义谋妥协, 使中国大多数民众仍被剥削

Après la grève économique des ouvriers de Shanghai et en raison de la couleur politique de la **révolution**, la **classe de la grande bourgeoisie** liées étroitement à l'impérialisme ont immédiatement anéanti le mouvement. Finalement, **ils** ont détruit la ligne de combat **révolutionnaire** et ont aidé l'impérialisme. Les congrès des douanes et des droits légitimes se sont tenus non pas parce que les impérialistes ont récompensé les **classes bourgeoises** qui avaient trahi la **révolution**, mais parce que le **prolétariat** et les forces droit-de-l'hommes les ont pris des mains des impérialistes au moyen du mouvement **révolutionnaire**.

Le programme politique du nationalisme de la **classe bourgeoise** dirigeante est d'unifier la Chine par la victoire militaire puis de comploter avec les impérialistes pour continuer à exploiter la majorité du peuple chinois.

Section 752 (congrès de 1969)

我们这次代表大会, 是在**毛主席**亲自发动和领导的**无产阶级文化大革命**取得了**伟大胜利**的时刻召开的. 这个**伟大的革命**风暴, 摧毁了以**叛徒 & 内奸 & 工贼刘少奇**为首的**资产阶级**司令部, 揭露了以**刘少奇**为总代表的党内一小撮**叛徒 & 特务 & 死不悔改的走资本主义道路的当权派**, 粉碎了**他们复辟资本主义的阴谋**, 大大地加强了我国的**无产阶级专政**, 大大地加强了我们的党, 从政治上 & 思想上 & 组织上为这次代表大会准备了充分的条件.

Ce congrès se tient au moment de la victoire **remarquable** de la **révolution culturelle prolétarienne** lancé et gouverné par le **président Mao**. Cette **grande tempête révolutionnaire** a détruit le commandant d'une garnison de la **classe bourgeoise**, **Liu shaoqi**, le dirigeant des **traîtres**, des espions, des voleurs.

Cette révolution a dénoncé le parti au pouvoir qui se dirige dans la voie du capitalisme dans lequel Liu Shaoqi est le représentant des traîtres, des agents secrets sans se repentir.

Elle a fait **échoué** des **complots** visant à restaurer le capitalisme, a fortement consolidé la **dictature du prolétariat** dans notre pays et a raffermi notre parti. Elle a permis, pour ce congrès, les conditions adéquates en matière politique, dans les esprits et du point de vue de l'organisation.

Section 928 (congrès de 1977)

我们党同王张江姚"四人帮"反党集团的**斗争**, 是我党历史上第十一次重大的路线斗争. 这次路线斗争, 是**无产阶级**同**资产阶级**的一次生死大搏斗, 关系我们党 & 我们国家的前途和命运. 这次路线斗争的**伟大胜利**, 应当归功于**伟大领袖毛主席**, 归功于**伟大的毛泽东思想**和**毛主席的革命路线**, 归功于我们**伟大的党 & 伟大的军队和伟大的人民**.

La **lutte** contre la « **Bande des quatre** » des Wang, Zhang, Jiang et Yao, opposés au parti, est la 11^{ème} grande lutte de phase dans l'histoire de notre parti. Cette lutte de phase est une lutte à mort de la **classe prolétaire** contre la **classe bourgeoise**. Il s'agit de notre parti, de notre avenir et du destin de notre pays.

La **grande** victoire de ce combat doit être attribué à notre **grand président Mao**, à sa pensée **remarquable** et à sa ligne **révolutionnaire**, et aussi à notre parti, à notre **grandiose** armée et à notre **grand** peuple.

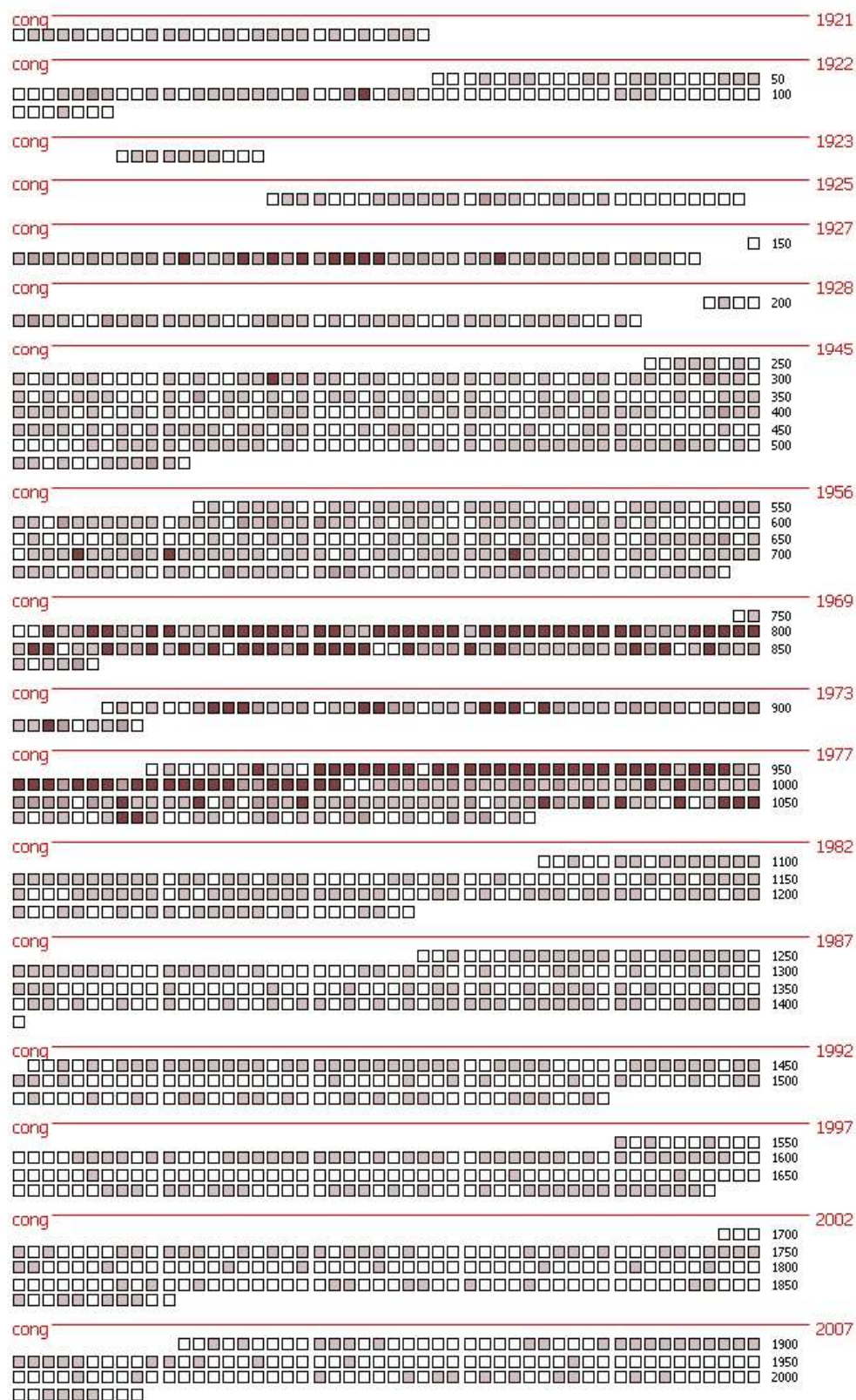


Figure 5 :
Chronique des spécificités positives majeures de la période 1969-1977
(indice ≥ 20) représentée sur une carte des paragraphes du texte.

4.3 Répétitions segmentales

De nombreux travaux textométriques réalisés dans le domaine de l'étude des textes politiques ont montré, au delà des études qui prennent pour objet l'évolution des formes isolées du vocabulaire, l'intérêt particulier qu'il y a à étudier les répétitions segmentales²⁵.

Tableau 8

Quelques segments répétés longs et fréquents

	Segments répétés	Équivalent français	Fréq.
1	在 过去 几 年 中	durant les dernières années	10
2	党 同 人 民 群 众 的	du parti et des masses populaires	10
3	建 设 有 中 国 特 色 的	construire ... à la chinoise	13
4	建 设 有 中 国 特 色 社 会 主 义	construire le socialisme à la chinoise	43
5	有 中 国 特 色 的 社 会 主 义	le socialisme à la chinoise	19
6	有 中 国 特 色 社 会 主 义 的	... du socialisme à la chinoise	34
7	以 经 济 建 设 为 中 心	centrer l'activité sur l'édification économique	12
8	第 二 个 五 年	le deuxième cinq ans	10
9	毛 主 席 的 无 产 阶 级 革 命 路 线	phase de la révolution prolétarienne du président Mao	10
10	改 革 开 放 和 现 代 化 建 设 的	de la réforme, de l'ouverture et de la mise en oeuvre de la modernisation	15
11	侵 略 者 和 建 设 新 中 国	les envahisseurs et édifier la chine nouvelle	10
12	建 设 有 中 国 特 色 的 社 会 主 义	construire le socialisme à la chinoise	11
13	建 设 有 中 国 特 色 社 会 主 义 的	construire ... du socialisme à la chinoise	24
14	建 设 有 中 国 特 色 社 会 主 义 理 论	construire la théorie du socialisme à la chinoise	11
15	走 资 本 主 义 道 路 的 当 权 派	le parti au pouvoir qui se dirige dans la voie du capitalisme	13
16	建 设 有 中 国 特 色 社 会 主 义 的 理 论	construire la théorie du socialisme à la chinoise	10

Le tableau 8 montre quelques-unes des séquences les plus longues et les plus fréquemment répétées dans le corpus *CongPCC*. La figure 6 montre que, loin d'être uniformément étalée dans le temps, la répétition de ces séquences se fait au contraire dans des périodes du corpus qui ne sont pas trop éloignées dans le temps.

Ainsi, par exemple, le segment :

建设 有 中国 特色 社会主义 的 (*construire ... du socialisme à la chinoise*)

qui est de longueur 5 et de fréquence 10, trouve toutes ses occurrences dans les congrès postérieurs à 1992, alors que le segment :

毛 主 席 的 无 产 阶 级 革 命 路 线 (*phase de la révolution prolétarienne du président Mao*)

n'apparaît que dans les congrès qui se sont tenus entre 1969 et 1977. On voit sur la figure 6 une ventilation des ces segments exprimée en spécificités à l'intérieur de l'ensemble du corpus.

²⁵ Sur les utilisations du calcul des segments répétés, on consultera, par exemple [Lebart et Salem, 1994].

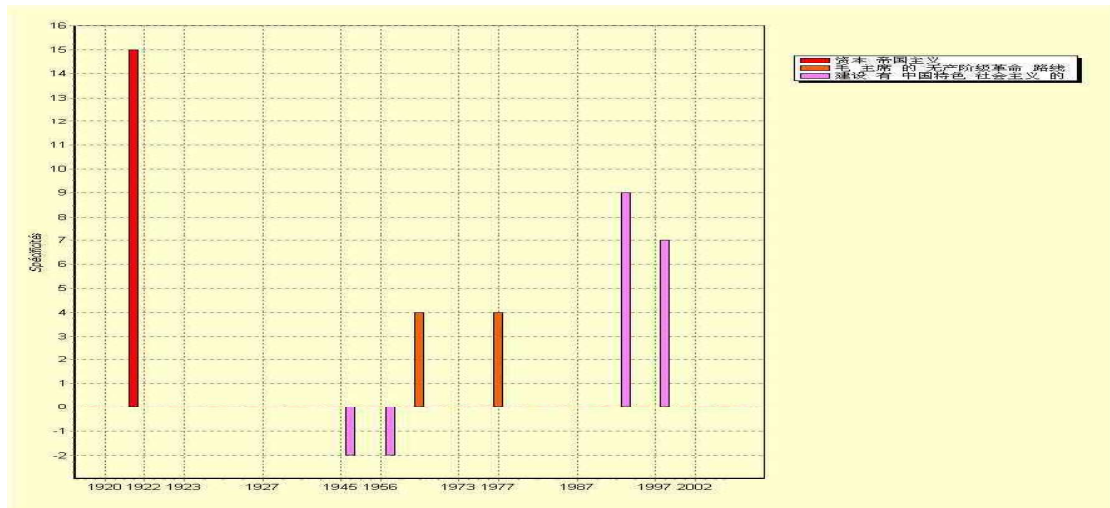


Figure 6

Ventilation au sein du corpus de quelques segments répétés de haute fréquence

- a) capitalisme ... impérialisme 资本主义 *ziben diguozhuyi*
- b) phase de la révolution prolétarienne du président Mao
毛主席的无产阶级革命路线 *mao zhuxi de wuchan jie jigeming luxian*
- c) construire ... du socialisme à la chinoise
建设有中国特色社会主义的 *jianshe you zhongguotese shehuizhuyi de*

==== Repères méthodologiques ====

Segmentation du texte chinois et calcul de segments répétés

- Le système d'écriture chinois constitue le texte sous forme de longues séquences de caractères (*hanzis*) qui ne sont pas interrompues par des espaces.
- Plusieurs logiciels permettent de segmenter les textes chinois en *mots* (à partir de critères qui ne sont pas toujours identiques). Une telle segmentation permet ensuite d'étudier la répartition des mots ainsi découpés dans le corpus de textes.
- Le calcul des *segments répétés* permet de repérer les suites d'unités textuelles qui trouvent plusieurs occurrences dans le texte.
- Ce calcul ne dépend pas totalement de la nature des unités de segmentation. Appliqué aux mots découpés par un segmenteur, il repèrera des séquences de mots qui auraient été repérées par le même algorithme appliqué aux suites de caractères isolés.
- Si le segment composé de deux mots : ABC DEF trouve x occurrences dans un corpus découpé en mots, le segment A B C D E F composé de 6 caractères isolés trouvera, **au moins**, le même nombre d'occurrences dans le corpus segmenté en caractères isolés.

4.4 Mesures comparatives de la répétition segmentale

La figure 7 permet d'établir une synthèse à propos de la ventilation de l'ensemble des segments composés d'au moins quatre formes, dont la fréquence est supérieure à dix dans le corpus. On voit sur cette figure que les périodes 1992 et 1997 contiennent un nombre relativement élevé de répétitions de ce type. Cette constatation nous incitera à vérifier l'hypothèse d'une répétition globalement plus importante des segments dans ces deux périodes du corpus.

Tableau 9

Spécificités positives majeures de la période 1992 dans le corpus *CongPCC*

	Forme	Équivalent français	FTot.	FPart	Coeff.
1	改革开放	réforme et ouverture	157	56	25
2	有 中国特色	à la chinoise	69	34	21
3	改革开放 和	réforme et ouverture et	54	29	20
4	建设	construction	857	135	18
5	现代化	modernisation	228	56	17
6	和 现代化 建设	et oeuvre de modernisation	40	22	16
7	改革开放 和 现代化	réforme et ouverture et modernisation	38	22	16
8	改革开放 和 现代化 建设	réforme, ouverture et oeuvre de modernisation	36	21	16
9	现代化 建设	oeuvre de modernisation	127	39	16
10	中国特色	à la chinoise	98	34	16
11	社会主义	socialisme	886	130	15
12	十四 年	quatorze ans	14	13	15
13	和 现代化	et modernisation	46	23	15
14	建设 有 中国特色	édification...à la chinoise	45	23	15
15	有 中国特色 社会主义	socialisme à la chinoise	57	25	15

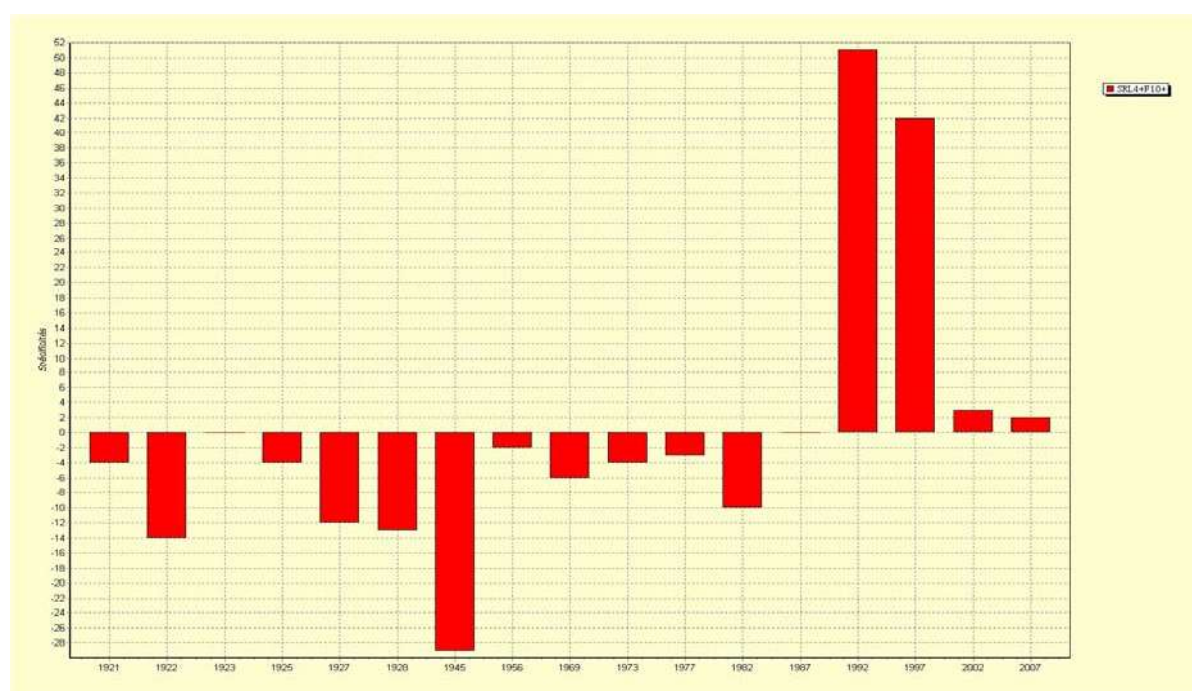


Figure 7

Ventilation dans les parties du corpus des segments répétés
de longueur ≥ 4 et de fréquence ≥ 10

Le tableau 9 présente un certain nombre de segments particulièrement répétés dans le congrès de 1992. Comme on peut le vérifier sur ce tableau, ces segments fortement répétés concernent avant tout : *les réformes, l'ouverture, la modernisation et l'édification du socialisme à la chinoise*. Le retour au texte permet de localiser des sections qui opèrent une utilisation privilégiée de ce type de vocabulaire. On trouve au tableau 10 deux sections caractéristiques

de ce dernier point de vue. On voit que les termes relatifs à l'*édification du socialisme à la chinoise* et aux *réformes économiques* sont, ici aussi, dominants.

Tableau 10

Deux sections de la période 1992 utilisant un grand nombre de segments répétés

Section 1515 (congrès de 1992) 党的基本路线 要 毫不动摇地 长期坚持下去, 社会主义的 改革开放 和 现代化建设 要 搞得更好更快, 国家要 长治久安 和 繁荣富强, 关键在于 我们党, 在于 坚持用 邓小平同志 建设有中国特色社会主义 的理论武装全党. La ligne fondamentale du parti devra persister à long terme sans se laisser ébranler. La réforme et l'ouverture du socialisme, ainsi que l'établissement de la modernisation devront se faire encore mieux et plus rapidement. Pour obtenir prospérité et stabilité à long terme, et pour que le pays s'épanouisse, la clé réside dans le parti et dans le renforcement de tout le parti grâce à la théorie de l'édification du socialisme à la chinoise de notre camarade Deng Xiaoping.
Section 1517 (congrès de 1992) 第一, 认真学习 建设有中国特色社会主义 的理论, 增强贯彻执行党的基本路线的 自觉性和坚定性. Premièrement, nous étudions consciencieusement la théorie de la construction du socialisme à la chinoise et appliquons, avec ardeur et animés d'une conscience et d'une détermination accrues, la ligne fondamentale du parti.

5 Faire varier les échelles de l'analyse : un corpus restreint (1982-2007)

Comme nous l'avons signalé plus haut, la question de l'homogénéité des textes qui composent une série textuelle chronologique doit être examinée sous plusieurs aspects. Une fois posée le principe d'homogénéité *constitutive* qui permet de rassembler des textes à partir de critères qui en font *a priori* un corpus homogène, il convient de vérifier que les textes réunis ont réellement été produits dans des conditions de productions comparables qui permettront d'observer, sans buter sur de nombreux artefacts, l'évolution du vocabulaire au cours de la période couverte par le corpus.

De ce point de vue, l'analyse d'un sous ensemble de congrès extraits du corpus précédent et limité aux seuls congrès tenus entre 1982 et 2007 va nous permettre de constituer une série chronologique beaucoup plus homogène que la série que nous venons d'étudier. Comme nous l'avons vu dans les analyses présentées aux sections précédentes, les congrès survenus à partir de 1982 ont donné lieu à des productions beaucoup plus homogènes, tant du point de vue de l'intervalle de temps écoulé entre chacun des congrès que de celui des caractéristiques quantitatives que l'on peut calculer à partir de chacun des textes.

Nous appellerons **Cong82-07** la série qui rassemble les 6 derniers congrès survenus entre 1982 et 2007. Dans cette nouvelle série textuelle chronologique, les congrès ont été tenus à des intervalles de temps comparables, ils présentent une homogénéité plus grande du point de vue de la longueur des textes produits. Les analyses sur l'ensemble **CongPCC** ont montré d'autre part, leur proximité relative au plan lexical ainsi qu'une évolution lexicale repérable au fil du temps. L'analyse de ce corpus restreint va nous permettre de préciser les premières constatations faites sur le corpus.

5.1 Analyse globale de l'évolution (1982-2007)

L'AFC réalisée à partir du corpus **Cong82-07** montre le caractère plus progressif des changements chronologiques dans ce second corpus²⁶. Les six congrès sont classés dans

²⁶ Comme lors de l'expérience précédente, nous avons analysé un tableau croisant les 6 congrès retenus avec les formes dont la fréquence est au moins égale à dix occurrences dans le second corpus.

l'ordre chronologique sur le premier axe, de la droite vers la gauche, à des intervalles très réguliers.

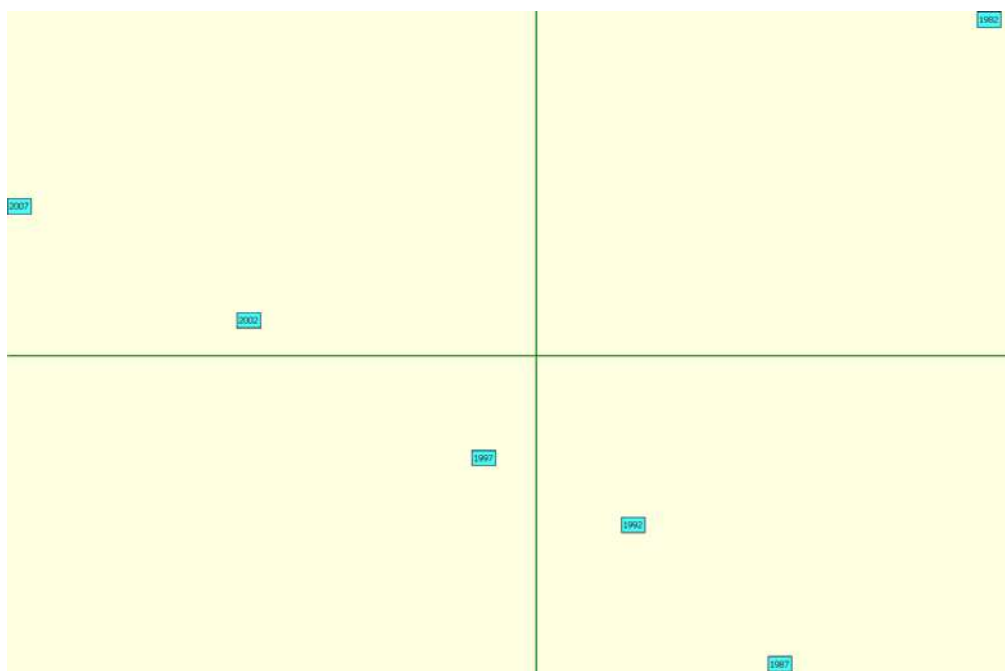
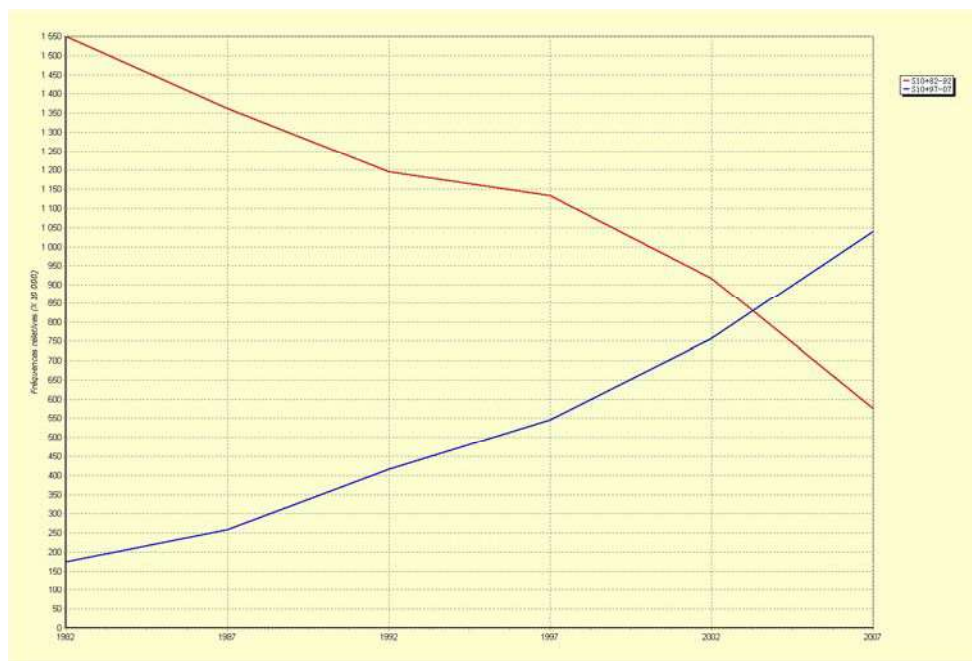


Figure 8

Analyse factorielle des correspondances de la série 1982-2007

L'analyse des chroniques effectuée sur ce second corpus montre, à l'instar de ce que nous avons obtenu pour l'ensemble du corpus *CongPCC*, une progression régulière de la chronique *S10+82-92*, constituée par la réunion des formes dont la spécificité dépasse l'indice de spécificité 10 dans les congrès tenus entre 1982 et 1992. De manière symétrique, la chronique *S10+97-07* est constituée à partir des formes les plus spécifiques dans le groupe des trois derniers congrès du corpus *Cong82-07*.

Le fait que les courbes prennent des valeurs plus élevées sur les parties qui ont servi à dégager les ensembles de formes spécifiques ne constitue pas une surprise en soi. La progression régulière de ces courbes au fil du temps couvert par le corpus apporte un éclairage interprétable sur l'évolution lexicale propre au corpus.

**Figure 9 :**

Évolution des chroniques spécifiques

S10+82-92 - formes de spécificités supérieure à 10 dans les congrès 1982, 1987 et 1992**S10+97-07** - formes de spécificités supérieure à 10 dans les congrès 1997, 2002 et 2007

5.2 Evaluation de l'apport spécifique de la période 2007

Comme nous l'avons vu à la section précédente, les chroniques spécifiques permettent d'esquisser une description de la variation lexicale survenue dans l'ensemble d'un corpus chronologique. Dans cette dernière section consacrée à l'analyse des chroniques spécifiques, nous nous concentrerons sur l'appréciation des transformations lexicales survenues par rapport aux congrès précédents, lors d'une période particulière. Pour illustrer cette démarche, nous avons choisi congrès de 2007, dernière période du corpus **Cong82-07**.

Comme plus haut, nous avons constitué une chronique spécifique **S10+07** en sélectionnant les formes dont l'indice de spécificité est supérieur à 10 dans la dernière partie du corpus. La ventilation de la chronique **S10+07** sur une carte des sections qui découpe le corpus en paragraphes permet de localiser les paragraphes qui contiennent une proportion particulièrement élevée²⁷ de formes appartenant à cette chronique. On repère également des paragraphes du congrès 2007 qui ne contiennent que peu ou pas du tout de formes appartenant à la chronique spécifique de la période 2007. On note que plusieurs des paragraphes appartenant à des congrès précédents contiennent également un grand nombre de ces formes. Ces données nous amèneront à étudier l'hypothèse que ces paragraphes constituent une première introduction de thèmes qui deviendront dominants pour le congrès de 2007.

La figure 11 présente, exprimées cette fois en terme de spécificités, les mêmes données relatives à l'apparition du vocabulaire spécifique pour la période 2007. Cette vue permet de confirmer le caractère extrêmement progressif de la mise en place du nouveau vocabulaire dont on trouve les principales formes spécifiques au tableau 12.

²⁷ Dans une métrique définie par le calcul des spécificités.

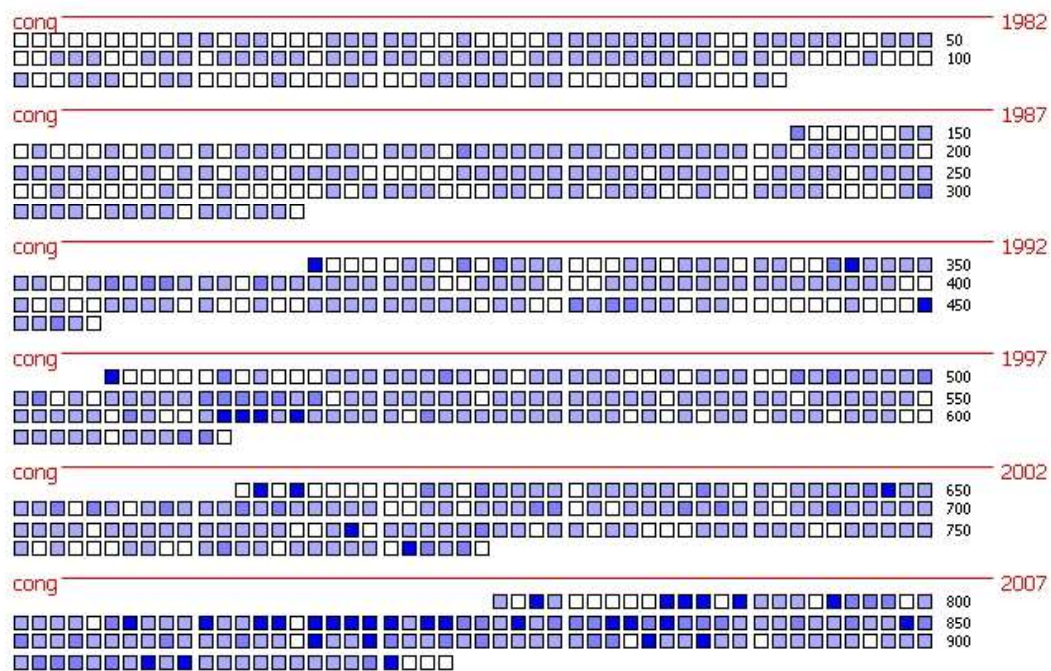


Figure 10 :
Corpus *Cong82-07* :
Les spécificités chronologiques positives majeures de la période 2007
sur une division en paragraphes

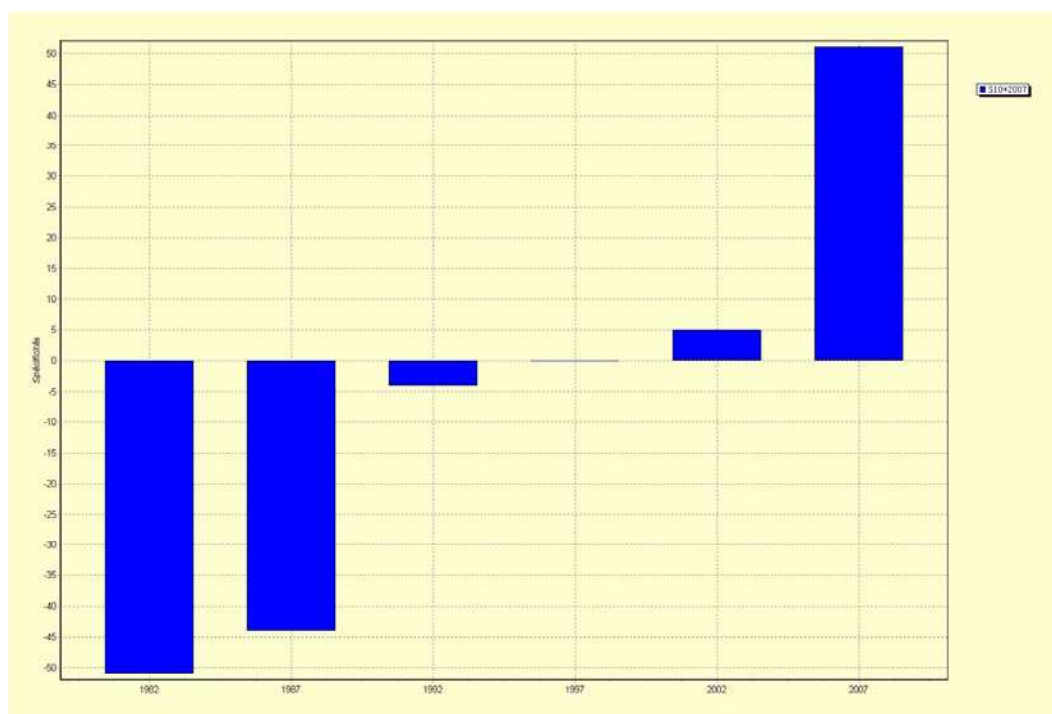


Figure 11 :
Corpus *Cong82-07* :
Les spécificités chronologiques positives majeures de la période 2007
ventilées sur l'ensemble de la période 1982-2007

Tableau 11 :
Corpus Cong82-07 :
 Quelques paragraphes spécifiques des congrès de 2002 et de 2007

Section 647 (annonce dans le congrès tenu en 2002) (六) 坚持稳定压倒一切的方针, 正确处理改革 发展 稳定的关系. 稳定是改革和 发展 的前提. 要把改革的力度 & 发展 的速度和 社会 可承受的程度统一起来, 把不断改善人民生活作为处理改革 发展 稳定关系的重要结合点, 在 社会 稳定中 推进 改革 发展, 通过改革 发展 促进 社会 稳定. # 6) Maintenir le principe selon lequel la stabilité sociale prime tout et traiter judicieusement les rapports entre la réforme, le développement et la stabilité sociale. Celle-ci étant la prémisse de la réforme et du développement, il importe de bien combiner les exigences en matière de vigueur de la réforme, de rythme de croissance et de capacité d'adaptation de la société à leurs effets, et de faire de l'amélioration continue de la vie du peuple le point de jonction incontournable de ces rapports et ce, de manière à promouvoir la réforme dans la stabilité sociale et à favoriser celle-ci par la réforme et le développement. #
Section 778 (congrès de 2007) 高举 中国特色 社会主义伟大旗帜为夺取 全面建设小康社会 新胜利而奋斗# lever bien haut l'étendard du socialisme à la chinoise et lutter pour réaliser de nouveaux succès dans le développement général d'une société de moyenne aisance. #
Section 806 (congrès de 2007) 改革开放是党在新的时代条件下带领人民进行的新的伟大革命, 目的就是要解放和 发展社会 生产力, 实现国家现代化, 让 中国 人民富裕起来, 振兴伟大的中华民族; 就是要推动我国社会主义制度自我完善和 发展, 赋予社会主义新的生机活力, 建设和 发展中国特色 社会主义; 就是要在引领当代 中国发展 进步中加强和改进党的建设, 保持和发展党的先进性, 确保党始终走在时代前列. # la réforme et l'ouverture s'inscrivent dans la grande révolution menée par le peuple sous la direction du parti dans le nouveau contexte historique. cette révolution a pour mission de libérer et de développer les forces productives sociales, de moderniser notre pays, de faire accéder la population chinoise à la prospérité et de redresser la grande nation chinoise. cette révolution a également pour but de pousser le régime socialiste de notre pays vers l'auto-perfection et le développement, de procurer de nouvelles vitalités au socialisme, ainsi que de construire et développer le socialisme à la chinoise. tout en menant le progrès et le développement de notre pays, elle vise aussi, pour que le parti reste toujours à la pointe de notre époque, à renforcer et améliorer l'édification du parti, ainsi qu'à conserver et développer sa nature progressiste. #
Section 920 (congrès de 2007) 让我们高举 中国特色 社会主义 伟大 旗帜, 更加 紧密 地 团结 在 党中央 周围, 万众一心, 开拓 奋进, 为夺取 全面建设小康社会 新胜利& 谱写人民美好生活新篇章而努力奋斗! # en brandissant bien haut l'étendard du socialisme à la chinoise, nous nous efforcerons, unis autour du comité central du parti, animés d'une seule et même volonté et dotés d'un esprit novateur, de réaliser de nouveaux succès dans le développement général d'une société de moyenne aisance et de composer un nouveau chapitre de la vie heureuse du peuple chinois ! #

Commentaire : Le thème général qui émerge de l'évolution du vocabulaire constatée en 2007 concerne, comme on peut le vérifier au tableau 11, le développement harmonieux de l'économie et la mise en place d'une société de moyenne aisance.

Tableau 12 :
Corpus *Cong82-07* :
Les spécificités chronologiques positives majeures de la période 2007

	Forme	Équivalent français	FTot.	FPart	Coeff.
1	体系	système	129	67	22
2	和谐	harmonie	40	33	21
3	创新	innovation	90	50	19
4	发展观	concept de développement	21	21	18
5	社会	société	550	164	18
6	发展	développement	1138	283	18
7	着力	effort inlassable	25	20	13
8	特色	caractéristique	185	65	12
9	公共	public	24	18	11
10	社会 和谐	harmonie sociale	12	12	11
11	推进	faire progresser	180	60	10
12	特色 社会主义	socialisme	138	51	10
13	理论 体系	système théorique	11	11	10
14	发展 中国 特色 社会主义	développement du socialisme à la chinoise	11	11	10
15	中国 特色 社会主义	socialisme à la chinoise	136	50	10
16	发展 中国	développer --- Chine	15	13	10
17	中国 特色	à la chinoise	177	60	10
18	保障	assurer	100	39	9
19	科学	science	198	62	9
20	完善	excellent	216	67	9
21	健全	perfectionner	97	38	9
22	文化	civilisation	270	74	8
23	自主 创新	innovation autonome	13	11	8
24	和平 发展	développement dans la paix	8	8	8
25	贯彻落实 科学	appliquer - appliquer - science	8	8	8
26	协调	coordination	56	26	8
27	能力	capacité	70	29	8
28	贯彻落实	appliquer - appliquer	9	9	8
29	落实	appliquer	33	18	8
30	小康社会	société de moyenne aisance	41	21	8
31	安全	sécurité	52	24	8
32	全党 全国 各族人民	tout le parti et l'ensemble de notre peuple multiethnique	8	8	8
33	中国	chine	508	123	8

On a rassemblé au tableau 13 les spécificités négatives majeures (i.e. formes dont l'emploi dans la dernière période est jugé particulièrement faible par le calcul des spécificités). Cette approche permet d'entreprendre l'étude de la partie du vocabulaire présent aux débuts du corpus considéré, à laquelle les rédacteurs du dernier texte feront moins appel que lors des congrès précédents. Ce vocabulaire mis à l'écart ou tombé en désuétude constitue une donnée

toute aussi importante pour analyser les évolutions de la politique d'une l'organisation politique telle que celle que nous étudions²⁸.

Tableau 13 :
Corpus *Cong82-07*
Les spécificités chronologiques négatives majeures de la période 2007

	Forme	Équivalent français	Frq. Tot.	Fréquence	Coeff.
1	对	pour	384	38	-4
2	我国	notre pays	283	25	-4
3	不	(négation)	414	43	-4
4	斗争	lutte	86	4	-4
5	纪律	discipline	53	1	-4
6	现在	maintenant	57	1	-4
7	两	deux	85	4	-4
8	于	(particule grammaticale)	109	6	-4
9	这个	ce	112	7	-4
10	的思想	pensé de	72	2	-4
11	民族的	du peuple	48	0	-4
12	在社会主义	dans le socialisme	42	0	-4
13	许多	multiple	59	1	-4
14	着	(particule grammaticale)	66	2	-4
15	严重	sévir	67	2	-4
16	他们	ils	57	1	-4
17	文明的	de la civilisation	36	0	-4
18	这样	ainsi	59	1	-4
19	情况	fait (nom)	58	1	-4
20	的领导	direction de	97	4	-4
21	建设有中国特色社会主义	édification du socialisme à la chinoise	44	0	-4
22	来	venir / pour	130	8	-4
23	的精神	élan de	39	0	-4
24	企业的	de l'entreprise	36	0	-4
25	这是	ceci est	79	3	-4
26	的经济	de l'économie	49	1	-4
27	一个	un	124	7	-4
28	条件	condition	131	8	-4
29	路线	phase	128	6	-5
30	也	aussi	152	9	-5

²⁸ Rappelons que des difficultés résiduelles de la segmentation automatique des textes chinois compliquent provisoirement cette dernière tâche. Des segmentations différentes d'une même séquence prise dans des contextes différents pouvant compliquer le repérage des occurrences d'une même séquence.

6 Conclusion

A partir d'une série textuelle chronologique de textes politiques chinois échelonnés sur près d'un siècle nous avons mis en évidence, à l'aide des méthodes de la textométrie chronologique, la variation progressive du vocabulaire employé dans ces textes.

Une première série d'outils textométriques (afc, spécificités, etc.) nous a permis de vérifier que la variation du vocabulaire au fil du temps constituait bien la caractéristique quantitative majeure du corpus. Ces mêmes méthodes nous ont permis de procéder à une typologie sommaire du corpus répartissant les congrès en quatre grandes périodes et de procéder à l'étude des spécificités lexicales de chacune de ces périodes.

L'*analyse des chroniques spécifiques* nous a permis de mettre en évidence des moments particuliers de l'évolution d'ensemble. La constitution d'ensembles de formes variant en liaison étroite avec la dimension chronologique du corpus nous a permis de préciser notre description et de mettre en lumière des sections particulières du corpus qui constituaient autant d'illustrations en contexte des phénomènes globaux que nous avons mis en évidence.

L'application des mêmes méthodes à un corpus réduit à des textes plus homogènes, les congrès survenus dans la toute dernière période (1982-2007), nous a permis de décrire avec une précision accrue les changements lexicaux survenus dans une période plus courte.

Les procédures de *textométrie chronologique* que nous avons employées pour décrire l'évolution de ce corpus de textes politiques chinois nous semblent constituer désormais un ensemble de méthodes mobilisables pour la description de tout corpus textuel chronologique.

7 Références

- Lamalle C., Salem A., "Types généralisés et topographie textuelle dans l'analyse quantitative des corpus textuels", *Actes des 6èmes Journées d'analyse des données textuelles*, St Malo, 2002
- Lebart L., Salem A., *Statistique textuelle*, Paris, Dunod, 1994, téléchargeable sur le site : <http://www.cavi.univ-paris3.fr/lexicometrica/livre/st94/st94-tdm.html>
- Miao J., Salem A., Comparaisons textométriques de traductions franco-chinoises, in *Explorations textométriques*, 2008.
- Salem A., "La lexicométrie chronologique", *Actes du colloque de lexicologie politique "Langages de la Révolution"*, collection "St. Cloud", Paris, Klincksieck, 1994.
- Salem A., "Les chroniques spécifiques", *Lexicometrica*, <http://www.lexicometrica.fr>, décembre 2008.