Mettre en évidence le temps lexical dans un corpus de grandes dimensions : l'exemple des débats du Parlement européen

Sascha Diwersy, Giancarlo Luxardo

Praxiling (Université Paul-Valéry Montpellier 3, CNRS),
sascha.diwersy@univ-montp3.fr, giancarlo.luxardo@univ-montp3.fr

Abstract

Within the framework of the French school of discourse analysis, two main methods borrowed from multivariate data analysis have been applied to the processing of text corpora: specificity analysis and correspondence analysis. In a complementary perspective, we present a classification technique specially dedicated to corpora ordered according to a chronological variable: variablility-based neighbour clustering (VNC), introduced by Gries & Hilpert (2008; 2012) as a periodisation tool in the area of diachronic corpus linguistics. This classification method is applied to the processing of a large corpus, i.e. the debates in plenary sitting of the European Parliament between 1996 and 2011. As an example, we study the chronological variation of the word *civilisation*, as it is shown through its collocational stocklist partitioned by year.

Résumé

L'École française d'analyse du discours a appliqué au traitement des corpus textuels deux principales méthodes empruntées à l'analyse de données multivariée : l'analyse des spécificités lexicales et l'analyse factorielle des correspondances (AFC). Dans une perspective complémentaire, nous présentons un procédé de classification spécifiquement adapté aux corpus ordonnés suivant une variable chronologique : la classification ascendante hiérarchique par contiguïtés (CAHC), introduite par Gries & Hilpert (variablility-based neighbour clustering, 2008 ; 2012) comme outil de périodisation dans le domaine de la linguistique de corpus diachronique. Cette méthode de classification est appliquée au traitement d'un corpus volumineux, celui des interventions en séance plénière au Parlement européen entre 1996 et 2011. A titre d'exemple, nous étudions la variation chronologique concernant l'usage du mot *civilisation*, tel qu'il se manifeste à travers ses inventaires collocationnels partitionnés par année.

Mots-clés : textométrie, analyse du discours politique, données diachroniques, classification hiérarchique

1. Introduction

Le discours parlementaire a toujours été un genre traité par la statistique textuelle. Dans ce contexte, il s'agit d'abord de fournir aux spécialistes en sciences politiques des outils construits autour de mesures quantitatives pour catégoriser ou comparer les positionnements de partis politiques.

Avec la disponibilité accrue de données ouvertes et massives dans le domaine politique, on peut prévoir le développement de nouvelles applications caractérisées par des volumes de textes toujours plus importants et une multiplicité de métadonnées les rendant plus accessibles à un large public. Nous nous intéressons ici au corpus des débats du Parlement Européen, lequel, outre son caractère massif, est aussi organisé en tant que corpus multilingue parallèle. Il peut également être traité comme ensemble de données diachroniques, couvrant une période de plus de quinze ans, qui justifie l'emploi de méthodes spécifiques.

Dans le cadre de l'École française d'analyse du discours, la plupart des méthodes de catégorisation proposées dérivent d'une utilisation de l'analyse de données multivariée : principalement, l'analyse factorielle des correspondances (AFC) et l'analyse des spécificités lexicales. De façon complémentaire, on peut y intégrer une dimension linguistique fournie suivant les cas par les cooccurrences, les segments répétés, les parties du discours ou la syntaxe. L'utilité de l'ensemble de ces méthodes a été démontrée également pour l'étude de données relevant d'un partitionnement de corpus suivant une variable temporelle, et ceci notamment dans le cadre des travaux portant sur le temps lexical (Salem, 1988, 1995 ; Mayaffre, 2000). Ces travaux insistent néanmoins sur le fait que le traitement lexicométrique de partitions constituées à partir d'une variable d'ordre chronologique impose une démarche méthodologique particulière, que ce soit dans la mise en œuvre des calculs de spécificités ou dans l'interprétation des résultats d'une AFC, laquelle se voit parfois complémentée par des méthodes de classification hiérarchique comme l'analyse arborée (cf. par exemple : de Sousa, 2012). On se propose ici de comparer l'approche suivie à partir des analyses factorielles avec une méthode de classification hiérarchique reposant sur la nature ordinale des variables chronologiques, le *variablility-based neighbour clustering*, jusqu'à présent expérimentée dans le domaine de la linguistique de corpus diachronique, dans une tradition britannique.

Notre article est structuré de la façon suivante : dans un premier temps, nous présentons la base textuelle (sa composition et son pré-traitement) qui nous a servi de corpus de travail pour les expériences décrites par la suite. La deuxième partie (sections 3 et 4) expose d'abord les enjeux méthodologiques liés au traitement du temps lexical et les approches adoptées de manière canonique dans le domaine de la textométrie avant d'introduire brièvement l'approche du *variability-based neighbour clustering* ou Classification Ascendante Hiérarchique par Contiguïtés. La troisième partie (sections 5 et 6) est consacrée à une étude exploratoire qui présente l'application des différentes méthodes exposées sur l'évolution de la combinatoire lexico-syntaxique du mot *civilisation* dans le Parlement européen au cours de la période allant de l'année 1996 à l'année 2011. La dernière partie (section 7) est consacrée à une conclusion d'ordre général.

2. Le corpus *Eurodisc-fr*

Le corpus *Europarl*[1] a été constitué par Philipp Koehn à partir des transcriptions des débats en séance plénière extraits du site du Parlement Européen. Il rassemble des textes datés entre avril 1996 et novembre 2011, traduits dans la plupart des langues officielles de l'UE (21 langues sur 23 pour la période la plus récente). Il représente un corpus parallèle aligné destiné à être exploité pour des travaux de traduction automatique statistique.

Le présent travail s'inscrit dans le cadre d'une réutilisation du corpus *Europarl* dans une perspective d'analyse de discours. Il dérive de l'intérêt représenté par un corpus volumineux de déclarations politiques prononcées par des élus de différentes nationalités et affiliations sur une période de seize ans. Le corpus est donc d'abord traité en tant que corpus monolingue (le français a été ici choisi, mais d'autres langues pourraient l'être pour des recherches ultérieures), tout en envisageant de traiter les questions de traduction par la référence au texte en langue originale (pour autant que celle-ci soit traçable).

Europarl décrit des sessions, regroupant un certain nombre d'interventions, auxquelles sont associées les métadonnées suivantes : la date de la session ; le locuteur (en général un député

[1] La dernière version en date du corpus est disponible à l'adresse http://www.statmt.org/europarl/.

européen, le nom étant quelquefois remplacé par une qualification comme : « Le Président ») ; son groupe politique au Parlement ; la langue de l'intervention (ces deux dernières métadonnées n'étant pas toujours présentes).

Avec des objectifs de classification et de mise en contraste de différents discours, un certain nombre de modifications ont été effectuées sur le corpus *Europarl* : d'une part, afin de compléter certaines métadonnées manquantes, d'autre part afin d'éliminer certains textes considérés comme non significatifs (par exemple, indications inhérentes aux procédures de transcription et ne pouvant pas être attribuées à un locuteur identifié).

Afin de corriger les métadonnées initiales du corpus *Europarl*, les informations associées à chaque député (actuel ou ancien) ont été extraites du site du Parlement Européen[2]. Elles permettent ainsi de décrire les propriétés suivantes de façon normalisée (ce qui n'est pas le cas avec *Europarl*) :

- le nom du député,
- son pays,
- son affiliation à un groupe politique (éventuellement plusieurs affiliations),
- les dates de début et de fin d'affiliation,
- son statut dans le groupe politique.

En vue de nos expériences, qui portent pour l'essentiel sur l'exploitation de cooccurrences lexico-syntaxiques, le volet français du corpus *Europarl* a également fait l'objet d'un étiquetage en parties du discours, d'une lemmatisation et d'une annotation en relations de dépendance au moyen de la chaîne de traitement Bonsaï (Candito et al., 2010) dans sa version qui fait appel à l'analyseur MaltParser (Nivre et al., 2006).

Le corpus monolingue ainsi modifié est appelé *Eurodisc-fr* dans la suite. Il s'agit d'un volume de données important constitué de : 66 millions de mots-occurrences, répartis sur 957 séances parlementaires, avec au total 205 757 interventions de 4 800 locuteurs.

3. Analyse chronologique et temps lexical

Dans le domaine de la lexicométrie, plusieurs démarches ont été proposées pour étudier les corpus structurés principalement au moyen d'une variable chronologique (corpus diachroniques).

(Salem, 1988), puis (Lebart & Salem, 1994, p. 217) caractérisent un type particulier de tels corpus en tant que « série textuelle chronologique » : (a) ils sont ordonnés suivant une variable chronologique, et (b) ils correspondent à des situations d'énonciation similaires qui assurent une certaine homogénéité (un seul locuteur, un collectif, etc...).

A partir de divers exemples (éditions d'un journal, discours historiques d'une personnalité ou d'une organisation...), il a été démontré que l'étude des variations du lexique au cours du temps (« l'évolution du stock lexical ») met le plus souvent en évidence un phénomène appelé « temps lexical ». Dans cette situation, l'enjeu de l'étude est double : analyser la périodisation de cette évolution et identifier les termes caractérisant les changements intervenant dans chaque période.

[2] http://www.europarl.europa.eu/meps/en/directory.html

Typiquement, une AFC appliquée sur de telles observations rend compte d'une évolution linéaire illustrée par un *effet Guttman* : le deuxième facteur correspond alors approximativement à une fonction quadratique du premier alors que les facteurs suivants ajoutent peu d'information. La disposition des périodes consécutives (points-colonnes) sur le graphique produit et les écarts par rapport à une représentation parabolique permettent dans certains cas de décrire une typologie de la périodisation.

Afin d'identifier les unités lexicales (formes, lemmes ou segments) responsables des variations du lexique, (Lebart & Salem, 1994) utilisent des variantes de l'analyse des spécificités (Lafon, 1981) en faisant des mesures particulières (cf. implémentation dans le logiciel Lexico 3):

- les spécificités évolutives : i.e. les accroissements spécifiques d'une partie par rapport aux parties précédentes (et non par rapport à l'ensemble du corpus),
- les spécificités chronologiques (ou spécificités connexes : Salem, 1988) associées à des groupes de parties contigües et décrivant au mieux les sur-emplois ou sous-emplois lexicaux.

Dans son étude de discours politiques historiques, (Mayaffre, 2000) montre que, même sans recourir à des analyses de spécificités mais en étudiant les fréquences relatives d'apparition de certains termes caractéristiques, il est possible d'identifier des « hiatus chronologiques » associés à des périodes (années) déterminées et ainsi préciser la typologie décrite par l'AFC.

Avec une approche similaire et en faisant intervenir également les résultats d'une analyse arborée de la distance intertextuelle, l'étude de la variation des parties du discours et de l'accroissement lexical (en écart réduit), (de Sousa, 2012) décrit une périodisation en deux niveaux (périodes divisées en plusieurs sous-périodes).

Les démarches précédentes présentent toutefois des limites dans certaines conditions :

- les critères d'homogénéité peuvent être plus ou moins bien satisfaits, par exemple si le corpus est constitué d'énoncés de plusieurs locuteurs,
- la diachronie envisagée peut correspondre à des temps longs (plusieurs décennies ou plusieurs siècles),
- d'une façon générale, quand le corpus étudié est de grande dimension,
- quand l'analyse concerne la cooccurrence (lexico-syntaxique ou textuelle) d'un mot donné, c'est-à-dire elle ne concerne pas l'ensemble du corpus mais un sous-ensemble (dont l'extraction n'est pas nécessairement aisée avec les propriétés définies sur le corpus, mais dépend plutôt du cotexte du mot-pôle).

Les caractéristiques du corpus *Eurodisc-fr* et les objectifs de cette étude amènent donc à envisager de nouvelles approches.

4. Données diachroniques, *variability-based neighbour clustering*

À partir de différentes études sur l'évolution de la langue anglaise, (Gries & Hilpert, 2008, 2012 ; Hilpert & Gries, 2009) ont développé une méthode de classification hiérarchique originale : *variability-based neighbour clustering* (VNC). Il s'agit d'un algorithme qui va chercher à regrouper par étapes successives les périodes adjacentes (contiguës) les plus proches du point de vue des fréquences relatives d'occurrence des termes objets de la recherche. Ici, la classification ne résulte donc pas d'une analyse factorielle. D'autre part, les

coefficients de variation obtenus d'une étape à l'autre permettent de déterminer le regroupement optimal des différentes périodes (de façon analogue aux facteurs résultant d'une analyse factorielle). L'interprétation consiste donc à identifier certaines tendances intervenant d'une classe de périodes à la suivante (par exemple : augmentation ou diminution des usages).

Nous proposons de mettre en œuvre l'approche VNC en l'appliquant, dans la lignée de l'étude de cas présenté par (Gries & Hilpert, 2008, pp. 65-71), non pas à une certaine lexie, mais à un ensemble de vecteurs représentant l'inventaire cooccurrentiel d'une lexie initiale. Nous appelons cette approche : *classification ascendante hiérarchique par contiguïtés* (CAHC).

5. *Eurodisc-fr*, la civilisation en débat

Nous prenons comme exemple d'application l'étude du lemme *civilisation* dans le corpus *Eurodisc-fr*. Il s'agit d'un terme qui est rencontré dans le discours politique, sans que l'on puisse le considérer comme banal : on suppose a priori à la fois une rareté d'utilisation et une instabilité sémantique. Il paraît donc intéressant de s'interroger sur son contexte d'apparition : sa fréquence en fonction des périodes ou selon certains locuteurs susceptibles de démontrer une prédilection pour son usage, en cherchant à discerner les événements pouvant expliquer son émergence ou sa disparition.

Le tableau 1 résume les fréquences d'observation du lemme *civilisation* pour chaque année et les dimensions respectives de chaque partie. Ce tableau confirme que le lemme est relativement peu fréquent ; une simple lecture de ces données ne suggérant pas de variations régulières dans le temps (pour des parties dont le volume varie dans un rapport de 1 à 2).

Total	**1996**	**1997**	**1998**	**1999**	**2000**	**2001**	**2002**	**2003**
1513	61	53	78	69	111	102	98	104
65851316	3140682	4129924	4623246	3656502	4688195	4423296	4089048	4325493
	2004	**2005**	**2006**	**2007**	**2008**	**2009**	**2010**	**2011**
	103	135	134	116	163	82	52	52
	3127041	4212904	4562322	3571944	5220716	4014466	4287911	3777616

Tableau 1. Fréquence du lemme civilisation dans les parties du corpus

Dans la suite, nous nous intéressons en particulier à l'évolution des emplois du mot *civilisation*, telle qu'elle se manifeste à travers les variations de son profil combinatoire[3] dans le discours du Parlement européen pendant la période représentée par le corpus *Eurodisc-fr*.[4] La méthodologie que nous avons mise en œuvre à cette fin repose sur le traitement des cooccurrences lexico-syntaxiques impliquant *civilisation*, que nous avons extraites de façon

[3] Pour la notion de profil combinatoire voir : (Blumenthal, 2002 ; 2006) et (Diwersy, 2012).

[4] L'intérêt primordial que revêt l'analyse des faits cooccurrentiels pour les études lexicométriques est démontré dans (Mayaffre, 2014). A cet égard, notre article se veut une contribution, certes modeste, aux réflexions méthodologiques concernant le traitement de données cooccurrentielles dans une perspective diachronique.

automatique en exploitant les annotations en relations de dépendance présentes dans le corpus (voir supra).

Dans un premier temps, nous avons ainsi établi l'inventaire cooccurrentiel du mot *civilisation* (représenté en tant que lemme) sous forme d'une liste répertoriant, pour chaque partie du corpus définie par année, ses collocatifs, représentés, quant à eux, en termes de triplets constitués par les propriétés <lemme + catégorie grammaticale + relation de dépendance syntaxique[5]>.

Cette liste a été soumise à un calcul de scores d'association que nous avons effectué, dans la lignée des méthodes implémentées par l'analyse dite collostructionnelle (Stefanowitsch & Gries, 2003), au moyen du test exact de Fisher-Yates. Les résultats obtenus pour les mots lexicaux (noms, verbes, adjectifs) retenus en tant que collocatifs du mot-pôle *civilisation* ont été enregistrés dans un lexicogramme (Tournier, 1987 ; Heiden & Tournier, 1998), dont le tableau 2 donne quelques extraits à titre illustratif :

Mot-pôle	**Collocatif**	**Sous-échantillon (année)**	**Fréquence**	**Score Fisher-Yates**
civilisation	européen__A__mod__D[6]	1996	15	16,2417
civilisation	valeur__N__dep__H	1996	5	14,5979
...	...	...	...	...
civilisation	enjeu__N__dep__H	1997	3	10,18
...	...	...	...	...
civilisation	conflit__N__dep__H	2001	6	15,7976
...	...	...	...	...
civilisation	choc__N__dep__H	2003	4	15,6064
civilisation	dialogue__N__dep__H	2003	5	14,5805
civilisation	guerre__N__dep__H	2003	4	10,1864
...	...	...	...	...
civilisation	fonder__V__obj__H	2005	3	8,7267
...	...	...	...	...
civilisation	européen__A__mod__D	2007	12	7,5429
...	...	...	...	...
civilisation	occidental__A__mod__D	2010	6	14,9721
...	...	...	...	...

Tableau 2. Lexicogramme de civilisation

À partir de ce lexicogramme, on génère un tableau de contingence des différents cooccurrents croisés par année[7]. Une analyse des correspondances est produite[8] et elle est illustrée par la figure 1 (les 32 points-lignes fournissant les contributions les plus élevées sont visualisés).

[5] Cette propriété englobe le statut recteur / régi (notés respectivement H (de l'anglais : *head*) et D (de l'anglais : *dependent*) du collocatif ainsi que la fonction syntaxique caractérisant la relation entre celui-ci et le mot-pôle.

[6] La notation « européen__A__mod__D » indique que le lemme *européen* avec la catégorie *adjectif* (A) est en position de *mot régi* (D) par *civilisation*, par rapport auquel il occupe la fonction de *modifieur* (mod).

[7] Le tableau est modifié par la suppression des lignes dérivées du terme *civilisation* (opération de *peeling*).

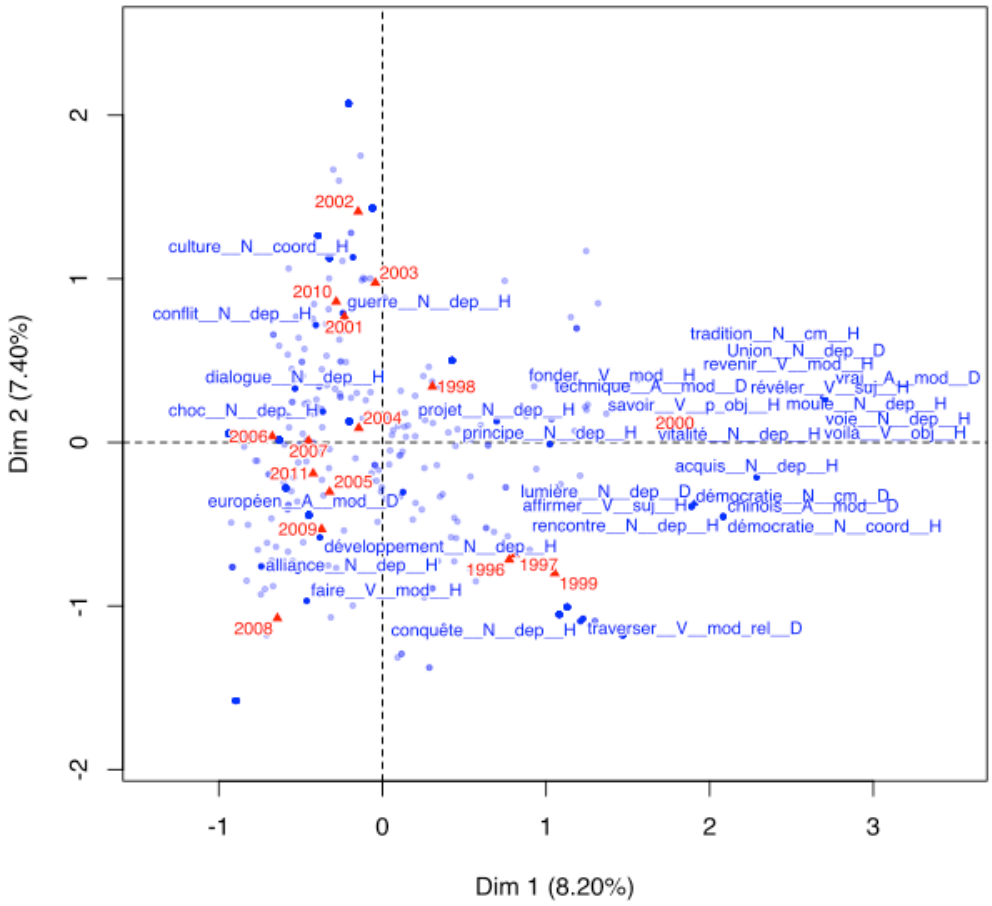

Figure 1. AFC dérivée du lexicogramme de civilisation

Ces résultats ne sont pas faciles à interpréter en raison du faible pourcentage d'inertie fourni (8,20% sur le premier axe, 7,40% sur le deuxième), qui peut s'expliquer par le nombre important de points-lignes. Toutefois, on peut remarquer que :

- le premier axe est construit principalement par une opposition entre les années 1999 et 2000 d'une part (sur des valeurs positives), 2006 et 2008 d'autre part,
- sur cet axe, on trouve d'une part en positif les termes : *acquis*, *fonder, rencontre, démocratie* ; d'autre part : *alliance*, *choc*,
- on ne reconnaît pas un effet Guttman permettant de décrire un temps lexical ; toutefois on peut situer les années 1996 – 2000 dans le demi-plan positif et les autres années dans celui négatif.

6. Évolution de l'inventaire cooccurrentiel de *civilisation*

L'approche CAHC peut s'appliquer au lexicogramme qui a été généré précédemment. Pour cela, on ordonne les cooccurrents suivant le score d'association donné par le test exact de Fisher-Yates. On note par exemple que pour l'année 1996 le cooccurrent de rang 1 est

[8] Calcul et graphique effectués au moyen du package R *FactoMineR* (http://factominer.free.fr/).

l'adjectif : « européen__A__mod__D », pour l'année 2011 il s'agit de : « choc__N__dep__H ».

L'algorithme permettant de produire une classification peut être résumé de la façon suivante :

- générer un tableau croisé indiquant le score d'association Fisher-Yates avec en lignes les cooccurrents et en colonnes les années
- pour chaque année jusqu'à l'avant-dernière :
  - calculer la similarité entre une année et l'année successive en termes du coefficient de corrélation de Pearson (Pearson product-moment correlation coefficient)
  - sélectionner la paire d'années les plus proches et les fusionner en une nouvelle période
  - calculer la moyenne des scores d'association par cooccurrent pour cette nouvelle période
- reprendre ces calculs avec la nouvelle période jusqu'à ce que toutes les périodes aient été fusionnées.

Le dendrogramme décrivant cette classification est reproduit par la figure 2[9].

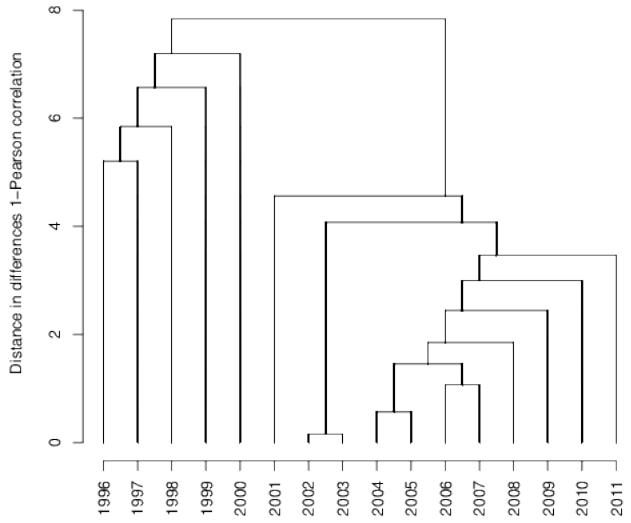

Figure 2. CAHC dérivée du lexicogramme de civilisation

Ce dendrogramme permet donc de visualiser les évolutions et les rapprochements ou écarts entre périodes successives :

- de 1996 à 2000 on observe des années très proches,

[9] La version originale du script R qui nous a permis d'effectuer le calcul CAHC a été créée par Stefan Th. Gries, que nous remercions pour l'avoir mise à notre disposition pour cet article.

- une rupture nette intervient en 2001,
- 2002 et 2003 représentent une évolution (mais sont aussi la paire d'années les plus semblables),
- 2004 – 2005 et 2006 – 2007 reproduisent le même cas de figure que les deux années précédentes,
- les années 2008 – 2011 sont homogènes.

Dans une étape successive, on cherche à identifier les éléments permettant d'expliquer cette périodisation. Avec le vocabulaire mis en évidence par l'AFC précédente, on peut reconnaître deux champs lexicaux principaux et émettre quelques hypothèses:

- jusqu'en 2001, le mot *civilisation* est associé à des cooccurrents véhiculant les notions de démocratie et de progrès en s'appliquant à un contexte européen,
- en 2001, le thème de la géopolitique conceptualisée en termes d'antagonisme intervient et bouleverse les usages,
- cette instabilité se poursuit en 2002 et 2003,
- on note une continuité dans les années 2004 – 2011, correspondant au retour des thèmes initiaux.

Afin d'expliquer de façon plus précise ces rapprochements, nous testons ces hypothèses sur un certain nombre de collocations inclues dans le lexicogramme analysé. En nous inspirant d'une méthode de calcul proposée par (Peirsman et al., 2010, p. 121) pour déceler les changements les plus saillants concernant la constitution de deux inventaires lexicaux ordonnées chronologiquement, nous avons mesuré, pour chaque cooccurrent du terme *civilisation*, la progression en rang (estimée par la différence des logarithmes décimaux des rangs) par paire d'années successives. Si l'on s'intéresse aux changements intervenus au sein de l'inventaire cooccurrentiel de *civilisation*, qui sont à l'origine de la rupture particulièrement prononcée mise en évidence par l'approche CAHC pour la période 2000-2001, on observe une progression marquée des collocatifs liés au champ lexical de la géopolitique (déclinée en divers degrés et formes d'antagonisme), comme l'atteste le tableau 3.

Collocatif	**Score Fisher-Yates**		**Rang**		**Progression en rang**
	2000	**2001**	**2000**	**2001**	**2000-2001**
choc__N__dep__H	3,6191	19,1087	21	1	3,0445
conflit__N__dep__H	0	15,7976	21	2	2,3514
supériorité__N__dep__H	0	11,9672	21	3	1,9459
guerre__N__dep__H	0	11,1451	21	5	1,4351
bataille__N__dep__H	0	7,0238	21	8	0,9651
différence__N__dep__H	0	4,8043	21	13	0,4796
clash__N__dep__H	0	4,6262	21	14	0,3704

Tableau 3. Principales progressions en rang des collocatifs de civilisation entre 2000 et 2001

Dans ce tableau, le score 0 indique l'absence de la collocation dans les lexicogrammes partiels établis par année. Nous avons choisi de limiter à 20 le classement des collocatifs : la valeur 1 correspondant à la principale collocation, la valeur 21 soit à un rang au-delà de 20, soit à son absence. Les graphiques inclus dans la figure 3 permettent de visualiser l'évolution globale des rangs absolus occupés par les collocatifs *choc*, *conflit*, *guerre* et *européen* par année.

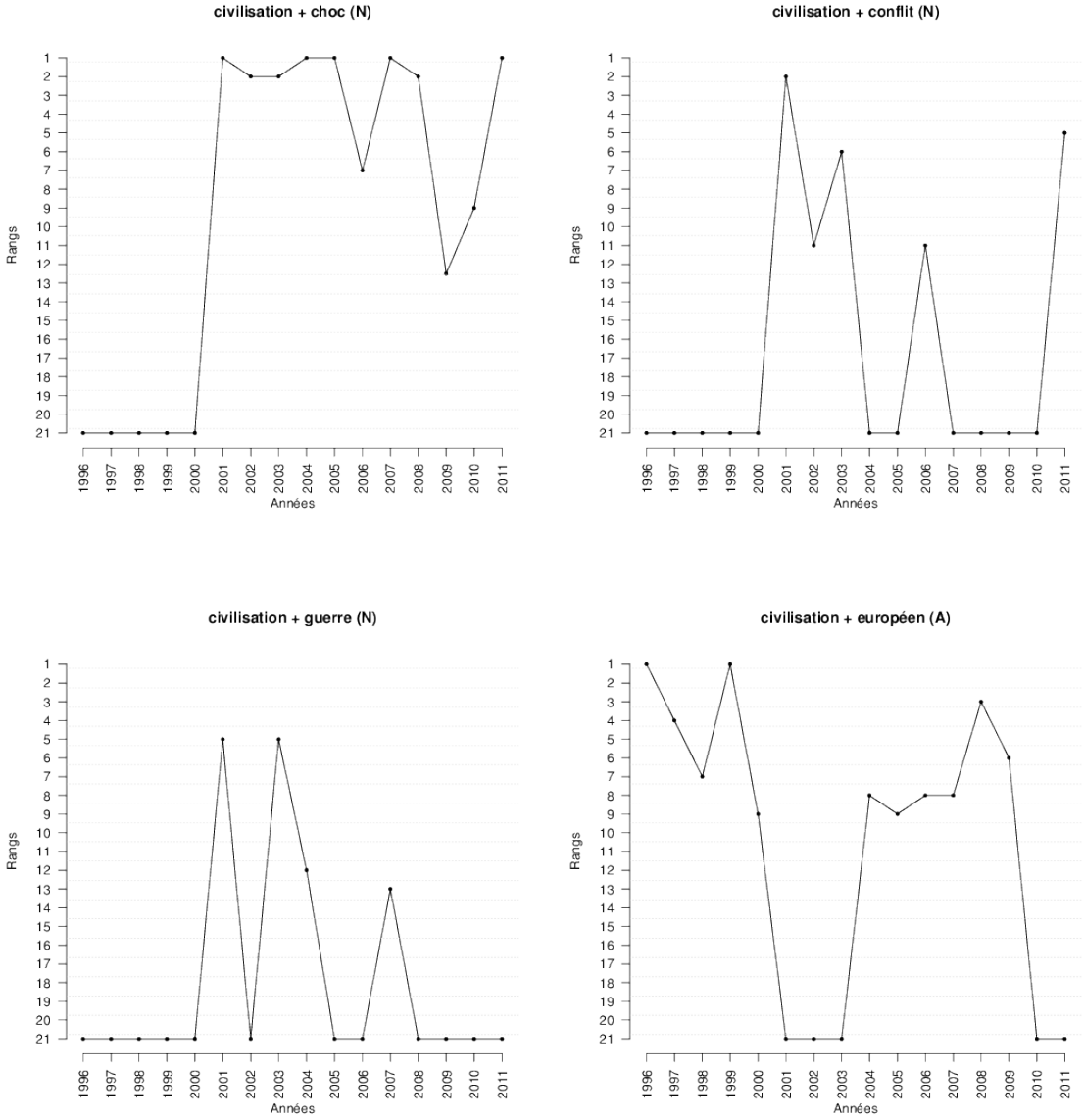

Figure 3. Évolution en rang de quelques collocatifs de civilisation

Les graphiques associés aux termes *choc*, *conflit* et *guerre* (représentant le champ lexical de la géopolitique) confirment qu'ils apparaissent seulement à partir de 2001 et marquent les périodes successives. Il est facile de vérifier que la référence dialogique[10] à l'hypothèse d'un « choc de civilisations » dans le discours parlementaire apparaît après les attentats du 11 septembre 2001 (l'ouvrage de Samuel Huntington publié en 1996 et traduit en français en 1997 n'était jusqu'alors que peu connu dans ce contexte).

A l'opposé, l'adjectif *européen*, bien présent de 1996 à 2000, tend à disparaître de 2001 à 2003 mais réapparaît de 2004 à 2009. S'agit-il du contexte des débats autour du traité établissant une constitution pour l'Europe (en 2003, début de la Convention sur l'avenir de l'Europe), puis du traité de Lisbonne (signé en 2007 et mis en œuvre en 2009) ? L'évolution de la collocation *civilisation européenne* peut aussi être expliquée dans le contexte du Conseil

[10] Pour la notion de dialogisme, voir Bres & Nowakoska (2006). Ici, il s'agit de dialogisme interdiscursif.

Européen d'Helsinki sur l'élargissement de l'Europe de décembre 1999, qui envisage une procédure d'adhésion de la Turquie à l'Union Européenne.

7. Conclusion

Quel bilan dresser de l'analyse outillée de la combinatoire lexicale dans le cadre d'une diachronie courte ? Les méthodes canoniques de la textométrie appliquées au temps lexical imposent à l'analyste – et ceci à juste titre – des contraintes d'utilisation très strictes, qui risquent souvent de n'être remplies que de façon imparfaite. C'est aussi le cas de l'étude présentée dans cet article, qui (a) est basée sur un corpus ne correspondant que partiellement à une Série Textuelle Chronologique (au sens de Salem, 1988) et qui (b), limitée à l'inventaire cooccurrentiel d'un seul mot, ne prend en considération qu'un sous-ensemble (très) restreint des données textuelles exploitables du corpus.

C'est donc par précaution méthodologique que nous avons été amenés à choisir avec la CAHC une méthode peu connue en textométrie, qui, contrairement à d'autres méthodes de classification proposées au delà de l'AFC pour le traitement du temps lexical, présente surtout l'avantage d'avoir été conçue en adéquation avec la nature ordinale de toute série de données chronologiques. C'est en cela que réside aussi une des particularités qu'impose la CAHC à la lecture des résultats qu'elle produit : là où l'AFC, appliquée à des données chronologiques, oriente l'analyste à observer la continuité (en conformité avec une configuration géométrique idéalisée, prenant, en l'occurrence, la forme d'une courbe parabolique), le CAHC incite plutôt à tenir compte, dès le départ, d'un découpage en périodes successives.

De ce fait, l'application, à titre illustratif, de la CAHC à l'inventaire cooccurrentiel du terme *civilisation* nous a permis de mettre en évidence le point de basculement majeur concernant l'évolution des usages de ce mot dans le discours du Parlement européen, phénomène lié à un événement discursif bien identifiable, à savoir « le 11 septembre 2001 ». Comme d'ailleurs dans le cas des AFC calculées à partir d'une Série Textuelle Chronologique, l'utilisation d'une CAHC doit être complétée par une méthode supplémentaire susceptible de cibler, à travers la diachronie entière en question, la dynamique (en termes de progression ou de régression) propre à des items particuliers de l'inventaire observé. Alors que pour l'AFC il s'agit en général soit des fréquences relatives, soit d'une analyse des spécificités, nous avons opté ici pour une méthode appliquée à des vecteurs : les rangs par paires d'années successives. Grâce à la différence entre ces rangs nous avons pu cerner l'émergence d'un champ lexical cooccurrentiel lié à *civilisation*, coïncidant avec une rupture dans les usages observée au moyen de la CAHC et représenté principalement par les termes : *conflit*, *choc*, *guerre*, qui l'associent, dans les périodes ultérieures, de façon plus ou moins stable au thème de la géopolitique, conceptualisée par diverses formes et degrés d'antagonisme.

Notre article, qui se veut avant tout une contribution exploratoire, reste certes lacunaire. Outre l'analyse approfondie des phénomènes discursifs observés, il aurait été souhaitable, d'un point de vue méthodologique, d'aborder aussi la relation entre évolution chronologique et orientations politiques. Il s'agirait là de croiser la méthodologie proposée avec des procédés permettant d'explorer de façon systématique l'interaction de ces deux variables (dont l'observation simultanée présente un grand intérêt pour l'analyse du discours politique), un exercice qui sera à entreprendre dans les contributions qui suivront.

Références

Blumenthal, P. (2002). Profil combinatoire des noms. Synonymie distinctive et analyse contrastive. *Zeitschrift für französische Sprache und Literatur* 112, pp. 115-138.

Blumenthal, P. (2006). De la logique des mots à l'analyse de la synonymie. *Langue française* 150, pp. 14-31.

Bres, J., Nowakowska, A. (2006). Dialogisme : du principe à la matérialité discursive. *Recherches linguistiques* 28, pp. 21-48.

Candito, M.-H., Nivre, J., Denis, P. & Henestroza Anguiano, E. (2010). Benchmarking of Statistical Dependency Parsers for French, in : *Proceedings of The 23rd International Conference on Computational Linguistics (COLING 2010)*. Beijing, China.

de Sousa, S. (2012). À l'épreuve des temps... Temps lexical et temps politique dans le discours de Fidel Castro (1959-2008), in : Dister, A., Longrée, D., Purnelle, G. (eds.) : *JADT'12. Actes des 11èmes Journées internationales d'Analyse statistique des Données Textuelles,* Liège, pp. 337-349.

Diwersy, S. (2012). *Kookkurrenz, Kontrast, Profil. Korpusinduzierte Studien zur lexikalisch-syntaktischen Kombinatorik französischer Substantive (mit ergänzenden Betrachtungen zum Deutschen)*. Berlin e.a. : de Gruyter (= Beihefte zur Zeitschrift für romanische Philologie, 373).

Gries, S.T., Hilpert, M. (2008). The identification of stages in diachronic data: variability-based neighbour clustering. *Corpora* 3 (1), pp. 59-81.

Gries, S.T., Hilpert, M. (2012). Variability-based neighbor clustering: a bottom-up approach to periodization in historical linguistics, in: Nevalainen, T., Traugott, E. (eds.), *The Oxford handbook of the history of English*, Oxford : Oxford University Press, pp. 134-144.

Heiden, S., Tournier, M. (1998). Lexicométrie textuelle, sens et stratégie discursive, in : *Actes I Simposio Internacional de Análisis del Discurso*, Madrid.

Hilpert, M., Gries, S.T. (2009). Assessing frequency changes in multi-stage diachronic corpora: Applications for historical corpus linguistics and the study of language acquisition. *Literary and Linguistic Computing* 24 (4), pp. 385-401.

Huntington, S. (1997). *Le Choc des Civilisations*. Paris : Éditions Odile Jacob.

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

Mayaffre, D. (2000). Temps lexical ou temps politique ? *Lexicometrica* 2, url : http://lexicometrica.univ-paris3.fr/article/numero2/mayaffre2000.PDF.

Mayaffre D. (2014). Plaidoyer en faveur de l'Analyse de Données co(n)textuelles : Parcours cooccurrentiels dans le discours présidentiel français (1958-2014), in : Néé, E., Valette, M., Daube, J.-M., Fleury, S. (eds.): *Actes des 12es Journées internationales d'Analyse Statistique des Données Textuelles (JADT 2014)*, Paris, Inalco-Sorbonne nouvelle, pp. 15-32.

Nivre, J., Hall, J., Nilsson, J. (2006). MaltParser: A Data-Driven Parser-Generator for Dependency Parsing, in : *Proceedings of the fifth international conference on Language Resources and Evaluation (LREC 2006)*. Genoa, Italy.

Peirsman, Y., Heylen, K., Geeraerts D (2010). Applying Word Space Models to Sociolinguistics. Religion Names before and after 9/11, in : Geeraerts, D., Kristiansen, G., Peirsman, Y. (eds.) : *Recent Advances in Cognitive Sociolinguistics*, Berlin: Mouton de Gruyter.

Salem, A. (1988). Approches du temps lexical. Statistique textuelle et séries chronologiques. *Mots* 17, pp. 105-143.

Salem, A. (1995), La lexicométrie chronologique. L'exemple du Père Duchesne d'Hébert, in: *Actes du colloque "Langages de la Révolution (1770-1815)"*, Paris : Klincksieck, pp. 313-328.

Stefanowitsch, A., Gries, S.T. (2003). Collostructions: Investigating the interaction between words and constructions. *International Journal of Corpus Linguistics* 8.2, pp. 209-43.

Tournier, M. (1987). Cooccurrences autour de travail (1971-1976). *Mots* 14, pp. 89-123.