L'analyse de données textuelles aujourd'hui : du corpus comme une urne au corpus comme un plan

Damon Mayaffre

▶ **To cite this version:**

Damon Mayaffre. L'analyse de données textuelles aujourd'hui : du corpus comme une urne au corpus comme un plan : Retour sur les travaux actuels de topographie/topologie textuelle. Lexicometrica, 2007, Spécial, pp.1-12. hal-00551468

HAL Id: hal-00551468
https://hal.science/hal-00551468

Submitted on 3 Jan 2011

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.

L'analyse de données textuelles aujourd'hui : du corpus comme une urne au corpus comme un plan. Retour sur les travaux actuels de topographie/topologie textuelle (partie I)

Fac-similé

Damon Mayaffre

CNRS-UMR 6039 Bases, Corpus et Langage

mayaffre@unice.fr

ABSTRACT. *Textual Data Analysis et Lexical Statistics try to consider, from now on, the text as an orderly structure and organized space. Softwares (Lexico and Hyperbase) can testify of the linear organization and the progress of texts and of corpora. So, the traditional Textual Linguistics and Textual Data Analysis move closer to their point of view.*

KEYWORDS : *Text Topology, Text Topography, Textual Statistics, Textual Data Analysis, Lexico, Hyperbase.*

RESUME. *L'Analyse de données textuelles se propose désormais de considérer le texte comme une structure ordonnée ou comme un espace organisé. Les logiciels d'ADT tels Lexico et Hyperbase peuvent en effet témoigner aujourd'hui de l'ordonnancement linéaire et de la progression du texte et du corpus. Ainsi, l'Analyse de données textuelles rejoint la Linguistique textuelle traditionnelle dans ses préoccupations fondamentales.*

MOTS-CLES : *topologie textuelle, topographie textuelle, statistique textuelle, ADT, Lexico, Hyperbase.*

Avertissement : cet article comporte III parties selon le sommaire décrit ci-dessous. Nous publions ici l'introduction et la partie I consacrée à la topographie textuelle.

Introduction
I. Topographie textuelle et segments répétés dans Lexico 3
I.1. Topographie textuelle : vision simple de la chaîne textuelle
I.2. Topographie textuelle : vision complexe de la chaîne textuelle
I.2.1. Topographie textuelle : seuillage
I.2.2. Topographie textuelle : représentation multiple
I.3. Topographie textuelle, segments répétés et stéréotypie
II. Topologie textuelle et profils co-occurrentiels dans Hyperbase
III. Topologie textuelle et mesure de voisinage dans Arboling
Conclusion
Références bibliographiques

Fac-similé

Introduction

La Linguistique textuelle [par exemple : Adam 1999] et l'Analyse de données textuelles [par exemple : Lebart et Salem 1994] se sont longtemps ignorées ; elles se courtisent aujourd'hui, se rapprochent, se fécondent dans une linguistique à finalité rhétorico-herméneutique [Rastier 2001] dont l'objet empirique et théorique est définitivement le texte.

1. Prioritairement concentrée sur la linéarité du texte, la Linguistique textuelle admet en effet désormais que la textualité n'est pas seulement affaire de *séquences*, *d'enchaînements* ou de *progression* mais aussi de *récurrences* graphiques, de *collocations* lexicales, de *réseaux* linguistiques. Ces réseaux linguistiques (typiquement des réseaux lexicaux), sous jacents aux textes et a-séquentiels, les logiciels d'Analyse de données textuelles (ADT) peuvent les objectiver mieux que le logiciel humain, pour peu que les textes ou corpus textuels analysés soient un peu importants. En effet, si la lecture humaine est avant tout syntagmatique, c'est-à-dire sensible, d'abord, au déroulement rectiligne du texte et à la *cohésion* intra- et inter-phrastiques, la lecture numérique des ordinateurs est une lecture paradigmatique susceptible, sur une échelle supra-phrastique (le paragraphe par exemple, la partie, le corpus) et dans une logique non séquentielle, de traiter des co-présences linguistiques attestées, d'étudier la microdistribution des termes (les affinités et répulsions lexicales locales), de mesurer les voisinages autour d'une unité linguistique pivot, de mettre à jour les isotopies ou isotropies[1]. En terme plus précis, [Viprey 2005-a] démontre que la lecture humaine est avant tout « *linéaire* » lorsque la lecture numérique est « *tabulaire* » et « *réticulaire* ». Partant, lorsque Jean-Michel Adam, l'un des meilleurs représentants de la Linguistique textuelle, admet devant la communauté internationale d'Analyse de données textuelles réunie en 2006 à Besançon que « **la textualité doit résolument être pensée comme la combinaison de parcours linéaires et réticulaires** » [Adam 2006 : 5, souligné par l'auteur], c'est pour réclamer la collaboration devenue inévitable entre tenants de l'approche naturelle du texte – approche linéaire donc, et souvent d'essence qualitative– et les tenants de l'approche assistée par ordinateur –approche réticulaire donc, et souvent d'essence quantitative :

[1] Sur le modèle des isotopies de François Rastier, Jean-Marie Viprey propose le concept d'isotropie [par ex : Viprey 2006 : 81 et ss]. L'isotropie a l'avantage d'opérer à un niveau infra-sémantique en pointant simplement la présence ou co-présence matérielle de formes à l'intérieur du corpus (collocation ou co-occurrence). Pour le débat qui va nous intéresser, il est vrai que les isotopies de Rastier, après celles de Greimas, se repèrent explicitement dans la chaîne syntagmatique ou linéaire du texte, lorsque les isotropies de l'ADT prétendent s'affranchir de cette chaîne pour raisonner de manière non-séquentielle ou réticulaire.

> « Nous avons, de toute évidence, besoin les uns des autres : tandis […] que nous mettons l'accent sur la définition des unités élémentaires, sur le traitement de la linéarité des textes, sur les enchaînements transphrastiques et sur la combinatoire d'unités de rangs de complexité supérieurs à la phrase, vos travaux insistent sur la structure non-séquentielle et réticulaire des textes. » [Adam 2006 : 4 ; propos tenu à la communauté ADT le 19 avril 2006, à Besançon, à l'occasion 8ème JADT]

2. Prioritairement concentrée sur le décompte puis le traitement statistique d'*occurrences*, l'ADT, quant à elle, admet aujourd'hui qu'un texte ou un corpus textuel n'est pas seulement une *urne* anarchique pleine de *données* linguistiques mélangées, mais aussi un *espace* ou un *plan* sur lequel ces données s'enchaînent (plus que s'additionnent) et s'organisent au fil du texte. C'est Etienne Brunet qui le concède en 2006 après une vie de recherche consacrée à l'ADT. Malgré les travaux pionniers de [Lafon 1984] sur les rafales, il regrette en effet que l'ADT se soit « surtout attachée jusqu'ici aux fréquences, sans trop s'occuper des séquences » [Brunet 2006 : 15]. Cette limite demande donc à être dépassée sans quoi l'ADT passera à côté d'un aspect essentiel de la textualité. Précisément, l'objectif de cet article est de montrer comment les travaux les plus novateurs d'ADT de [Lamalle et Salem 2002, Salem 2004 et 2006, Luong, Longrée et Mellet, 2004 et 2006, ou Brunet 2006] visent à compléter l'approche statistique paradigmatique ou non-séquentielle originelle de la lexicométrie par un traitement plus global de la surface des textes et des corpus, à même de rendre compte de leur organisation spatiale, linéaire ou continue : ce que l'on appellera désormais leur organisation *topographique* ou *topologique*[2].

La démonstration s'appuiera directement sur les travaux sus-nommés auxquels le lecteur sera souvent renvoyé. Elle sera outillée, respectivement pour les trois parties de l'article, par trois logiciels qui intègrent des fonctions de topographie/topologie textuelle (Lexico 3, Hyperbase et Arborling) ; trois logiciels, avec quelques autres comme Weblex ou Astartex, qui se trouvent aujourd'hui au cœur d'un projet ANR de textométrie en charge de la refondation de l'analyse de texte assistée par ordinateur.
Enfin, la démonstration se fera sur deux corpus de textes politiques que nous connaissons bien [Mayaffre 2004] : le premier est composé de 816 discours « grand public » des présidents français successifs de la Vème République, entre 1958 et 2002. Le second rassemble 152 discours de Chirac soit la quasi-exhaustivité des interventions télévisées (interviews et allocutions) du président entre 1995 et 2007.

I. Topographie textuelle et segments répétés dans Lexico 3

Adossés au dictionnaire de fréquences, les traitements lexicométriques s'attachent avant tout à décompter le nombre d'occurrences d'un terme ou d'une unité linguistique quelconque (lemme, code grammatical, etc.) dans le corpus, et, sous condition que celui-ci soit découpé en parties, à étudier la distribution fréquentielle du terme ou de l'unité linguistique à l'intérieur des sous-parties constituées. La plupart des calculs –dont le plus célèbre reste celui des spécificités– repose sur cette approche du corpus ; approche contrastive mais statique car uniquement animée par une partition préalable et toujours grossière (en 2, 3, 10,.. parties) du corps de textes étudiés.
Si un certain nombre de traitements prétendent, certes, saisir la dynamique du corpus –la *corrélation chronologique* par exemple ou l'*accroissement du vocabulaire*– constatons donc

[2] On trouvera aussi le terme « cartographie » textuelle (Lamalle et Salem 2002). Topographie, topologie, cartographie textuelles : au-délà des différents termes, on aura compris en tout cas que la prise en compte de l'organisation linéaire ou spatiale du texte dont nous allons traiter n'a rien à voir avec l'analyse syntaxique de la phrase. L'approche des analyseurs syntaxiques, désormais très performante, relève d'une autre philosophie.

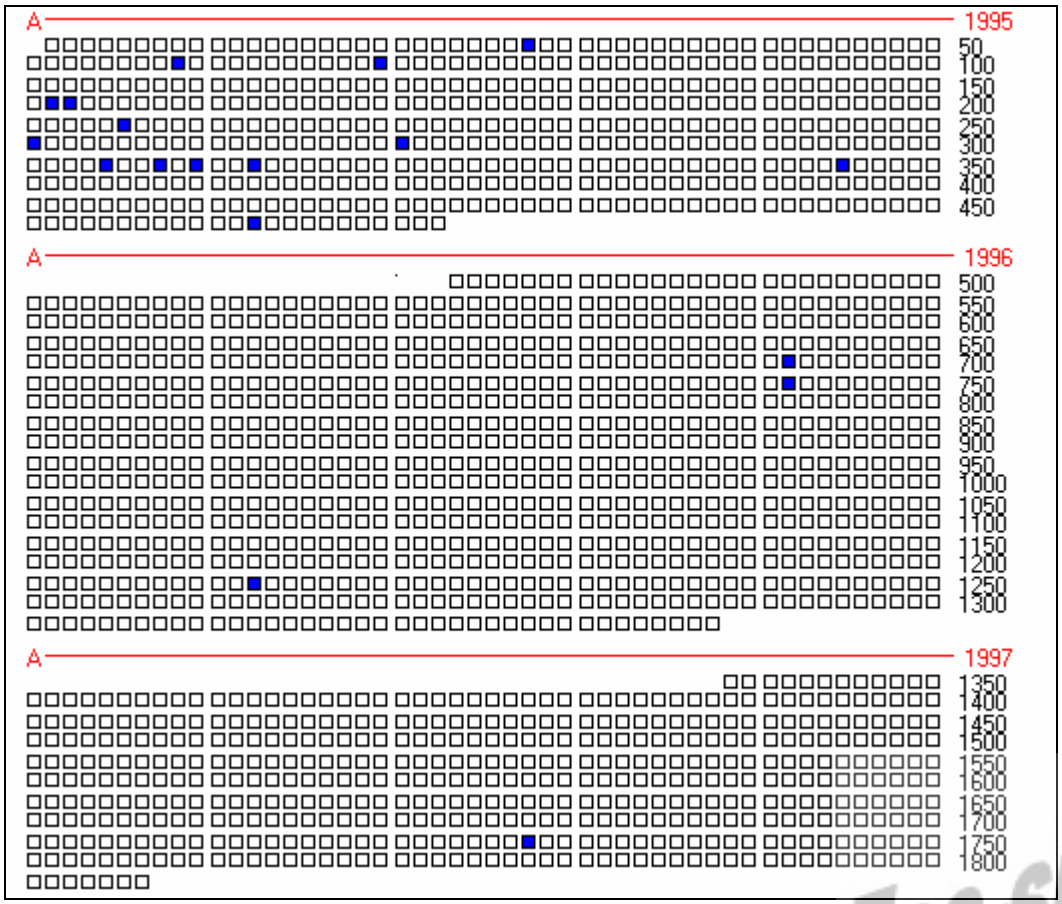
que nos études peinent à prendre en compte sa *chaîne*, ou ce faisant partiellement, restent assujetties au rythme imposé *a priori* par la partition proposée[3].
Bref, la lexicométrie permet de saisir les contrastes d'un corpus partitionné mais non sa progression ou sa séquentialité. On objectera encore que faire contraster deux, trois, quatre… parties contiguës apporte des réponses sur le déroulement du corpus, mais ces réponses seront ponctuelles et toujours conditionnées par une partition, lorsque l'on voudrait réfléchir plutôt en terme de progression linéaire dans le continuum logique que constitue, pour la linguistique textuelle, un texte. (ie : « une **suite** d'énoncés […] posés […] comme un ensemble cohérent **progressant** vers une **fin** » [Détrie, Siblot, Verine 2001 : 349, souligné par nous].

I.1. Topographie textuelle : vision simple de la chaîne textuelle

C'est pour pallier cette faiblesse qu'André Salem propose dans son logiciel Lexico 3, depuis 2001, un outil d'autant plus suggestif d'un point de vue graphique qu'il est économique du point de vue statistique[4].
L'apparition des unités linguistiques (les mots, en l'occurrence) est notée désormais au fil du corpus tels que l'ont montré pour la première fois [Lamalle et Salem 2002 : 409]. Le corpus est représenté, à l'écran, phrase à phrase, ou paragraphe après paragraphe par autant de carrés successifs[5]. Et ces carrés se colorent ou restent vierges selon la présence ou non de la forme linguistique recherchée. L'outil permet donc de localiser et visualiser des formes *dans la suite continue* du corpus. La linéarité du texte n'est plus ignorée mais au contraire restituée comme l'illustre la figure 1 représentant les occurrences du mot « fracture » au fil du discours de Jacques Chirac entre 1995 (l'année commençant en mai après l'élection) et 1997.

Fac-similé

[3] On notera néanmoins les tentatives de segmentations nombreuses et aléatoires, par tranche de 1000 mots par exemple, pour le calcul de l'accroissement lexical. Voir par exemple Labbé 2002 ou Arnold 2005.
[4] Et l'on notera au passage que Salem dont la thèse d'Etat et les ouvrages de référence ont construit un appareil statistique perfectionné, s'oriente aujourd'hui vers toujours plus d'épure mathématique pour donner simplement à voir et laisser librement à penser.
[5] Sur un écran traditionnel, ce sont quelque 3000 carrés-pargragraphes qui peuvent être ainsi embrassés d'un seul coup d'œil, soit plusieurs centaines de pages d'un ouvrage ou d'une oeuvre.



Facsimilé

Figure 1 : "Fracture" dans le discours de J. Chirac (1995-1997). Les carrés représentent les paragraphes successifs du corpus. Ceux qui sont colorés contiennent une occurrence du mot « fracture ».

Le pouvoir descriptif d'une telle représentation parait évident. Il nourrit la compréhension-interprétation des corpus textuels en restituant leur déroulement. Ainsi ici, pouvons nous voir combien le terme ou le thème de la « fracture » sociale, sur lequel Jacques Chirac a été élu, résiste seulement quelques mois après l'élection en 1995, pour ensuite s'évanouir dès la fin de cette année, puis disparaître (3 occurrences survivent seulement en 1996 ; 1 seule en 1997 ; il y en aura 0 en 1998). Le retour aux textes du corpus –dont nous verrons qu'il est facilité par le logiciel– permet de définir précisément un tournant politico-linguistique important de la vie politique française récente. C'est en effet le 5 septembre 1995 à la télévision que Jacques Chirac suggère, à mots encore couverts, une réorientation radicale de la politique qu'il avait préconisée durant la campagne électorale. Quelques jours après, le 26 octobre 1995, par le même média, il annonce clairement une politique d'austérité budgétaire et présente la lutte contre les déficits comme le nouvel axe gouvernemental[6] : c'est à cette période que le terme de « fracture », emblématique des préoccupations sociales, enregistre les dernières occurrences pour *quasi* disparaître du discours.

A l'image de l'exemple qui vient d'être développé, notons que la représentation topographique trouve une pertinence particulière sur les *séries textuelles chronologiques* (ou corpus diachroniques) dont André Salem a décrit les caractéristiques [Habert, Nazarenko, Salem 1997 : 207 et ss]. Les séries textuelles chronologiques sont en effet des corpus

[6] « Et donc la priorité [désormais], c'est la réduction des déficits » (J. Chirac, 26 octobre 1995, entretien télévisé). Ce changement de politique, que symbolisera le « plan Juppé », provoquera immédiatement surprise et mécontentement avec un mouvement de grève de grande ampleur en décembre.

particuliers qui peuvent à juste titre être assimilables à des textes : une des clefs de leur lecture est, précisément, le déroulement, la progression, la séquentialité. On voit à l'œuvre sur ce type de corpus un « temps lexical » –le plus souvent un continuum lexical– peser sur les pratiques discursives. Et la topographie textuelle permet de mesurer, en général en trois temps successifs, l'apparition graduelle, l'affirmation massive, puis la disparition progressive des termes clefs des discours. Dans un premier temps en effet les termes sont diffus, lancés comme des ballons d'essai. Dans un deuxième temps les termes se constituent en thématique majeure du discours comme l'illustrera par exemple, à propos de l'Europe chez Chirac, la figure 2 : la récurrence, et les récurrences en rafales, apparaissent alors comme le meilleur indice de cette thématisation. Enfin, dans un dernier temps, le thème s'épuise et la fréquence des termes s'étiole, les occurrences s'espacent puis disparaissent.

I.2. Topographie textuelle : vision complexe de la chaîne textuelle

La représentation topographique dans Lexico 3 peut être affinée au moins de deux manières. Ceci nourrit la question discutée dans cet article de la construction du sens par la fréquence et la réitération d'occurrences et, plus généralement, de la dimension cohésive des textes que le traitement quantitatif des données textuelles, se propose désormais d'explorer en traitant la chaîne syntagmatique.

Fac-similé

I.2.1. Topographie textuelle : seuillage

L'analyse va au-delà de la notion de présence/absence jusqu'ici mentionnée. Celle-ci est en effet vite insuffisante pour les termes très employés dont la présence se trouve partout dans le corpus. Lexico indique alors, par un jeu de teinte, l'intensité d'utilisation des mots recherchés dans les paragraphes. La figure 2 l'illustre par le terme « Europe » dans le discours chiraquien.

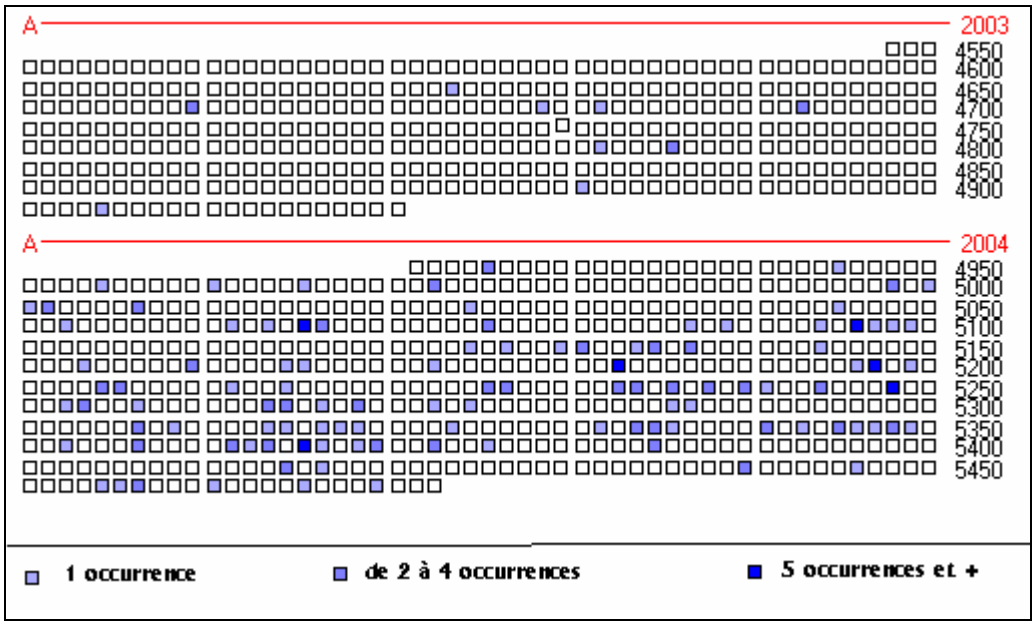


Figure 2 : « Europe » dans le discours de Chirac (2003-2004) avec seuillage

Sans surprise, la campagne électorale pour les élections européennes puis surtout pour le référendum sur le traité constitutionnel voit Chirac marteler le thème dans des émissions spécialement dédiées dans la deuxième partie de 2004 (par exemple le 15 décembre 2004 sur TF1). Les carrés colorés se multiplient et la teinte se fonce pour indiquer que le terme « Europe » est non seulement présent (bleu clair) dans les paragraphes mais parfois répété 2 à

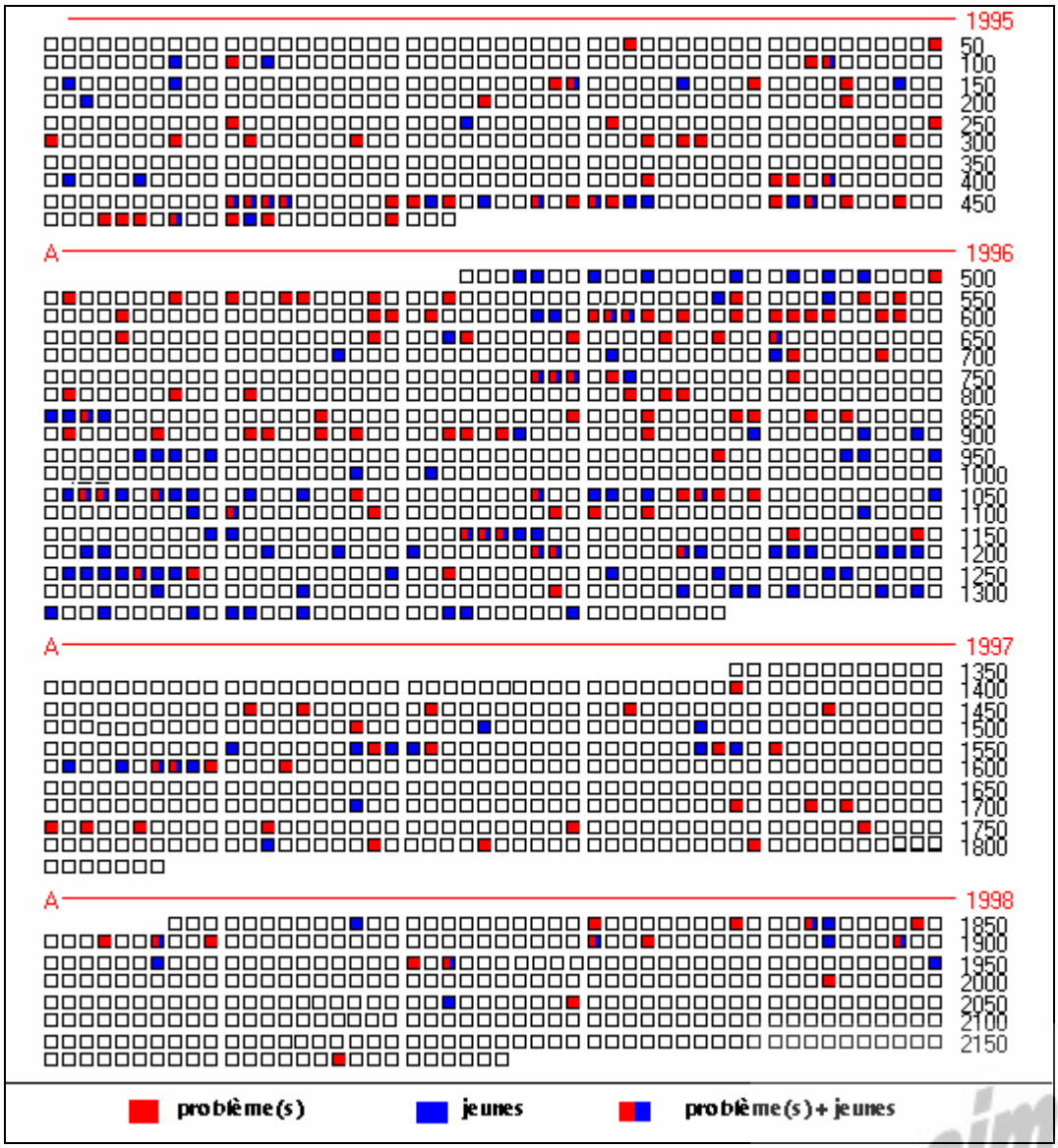
4 fois (bleu roi) voire 5 fois et plus (en bleu foncé). Le recul historique permet de conclure que lors de son premier mandat entre 1995 et 2002, Chirac n'affiche pas des préoccupations européennes très marquées malgré le passage réussi à la monnaie unique [Mayaffre 2004 : 28]. Et la campagne présidentielle de 2002 du candidat Chirac, concentrée sur le thème de l'insécurité, fait en partie l'impasse sur ces préoccupations communautaires. Dans cette logique, durant les premiers mois du second mandat, l'engagement apparaît d'abord diffus comme l'atteste ici la faible récurrence d'« Europe » encore en 2003. Il faudra donc attendre les circonstances particulières du référendum pour que l'Europe devienne une thématique majeure du discours chiraquien, en 2004 donc et plus encore durant le premier semestre 2005. Dans ces conditions, il n'est pas interdit de penser que cette thématisation subite de l'Europe dans le discours grand public du président est apparue contrefaite. Si le discours de Jacques Chirac, comme celui des principaux partis de gouvernement, n'a pas su emporter l'adhésion des citoyens sur le Traité constitutionnel européen c'est sans doute car le texte proposé était politiquement inadéquat, mais aussi parce que les Français n'avaient pas été, assez tôt, ni assez sincèrement, préparés à la thématique européenne.

I.2.2. Topographie textuelle : représentation multiple

Option pertinente, la représentation topographique peut concerner non pas un terme mais deux. Ce croisement n'est pas seulement deux fois plus intéressant en doublant la vision que l'on peut avoir des mots dans le fil du corpus : il nous fait changer de paradigme en passant d'une vision occurrentielle à une vision co-occurrentielle –c'est-à-dire déjà sémantique– du corpus.

Sans qu'une mesure de voisinage au sens mathématique ne soit ici mise en place comme dans les travaux de [Longrée, Luong et Mellet 2004 et 2006] (*cf. infra*), les co-présences (ou au contraire les répulsions) linguistiques peuvent être constatées. Les profils topographiques des mots sont visuellement comparés comme le montre la figure 4 qui superpose la distribution linéaire de « problème(s) » et de « jeunes » au fil du corpus Chirac.

Fac-similé



fac-similé

Figure 3 : « Problème(s) » et « jeunes » dans le discours de Chirac (1995-1998)

Inutile, ici, d'insister sur le fait que « jeunes » et « problème » apparaissent en nombre au même moment dans le texte du corpus, fin 1995, puis en 1996, pour disparaître ensemble, comme si leurs destins lexico-politiques étaient liés dans la bouche du président. Et nous laissons le lecteur formuler des interprétations socio-linguistiques sur une société démographiquement vieillissante ou sur un président âgé, qui problématisent la question de la jeunesse.

Précisons simplement, de manière générale, que l'étude des co-occurrents au sens de [Lebart et Salem 1994 : 312], des collocations au sens de [Hausmann 1979 : 187], ou du profil microdistributionnel des termes (ou co-occurrences généralisées) au sens de [Viprey 2005-b : 258] est un des éléments moteurs de l'analyse de données textuelles. Elle est le premier mouvement pour pointer les réseaux sémantiques qui se forgent dans un texte, ou plus précisément *qui forgent un texte* ; le premier mouvement pour toucher à l'essentiel de ce qu'est la textualité (i.e : ce qui fait d'un texte une suite linguistique signifiante (« une complétude de sens » [Détrie, Siblot, Verine 2001 : 349]) et un assemblage de mots à la fois cohérent et cohésif).

Ici le mouvement pour embrasser la « trame » ou le « tissage » du texte –pour reprendre l'étymologie– est double : l'axe syntagmatique ou fil horizontal (le fil du texte) est considéré

grâce à la représentation topographique, mais se trouve croisé ou tramé avec l'axe paradigmatique ou fil vertical (les entrées lexicales habituellement considérées isolément dans l'index alphabétique ou le tableau des fréquences) grâce à la superposition de plusieurs entrées. Bien sûr le logiciel permet pour l'heure de ne superposer que deux termes dans des paragraphes-carrés bicolores, mais la représentation de trois ou quatre mots (et trois ou quatre couleurs) est envisageable. Surtout, c'est ici qu'il est possible de renouer avec le traitement statistique connu des co-occurrences [Lafon 1981] : dans la fenêtre ou focale choisie (ici le paragraphe représenté par un carré), et à partir d'un mot-pôle recherché (ici, le mot « jeunes »), le logiciel propose de calculer systématiquement les mots associés à ce dernier. Nous trouvons alors à proximité de « jeunes », le mot « problème » donc, mais encore « chômage » ou « banlieue ». A n'en pas douter, une thématique selon les littéraires, une isotopie selon les linguistes se dessinent formellement.

I.3. Topographie textuelle, segments répétés et stéréotypie

Enfin, comme souvent dans les travaux d'André Salem, l'outil prend toute sa valeur lorsque l'unité linguistique traitée n'est pas le mot, mais le segment répété (suite de mots d'une longueur 2, 3, 4, 5) [Salem 1987]. Moins encore que la fréquence d'un mot, la récurrence de segments ne peut être naïvement attribuée au hasard : soit elle pointe une contrainte syntaxique, soit elle indique une détermination ou option sémantique. Dit rapidement, le mot est une unité graphique, le plus souvent ambiguë, sans sens explicite, pas même doté de signification. Le segment, lui, devient une unité linguistique porteuse de sens. La forme « classe » n'a pas de sens, « classe ouvrière » en est doté. Le mot « parti » est ambigu (le substantif vs. le verbe au participe passé) ; et une fois désambiguïsé en substantif par exemple, il reste polysémique (un bon parti, prendre son parti,…). Les contours sémantiques du segment « le parti » ou du segment « le parti communiste », sans parler de « vive le parti communiste français », etc. sont eux plus précis.

Ainsi, l'étude des segments répétés offre-t-elle une alternative à la lemmatisation. Elle permet de désambiguïser les termes de manière formelle et surtout de manière endogène, en corpus et non en référence (arbitraire) au dictionnaire ou à la langue.

Pour le débat qui nous intéresse, il faut remarquer que l'étude des segments répétés restitue, localement, une partie de la chaîne syntagmatique du texte là où l'approche des mots ne le fait pas. Nous touchons là souvent à des unités phraséologiques, riches sémantiquement et intermédiaires entre le lexique et la syntaxe, dont l'inscription se situe bien dans la chaîne du texte.

De la même manière que les occurrences de « fracture » (figure 1) ou d' « Europe » (figure 2) ont pu être cartographiées dans le fil du corpus, nous pourrions donc représenter les occurrences de tel ou tel segment répété («fracture sociale », « la construction de l'Europe », « le problème des jeunes de banlieue »).

L'expérience proposée ci-après est autre. Elle apparaît très suggestive de l'évolution du discours de Chirac.

L'ensemble des segments répétés de longueur importante (6 mots ou plus) ont été rassemblés et peuvent être visualisés dans le discours au cours des trois premières années du deuxième mandat de Chirac (2002-2003-2004).

Fac-similé

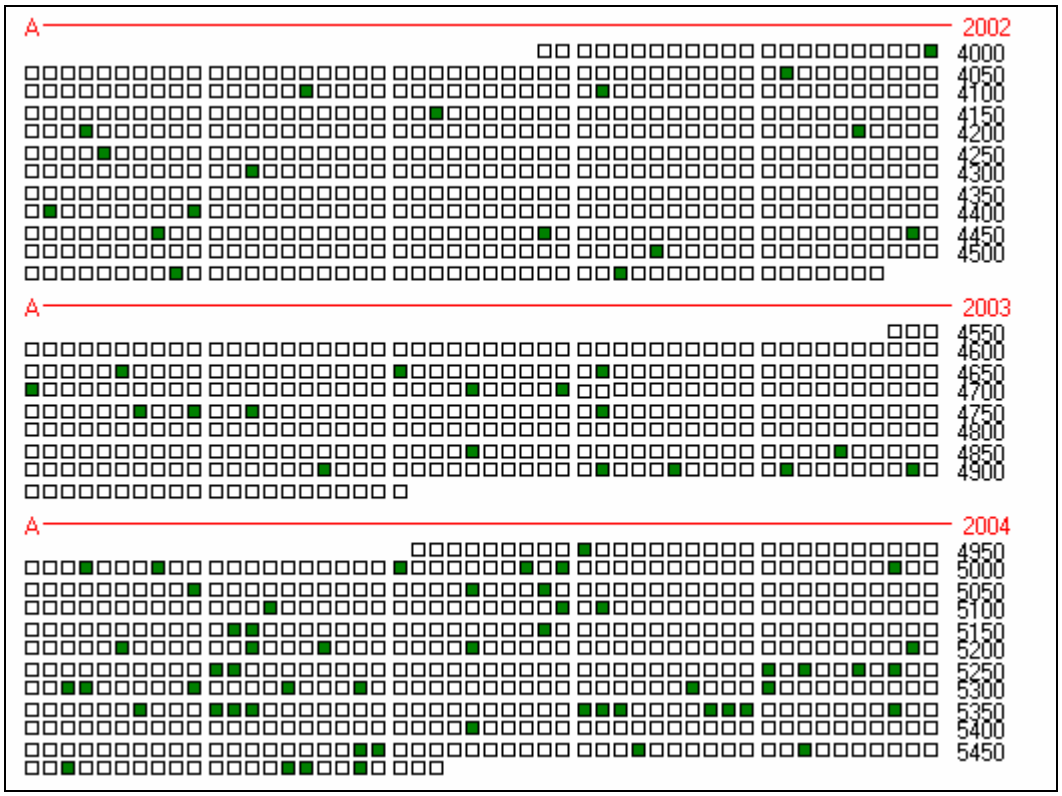


Figure 4 : Segments répétés de longueur 6 et + dans le discours de Chirac (2002-2004)

Les segments répétés de longueur importante sont des tunnels linguistiques dans lesquels la créativité du locuteur recule au profit d'une forme de récitation. Ils apparaissent de plus en plus présents dans le discours grand public de Chirac au fil du temps. Le discours se stéréotypie au cours du second mandat. Le président, lors d'interviews télévisées en partie improvisées (l'interview annuelle du 14 juillet par exemple) est en mal, au fil des années, de créations linguistiques et meuble, de plus en plus, son discours de formules lexico-syntaxiques toutes faites : « mes chers compatriotes de France et d'outre-mer » (11 occurrences), « ce que je peux vous dire c'est que » (17 occurrences), « je vous l'ai dit tout à l'heure » (12 occurrences), « à chacune et à chacun d'entre nous » (11 occurrences), « c'est la raison pour laquelle, je », « je vais vous dire une chose » etc.

Dans l'impossible définition de la langue de bois, qui ne (dis)qualifie souvent que la langue de l'adversaire, peut-être tenons-nous ici un élément formel. Un discours fait de formules reprises et convenues qui finissent par occuper l'essentiel de l'espace du texte au détriment de constructions libres. Un discours phraséologique qui s'appuie d'avantage sur les compétences élémentaires du locuteur que sur la performance créative de l'orateur. Un discours creux où les formules les plus lourdes sont les moins chargées de sens.

*

Fac-similé

Les outils topographiques s'avèrent des outils précieux dans la représentation cartographique du corpus et l'appréhension du texte dans sa continuité. Ils complètent les outils traditionnels de la statistique textuelle, performants dans la description fréquentielle d'occurrences ou de co-occurrences.

Si l'on admet que ce qui fait d'un recueil de mots un texte est la cohésion/cohérence de l'ensemble, ou que l'organisation logico-sémantique est l'aspect essentiel de la textualité ; si l'on admet, en suivant, avec la plupart des auteurs, que les effets cohésifs doivent d'abord se repérer sur la chaîne du texte et que les effets de sens doivent se penser d'abord en terme de

progression thématique ; si l'on admet enfin que cette progression thématique repose sur la récurrence ou l'itérativité de termes au fil du corpus et leur combinaison avec d'autres termes dans un entrelacs linéaire et réticulaire ; alors les programmes de Lexico, tels qu'on a pu les décliner et tels qu'ils sont regroupés dans le logiciel sous la fonction « carte de section », instruisent le débat.
Reste qu'aucune analyse textuelle ne peut faire l'économie d'un retour au texte, sans lequel les unités linguistiques, dé-con-textualisées, ne donnent pas accès au sens. C'est pour cette raison que l'ergonomie du logiciel permet par simple clic sur le carré-paragraphe de la carte de section d'être projeté dans le texte même et de le lire.

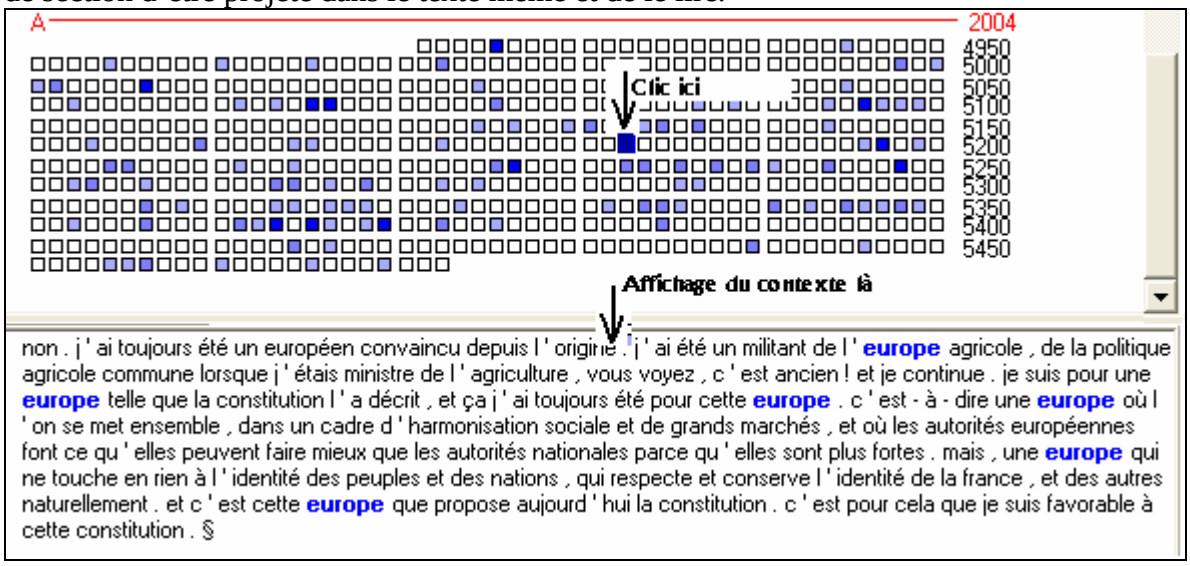


Figure 5 : « Europe » avec retour au (con)texte dans le corpus Chirac. (Rappel : le bleu foncé du carré indiquait 5 occurrences et + du mot « Europe » dans le paragraphe)

Moins que jamais, l'opposition entre quantitativistes et qualitativistes, entre lecture numérique et lecture naturelle, entre approche tabulaire et approche linéaire n'a de pertinence. Appuyés, certes, sur un repérage fréquentiel des occurrences, assistés certes par la statistique, outillés certes par l'hypertextualité informatique, les chercheurs en ADT entendent partager avec la Linguistique textuelle une posture philologique : les efforts pour prendre en compte désormais la chaîne du texte, comme le retour ultime vers le texte dans son ensemble, en sont le témoignage.

Fac-similé

Références bibliographiques

Adam J.-M. (1999). *Linguistique textuelle. Des genres de discours aux textes*. Paris : Nathan.
Adam J.-M. (2006). « Autour du concept de texte. Pour un dialogue des disciplines de l'analyse de données textuelles », conférence d'ouverture aux JADT 2006 [texte en ligne sur Lexicométrica (http://www.cavi.univ-paris3.fr/lexicometrica/jadt/JADT2006-PLENIERE/JADT2006_JMA.pdf)].
Arnold E. (2005). « Le discours de Tony Blair », *Corpus*, 5, pp. 55-78.
Brunet Ét. (2006). « Navigation dans les rafales » *in* J.-M. Viprey (textes réunis par), *JADT' 06*. Besançon : Presses universitaires de Franche-Comté, vol. 1, 15-29.
Habert, B., Nazarenko A. et Salem, A. (1997). *Les linguistiques de corpus*, Paris : Colin.
Haussmann F. (1979). « Un dictionnaire des collocations est possible ? », *Travaux de linguistique et de littérature*, 17-1, 187-195.
Labbé C. Labbé D. & Hubert P. (2002). « Segmentation automatique des corpus » *in* A. Morin et P. Sébillot (éds), *JADT 2002*. Saint-Malo : IRISA-INRIA, vol. 1, 359-349.
Lafon P. (1981). « Analyse lexicométrique et recherche des cooccurrences », *Mots* 3, 95-148.

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

Lamalle C. et Salem A. (2002). « Types généralisés et topographie textuelle dans l'analyse quantitative des corpus textuels » *in* A. Morin et P. Sébillot (éds), *JADT 2002*. Saint-Malo : IRISA-INRIA, vol. 1, 403-411.

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

Longrée D., Luong X. et Mellet S. (2004). « Temps verbaux, axe syntagmatique, topologie textuelle : analyses d'un corpus lemmatisé » *in* G. Purnelle, C. Fairon, A. Dister (éds), *JADT04*. Louvain : Presses universitaires de Louvain, vol. 2, 743-752.

Longrée D., Luong X. et Mellet S. (2006). « Distance intertextuelle et classement des textes d'après leur structure : méthodes de découpage et analyses arborées » *in* J.-M. Viprey (textes réunis par), *JADT' 06*. Besançon : Presses universitaires de Franche-Comté, vol. 2, 643-654.

Mayaffre D. (2004). *Paroles de président. Jacques Chirac (1995-2003) et le discours présidentiel sous la V*^{ème} *République*. Paris : Champion.

Rastier, F. (2001). *Arts et sciences du texte*. Paris : Puf.

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

Salem A. (2004). « Introduction à la résonance textuelle » *in* G. Purnelle, C. Fairon, A. Dister (éds), *JADT04*. Louvain : Presses universitaires de Louvain, vol. 2, 986-992.

Salem A. (2006). « Proximité segmentale » *in* J.-M. Viprey (textes réunis par), *JADT' 06*. Besançon : Presses universitaires de Franche-Comté, vol. 2, 843-853.

Viprey J.-M. (2005-a). « Philologie numérique et herméneutique intégrative », in Adam J.-M. et Heidmann U. (éds.), *Sciences du texte et analyse de discours*. Genève : Slatkine, 51-68.

Viprey J.-M. (2005-b). « Corpus et sémantique discursive : éléments de méthode pour la lecture d corpus », *in* A. Condamines (dir.), *Sémantique et corpus*. Paris : Lavoisier, pp. 245-276.

Viprey J.-M. (2006). « Structure non-séquentielle des textes », *Langages*, 163, 71-85.

Fac-similé