



Le texte se transforme... Analyse textométrique des rapports d'ouverture présentés aux congrès du Parti Communiste Chinois (1982-2017)

André Salem, Jun Miao

► To cite this version:

André Salem, Jun Miao. Le texte se transforme... Analyse textométrique des rapports d'ouverture présentés aux congrès du Parti Communiste Chinois (1982-2017). 2019. hal-02119927

HAL Id: hal-02119927

<https://hal.science/hal-02119927>

Preprint submitted on 4 May 2019

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Le texte se transforme ...

Analyse textométrique des rapports d'ouverture présentés aux
congrès du Parti Communiste Chinois (1982-2017)

André Salem¹, Jun Miao²

¹ Université de la Sorbonne nouvelle - Paris 3, France – salem@msh-paris.fr

² Université Lumière de Lyon 2, France – miaojun@miaojun.net

Abstract: Speeches produced by political parties are often considered as "repetitive and non-informative". Yet, over time such speeches can yield extremely valuable information as we show here by applying statistical methods to Chronological text series. When examined, a compilation of reports presented during the eight congresses of the Chinese Communist Party (CCP), held between 1982 and 2017 (§1), reveals variations of vocabulary over time in the form of Textual highlights which point out key moments in lexical evolution (§2). Textometric analysis provides a critical perspective towards the tokenization of texts into words as operated by automatic tokenizers of Chinese. More efficient methods for the observation of lexical variations, linked to the political changes in progress, are then presented (§3). Finally, a constant evolution of the enunciation style during the observed period, is highlighted (§4). Such evolution is particularly relevant to frequent word variations, which are often discarded by discourse researchers, under pretext that they convey less information. The methodology described in this study is applicable to a very wide range of textual corpora.

Keywords: Textual series, Chinese political speeches,

Résumé : Les discours produits par les partis politiques sont souvent considérés comme "répétitifs et dépourvus de caractère informatif". Nous montrons, sur l'exemple d'une *série textuelle chronologique*, que ces corpus peuvent au contraire contenir des renseignements extrêmement précieux sur l'évolution des textes dans le temps. Notre étude concerne une série de rapports présentés durant les huit congrès du Parti Communiste Chinois (PCC), survenus entre 1982 et 2017 (§1). Les méthodes de la *veille textuelle textométrique* mettent en évidence des variations du vocabulaire au fil du temps et des moments-clés dans cette évolution (§2). Le dépouillement textométrique permet de porter un regard critique sur la segmentation des textes en *mots*, opérée par les segmenteurs automatiques du chinois, et de présenter des méthodes plus efficaces pour l'observation des variations lexicales, liées aux changements politiques en cours (§3). Durant la période observée, le style d'énonciation subit une évolution profonde et continue. Cette dernière transformation concerne particulièrement les variations de la fréquence de "mots-outils", souvent négligés dans les recherches quantitatives sur le discours sous le prétexte qu'ils véhiculeraient moins d'information (§4). L'ensemble des méthodes décrites dans la présente étude est applicable à une très large gamme de corpus textuels.

Mots-clés: Séries textuelles chronologiques, discours politique chinois.

摘要: 政党讲稿通常被认为“重复多、信息弱”，然而，我们发现此类语料库恰恰包含了文本在历时中演变的重要信息。本研究涉及 1982 年至 2017 年间，中国共产党代表大会开幕上的八次报告 (§1)。通过词量学的新词观测法，可以观测到语料库中词汇的历时变化，并自动显示出“关键时期” (§2)。借助词量学的文本分析法，不仅对中文自动切分有较为客观的评估，而且可以对与政治事件相关联的词汇变化进行有效的考察 (§3)。虚词因所含信息少，通常被其他研究者在进行量化研究时所忽略，而我们注意到，讲稿文本的话语风格在这一时期经历了深刻的变化。这些变化事实上与“虚词”的使用有着密不可分的关系 (§4)。本研究所使用的多种方法均适用于其它各类型的语料库文本研究。

关键词: 历史性文本，中国政治讲稿。

Les dernières décennies ont vu se multiplier les analyses quantitatives portant sur de vastes ensembles de textes produits par des institutions diverses (partis politiques, organismes nationaux ou internationaux, etc.), durant des périodes plus ou moins

étendues. Ces analyses s'appuient sur des dépouillements automatisés de très gros corpus et mobilisent parfois un appareil d'analyse statistique plus ou moins élaboré. La première raison de cette évolution tient au fait que les textes sont devenus aisément accessibles, souvent mis à la disposition des internautes par les institutions qui les ont produits¹. Une autre raison tient au fait que les communautés de chercheurs qui analysent les discours socio-politiques, rompant de plus en plus avec une tradition qui considérait ces textes comme "répétitifs et dépourvus de caractère informatif", entreprennent désormais de les étudier de près et de s'intéresser aux variations du vocabulaire qu'ils présentent au fil du temps.

De plus en plus nombreux, des chercheurs se penchent sur l'analyse du discours politique institutionnel rédigés dans différentes langues : en français (Deroubaix 1995, Labbé et Monière 2003, Mayaffre 2004, Marchand 2007), en chinois (Salem et Wu 2008, Zhou et Li 2015, Tang et Tao 2016, Qian 2017) et dans d'autres langues (De Souza 2014). Ainsi, les textes produits par le Parti au pouvoir en Chine constituent-ils un matériau d'étude incontournable pour la compréhension des changements actuellement à l'œuvre dans ce pays. Cet intérêt tranche avec la période précédente au cours de laquelle ce type de document était considéré comme un simple tissu de propagande à partir duquel il était inutile de chercher à extraire du sens².

C'est dans cet esprit que nous avons entrepris l'analyse d'un corpus de textes rassemblant les principaux documents produits lors des derniers congrès du Parti communiste chinois (désormais PCC). Le PCC (中国共产党; zhōngguó gòngchǎndǎng) a été fondé en 1921. Depuis la révolution survenue en 1949, c'est le parti dirigeant de la République Populaire de Chine (RPC). Le Congrès national constitue la plus haute instance de cette organisation. C'est à l'occasion de ces congrès que sont annoncées les décisions importantes et que sont fixés les objectifs pour les années à venir. Depuis quelques décennies, le Congrès national du PCC a lieu régulièrement une fois tous les cinq ans. Les textes publiés lors des congrès revêtent désormais une forme relativement stable qui les rend tout à fait comparables, tant du point de vue de leur longueur que par la manière dont ils sont structurés³.

Notre premier souci était de rassembler un corpus de textes relativement *homogène*, tant du point de vue des sujets abordés que de celui du genre et de la volumétrie. Le corpus **PCC1982-2017** nous a semblé parfaitement adapté à cet ensemble de préoccupations. Les analyses textométriques ont montré que le corpus présentait toutes

¹ Les analyses dont nous rendons compte ci-dessous, ont été effectuées à l'aide du logiciel *Lexico5*. Cedric Lamalle, William Martinez, Serge Fleury ont largement contribué au développement des fonctionnalités de ce logiciel. Les auteurs tiennent à les en remercier vivement.

² Ainsi, dans un billet publié sur le site du *Centre d'études prospectives et d'informations internationales* (CEPII), placé auprès du Premier ministre français, l'économiste Michel Aglietta insiste sur la nécessité de lire attentivement les documents publiés lors d'un plenum du PCC, afin de prendre conscience des éléments nouveaux qu'il contient. *L'enjeu politique du troisième plenum du 3 décembre 2013*, par Michel Aglietta : Le Parti Communiste Chinois a produit un texte d'une grande portée, dont la mise en œuvre va transformer profondément la société chinoise. /.../ <http://www.cepii.fr/blog/bi/post.asp?IDcommuniqu=263>, consulté le 10 juillet 2018

³ La périodicité des congrès n'a pas toujours été la même, dans l'histoire du PCC. Le congrès n'a pas été convoqué pendant les périodes de guerre, ni durant les périodes de troubles politiques.

les caractéristiques d'une *série textuelle chronologique* dans laquelle l'évolution du vocabulaire autour du temps constitue le principal axe de variation du corpus.

Laissant aux spécialistes qui disposent de connaissances plus approfondies sur la vie politique chinoise le soin de tirer profit des méthodes et des résultats que nous présentons, nous ne tenterons pas d'évaluer, dans ce qui suit, la pertinence ou encore l'adéquation à la réalité des options affichées dans les textes qui font l'objet de notre étude. Celle-ci aura pour objet principal l'analyse des variations du vocabulaire employé dans les rapports de congrès. Le corpus que nous allons étudier est constitué de l'ensemble des rapports produits lors des huit congrès du PCC qui se sont tenus entre 1982 et 2017. Pour des raisons que nous avons analysées par ailleurs (Miao & Salem 2018), les textes produits durant cette dernière période présentent une grande homogénéité, tant du point de vue de leur taille que de celui des thèmes qu'ils abordent, de leur structure et de leur style de rédaction⁴. Nous commencerons par étudier la série chronologique *PCC1982-2017* divisée en *congrès* afin de mettre en évidence les principales variations historiques dans l'emploi du vocabulaire.

1 Questions de méthodes

La révolution informatique des années 70 a complètement bouleversé les possibilités d'exploration et d'analyse des textes. À l'analyse des corpus basée sur la lecture cursive par des spécialistes du domaine, succède une période dans laquelle les possibilités de traitements automatisés permettent aux chercheurs un accès plus efficace aux corpus de textes pris dans leur ensemble. Certains de ces traitements constituent de simples automatisations de méthodes d'analyse textuelle, utilisées auparavant sans l'aide de l'ordinateur. D'autres font intervenir des méthodes d'analyses plus ou moins formalisées et plus ou moins explicites.

1.1 Méthodes et illusions

Commentant le rapport présenté au 19^e congrès du PCC (octobre 2017), plusieurs commentateurs ont signalé que l'expression 新时代 (xīn shídài, *nouvelle ère*) était souvent employée dans ce texte, mentionnant parfois sa fréquence dans ce tout dernier congrès (36 occurrences). Ce décompte particulier ne constitue cependant pas une analyse, à lui seul. La répétition constatée entraîne une série d'autres questions : *Peut-on produire la liste de tous les termes dont la fréquence s'accroît de la même manière lors du dernier congrès ? Y a-t-il, à l'inverse, des termes dont la fréquence diminue fortement lors des derniers congrès ? Cette expression a-t-elle déjà été utilisée*

⁴ L'article de Salem et Wu (2008) constitue une étude chronologique portant sur l'intégralité des congrès du PCC survenus depuis sa fondation 1921 jusqu'à l'année 2012. Au-delà des évolutions chronologiques qu'elle avait permis de mettre en évidence, cette étude avait également montré le caractère très hétérogène de la forme *congrès* considérée sur une échelle de temps aussi large.

lors des congrès précédents ? ⁵

La lecture des travaux produits par les chercheurs qui utilisent des méthodes informatisées pour analyser les textes laisse parfois transparaître des postulats qui mériteraient d'être soumis à discussion. Malheureusement très répandue de nos jours, la technique des *nuages de points*, consiste à représenter les mots du corpus avec des polices, des positions et des couleurs variables sur un graphique plan, selon des règles dont l'explicitation n'est souvent que très partielle. Le postulat implicite est qu'un corpus de discours peut être représenté par une simple figure qui permettra au lecteur de reconstruire des liens sémantiques inhérents au texte⁶. On ne peut être que très sceptique sur l'efficacité de telles procédures qui ne servent, au mieux, que de support à l'expression de commentaires qui ne découlent que très rarement des données textuelles soumises à l'analyse.

Un autre postulat que l'on retrouve couramment dans des études sur le même sujet⁷ consiste à poser que les *mots* contenus dans un texte peuvent être divisés en deux catégories : une catégorie de *mots pleins*, pourvus d'un contenu sémantique et une catégorie complémentaire de *mots-outils* qui ne seraient pas porteurs de sens⁸. A partir de cette croyance, de nombreux chercheurs se désintéressent systématiquement d'une liste de formes qu'ils rangent dans la seconde catégorie, avant toute expérience, pour se concentrer sur la description de l'utilisation des formes qu'ils jugent seules capables de véhiculer du sens⁹. Nous montrerons plus loin que cette croyance est profondément erronée. Nous montrerons également (§3) que le *mot* n'est pas toujours la seule unité pertinente pour l'étude des textes. La prise en compte d'unités d'analyse textuelle plus complexes comme les *segments répétés*, les *cooccurrences*, etc. permet de faire des observations qui auraient été impossibles autrement.

1.2 La méthodologie textométrique

Les *méthodes de textométrie chronologique* que nous présentons ici ont été spécifiquement élaborées pour traiter les séries textuelles chronologiques (Salem 1991, Miao et Salem 2018). Le but de ces méthodes est de mettre en évidence les

⁵ La réponse à la dernière question est positive, l'expression a également été employée (une seule fois), lors du congrès précédent.

⁶ Les méthodes mises en œuvre par ces logiciels conduisent à donner à chaque mot une taille qui correspond plus ou moins à sa fréquence dans le texte. La disposition des mots sur le graphique et la couleur choisie sont *aléatoires* ou dirigées par des considérations esthétiques.

⁷ On trouve, par exemple, un exemple de cette attitude face aux mots-outils dans l'article (Tang et Tao 2016).

⁸ A ce problème difficile vient s'ajouter celui, non moins délicat, de la délimitation effective des limites pratiques de chacun des groupes de formes dans chaque cas particulier.

⁹ Parmi les chercheurs qui se sont consacrés à l'étude des discours officiels chinois, on signalera : Yanhua Di et Zhong Yang (2010 : 69-72). Ces chercheurs ont réalisé une analyse sur des mots-clés (qu'ils définissent par ailleurs) dans les rapports d'activités du gouvernement chinois (1978-2010). Les résultats montrent que le gouvernement chinois met l'économie (经济) et le développement (发展) au premier plan. Jingsheng Zhou et Tianhao Li (2015) s'appuient également sur des mots-clés pour étudier les rapports des 17^e et 18^e congrès du PCC. Hangfeng Tang et Jianwu Tao (2016) ont analysé les mots fréquents (*Word Frequency Analysis*) dans les congrès du PCC (1956-2015). Ces auteurs distinguent trois types d'évolution lexicale : stabilité, dégradation, apparition de nouveaux termes.

changements lexicaux qui surviennent dans le temps et de signaler les périodes particulières au cours desquelles ces changements se manifestent le plus fortement. La démarche que nous présentons, dans ce qui suit, a été largement validée par de nombreuses études réalisées sur des corpus, constitués de textes écrits dans des langues diverses.

1.3 L'écriture chinoise et ses traitements automatisés

Le traitement textométrique des corpus de textes en langue chinoise présente, au premier abord, quelques difficultés qui peuvent être aisément surmontées. Dans les textes chinois, les caractères sont disposés les uns à la suite des autres, sans intervalles qui permettent de séparer des *mots-graphiques*. Ce type d'écriture s'appelle *scriptio continua*¹⁰. Depuis le début du XXe siècle, des caractères que l'on peut considérer comme des *signes de ponctuation* apparaissent couramment dans les textes chinois. De son côté, l'analyse textométrique implique que le texte soit segmenté en unités distinctes qui serviront de base aux décomptes ultérieurs¹¹.

Les segmenteurs automatiques du chinois répondent à cette nécessité pratique de l'analyse textométrique¹². Ces procédures informatisées permettent, en effet, de décomposer les phrases en segments disjoints. Au-delà de la pertinence du découpage ainsi réalisé, la qualité principale de la segmentation automatique, pratiquée sur de grands corpus de textes, réside dans *l'homogénéité des traitements* mis en œuvre pour effectuer la segmentation du texte : si les décisions prises par le segmenteur peuvent être contestables dans certains cas particuliers, l'automatisme du dépouillement garantit, en revanche, une forte homogénéité de l'ensemble du dépouillement.

Il faut garder à l'esprit que, dans chaque cas particulier, les décisions de segmentation prises par les automates s'appuient nécessairement sur l'existence de considérations d'ordre linguistique, prises en compte par des chercheurs, en amont de la segmentation effective des textes, souvent indépendamment du corpus étudié. Ces décisions sont rassemblées dans un *dictionnaire opérationnel du segmenteur*, accessible pour certains des logiciels.

texte original	现在, 我代表第十七届中央委员会向大会作报告
texte segmenté	现在 , 我 代表 第十七 届 中央 委员会 向 大会 作 报告
traduction	<i>Au nom du 17e Comité central, j'ai l'honneur de présenter maintenant le rapport d'activité.</i>

Une fois la segmentation d'un corpus réalisée, l'analyse des résultats permet de dégager,

¹⁰ Rappelons que cette manière d'écrire constituait la norme des systèmes d'écriture, y compris dans le monde occidental, avant le 6ème siècle de notre ère.

¹¹ Dans les langues occidentales ces unités sont plus ou moins assimilées à des *mots* de la langue. Cette approximation, efficace au plan des traitements informatisés, ne règle cependant pas l'ensemble des problèmes de flexions, de conjugaison, d'apposition de suffixes, etc. propres à chaque langue particulière.

¹² Pour ce travail, nous avons utilisé le segmenteur automatique NLPPIR (Zhang, 2016), très largement utilisé dans le monde sinophone.

a posteriori, certaines des règles qui ont présidé au découpage du texte. Cette phase fait nécessairement apparaître le fait que certaines séquences de caractères chinois ont été considérées comme des *mots* ne devant pas être divisés alors que d'autres, qui présentent, au plan linguistique, des similitudes avec les premières, ont été scindées. Nous reviendrons sur ces problèmes de segmentation du texte au paragraphe (§3.2).

2 L'analyse textométrique du corpus *PCC1982-2017*

Le corpus *PCC1982-2017*, constitué par la réunion des textes des huit congrès a donc été soumis aux analyses textométriques. Le dépouillement montre que le corpus contient 115 138 occurrences ventilées sur 7 365 formes différentes. Comme dans la plupart des corpus chinois que nous avons dépouillés, la forme la plus fréquente, avec 7 664 occ., est la forme 的 (DE1) qui correspond grosso modo à la préposition *de* en français. Le corpus contient 2 776 *hapax* (formes qui n'ont qu'une seule occurrence dans l'ensemble du corpus). La division en huit parties, dont chacune correspond à l'un des congrès survenus au cours de cette période, aboutit à un découpage équilibré dont on pourra consulter les caractéristiques sur le tableau 1.

Tableau 1
Principales caractéristiques textométriques du corpus
PC1982-2017, divisé en 8 congrès

	N	Partie	Occurrences	Formes	Hapax	Fmax	Forme	Occurrences
✓	1	1982	16379	2841	1423	1556	的	
✓	2	1987	16007	2787	1307	1375	的	
✓	3	1992	12796	2606	1347	985	的	
✓	4	1997	13765	2614	1308	1023	的	
✓	5	2002	13536	2538	1268	946	的	
✓	6	2007	13105	2590	1285	570	的	
✓	7	2012	13800	2474	1214	513	的	
✓	8	2017	15750	2904	1459	696	的	
	T	Corpus	115138	7365	2776	7664	的	

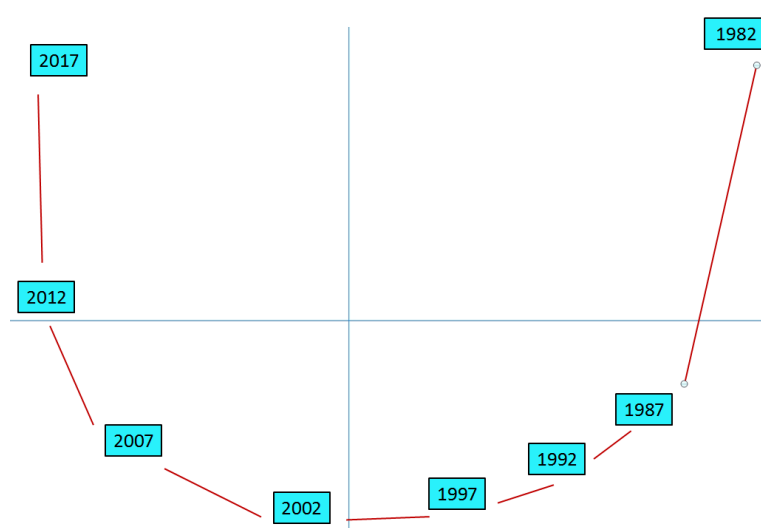
On vérifie que les longueurs des textes produits lors de chaque congrès sont comparables. On remarque que la proportion des occurrences de la forme de fréquence maximale, toujours la forme 的 (DE1, *de*), tend à diminuer entre le congrès de 1982 et celui de 2017, ce qui peut être vérifié sur la figure 1, au paragraphe suivant. Nous verrons plus loin (§4.1) comment cette diminution peut être mise en rapport avec l'évolution du style d'écriture.

2.1 Typologie des congrès

Afin de dresser une typologie des congrès à partir du stock distributionnel du vocabulaire qu'ils utilisent, nous avons soumis à l'Analyse Factorielle des Correspondances (AFC)¹³ le tableau constitué par le décompte des occurrences, dans

¹³ Pour un exposé complet sur l'Analyse Factorielle des Correspondances (AFC), on se reportera à l'ouvrage (Benzecri 1973). Pour des applications à l'étude des corpus de textes, on consultera également (Lebart et Salem

chacun des huit congrès, des 1 529 formes dont la fréquence dépasse 10 occurrences dans l'ensemble du corpus. La figure 1 montre des résultats issus de cette analyse. On reconnaît immédiatement sur cette figure une situation que l'on rencontre couramment lorsqu'on analyse une série textuelle chronologique¹⁴ : les congrès successifs se trouvent relativement rapprochés sur la figure, par rapport aux congrès survenus à des intervalles de temps plus longs. Cette proximité s'explique par le fait que le principal axe de variation du corpus résulte de l'évolution progressive du vocabulaire employé dans les rapports de congrès. Le congrès de 1982, tout en ne dérogeant pas à ce schéma d'ensemble, s'éloigne assez nettement des autres. Une telle position suggère que, dans les congrès qui ont suivi, le renouvellement a été important. Si l'on examine de plus près la figure 1, on peut également y noter des écarts importants (entre le congrès de 1982 et celui de 1987, par exemple, ou encore entre les congrès de 2002 et de 2007).¹⁵



Guide de lecture

Sur le plan sous-tendu par les deux premiers facteurs de l'analyse du tableau (8 congrès 1529 formes), les différents congrès se disposent selon une parabole qui respecte leur échelonnement dans le temps. Cette disposition traduit un renouvellement progressif du vocabulaire au fil des congrès.

Figure 1 :
AFC sur les huit rapports du corpus *PCC1982-2017*
1529 formes de fréquence supérieure à 10.

La typologie obtenue de manière automatique confirme donc que le vocabulaire des rapports évolue progressivement avec le temps. Il nous reste encore à mieux cerner la nature de cette évolution lexicale en étudiant plus précisément le vocabulaire spécifique de chacune des périodes.

1994). Lors de l'analyse du tableau (8 congrès x 1529 formes), le pourcentage d'inertie correspondant au premier facteur de l'AFC est supérieur à 48%, ce qui peut être interprété comme une confirmation que l'évolution du vocabulaire au fil du temps constitue la principale caractéristique du corpus.

¹⁴ Cf. Salem A. (1991), pour un exposé plus complet sur les *séries textuelles chronologiques*.

¹⁵ Les trois congrès 1992 et 1997 et 2002 se sont déroulés sous la présidence de Jiang Zheming. Celui de 2002 assume explicitement une évolution qui se prolongera dans les années qui suivent.

2.2 Contrastes lexicaux entre 1982-1997 et 2002-2017

Pour chaque congrès, et pour chaque groupe de congrès que l'on peut constituer dans le corpus, la *méthode des spécificités*¹⁶ permet de mettre en évidence les formes qui sont sur-représentées dans le fragment que l'on considère ainsi que celles qui y sont particulièrement sous-représentées. La méthode des *spécificités chronologiques* constitue une adaptation aux séries textuelles chronologiques qui souligne les moments marquants de ces évolutions.

En nous appuyant sur la typologie obtenue, nous avons commencé par opérer un calcul de spécificités portant sur la comparaison des deux groupes de périodes situés de part et d'autre de l'axe vertical : le premier groupe rassemble les congrès survenus entre 1982 et 1997; le second, ceux qui se sont tenus durant la période 2002-2017. On peut voir, pour chacun de ces deux groupes, les principaux diagnostics de spécificité positive (formes particulièrement sur-représentées) obtenus à partir de ce découpage¹⁷.

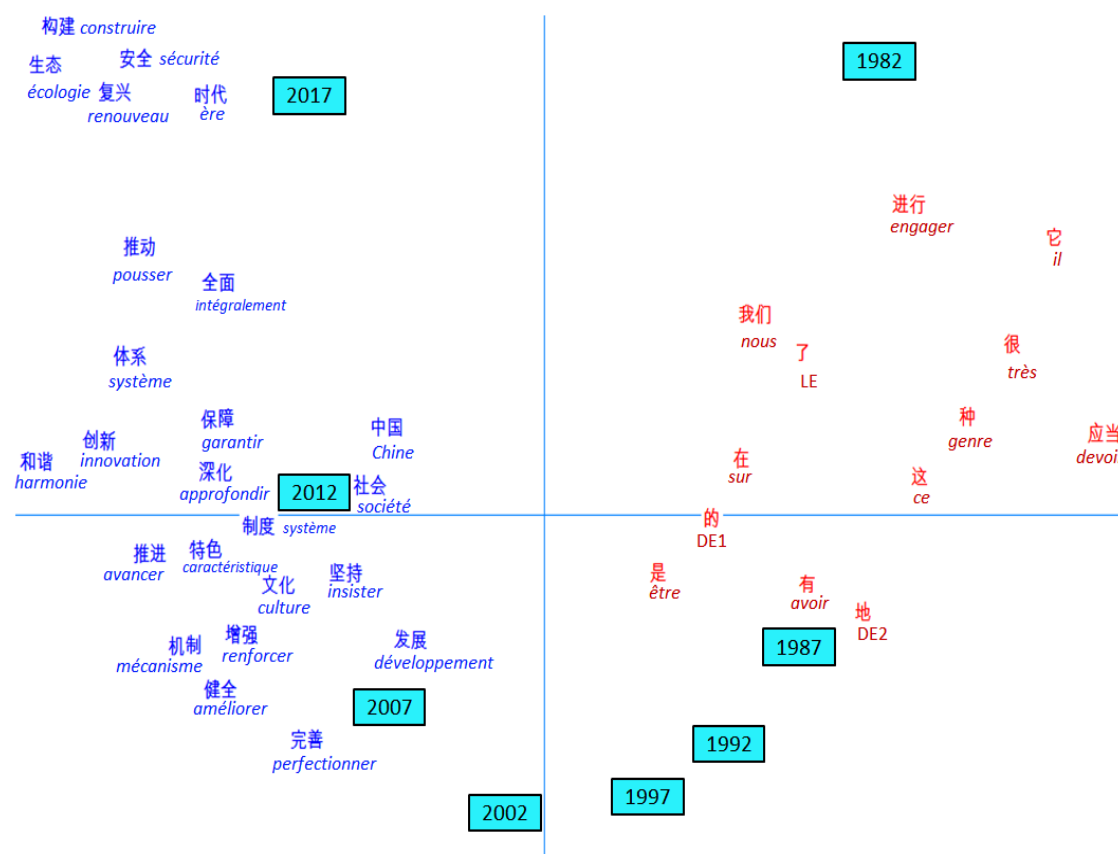


Figure 2:

AFC sur les 8 congrès du corpus **PCC1982-2017**. Formes de fréquence ≥ 10 .
Les spécificités ($S \geq 15$) de chacune des deux divisions du corpus
(1982-1997 et 2002-2017) ont été placées en éléments illustratifs.

¹⁶ Sur la méthode des spécificités et sur les spécificités chronologiques cf. (Lafon 1981) et (Lebart et Salem 1994).

¹⁷ Lors du découpage d'un corpus en deux parties complémentaires, les diagnostics de spécificité positive de chacune des parties correspondent aux diagnostics de spécificité négative de la partie complémentaire. Pour cette raison nous ne considérerons ici que les diagnostics positifs de chacune des parties.

Le calcul des spécificités pour chacune de ces grandes périodes amène un constat inattendu : du côté des premières périodes (1982-1997), sur la gauche du tableau, les formes les plus caractéristiques sont essentiellement des formes que l'on pourrait qualifier de *grammaticales*, telles que 的 (de, DE1), 地 (de, DE2), 了 (le, LE), etc. A l'inverse, du côté de la période (2002-2017), à droite sur le tableau, les formes caractéristiques sont des mots à caractère plus nettement *lexical*. Par exemple, des formes comme : 创新 (chuàngxīn, *innovation*), 体系 (tǐxì, *système*), 发展 (fāzhǎn, *développement*).

Tableau 2:

Extrait du calcul de spécificités sur deux grandes périodes du corpus PCC

Sous-corpus 1982-1997					Sous-corpus 2002-2017				
Terme	Eq.F	F	f	S	Terme	Eq.F	F	f	S
的	DE1	664	4939	***	创新	<i>innovation</i>	200	190	47
这	<i>ce</i>	501	427	***	体系	<i>système</i>	280	247	45
有	<i>avoir</i>	416	332	34	推进	<i>avancer</i>	337	274	35
了	LE	552	413	31	全面	<i>intégralement</i>	326	264	34
在	<i>sur</i>	1096	749	31	发展	<i>développement</i>	1467	934	31
地	DE2	373	288	26	文化	<i>culture</i>	397	306	31
应当	<i>devoir</i>	80	80	24	生态	<i>écologie</i>	90	90	29
种	<i>genre</i>	131	115	19	社会	<i>société</i>	616	430	27
它	<i>il</i>	66	65	18	制度	<i>système</i>	429	313	25
我们	<i>nous</i>	771	510	18	坚持	<i>insister</i>	628	434	25
是	<i>être</i>	1385	862	17	特色	<i>caractéristique</i>	359	270	25
很	<i>très</i>	94	85	16	推动	<i>pousser</i>	162	140	24
进行	<i>engager</i>	184	146	16	安全	<i>sécurité</i>	123	112	24
Guide de lecture La partie gauche du tableau contient formes sur-représentées ($S \geq 15$) dans le sous-corpus 1982-1997. On trouve sur la droite, les formes sur-représentées dans le sous-corpus 2002-2017. A la droite de chaque forme spécifique, on trouve : <i>F</i> - la fréquence dans le corpus, <i>f</i> la fréquence dans le sous-corpus ainsi qu'un diagnostic de spécificité. L'équivalent français est une traduction de la forme hors contexte, fournie pour faciliter la lecture du tableau.					和谐	<i>harmonie</i>	76	75	23
					完善	<i>perfectionner</i>	328	244	22
					中国	<i>Chine</i>	771	503	21
					复兴	<i>renouveau</i>	61	59	17
					健全	<i>améliorer</i>	163	131	17
					增强	<i>renforcer</i>	198	153	17
					保障	<i>garantir</i>	158	128	17
					时代	<i>ère</i>	132	109	16
					深化	<i>approfondir</i>	156	124	16
					机制	<i>mécanisme</i>	141	114	15
					构建	<i>construire</i>	45	45	15

Cette division du vocabulaire employé de manière privilégiée par l'un et l'autre des groupes de partie du corpus peut être illustrée de manière encore plus lisible. Sur la figure 2, nous avons reproduit la typologie présentée sur la figure 1, en l'enrichissant

cette fois de la projection des formes les plus spécifiques pour chacun des groupes¹⁸. Cette seconde présentation des données confirme les constatations que nous avons esquissées plus haut.

La figure 2 permet, en effet, d'apprécier l'évolution du vocabulaire survenue dans la dernière période en combinant une représentation factorielle sur l'ensemble des congrès et les spécificités calculées pour le dernier congrès. Une analyse réalisée sur l'ensemble des congrès met en évidence la progressivité des changements lexicaux. On a projeté en qualité d'éléments supplémentaires les formes *spécifiques positives* de la dernière partie. Ce type de représentation peut être articulé avec la *carte des sections*, présentée plus bas (figure 3) pour illustrer les changements lexicaux. Le contraste lexical entre les deux périodes apparaît très nettement. Du côté droit, les spécificités des années 90 (en rouge), sont pour la plupart, des formes fonctionnelles 的 (de, DE1), 这 (zhè, ce), 有 (yǒu, avoir), 了 (le, LE), 地 (de, DE2).

Du côté gauche, qui correspond aux années 2002-2017, on trouve au contraire des formes dont la signification est plus directement liée à des concepts socio-politiques :

创新 (chuàngxīn, innovation), 体系 (tǐxì, système), 推进 (tuījìn, avancer), 全面 (quánmiàn, intégralement), 发展 (fāzhǎn, développement), 文化 (wénhuà, culture), 生态 (shēngtài, écologie).

A la différence de ce qui se passe dans le cas des *nuages lexicaux* que nous avons évoqués plus haut, les positions des points portés sur le graphique sont ici directement calculées en fonction des profils de répartition de chacun des termes dans l'ensemble du corpus.

3 Exploration des séries textuelles chronologiques

A partir d'un corpus constitué de documents distincts, les méthodes d'analyse textométrique mettent en évidence les spécificités lexicales de chacun des documents (*méthode des spécificités*). Elles permettent également de rapprocher les groupes de textes qui présentent des similarités entre eux et d'en opposer d'autres, sur la base du vocabulaire qu'ils emploient (*analyses factorielles, classifications de textes*).

Dans le cas où le corpus analysé est une *série textuelle chronologique*, des questions relatives à l'évolution du vocabulaire dans le temps se posent immédiatement. En sus du texte contenu dans chacun des documents, l'utilisateur dispose en général de *métadonnées* portant sur chacun des documents : sa date, son lieu d'édition, son auteur, etc. Ces métadonnées constituent des compléments précieux au moment de l'interprétation des résultats calculés automatiquement à partir des données textuelles.

¹⁸ Nous avons utilisé pour produire ce graphique une fonctionnalité disponible dans le logiciel Lexico5 qui permet de projeter des formes, des groupes de formes et des segments répétés sur le plan des facteurs issus d'une AFC.

3.1 La veille textométrique

Lorsque survient un congrès, qui complète une série chronologique pré-existante (dans notre cas, l'inclusion du congrès de 2017 à la série **PCC1982-2017**), les premières questions qui se posent sont : *Quelles sont les principales évolutions lexicales survenues lors du dernier congrès? Les changements survenus confirment-ils l'évolution lexicale constatée lors des périodes précédentes ou marquent-ils, au contraire, une inflexion dans cette évolution ?*

Ces questions font partie de l'activité de *veille lexicale* qui se pratique couramment, à propos des textes mis en ligne sur le web, dans plusieurs domaines d'activité industrielle. Elles sont centrales dans l'étude de l'évolution des textes politiques, pour lesquels les changements formels les plus minimes peuvent parfois se révéler lourds de signification. Le *calcul des spécificités*, appliqué au congrès de 2017, signale des spécificités positives importantes dont le contenu possède un caractère nettement lexical : 时代 (shídài, ère, S +24), 治理 (zhìlǐ, gouvernance, S +21), 生态 (shēngtài, écologie, S +15), 梦 (mèng, rêve, S +14)¹⁹. A l'inverse, les formes de spécificités négatives, pour cette même période, sont plutôt des formes grammaticales, telles que 的 (de, DE1, S -38), 这 (zhè, ce, S -22), 地 (de, DE2, S -14). Le même calcul appliqué aux segments répétés du corpus permet de préciser les modifications lexicales survenues lors de ce même congrès. Comme nous l'avons signalé, la mise en vedette du terme 新时代 (xīn shídài, nouvelle ère), employé 36 fois lors du congrès de 2017, a été largement perçue et commentée par les analystes qui se sont penchés sur ce texte²⁰. Le recensement systématique des segments fortement spécifiques pour cette même période permet de mettre en évidence d'autres séquences dont la forte répétition n'a cependant pas retenu l'attention des commentateurs et dont la répartition montre qu'elles constituent des néologismes par rapport aux congrès précédents :

新时代 中国特色社会主义

le socialisme à la chinoise dans la nouvelle ère (13 occ., S +12)

治理 体系

le système de gouvernance (13 occ., S +12)

Plus remarquable, à nos yeux, certaines expressions extrêmement courantes dans les périodes précédentes ont complètement disparu du texte du dernier congrès. Tel est le cas, par exemple pour des segments comme :

有 中国特色

posséder des caractéristiques chinoises, (0 occ., S -7)

有 中国特色社会主义

avoir un socialisme à la chinoise, (0 occ., S -5)

Les mêmes unités lexicales, ou des fragments de ces dernières, sont désormais utilisées dans un contexte qui a été modifié. Ainsi, le *socialisme à la chinoise* précédemment

¹⁹ Dans nos exemples, la forme *native* chinoise est suivie de sa transcription en *pinyin*, puis d'un équivalent français (lequel ne peut prétendre au statut de traduction satisfaisante, pour chacune des occurrences du terme). Un *coefficient de spécificité*, positive ou négative, de forme S +/- xx indique enfin le degré de spécificité de la forme dans la partie du texte considérée.

²⁰ De nombreux articles publiés à cette occasion ont explicitement mentionné la fréquence de 36 occurrences pour la formule 新时代 (xīn shídài, nouvelle ère), par exemple : Vandepitte (2017).

envisagé comme une perspective est désormais présenté comme une réalité sur laquelle on peut s'appuyer.

3.1.1 Repérage des transformations

Une fois repérées les formes spécifiques pour un texte de congrès, on peut se fixer l'objectif de préciser les causes de l'augmentation de leurs fréquences. Pour chaque unité textuelle spécifique de la dernière partie, deux situations opposées peuvent se présenter :

- l'augmentation peut résulter d'une utilisation accrue de ces formes sur l'ensemble du texte.
- elle peut résulter au contraire de l'apparition de passages particuliers dans lesquels ces formes sont fortement représentées.

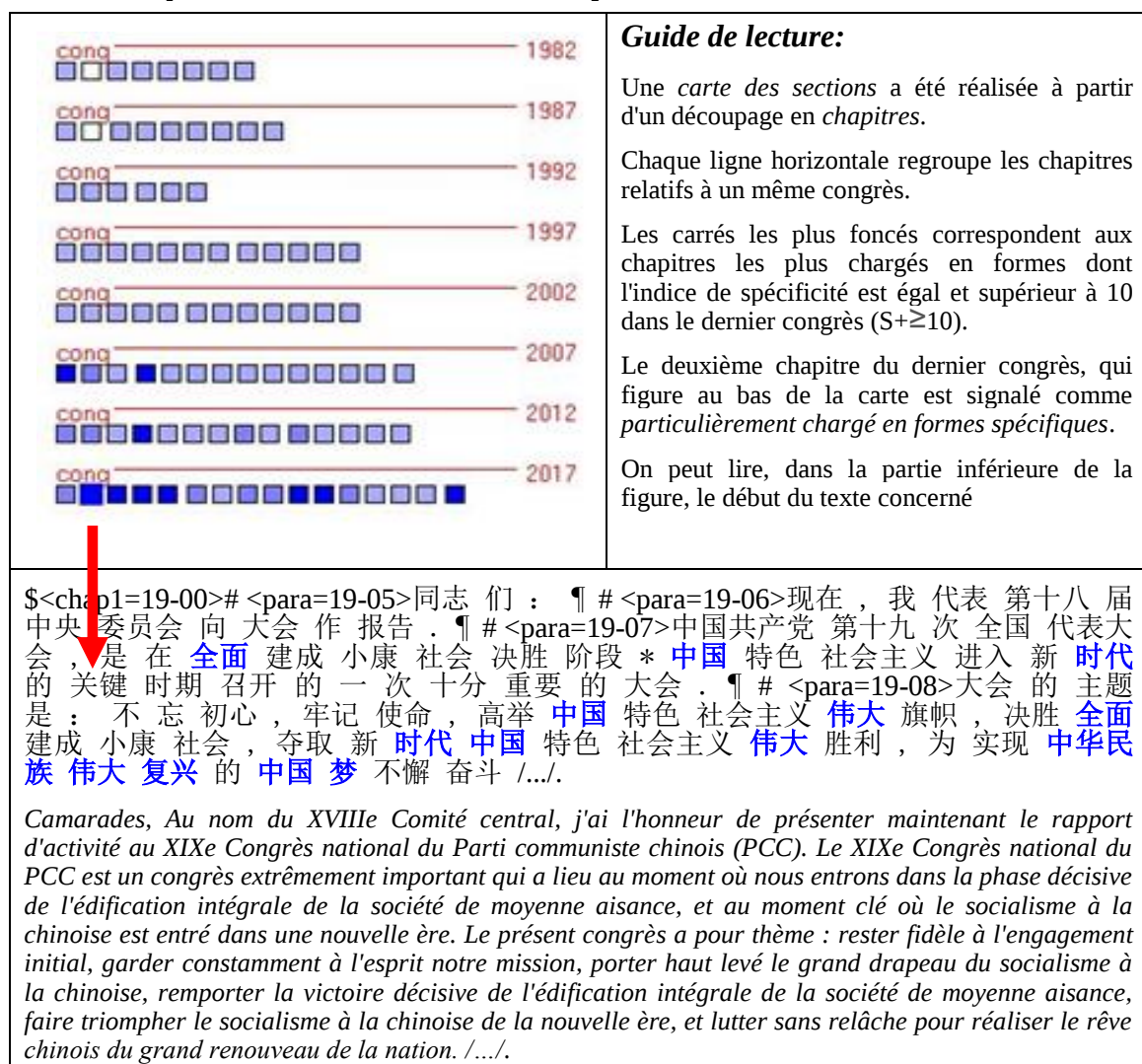


Figure 3 :
Repérage des sections caractéristiques des spécificités ($S+ \geq 10$)
pour le dernier congrès (2017)

L'analyse des spécificités permet de localiser les parties du texte dans lesquelles le renouvellement lexical se révèle particulièrement important. Sur la figure 3, une carte des sections a été établie pour l'ensemble du corpus divisé en chapitres. Les différents

congrès sont séparés par des lignes horizontales. Les sections représentées sur cette carte apparaissent d'autant plus sombres qu'elles renferment de nombreuses occurrences de termes spécifiques pour le dernier congrès. Cette représentation permet de vérifier que le renouvellement du vocabulaire ne se fait pas de manière uniforme. La carte permet de localiser avec précision les chapitres du dernier congrès dans lesquels le renouvellement lexical se manifeste avec le plus de force.

Cette carte permet de vérifier qu'une partie du vocabulaire spécifique du congrès de 2017 est déjà largement présente dans les deux congrès précédents. Ainsi, la notion du *grand renouveau de la nation* (中华民族伟大复兴), initialement lancée lors du 13e congrès de 1987 constitue l'un des points importants évoqués au 16e congrès de 2002. Elle reprend une place importante dans le cadre de l'évocation du *rêve chinois* (中国梦) au cours du congrès de 2017.

16e congrès-2002:<para=16-157>同志们!全面建设小康社会,开创中国特色社会主义事业新局面,就是要在中国共产党的坚强领导下,发展社会主义市场经济*社会主义民主政治和社会主义先进文化,不断促进社会主义物质文明*政治文明和精神文明的协调发展,推进中华民族的伟大复兴。 /.../

Camarades! En vue d'assurer à toute la société une vie relativement aisée et de créer un nouveau contexte pour la cause du socialisme à la chinoise, c'est sous la ferme direction du Parti communiste chinois, en développant l'économie socialiste de marché, la démocratie et la culture avancée socialistes, /.../ en favorisant continuellement le développement harmonieux de la civilisation sous ses aspects matériel, politique et spirituel, que nous parviendrons à assurer le magnifique renouveau de la nation chinoise. /.../

19e congrès-2017:<para=19-08>大会的主题是: /.../为实现 中华民族伟大复兴的中国梦不懈奋斗。

Le présent congrès a pour thème : /.../ et lutter sans relâche pour réaliser le rêve chinois du grand renouveau de la nation.

3.1.2 Nouvelles formes vs prolongement d'une augmentation en cours

Au point précédent nous avons mis en évidence les formes dont la fréquence augmentait tout particulièrement au cours du dernier (19e) congrès. Parmi ces formes, il est utile de distinguer des formes employées pour la première fois lors de ce congrès, comme par exemple : 梦 (mèng, *rêve*, S +14) ou 梦想 (mèngxiǎng, *rêve*, S +12). Pour les formes 时代 (shídài, *ère*, S +24) et 治理 (zhìlǐ, *gouvernance*, S +21) l'accroissement constaté confirme seulement une tendance amorcée lors des congrès précédents. Le segment 新时代 (xīn shídài, *nouvelle ère*), dont la fréquence a été particulièrement commentée (36 occ. dans le 19e congrès), est également attesté une fois dans le congrès précédent²¹.

La méthode des *spécificités chronologiques*²² permet de dresser la liste des formes dont la fréquence s'est notablement accrue lors d'une série de congrès consécutifs. Sur le tableau 3a, on trouve la liste des formes plus spécifiquement employées lors des

²¹ La phrase contenue dans le texte du 18e congrès est : 我们一定要更加自觉地珍爱自然,更加积极地保护生态,努力走向社会主义生态文明新时代 / En veillant à chérir la nature en toute connaissance de cause et à prendre une part plus active à la protection des écosystèmes, nous nous efforcerons de passer à une nouvelle ère de développement d' une civilisation écologique socialiste.

²² La méthode des *spécificités chronologiques* constitue une adaptation de la méthode des spécificités aux séries textuelles chronologique. Elle permet d'étudier les variations de fréquence des termes au sein de groupes de périodes consécutives.

dernières périodes. La dernière colonne indique le groupe de période pour lequel la spécificité est maximale. On vérifiera, par exemple, sur le tableau 3a que les notions 创新 (chuàngxīn, *innovation*), 体系 (tǐxì, *système*), 生态 (shēngtài, *écologie*) ne deviennent spécifiques que pour le groupe des congrès 2002-2007, et 2012 pour la forme *écologie*. L'utilisation de ces formes introduites lors de congrès précédents ne va cesser de croître dans les congrès qui suivent.

Tableau 3a :

Les spécificités chronologiques majeures de la partie (1997-2017).

Terme	Eq. F	F	f	S	Période
创新	<i>innovation</i>	200	190	47	2002 - 2017
体系	<i>système</i>	280	60	46	2007 - 2017
生态	<i>écologie</i>	90	8	40	2012 - 2017
推进	<i>avancer</i>	337	28	38	1997 - 2017
文化	<i>culture</i>	397	45	36	1997 - 2017
全面	<i>intégralement</i>	326	264	34	2002 - 2017
中国	<i>Chine</i>	771	81	33	1992 - 2017
特色	<i>caractéristiques</i>	359	16	32	1992 - 2017
发展	<i>développement</i>	1467	533	31	2002 - 2017
社会	<i>société</i>	616	249	30	2007 - 2017
安全	<i>sécurité</i>	123	20	27	2007 - 2017
坚持	<i>insister</i>	628	68	27	1992 - 2017
制度	<i>système</i>	429	313	25	2002 - 2017
推动	<i>pousser</i>	162	38	25	2007 - 2017
完善	<i>perfectionner</i>	328	284	25	1997 - 2017
时代	<i>ère</i>	132	66	24	2017
和谐	<i>harmonie</i>	76	69	23	2007 - 2017
发展观	<i>concept du développement</i>	39	37	22	2007 - 2012
治理	<i>gouvernance</i>	58	19	21	2017

Le tableau 3b présente des termes dont l'utilisation a au contraire tendance à se raréfier avec le temps. Les spécificités les plus marquées concernent des formes grammaticales dont l'étude fera l'objet du paragraphe §4.

Tableau 3b :

Les spécificités chronologiques majeures de la partie (1982-1997).

Terme	Equivalent français	F	f	S	Période
的	<i>DE1</i>	7664	5885	***	1982 - 2002
这	<i>ce</i>	501	462	***	1982 - 2002
了	<i>LE</i>	552	139	31	1982 - 1997
应当	<i>devoir</i>	80	78	30	1982 - 1992
邓小平理论	<i>théorie de Deng ...</i>	60	54	28	1997 - 2002
地	<i>DE2</i>	373	47	27	1982 - 2002
和	<i>et</i>	3584	1538	25	1992 - 2002
三个代表	<i>triple représentativité</i>	35	35	23	2002 - 2007
我们	<i>nous</i>	771	213	23	1982
商品经济	<i>économie marchande</i>	34	34	21	1987 - 1992

On trouve aussi, parmi ces termes dont l'utilisation se raréfie, des concepts importants qui ont marqué l'histoire du PCC comme : *la théorie de Deng Xiaoping* (邓小平理论, 1997-2002), le concept *de la triple représentativité* (三个代表²³, 2002-2007). Les unités textuelles correspondant à ces concepts se signalent par leur forte spécificité dans la première période du corpus. Le concept d'*économie marchande* (商品经济) est, au contraire, signalé comme appartenant plus spécifiquement à la période 1987-1992. Cette circonstance nous conduira à étudier dans ce qui suit (§3.2.2) l'ensemble des séquences dans lesquelles se trouve impliqué le terme 经济 (économie).

3.2 Fouille autour d'un motif donné : le terme 经济 (économie)

L'analyse de la répartition des termes spécifiques de la première période nous amène à nous pencher à nouveau sur un aspect important de la segmentation des textes que nous avons négligé dans un premier temps.

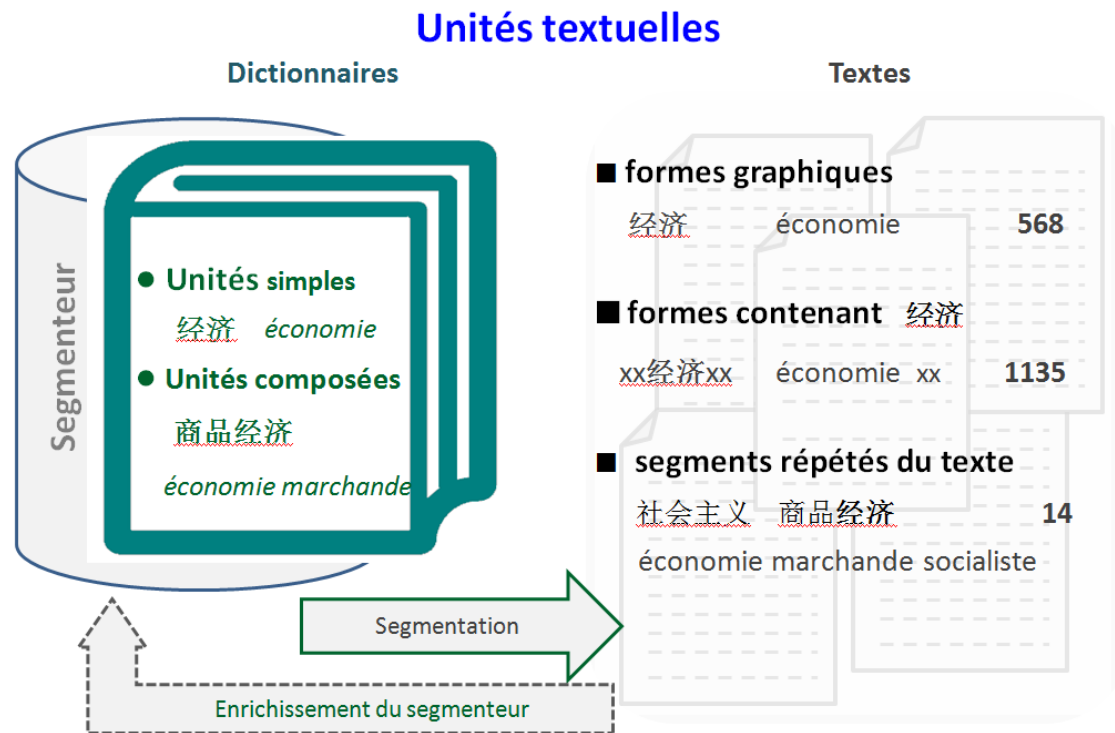
On note par exemple que le terme 商品经济 (économie marchande) apparaît comme très spécifique de la période 1987-1992. Comme nous l'avons signalé plus haut, le segmenteur NLPPIR-ICTCLAS n'a pas décomposé ce mot quadrisyllabique en deux unités (商品+经济, marchand + économie). On en conclut que cette unité appartient au *dictionnaire opérationnel* du segmenteur, alors qu'elle ne figurera pas nécessairement dans les dictionnaires de langue, produits à différentes époques. Les raisons de ce choix résident probablement dans le fait que cette association, fréquente dans la société chinoise des années 80, a fini par être perçue comme une entité lexicale dotée d'un sens propre et pourvue d'une certaine autonomie. Elle a été incluse, de ce fait, comme un *mot* dans le dictionnaire du segmenteur. On comprend que, dès que l'on dépasse le cercle des unités qui figurent traditionnellement dans les dictionnaires de langue, les choix entre les unités que l'on considérera comme des unités lexicales et celles qui seront considérées comme des compositions lexicales font intervenir des considérations socio-politiques complexes et ne peuvent prétendre au statut de règles purement linguistiques.

Un exemple : la séquence des deux caractères 经济: (économie) apparaît 1135 fois dans l'ensemble du corpus. L'analyse des résultats produits par le segmenteur montre que cette séquence a été considérée 568 fois comme un mot isolé alors qu'elle participe à la composition de plusieurs autres formes graphiques : 市场经济 (économie de marché), 86 occ., 经济发展 (développement économique), 62 occ., elles mêmes très fréquentes et qui n'ont pas été segmentées. Les choix linguistiques qui sont à l'origine de ces décisions de paramétrage du segmenteur ne s'imposent pas indépendamment des domaines de communication, des périodes et des idéologies. La fréquence d'une unité

²³ Le concept des « trois représentativités », introduit par Jiang Zemin lors du 16e congrès de 2002, intégrera rapidement le corps de doctrine du PCC. Il s'agit d'une justification a posteriori de la légitimité du pouvoir exercé par ce Parti. Ce dernier exerce le pouvoir parce qu'il est le mieux à même de représenter, à la fois :

- le développement des forces productives les plus avancées en Chine ;
- le progrès de la culture chinoise ;
- les intérêts de la grande majorité du peuple chinois.

particulière (ex : 市场经济, *économie de marché*), à une époque donnée, entraîne, bien entendu, son inclusion dans le *dictionnaire opérationnel* du segmenteur. L'inclusion des néologismes et des expressions appelées à devenir des concepts récurrents constitue un travail permanent pour les concepteurs du dictionnaire opérationnel, les décisions prises au cas par cas pouvant toujours être remises en cause.



Guide de lecture:

Sur la partie gauche, nous avons représenté les dictionnaires (dictionnaires de langues et/ou ressources dictionnaires informatisées) qui renferment les unités lexicales connues au départ de la segmentation automatique.

A droite, on trouve un exemple d'unité, composée de plusieurs formes produites par le découpage automatique initial, dont la fréquence suggère cependant qu'elle constitue une unité à part entière et qu'elle mériterait d'intégrer le dictionnaire qui sert à la reconnaissance des lexies.

Figure 4:

Unités dictionnaires et unités textuelles

Dès lors, pour tenter une étude complète sur ce que l'analyste considère comme le concept de l'*économie* (经济), il apparaît indispensable d'étudier à la fois les formes graphiques qui incluent les deux caractères en question (ex: 商品经济, *marchand + économie*) mais aussi les segments qui contiennent ces mêmes caractères (ex: 社会主义 商品经济, *économie marchande socialiste*). Nous proposons ci-après deux méthodes de localisation de ces unités : création d'unités lexicales complexes (figements) et répartition des unités contenant un *motif*.

3.2.1 Création d'unités lexicales complexes (figements)

Comme nous allons le voir, les séquences lexicales découpées par le segmenteur trahissent inévitablement des jugements d'ordre linguistique, intervenus en amont du processus de segmentation. Ces jugements concernent le degré de lexicalisation de séquences composées de plusieurs formes simples. A partir d'un *motif*, défini au départ de l'expérience (nous avons choisi ici le couple de caractères 经济 *économie*), il est possible de repérer des formes découpées par l'automate de segmentation dont la taille dépasse deux caractères et qui n'ont pas été séparées en unités distinctes.

Tableau 4:

Termes et segments répétés contenant les caractères 经济 (*économie*)
dans le corpus PCC

<i>formes</i>	<i>équivalent français</i>	<i>fréquence</i>
经济	<i>économie</i>	568
市场经济	<i>économie de marché</i>	86
经济发展	<i>développement économique</i>	62
国民经济	<i>économie nationale</i>	56
商品经济	<i>économie marchande</i>	34
经济体制改革	<i>réforme du système économique</i>	29
经济效益	<i>bénéfices économiques</i>	20
经济结构	<i>structure économique</i>	20
社会主义经济	<i>économie socialiste</i>	17
经济体制	<i>système économique</i>	15
计划经济	<i>économie planifiée</i>	13
集体经济	<i>économie collective</i>	13
经济全球化	<i>mondialisation économique</i>	12
农村经济	<i>économie rurale</i>	11
地区经济	<i>économie régionale</i>	11
经济增长	<i>croissance économique</i>	11
经济制度	<i>système économique</i>	10
经济社会	<i>société économique</i>	10
<i>segments</i>	<i>équivalent français</i>	<i>fréquence</i>
社会主义 商品经济	<i>économie marchande socialiste</i>	14
社会主义 市场经济 体制	<i>système d'économie de marché socialiste</i>	41
社会主义 市场经济	<i>économie de marché socialiste</i>	65
经济 发展	<i>développement économique</i>	20
经济 实力	<i>force économique</i>	11
经济 建设	<i>édification économique</i>	77
经济 特区	<i>zone économique spéciale</i>	11
国际 经济	<i>économie internationale</i>	10
市场经济 体制	<i>système d'économie de marché</i>	43
国有 经济	<i>économie publique</i>	14
公有制 经济	<i>économie de propriété publique</i>	18
所有制 经济	<i>économie de propriété</i>	16
非公有制 经济	<i>économie non publique</i>	14
开放型 经济	<i>économie ouverte</i>	12

La fonction *segments répétés*²⁴ qui permet de recenser toutes les suites d'unités lexicales qui apparaissent plusieurs fois dans le même ordre au sein d'un corpus permet, d'autre part, de repérer des suites d'unités récurrentes qui auraient peut-être mérité d'intégrer le dictionnaire du segmenteur.

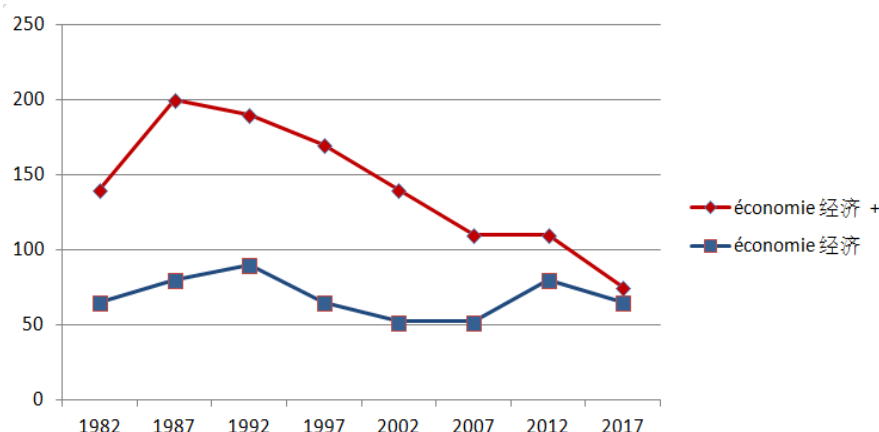


Figure 5:

Ventilation dans les congrès des termes contenant les caractères 经济: (économie)
 - en haut (rouge) les formes **contenant** les deux caractères
 - en bas (bleu) les formes 经济 isolées par le segmenteur

Sur la figure 5, la courbe inférieure représente la distribution des seules occurrences de la forme 经济 (économie) isolées par le segmenteur. Comme nous l'avons vu, la forme isolée apparaît 568 fois dans le corpus. La courbe rouge (en haut) tracée à partir du décompte des occurrences de toutes les formes contenant la séquence des deux caractères montre que le total de ces dernières s'élève à 1135 occurrences, soit presque le double des formes décomptées isolément par le segmenteur. On voit à partir de ces données que les occurrences de la séquence des deux caractères (économie) qui ont donné lieu à une segmentation ne constituent qu'une moitié du nombre total des occurrences de ce couple de caractères. Les autres occurrences de ce couple ont été considérées comme faisant partie de formes composées à partir de ce mot.

On peut avancer, à ce stade, une première interprétation de ce phénomène : la forme *économie* connaît un usage relativement stable au cours de la période 1982-2017, que nous considérons. Dans les premières périodes, en plus des occurrences du mot isolé, on trouve fréquemment des formes construites à partir de la forme *économie* (cf. tableau 4). Dans les dernières périodes, à l'inverse, le concept de l'économie devient une entité qu'il n'est plus indispensable de qualifier, chaque fois que l'on y fait référence. Dans les paragraphes suivants, nous étudierons la répartition de ces différentes unités parmi les différents textes de congrès, en liaison avec les séquences répétées plus longues qui contiennent la forme *économie*.

²⁴ Il s'agit d'une méthode implémentée dans l'outil textométrique *Lexico 5* qui permet de repérer les séquences de formes qui se trouvent répétées dans le corpus.

3.2.2 Répartition des unités contenant un motif

Pour tenter de classer les unités contenant un même motif, il est possible de les projeter en qualité d'*éléments supplémentaires* sur le plan des deux premiers facteurs obtenu lors de l'analyse du tableau croisant les congrès et les formes (§2.1 - figure 1)²⁵.

Cette représentation confirme l'évolution, dans le temps, des principaux concepts mobilisés pour parler de l'économie de la Chine. On voit sur la figure 6 que les termes que l'on peut rattacher au concept de l'économie occupent des places très variées sur le plan des deux premiers facteurs. *Les bénéfices économiques, l'économie socialiste, l'économie nationale, l'économie planifiée* des premières périodes cèdent progressivement le pas à des lexies comme : *la capacité économique, le développement économique, la mondialisation économique*. Les rapports du PCC mettent de plus en plus l'accent sur le choix et la définition du système économique. Les termes de la première période sont prioritairement liés aux *réformes du système économique* (经济体制改革, 1987-1992) en cours. L'économie de marché apparaît dans les périodes suivantes : *établir un système d'économie de marché socialiste* (建立社会主义市场经济体制, 1992-1997). Une fois ce système installé autour des années 2002, la question est posée de *l'amélioration du système d'économie de marché socialiste* (完善社会主义市场经济体制, 2002-2007), ainsi que celle des rapports entre l'économie de propriété privée et l'économie collective (所有制经济, 集体经济). Au fur et à mesure que l'économie chinoise s'intègre à *la mondialisation économique* (经济全球化), *l'économie non-publique* et *l'économie de type ouvert* peuvent prendre place (2012). *La puissance économique* (经济实力) et le *développement économique* (经济发展) deviennent, vers l'année 2017, des objectifs prioritaires pour le pays. On trouve également la trace, dans ces changements lexicaux, d'un accroissement constant de la préoccupation d'une économie tournée vers les marchés extérieurs.

Pour mieux comprendre l'évolution de ces termes, il est nécessaire de retourner au contexte. Ainsi, on notera que les deux lexies *économie de marché* (市场经济) et *économie socialiste* (社会主义经济), qui semblent s'opposer radicalement sur un plan conceptuel, trouvent simultanément la majorité de leurs occurrences dans les premières périodes du corpus (1992). Mobilisées lors d'un même débat, elles seront toutes deux abandonnées dans les périodes qui suivront.

²⁵ En AFC, on appelle *éléments supplémentaires* des éléments qui n'ont pas participé aux calculs initiaux et que l'on décide de projeter sur les plans factoriels, après l'analyse, pour illustrer les représentations obtenues.

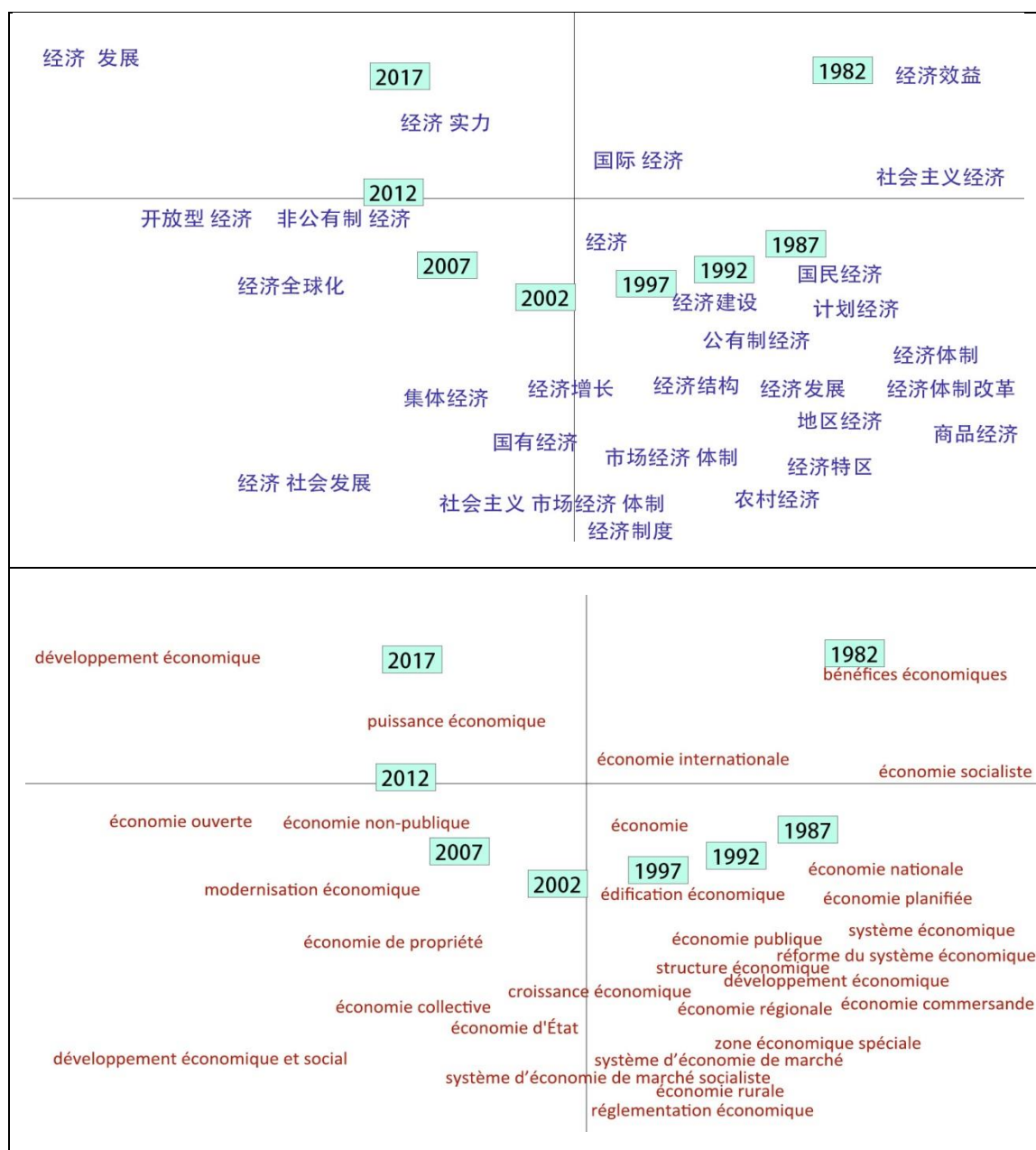


Figure 6 :

Les principaux termes qui contiennent les caractères 经济(économie) sur le plan des deux premiers facteurs issus de l'analyse du tableau (8 congrès x formes de fréquence supérieure à 10)
En haut : le texte original chinois; en bas : un équivalent traductionnel français pour chacun des termes²⁶

4 Evolution des formes de l'énonciation

Comme nous venons de le voir, par-delà l'apparition de nouvelles thématiques qui retiendront prioritairement l'intérêt des politologues, l'une des caractéristiques marquantes de l'évolution lexicale dans l'ensemble du corpus est constituée par le fait qu'une série de formes grammaticales, relativement abondantes dans les premières

²⁶ Nous avons allégé cette figure en écartant certains des termes, que nous avons jugés redondants ou moins importants, afin de lui donner une plus grande lisibilité. Dans ce même but, d'autres termes ont été très légèrement déplacés, par rapport à la position qui leur avait été attribuée lors de l'analyse.

périodes du corpus, a tendance à disparaître, à mesure que l'on avance dans le temps. Nous allons tenter de décrire ce phénomène dans la présente section.

L'évolution que nous venons d'évoquer est sous-tendue par quatre types de phénomènes qui touchent à autant de niveaux du processus d'énonciation :

- une amélioration du style et de la clarté de l'énoncé ;
- une atténuation progressive du caractère injonctif de l'énoncé ;
- une évaluation de l'activité propre ;
- une modification progressive de l'identité de l'énonciateur formel.

4.1 Amélioration du style et de la clarté de l'énoncé

Comme c'est le cas dans la plupart des textes chinois (Chen 1955, Miao 2012), la particule 的 (DE1, *de*) est la forme la plus utilisée dans chacun des volets de notre corpus. Sa principale fonction consiste à relier le déterminant au déterminé : *déterminant + 的 + déterminé* (ex : 党+的+工作, *dǎng de gōngzuò, travail du Parti*). La forme remplit avant tout une fonction grammaticale et, pour cette raison, elle a peu retenu, à notre connaissance, l'attention des chercheurs qui s'intéressent à l'analyse du discours politique chinois. L'observation des variations de la fréquence de cette forme apporte, cependant, des enseignements précieux sur le style du discours adopté lors de chacune des périodes. Ces considérations s'appliquent également à la répartition des formes 地 (DE2) et 了 (LE), que l'on présentera plus bas.

Les analyses statistiques montrent que la forme 的 (DE1) possède plus d'occurrences dans les congrès de la première période (1982-1997). Le retour systématique aux contextes montre que l'omission de 的 (DE1) rend l'écriture plus concise. Ainsi, le congrès de 1982 utilise systématiquement 的 (DE1) pour relier le déterminant aux déterminés *civilisation* et *édification*.

<para=12-59>/.../物质文明 的 建设 是 社会主义 精神文明 的 建设 不可缺少 的 基础。社会主义 精神文明 对 物质文明 的 建设 不但 起 巨大 的 推动作用，而且 保证 它 的 正确 的 发展 方向。两 种 文明 的 建设，互为 条件，又 互为 目的。

L'édification d'une civilisation matérielle constitue la base indispensable pour l'édification de la civilisation spirituelle du socialisme. Et celle-ci donne non seulement une immense impulsion à l'édification de la civilisation matérielle, mais encore, elle en assure la juste orientation. Ces deux civilisations se conditionnent mutuellement et elles sont le corollaire l'une de l'autre.

Le congrès de 2012 parle encore de la *civilisation* et de l'*édification*, mais la particule 的 (DE1) ne figure plus dans cette expression. Ici encore, l'absence de DE1 permet de renforcer la cohésion du syntagme, et de produire une écriture plus concise²⁷.

<para=18-105>/.../把 生态 文明 建设 放 在 突出 地位，融入 经济 建设

²⁷ La notion de déterminant couvre en chinois un champ beaucoup plus large que celui qui est couvert en français. En plus des articles (définis, indéfinis, partitifs) et des adjectifs (possessifs, démonstratifs, indéfinis, numéraux, interrogatifs, relatifs), le déterminant chinois englobe: les épithètes, les syntagmes nominaux compléments et les propositions relatives. Dans notre étude comparative, nous entendrons par *déterminants* l'ensemble des mots, syntagmes ou propositions qui ont pour objectif de transformer un nom (ou pronom) du discours

* 政治 建设 * 文化 建设 * 社会 建设 各 方面 和 全 过程 , 努力 建设 美丽 中国 , 实现 中华民族 永续 发展 .

En accordant une attention particulière au développement d'une civilisation écologique, nous ferons en sorte que celui-ci fasse partie intégrante des programmes de l'édification du pays sur les plans économique, politique, culturel et social. Nous nous efforcerons ainsi de construire une Chine magnifique et de réaliser un développement perpétuel de la nation chinoise.

Les mêmes remarques peuvent être faites au sujet des formes 地 (DE2) et 了 (LE). La forme 地 (DE2) a pour fonction de transformer un mot (nom, adjectif, verbe), ou un groupe de mots, en complément circonstanciel du verbe. Elle s'insère dans des séquences du type : groupe circonstanciel + 地 + verbe (voir, plus loin, § 4.2). On peut dire, de manière approximative, que ce procédé assume la même fonction que l'adverbe français qui modifie le sens du verbe ou de l'adjectif. Mais l'emploi de 地 (DE2) n'est pas obligatoire dans ce genre d'expression. Son omission traduit également une volonté de concision dans l'énonciation du texte.

La correspondance entre alternances de ce type, qui constituent des séquences dont le sens est identique, est fréquente dans notre corpus.

1987 : 更好+地+发挥[...]作用, *mieux + DE2 + remplir la fonction de...*

2007 : 更好发挥[...]作用, *mieux + remplir la fonction de...*

Dans notre corpus, la particule aspectuelle 了 (LE) est principalement employée comme suffixe verbal. Placée immédiatement après le verbe, elle exprime le caractère accompli de l'action (LE1). Utilisée en fin de phrase, cette même particule permet d'exprimer un changement d'état ou une actualisation (LE2)²⁸. L'emploi massif de 了 (LE2) trahit le genre du compte rendu d'activités. Lorsqu'elle n'entraîne pas d'ambiguïté, son omission produit une écriture plus élégante, car elle permet de rapporter les faits avec plus de simplicité.

1987 : 代表 + 了 + 广大人民群众的最大利益

Représenter + LE + l'intérêt supérieur des grandes masses du peuple

2002 : 代表+中国最广大人民的根本利益

Représenter + les intérêts fondamentaux de l'écrasante majorité de la population en Chine.

L'omission de 了 (LE) rend la séquence plus facile à énoncer. L'amélioration de la clarté de l'énoncé influe à son tour sur la longueur des rapports qui ont tendance à être plus courts dans la seconde période.

4.2 Atténuation du caractère injonctif de l'énoncé

L'analyse chronologique nous conduit à remarquer un second type d'évolution dans le style d'énonciation : la forte atténuation du caractère injonctif des énoncés. Cette transformation résulte à la fois de la diminution de l'emploi de la particule 的 (DE1) et

²⁸ Ces deux fonctions ont été décrites par de nombreux spécialistes du chinois, cf., par exemple, (Darrobers et Xiao Planes 1998).

de celle des auxiliaires liés à l'obligation, en particulier 应当 (yīngdāng, *devoir*).

En plus de la fonction que nous venons d'évoquer, la forme 的 (DE1) assume, dans la langue chinoise, une partie de la fonction de prédicat. Cette fonction peut apparaître sous deux formes :

déterminant + 的 (DE1) ou 是 + déterminant + 的 /SHI (être)...DE1...

La construction du prédicat avec 的 (DE1) met plutôt en valeur l'information donnée sur le sujet ; l'utilisation de la forme SHI...DE1 (是...的) introduit une nuance qui centre le propos sur l'ensemble de l'information fournie. Dans une certaine mesure, SHI...DE1 peut être considéré comme l'équivalent sémantique du français : *c'est X qui/que*.

Les exemples ci-dessous illustrent la manière dont la construction SHI...DE1 véhicule des sortes d'injonctions, dans le congrès de 1982.

[...]方向是**正确**的，收效也很明显。

orientation-être-bon-DE, résultat-aussi-très-évident.

C'est cette orientation qui est la bonne et ses résultats sont évidents.

La co-présence de 是 (SHI, *être*) et du pronom indicatif 这 (ZHE, *ce/ceci*) dans les spécificités des congrès des années 80 résulte directement du recours fréquent à ce type d'écriture qui sera progressivement abandonné au fil des années. Il s'ensuit que l'énoncé tend à devenir moins injonctif, plus persuasif.

L'évolution de cette forme d'écriture se manifeste également dans l'emploi de l'auxiliaire d'obligation 应当 (yīngdāng, *devoir*), terme spécifique du congrès de 1987. Nous avons pu confirmer cette observation en étudiant les fréquences d'un ensemble de formes qui expriment l'obligation et la nécessité dans la langue chinoise :

应当 (yīngdāng), 应该 (yīnggāi), 应 (yīng), 必须 (bìxū), 必 (bì),

务必 (wùbì), 须 (xū), 得 (děi), 要 (yào), 需要 (xūyào) et 需 (xū)²⁹.

La figure 7 illustre l'évolution de ces termes, tout au long de la période couverte par notre corpus. Les deux premiers congrès de 1982 et 1997 utilisent particulièrement les formes : 应当 (yīngdāng, *falloir*), 应该 (yīnggāi, *devoir*), 必须 (bìxū, *devoir*) et 需要 (xūyào, *avoir besoin de*), tandis que le congrès de 2002 leur préfère la forme 应该 (yīnggāi, *devoir*) et 应 (yīng, *devoir*). Pour les congrès 2012 et 2017, le retour au contexte montre qu'ils utilisent un mélange de termes liés à l'obligation, principalement 要 (yào). Ici encore, un ensemble de constats particuliers nous amène à considérer que le caractère injonctif de l'énoncé s'atténue au fur et à mesure que l'on avance dans le temps des congrès.

²⁹ Tous ces termes peuvent plus ou moins être traduits par l'expression *être dans l'obligation de faire quelque chose*. (cf. Alleton 1984) :

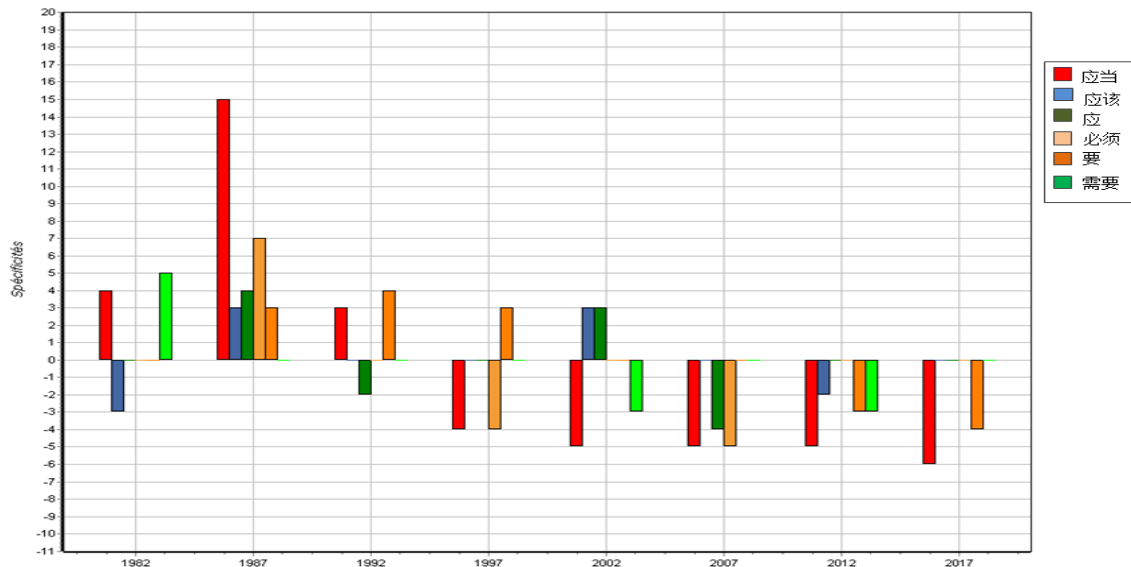


Figure 7

Répartition de formes liées au caractère d'obligation dans le corpus *PCC1982-2017*

4.3 Evaluation de l'activité propre

Chacun des rapports de congrès contient traditionnellement un compte rendu d'activités détaillé pour la période écoulée. On peut noter que l'évaluation de l'activité propre se transforme au fil du temps : des *problèmes rencontrés* vers les *réalisations atteintes*, des *insuffisances* vers les *succès* et les *triumphes*. L'emploi de l'adverbe de degré 很 (hěn, très) peut être considéré comme un indice qualifiant le degré des difficultés rencontrées dans l'action. Cet adverbe apparaît souvent lors des deux premiers congrès (1982 et 1987), la plupart du temps associé avec des adjectifs péjoratifs :

中的一部分低产地区和受灾地区,农民还是很很贫困,要积极扶助他们发展生产,增加收。是我们已经做的比我们应当做的还差得很很远。我们要作出更大的努力,加强自己的。事业单位的领导问题,是机构改革中一个很重要的问题。党的工作和政府的工作,企展水平总的说来还比较低,又很不平衡,在很长时期内需要多种经济形式的同时并存。粉碎江青反革命集团之后,人民对党寄以很高的希望。人民要求拨乱反正,要求安定团结。

Partie : 1987, nombre de contextes : 30

社会主义政治所必需的一系列经济文化条件很不充分,封建主义*资本主义腐朽思想和资本主义社会化程度还很低,商品经济和国内市场很不发达,自然经济和半自然经济占相当比重。在改革中,保持安定团结是非常必要的,也是很不容易的。社会主义民主和法制的建设,社会主义和腐败现象,全党同志和广大群众是很不满意的。我们一定要正视它,努力去消灭它。同我们正在进行伟大事业相比,是很不相称的。改革开放已进行多年,改革开放很不稳定,加强农业建设尤为迫切和重要。其他经济成份,不是发展得太多了,而是还很不够。对于城乡合作经济*个体经济和私营经济,必须实行改革开放,增强活力,坚持按劳分配为主体,实行多种分配方式并存的制度。随着生产力的不断发展,逐步消灭剥削,实行按需分配。坚持和完善人民代表大会制度,保证人民当家作主,加强法制建设,维护社会安定团结,保证改革开放和社会主义现代化建设的顺利进行。我们的任务是很艰巨的,我们的信心是很坚定的。我们一定要坚持四项基本原则,坚持改革开放,自力更生,艰苦创业,为把我国建设成为富强、民主、文明的社会主义现代化国家而奋斗。

Dans les exemples ci-dessus, les textes de rapport abordent divers problèmes : les *paysans sont encore très pauvres* (农民还很很贫困), *une série de conditions économiques et culturelles sont (très) insuffisantes* (一系列经济文化条件很不充分). L'adverbe 很 est souvent combiné avec la particule négative 不 (bù, non, pas) : 很不充分 (hěn bù chōngfèn, insuffisant), 很不发达 (hěn bù fādá, non développé),

很不容易的 (hěn bù róngyì de, *pas facile*), 很不满足的 (hěn bù mǎnzú de, *insatisfaisant*), 很不稳定 (hěn bù wěndìng, *instable*), 很不相称的 (hěn bù xiāngchèn de, *incohérent*).

A partir du congrès de 1992, les syntagmes intégrant 很 (hěn, *très*) deviennent moins fréquents (33 occ. en 1982 vs. 9 occ. en 1992). Les formes qui apparaissent dans l'entourage de 很 (hěn, *très*) changent, elles aussi. On peut y voir l'émergence d'une volonté nouvelle d'établir un nouveau rapport avec le destinataire du message sur la base de considérations positives. Dans les premiers congrès, l'emploi de 很 (hěn, *très*) apparaît dans des considérations critiques portant sur des problèmes intérieurs à la Chine ; les quelques occurrences de cette particule qui subsistent dans les derniers congrès portent au contraire sur des problèmes extérieurs (ex : 世界仍然很不安宁, *le monde reste encore très instable*).

Comme nous l'avons vu plus haut, la particule 地 (DE2) sert à l'articulation entre le groupe circonstanciel et le verbe. Les formes les plus utilisées devant cette particule sont les suivantes :

Terme	Eq.fr	F	Terme	Eq.fr	F
更好	<i>mieux</i>	30	创造性	<i>créatif</i>	7
坚定不移	<i>persévérant</i>	15	有效	<i>efficace</i>	7
毫不动摇	<i>inébranlable</i>	11	正确	<i>correct</i>	7
清醒	<i>clair</i>	11	紧密	<i>intense</i>	6
经常	<i>souvent</i>	10	更加自觉	<i>plus conscient</i>	5
坚持不懈	<i>sans relâcher</i>	8	始终不渝	<i>constant</i>	5
有力	<i>fort</i>	8	有步骤	<i>avoir des étapes</i>	5
自觉	<i>conscient</i>	8			

Dans l'environnement qui précède la forme 地 (DE2), l'adverbe 更好 (gènghǎo, *mieux*) est la forme la plus utilisée (30 occ.) Cet adverbe constitue souvent une incitation à l'amélioration du travail. D'autres expressions, également très fréquentes, peuvent être rapprochées de ce type de préoccupation : des expressions figées quadrisyllabiques : 坚定不移 (jiāndìngbùyí, *persévérant*), 毫不动摇 (háobùdòngyáo, *inébranlable*), 坚持不懈 (jiānchíbùxiè, *sans relâche*), 始终不渝 (shǐzhōngbùyú, *constant*), ainsi que des adjectifs : 清醒 (qīngxǐng, *clair*), 自觉 (zìjué, *conscient*) et 更加自觉 (gèngjiāzìjué, *plus conscient*). Les segments tels que 经常 (jīngcháng, *souvent*), 有力 (yǒulì, *fort*), 创造性 (chuàngzàoxìng, *créatif*), 有效 (yǒuxiào, *efficace*), 正确 (zhèngquè, *correct*), 紧密 (jǐnmì, *intense*), 有步骤 (yǒubùzhòu, *prévoir des étapes*), participent également à la qualification positive des objectifs fixés par les organes de direction.

A partir du congrès de 2007, l'adverbe 更加 (gèngjiā, *plus*) apparaît de plus en plus souvent dans le groupe circonstanciel introduit par 地 (DE2). Notons, par exemple 更加自觉 (gèngjiā zìjué, *plus conscient*), présent une seule fois dans le congrès de 2007 et cinq fois dans celui de 2012. La combinaison entre 更加 (gèngjiā, *plus*) et d'autres segments se fait de plus en plus fréquente.

2007			2012		
Terme	Eq.fr	F	Terme	Eq.fr	F
更加紧密	<i>plus serré</i>	1	更加自觉	<i>plus conscient</i>	5
更加主动	<i>plus d'initiatives</i>	1	更加积极	<i>plus actif</i>	1
更加自觉	<i>plus conscient</i>	1	更加尖锐	<i>plus aigu</i>	1
			更加紧密	<i>plus serré</i>	1
			更加科学有效	<i>plus d'efficacité scientifique</i>	1

4.4 Accent sur l'identité nationale

L'utilisation des pronoms personnels qui désignent les acteurs du contexte, dans les rapports de congrès, permet également d'observer une évolution du style de l'énonciation. L'emploi du pronom 我们 (wǒmen, *nous*) diminue progressivement alors que le nom propre 中国 (zhōngguó, *Chine*) voit sa fréquence augmenter. Les exemples ci-dessous vont nous permettre d'illustrer cette évolution.

1982 <para=12-25>同志 们！**我们** 在 过 去 六 年 中 取 得 的 伟 大 胜 利 是 有 目 共 睹 的 。 但 是 ， **我们** 不 仅 不 能 满 足 于 已 经 取 得 的 胜 利 ， 而 且 必 须 看 到 ， **我们** 党 的 工 作 还 有 不 少 缺 点 ， 还 有 很 多 困 难 ， 还 有 种 种 不 如 人 意 的 地 方 。 **我们** 一 定 要 进 一 步 振 奋 革 命 精 神 ， 埋 头 苦 干 ， 为 夺 取 新 的 更 大 胜 利 而 努 力 奋 斗 。

#<para=12-25>Camarades! Les grandes victoires que **nous** avons remportées durant les six années écoulées sont connues de tous. Cependant, loin de pouvoir **nous** en contenter, **nous** devons être conscients que le travail de **notre** Parti présente encore bon nombre d'insuffisances, qu'il y a encore beaucoup de difficultés et beaucoup de choses qui laissent à désirer. **Nous** devons rehausser l'esprit révolutionnaire, travailler dur et lutter de toutes nos forces pour remporter de nouvelles victoires encore plus grandes.

Dans le congrès de 1982, le recours massif au pronom *nous* pose une certaine proximité entre l'orateur et le destinataire du message. Progressivement, l'emploi du nom propre *Chine*, tend, au contraire, à manifester un point de vue moins subjectif entre l'énonciateur et le destinataire. On peut avancer l'hypothèse que, le développement économique du pays ouvre progressivement une possibilité d'affirmation nationale plus directe.

2012 #<para=18-07>**中国** 共产党 第 十 八 次 全 国 代 表 大 会 ， 是 在 我 国 进 入 全 面 建 成 小 康 社 会 决 定 性 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 * " 三 个 代 表 " 重 要 思 想 * 科 学 发 展 观 为 指 导 ， 解 放 思 想 ， 改 革 开 放 ， 凝 聚 力 量 ， 攻 坚 克 难 ， 坚 定 不 移 沿 着 **中国** 特 色 社 会 主 义 道 路 前 进 ， 为 全 面 建 成 小 康 社 会 而 奋 斗 。

#<para=18-07>Le XVIIIe Congrès du PCC est un congrès particulièrement important étant donné qu'il a lieu au moment où notre pays est entré dans une phase décisive pour mener à terme la construction in extenso de la société de moyenne aisance. 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, la pensée importante de la "Triple Représentation" et le concept de développement scientifique, libérons notre pensée, persévérons dans la réforme et l'ouverture, rassemblons toutes les forces du pays, vainquons toutes les difficultés, poursuivons résolument la voie du socialisme à la **chinoise** et luttons pour la réussite de l'édification intégrale de la société de moyenne aisance.

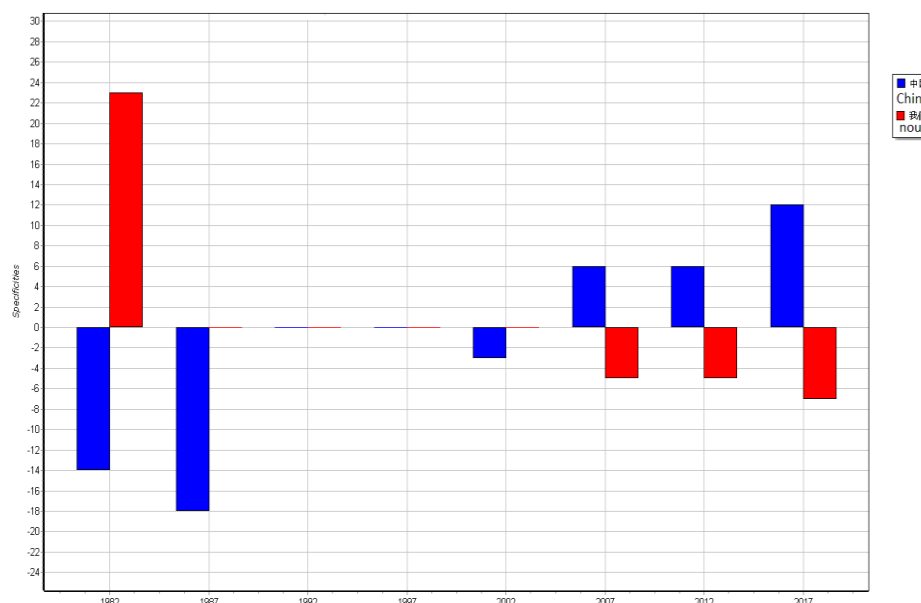


Figure 8

Répartition des formes *nous* et *Chine* dans le corpus *PCC1982-2017*

Conclusion

Comme le lecteur aura pu le vérifier tout au long de cette étude, l'exploration du corpus *PCC1982-2017* à l'aide des méthodes de la textométrie chronologique nous a permis de faire des observations importantes sur l'évolution du vocabulaire et du style d'énonciation employés dans les rapports présentés lors des congrès du PCC.

A chacune des étapes des analyses que nous avons mises en œuvre, nous avons été confrontés au même genre de choix : accepter sans les remettre en cause les idées couramment véhiculées dans l'analyse automatique des textes ou, au contraire, tenter de les confronter aux textes que nous étions en train d'analyser. Dans la plupart des cas, il nous est rapidement apparu que les résultats des analyses que nous faisons nous entraînaient vers des conclusions opposées aux idées que l'on rencontre dans les publications sur l'analyse automatique des textes politiques chinois.

Nous avons analysé les résultats produits par la segmentation en *mots graphiques*, réalisée automatiquement par le segmenteur NLPIR-ICTCLAS. La principale qualité de la segmentation produite par un segmenteur automatique réside dans l'homogénéité des résultats qu'il produit. Cela dit, il apparaît clairement que toute segmentation de ce type s'appuie sur l'existence d'un *dictionnaire opérationnel* dont les limites ont été fixées préalablement. Comme c'est le cas pour tous les dictionnaires, les entrées ont été choisies en fonction de considérations qui ne sont pas toutes d'ordre strictement lexical. Certaines unités composées de plusieurs caractères ont été segmentées en unités distinctes alors que d'autres unités présentant des caractéristiques très semblables aux plans lexicaux et discursifs ont été considérées comme des mots non-segmentables. Les raisons de ces variations sont sans doute à chercher dans les différences de

fréquence qui incitent les auteurs des dictionnaires à inclure certaines unités et à en écarter d'autres, ce que nous avons pu montrer en détail sur l'exemple de la forme 经济 (*économie*). Ces circonstances renforcent notre conviction que, si les unités découpées par le segmenteur constituent une base solide pour l'établissement de typologies, l'exploration textométrique des textes ainsi segmentés ne doit pas se borner à l'étude des seules unités isolées par le segmenteur. Dans les corpus de textes chinois, les variations lexicales doivent être recherchées à la fois en deçà de la segmentation en *mots* (*motifs* répétés dans plusieurs unités différentes : 经济, *économie*, 市场经济, *économie de marché*, 经济发展, *développement économique*) et au delà des limites des *mots* découpés par le segmenteur (社会主义 商品经济, *économie marchande socialiste*).

L'analyse textométrique a mis en évidence la fréquence élevée de nombreux segments, répétés dans le texte. Ces segments correspondent à des tournures plus ou moins ritualisées. Les occurrences de ces segments se retrouvent souvent dans des périodes du corpus qui sont rapprochées dans le temps.

L'analyse chronologique du corpus des congrès nous a permis de mettre en évidence des sous-ensembles de formes dont la fréquence variait fortement au fil du temps. Nous avons tout d'abord noté l'augmentation continue, de formes au contenu sémantique nettement marqué (ex : 创新, *innovation*, 体系, *système*, 生态, *écologie*). D'autres formes sont apparues plus caractéristiques des périodes intermédiaires.

La surprise principale a été la découverte de la diminution, au fil du temps, d'un groupe de formes de caractère plutôt grammatical : 的 (DE1, *de*), 地 (DE2) et 了 (LE) que certains chercheurs ont l'habitude de considérer comme des *mots-outils* incapables de véhiculer un sens quelconque et qu'ils écartent, pour ces raisons, de leurs analyses quantitatives. La synthèse de l'ensemble des phénomènes de ce dernier type nous a permis de conclure que cette diminution simultanée était le signe d'une modification profonde de l'énonciation, entre les premiers congrès et les derniers. Au fil des congrès, le texte devient moins injonctif. Parallèlement, on note un effort de clarté et de concision qui s'accroît dans les derniers textes.

Les embaseurs de l'énonciation varient également au cours du temps. Si dans les premiers textes, le sujet de l'énonciation est souvent le pronom 我们 (*wǒmen*, *nous*), l'emploi de cette forme diminue graduellement au profit du nom propre 中国 (*zhōngguó*, *Chine*) qui prend progressivement en charge le rôle de sujet de l'énonciation. Les premiers textes énoncent des affirmations que le destinataire est incité à prendre en charge, aux côtés de l'énonciateur ; les derniers mettent plutôt le pays tout entier au centre de l'énonciation.

En résumé : les expériences menées le corpus **PCC1982-2017** nous ont permis de progresser dans deux directions différentes. Tout d'abord elles nous ont permis d'explorer un corpus de textes politiques chinois et de faire apparaître les grandes évolutions qui se sont manifestées durant la période couverte par le corpus. Sur le plan méthodologique, l'apport nous paraît encore plus important. Par delà la confirmation de

la validité de l'application des méthodes de la textométrie chronologique aux corpus de textes politiques chinois, cette expérience nous a permis de nous livrer à une réflexion utile sur le statut des segmenteurs automatiques utilisés dans les études automatisées de textes en langue chinoise.

La démarche et les méthodes décrites dans le cadre de notre étude nous paraissent adaptables à une large gamme de corpus écrits dans des langues dont les systèmes d'écriture sont très éloignés entre eux.

Références

- Aglietta M. (2013). « Chine : l'enjeu politique du troisième plenum ». <http://www.cepii.fr/blog/bi/post.asp?IDcommuniqu=263>, consulté le 10 juillet 2018.
- Alleton V. (1984). *Les auxiliaires de mode en chinois contemporain*. Paris : Éditions de la Maison des Sciences de l'Homme.
- Benzécri J.-P. & coll. (1973). *La taxinomie, Vol. I ; L'analyse des correspondances, Vol. II*, Dunod, Paris.
- Darrobers R., Xiao Panes X. (1998). *Éléments fondamentaux de la phrase chinoise*. Paris : éditions You-Feng.
- De Souza S. (2014). *De la statistique textuelle à l'analyse des idéologies : l'exemple du discours révolutionnaire en Amérique latine (1810-2010)*, Thèse de doctorat, Université de Besançon.
- Deroubaix J.-C. (1995) « Quand la Commission se présente devant le Parlement...Étude du vocabulaire des discours de présentation de la Commission européenne et de la Haute Autorité de la CECA. 1958-1993 », in *Actes des 3èmes journées internationales des JADT 1995*, Analisi statistica dei dati testuali , a cura di S. Bolasco, L. Lebart, A. Salem, Editeur CISU (Centro d'Informazione e Stampa Universitaria), Rome, p. 209-21
- Di Y.h. (狄艳华), Yang Z. (杨忠). (2010) « 基于语料库的中国政府工作报告核心主题词研究 » (Analyse des mots-clés basée sur le corpus des rapports de travail du gouvernement chinois). *Foreign language research (外语学刊)*, 6 :69-72
- Labbé D., Monière D. (2003). *Le discours gouvernemental. Canada, Québec, France (1945-2000)*. Honoré Champion, Paris.
- Lafon P. (1981). *Dépouillements et statistiques en lexicométrie*, Slatkine-Champion, 1984, Paris.
- Lebart L., Salem A. (1994). *Statistiques textuelles*, Dunod, Paris.
- Lebart L., Salem A., Berry E. (1991). « Recent development in the statistical processing of textual data », *Applied Stoch. Model and Data Analysis*, 7, p 47-62.
- Marchand P. (2007). *Le grand oral. Le discours de politique générale de la V^e République*. Editions De Boeck Université, Bruxelles.
- Mayaffre D. (2004). *Paroles de Président. Jacques Chirac (1995-2003) et le discours présidentiel sous la Ve République*. Honoré Champion, Paris.

- Miao J. (2012). *Approches textométriques de la notion de style du traducteur - Analyses d'un corpus parallèle français-chinois : Jean-Christophe de Romain Rolland et ses trois traductions chinoises*. Thèse doctorale, Paris 3.
- Miao J., Salem A. (2018). « Série textuelles homogènes ». *Actes JADT-2018* (14ème Journées internationales d'Analyse statistique des Données Textuelles), 12-15 juin 2018, à Rome, UniversItalia : 419-500.
- Qian G. (2017). « 中共历届党代会报告语象分析 » (Analyses lexicales des rapports de tous les congrès du Parti communiste chinois). *Lianhe Zaobao* du 19 novembre 2017.
- Salem A. (1991). « Les séries textuelles chronologiques ». *Histoire & Mesure Année*, Vol. (6) : 149-175.
- Salem A., Wu L.C. (2008). « Essai de textométrie politique chinoise ». In André Salem et Serge Fleury, éditeurs, *Lexicometrica – Explorations textométriques*, Vol. (1). URL : <http://lexicometrica.univ-paris3.fr/numspeciaux/special8.htm> (consulté le 5 février 2017).
- Tang H.f. (唐皇凤), Tao J.w. (陶建武). (2016) « 建国以来中国共产党执政理念的现代演进——基于历届党代会工作报告的词频分析 » (L'évolution moderne du concept de gouvernance du Parti communiste chinois depuis la fondation de la République populaire de Chine – Analyse basée sur la fréquence des mots dans les rapports de travail des Congrès du Parti communiste chinois). *Zhejiang social sciences* (浙江社会科学). N°4 : 27-35.
http://www.cssn.cn/mkszy/mkszyzgh/201610/t20161014_3234964.shtml
(consulté le 8 mars 2018)
- Vandepitte M. (2017). « Quatre choses à savoir sur la Chine – dans le cadre du XIXème congrès du Parti ». Traduit par Anne Meert en français du néerlandais. *Investig'Action* du 15 novembre 2017. URL : goo.gl/8fgSkq (consulté le 25 novembre 2017).
- Zhou J.s. (周景升), Li H.t. (李昊天). (2015) « 基于自建语料库的政治文献主题词研究 » (Analyse des mots-clés dans un corpus des discours politique construit par nous-même). *Time Education* (时代教育杂志) 3 :150-151

Logiciels utilisés :

- NLPIR (ICTCLAS) : segmenteur automatique chinois, développé par Huaping Zhang (张华平), à l' Institut de technologie informatique de l'Académie des Sciences de Chine. <http://ictclas.nlpir.org/>
- Lexico5 : logiciel de textométrie réalisé par l'équipe universitaire SYLED-CLA2T, à l'ENS de St. Cloud et à l'Université de la Sorbonne nouvelle - Paris 3. <http://lexi-co.com/>.