Les manifestations textométriques de la saillance lexicale. Expérimentations et tentative de caractérisation

Marie Veniard[1], Serge Fleury[2]

[1]Université Paris Descartes, EDA, Paris – France

[2]Université Sorbonne nouvelle, Paris – France

Abstract

Salient words are remarkable by their frequency, but both qualitative and quantitative studies underlined other characteristics such as dialogism or collocations revealing ideology. In this study, we investigate three salient words in a press corpus about immigration. We combine an evolutive collocational analysis to a phraseology analysis ("segments répétés"). Our analysis reveal that as frequency rises, so does the number of collocates. On the one hand, they become more diverse, but on the other hand, they stabilize in *ad hoc* chunks. Discourse around the keyword is at the same time more diverse and more repetitive.

Résumé

Nous cherchons à décrire par des outils textométriques (avec *Le Trameur*) les manifestations discursives nées de la saillance lexicale, à savoir l'augmentation de la fréquence d'un mot suite à un évènement ou un fait social donné. A partir d'une analyse des cooccurrents de trois mots qui présentent des pics de fréquence dans un corpus chronologique de discours médiatique sur l'immigration, nous montrerons que la fréquence s'accompagne, au fil du temps, de la diversification des cooccurrents, de la constitution d'un réseau de phraséologie *ad hoc* et de la prégnance plus forte des mots lexicaux sur les mots grammaticaux dans les segments répétés.

Key words : textometry – corpus linguistics – media discourse – frequency – keywords – discourse analysis.

Cet article se propose d'interroger les interactions possibles entre analyse de discours, sémantique et textométrie à partir d'une analyse des manifestations discursives nées de la saillance lexicale, qui se manifeste notamment par l'augmentation de la fréquence d'un mot à une époque donnée suite à un évènement ou un fait social donné. La montée, progressive ou non de l'usage d'un mot, suivie d'une période d'emploi dense et de sa disparition, brutale ou non, peut souvent être rapportée à un évènement externe au corpus, Cette saillance lexicale s'accompagne-t-elle de manifestations spécifiques, que l'on pourrait saisir par la textométrie ? Nous nous intéressons spécifiquement à la cooccurrence en posant la question suivante : la fréquence d'un mot s'accompagne-t-elle de variations sur son environnement cooccurrentiel ? Une augmentation de fréquence (dans des discours publics ou privés) est nécessairement suscitée par une production discursive plus importante, avec tout ce que cela implique comme la multiplication des énonciateurs, la nature éventuellement problématique du référent en question, son axiologisation, la circulation du discours par exemple. La cooccurrence étant étroitement corrélée au texte (Mayaffre et Viprey, 2012), on peut penser qu'une variation dans le nombre et surtout dans la nature des cooccurrents va s'observer en parallèle de l'augmentation de la fréquence. Nous testons cette hypothèse sur un corpus de discours médiatiques dont la nature et l'accessibilité offrent de bonnes conditions pour cette expérimentation. Nous étudierons l'évolution de la cooccurrence de mots qui connaissent un pic de fréquence dans un corpus de discours médiatiques autour de l'immigration (articles

concernant ce sujet, publiés entre 1998 et 2012 dans *Libération* et *Le Figaro*). L'analyse est menée avec le logiciel *Le Trameur*[1] (Fleury et Zimina, 2014).

1. Saillance lexicale et cooccurrence

Nous définissons la saillance lexicale comme l'apparition notable d'une forme lexicale (ou de plusieurs) dans une période de temps donnée, ce qui se traduit par une hausse de fréquence ainsi que par des phénomènes discursifs (énonciatifs, lexicaux, sémantiques notamment). Ce questionnement s'inscrit dans une réflexion plus large sur l'articulation lexique et évènement (Veniard 2013). La saillance se mesure par rapport aux phases « calmes » qui la précèdent et/ou la suivent. Elle est au fondement à la fois des travaux en textométrie et des travaux en analyse de discours « non statistique », chacune de ces traditions ayant mis en évidence des fonctionnements discursifs qui accompagnent l'augmentation de la fréquence et qui dépendent de la nature du discours examiné. Nous ne cherchons pas tant, pour l'instant, à donner une définition statistique de la saillance qu'une définition discursive.

1.1. Saillance et évènement

Le travail d'E. MacMurray (2012) se singularise par l'articulation d'une réflexion et de méthodes textométriques avec un arrière-plan d'analyse de discours et des lectures sur l'herméneutique de l'évènement. Dans l'optique d'exploiter des informations fréquentielles pour mettre au jour des évènements économiques (corpus : *The New York Times*), MacMurray analyse les caractéristiques du *buzz*, supposé marquer un évènement, avec l'ambition de mettre au jour des indicateurs d'évènementialité, qui seraient susceptibles d'être utilisés dans le cadre d'une veille. Le phénomène est abordé dans sa dimension évolutive : le buzz de l'occurrence est toujours contrasté avec le calme de l'avant/après-évènement (MacMurray 2012 : 263). E. MacMurray cherche à montrer « la significativité de la production cooccurrentielle comme indicateur d'un évènement potentiel » (2012 : 222). Elle conclut sur la nécessité d'explorer de manière plus méthodique le rapport entre densité cooccurrentielle et fréquence, mais souligne tout de même qu'il s'agit « certainement d'un indicateur très intéressant pour la détection de *buzz*, surtout lorsque d'autres traitements classiques de la fréquence font défaut » (2012 : 264).

1.2. Questions de recherche

Pour notre part, nous avons cherché à caractériser le phénomène de saillance lexicale né non pas d'un évènement mais d'un fait social, les questions de société autour de l'immigration. L'augmentation de la fréquence d'un mot s'accompagne-t-elle de changements dans le réseau cooccurrentiel qui en détermine l'emploi ? Dans cet article, on fait l'hypothèse que la cooccurrence traduit la routinisation du discours, c'est-à-dire une stabilisation, durable ou non, d'associations lexicales de divers types (voir (Fiala, 1987)). Il peut s'agir d'associations très stables venues du système de la langue (déterminants par exemple). Dans ce cas, l'expression de « routinisation » vise à décrire le fait que l'association déterminant + nom est significative statistiquement, et non le fait linguistique qu'un nom est le plus souvent accompagné d'un déterminant. Il s'agit souvent d'associations qui révèlent des lexies complexes figées (syntagmes nominaux) à une époque donnée (« jeunes issus de l'immigration », « diversité culturelle » ou « l'échec de l'intégration »). Les liens peuvent également être plus lâches et relever d'associations thématiques. Les cas les plus intéressants pour l'analyse de discours sont les cas dans lesquels la relation entre le mot-cible et le

[1] http://www.tal.univ-paris3.fr/trameur/

cooccurrent est encore plus lointaine, symptomatiques d'associations idéologiques qui s'inscrivent de manière quasi subreptice dans le discours (Tournier, 1985 ; Mayaffre, 2008). La mise au jour de ces différentes relations de cooccurrence dépend pour une grande part de l'empan choisi : des cooccurrents proches, dans un empan de quelques mots ou d'une phrase, tendront plutôt à révéler des relations linguistiques ou des lexies figées tandis qu'un calcul sur un empan plus large (le paragraphe par exemple) tendra à révéler des relations plus idéologiques.

Nous ne nous intéressons pas seulement au dernier cas de relations cooccurrentielles mais à toutes les relations en faisant l'hypothèse qu'elles révèlent toutes la présence d'une phraséologie *ad hoc* (une routinisation) qui accompagne l'usage du mot. Il s'agit pour nous de passer d'un calcul sur les formes à une interprétation sur les mots. Dans l'analyse de la cooccurrence, il est d'usage, dans une première phase de l'analyse, de regrouper les listes de cooccurrents en catégories méta-informationnelles de type thématique. Nous proposons de faire précéder cette analyse du contenu des cooccurrents d'une analyse de la nature des cooccurrents : leur nombre, leur récurrence etc. Sans aller trop loin dans une conception mécaniste des rapports entre fréquence et densité de cooccurrents, on peut observer un parallèle entre la fréquence du mot-cible et le nombre de cooccurrents. Les trois images qui suivent illustreront le point de départ de notre questionnement. Le corpus de référence est l'année concernée et le cotexte la phrase.

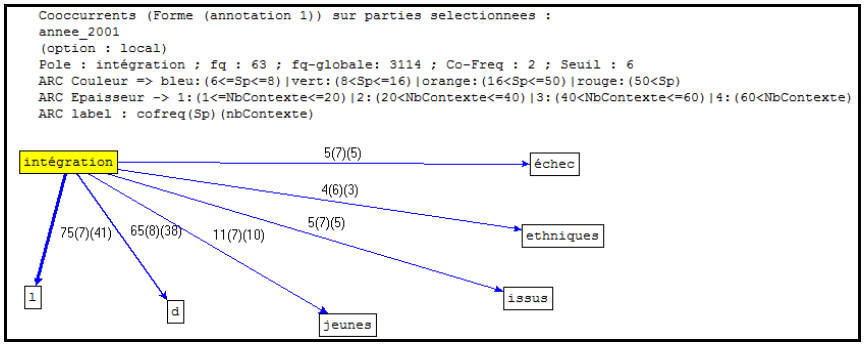

Figure 1 : Cooccurrents de *intégration* (*Le Figaro*) en 2001 (63 occurrences de la forme)

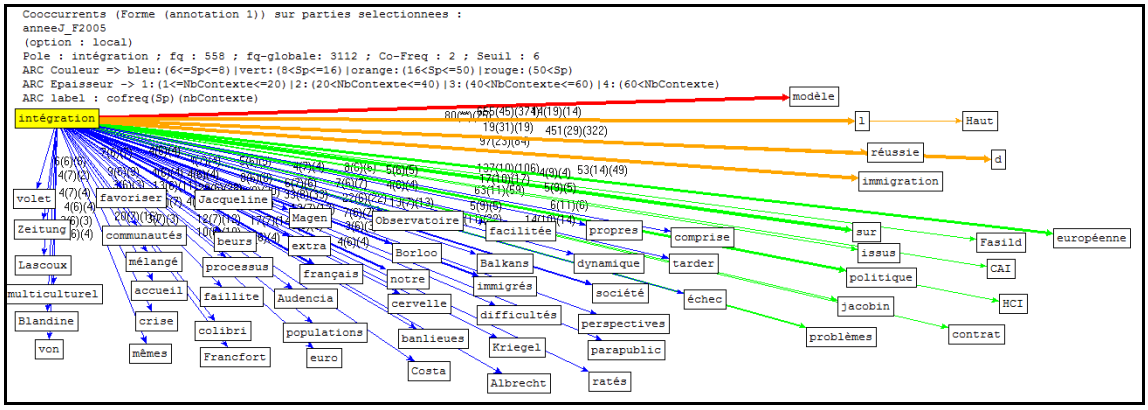

Figure 2 : Cooccurrents de *intégration* (*Le Figaro*) en 2005 (558 occurrences de la forme)

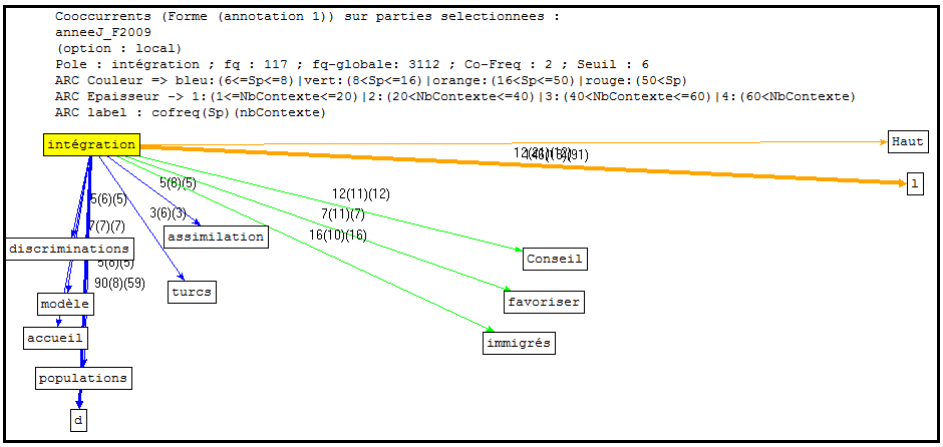

Figure 3 : Cooccurrents de *intégration* (*Le Figaro*) en 2009 (117 occurrences de la forme)

Que peut-on dire de ce foisonnement de cooccurrents observé lors du pic de fréquence (2005) ? Il est probablement corrélé au pic de fréquence, mais la relation entre les deux est-elle strictement mécanique ? Peut-on approfondir l'analyse ? Pour décrire les effets éventuels de la saillance sur l'environnement cooccurrentiel d'une forme, nous poserons les questions suivantes : pendant la période de saillance et par rapport aux périodes de hors-saillance (avant ou après), (1) observe-t-on plus de cooccurrents, à seuil constant ? (2) les attractions cooccurrentielles se renforcent-elles ? (3) ces cooccurrents sont-ils différents de ceux observés dans les périodes de hors-saillance ? Autrement dit, la saillance affecte-t-elle les cooccurrents ?

2. Le corpus

Le corpus est constitué de textes médiatiques autour de l'immigration dans deux journaux d'orientation politique différente, *Libération* (gauche) et *Le Figaro* (droite), ces deux volets du corpus sont parfois notés respectivement LIB et FIG dans ce qui suit. Les articles ont été sélectionnés à partir d'une série de mots-clés thématiques (*immigration*, *immigrant(s)*, *immigré(s)*). De par l'objectif de la recherche et les conditions de recueil du corpus, celui-ci est un corpus thématique. Il vise à donner une représentation des discours auxquels sont exposés les lecteurs d'un journal dans la mesure où ceux-ci lisent/parcourent différents articles, ressortissant de différents genres. Ce corpus prend place dans un projet plus large de comparaison des discours de la presse sur l'immigration dans différents pays européens (Grande-Bretagne, Allemagne, Italie, France), pour lesquels des corpus comparables ont été élaborés[2]. Le corpus a été construit sur la base de données Factiva (https://global.factiva.com), qui présente l'inconvénient de proposer certains articles en plusieurs exemplaires (un nettoyage de ces doublons a été réalisé). Les données récoltées sur Factiva ont été nettoyées, réorganisées avec mise en œuvre d'un balisage classique permettant de construire un corpus chronologique (1998➔2012). Au final, les caractéristiques quantitatives du corpus sont résumées dans les tableaux qui suivent :

[2] Le projet s'intitule « Discourse Keywords in media discourses about migration ». Financé par le programme *Sociétés Plurielles* (Comue Universités Sorbonne Paris Cité) pour l'année 2015, ce projet réunit C. Taylor (University of Sussex), M. Schröter (University of Reading), A. Blätte (University of Duisburg) et Marie Veniard (Université Paris Descartes). Les corpus sont hébergés sur la plateforme The Corpus Workbench.

Part	Tokens	Types	Hapax	Fmax	Type	Tokens
"1998"	466604	34534	15945	22877	de	
"1999"	424400	32729	15032	20632	de	
"2000"	480114	34480	15535	23706	de	
"2001"	432972	32070	14717	21158	de	
"2002"	655069	37549	16294	31521	de	
"2003"	402240	29880	13318	20001	de	
"2004"	472676	33564	15395	22508	de	
"2005"	532860	33785	14858	26204	de	
"2006"	684744	36604	14346	33662	de	
"2007"	761949	38839	15437	37268	de	
"2008"	551909	34006	14253	26557	de	
"2009"	477673	32920	14920	23353	de	
"2010"	447930	32436	14802	21907	de	
"2011"	459111	32220	14507	22071	de	
"2012"	535995	34126	15148	25661	de	
Corpus	7786246	119992	43190	379086	de	

(Libération / LIB)

N	Partie	Occurrences	Formes	Hapax	Fmax	Forme	Occurrences
1	"1998"	51739	9019	5123	2532	de	
2	"1999"	29887	6492	3969	1499	de	
3	"2000"	40982	7879	4657	1907	de	
4	"2001"	220858	21680	10505	11123	de	
5	"2002"	876293	41439	16973	44049	de	
6	"2003"	699450	38033	15928	35609	de	
7	"2004"	1054049	42367	14090	53468	de	
8	"2005"	1171493	38136	7621	60047	de	
9	"2006"	841382	36352	9934	42791	de	
10	"2007"	1020325	40915	12528	50520	de	
11	"2008"	614814	37660	15041	31156	de	
12	"2009"	461085	33562	15330	23523	de	
13	"2010"	430962	30727	14507	22889	de	
14	"2011"	425036	31060	14937	22097	de	
15	"2012"	416745	30315	14651	21550	de	
T	Corpus	8355100	118579	40487	424760	de	

(Le Figaro / FIG)

	Libération	Le Figaro
nombre d'articles	10 797 (chiffre CWB)	11 827 (chiffre CWB)

Figure 4 : Caractéristiques quantitatives du corpus

Dans *Le Figaro*, les parties de 1998, 1999 et 2000 sont peu conséquentes, ce qui peut être imputé soit à un défaut de Factiva, soit à moindre traitement du sujet par le journal. L'expérimentation sera testée sur 3 mots (*intégration, immigration, diversité*). Dans un premier temps, nous avons vérifié que les mots choisis présentaient un profil de saillance du point de vue de la fréquence. Nous présentons les résultats en fréquence absolue pour observer la fréquence du mot lui-même (voir aussi les tableaux en annexe).

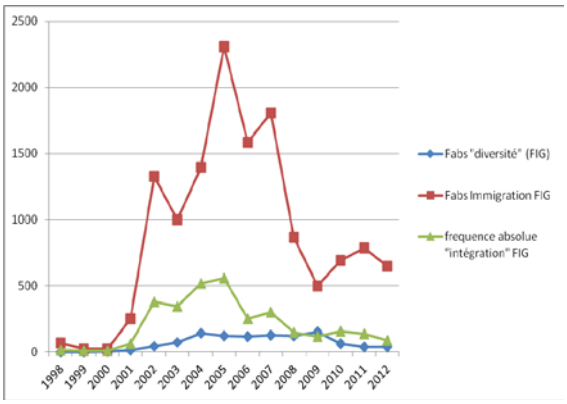

Figure 5 : Fréquence absolue des mots-cibles dans *Le Figaro*

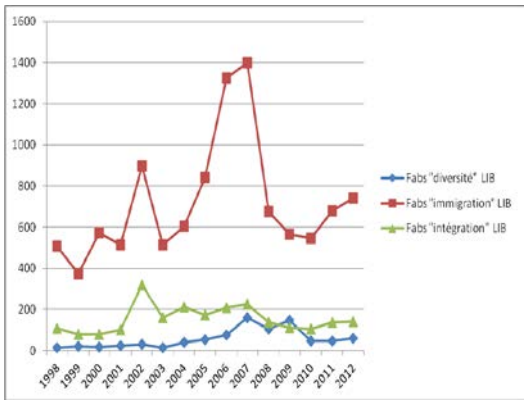

Figure 6 : Fréquence absolue des mots-cibles dans *Libération*

En dépit des différences de fréquence, les trois formes présentent, dans les deux journaux, des variations de fréquence au cours de la période. La forme *immigration* est la plus fréquente, parce qu'elle identifie le thème mais surtout parce qu'elle a servi de mot-clé pour recueillir le

corpus. Elle montre trois pics : en 2002, 2005-2007 et un dernier pic, plus modeste, en 2011-2012, selon les journaux. Son profil fréquentiel est proche dans les deux publications. La forme *intégration* connaît une hausse de fréquence entre 2002 et 2005 dans FIG, et entre 2002 et 2007 dans LIB, phase qui est suivie d'une décroissance moins nette que dans FIG. La dernière forme, *diversité*, connaît des hausses moins marquantes, entre 2007 et 2009 dans LIB et entre 2004 et 2009 dans FIG. Nous ne chercherons pas ici à expliquer ces hausses, ce qui nécessite un retour au texte et au contexte année par année (voir Schröter et Veniard (à paraître en 2016) pour la description de la ventilation de *intégration*). A des degrés divers, les trois formes présentent un profil de saillance du point de vue de la fréquence au sein du corpus. L'expérimentation a pour objectif d'évaluer si la saillance s'accompagne d'autres manifestations que la fréquence, en particulier des phénomènes touchant l'environnement cooccurrentiel du mot.

3. Analyse de la cooccurrence

La recherche s'oriente dans deux directions : une piste quantitative, qui correspond à la description d'une éventuelle augmentation et une piste qualitative, celle des changements éventuels dans les formes cooccurrentes. Nous avons travaillé en gardant des seuils de co-fréquence et d'indice de spécificité constants afin de pouvoir comparer les résultats. Suivant MacMurray (2012 :189), nous avons cherché à trouver un équilibre entre « la lisibilité » des cooccurrents, qui demande des seuils relativement élevés et la « richesse » en cooccurrents, qui demande des seuils relativement bas. Nous avons privilégié la singularité du fonctionnement de chacune des formes saillantes, en adaptant les seuils pour chaque mot.

Nous présentons l'expérimentation mise en place pour les trois mots à partir du cas de *intégration*, qui est assez exemplaire (voir annexe 2, présentation synthétique des résultats). Cette forme est saillante dans le corpus, en raison de divers facteurs externes : augmentation des flux migratoires, politique européenne d'intégration des immigrants (Veniard, 2015), contexte politique national qui a vu le vote de nombreuses lois concernant l'immigration ainsi que des mesures et décisions politiques : création du Haut Conseil à l'Intégration (HCI, 1989), d'un ministère de l'immigration, de l'intégration et de l'identité nationale (2007), etc.

3.1. Observe-t-on plus de cooccurrents pendant la saillance ?

Sans aller trop loin dans une conception mécaniste des rapports entre fréquence et densité de cooccurrents, on peut observer un parallèle entre la fréquence du mot-cible et le nombre de cooccurrents. Dans les 6 cas analysés, nous avons pu observer un parallélisme entre la courbe du nombre d'occurrence et celle du nombre de cooccurrents. Il est particulièrement visible pour la forme *intégration*.

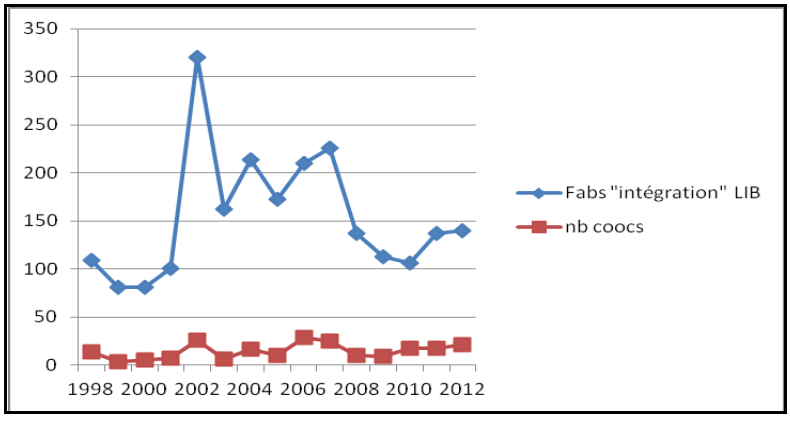

Figure 8 : Fréquence absolue et nombre de cooccurrents de *intégration* dans LIB

Le graphique fait apparaître deux pics, en 2002 et 2006-2007. Ces deux pics sont présents à la fois pour la fréquence absolue et pour le nombre de cooccurrents. Par ailleurs, les deux courbes ont un comportement similaire, même en ce qui concerne le pic plus petit de 2004 et la hausse en fin de corpus. Des décalages sont tout de même observés : la tendance à la hausse en fin du corpus commence en 2011 (106 occurrences en 2010 et 137 en 2011) pour la fréquence mais le nombre de cooccurrents augmente dès 2010 (18 cooccurrents en 2010 et 2011). Si l'augmentation du nombre de cooccurrents en parallèle avec la fréquence semble intuitivement normale et prévisible, nous n'avons à notre connaissance que des indices imparfaits pour confirmer quantitativement cette intuition. Après expérimentation, la proposition de Martinez (2012), qui consiste à diviser le nombre d'occurrences par le nombre de cooccurrents s'est révélée intéressante mais trop délicate à interpréter. Cela nous a amené à faire le choix d'analyser la variation des cooccurrents sous l'angle qualitatif. En effet, s'il est normal et prévisible que le nombre de cooccurrents augmente avec le nombre d'occurrences, rien ne dit que ces cooccurrents doivent se diversifier. Nous faisons l'hypothèse que la saillance liée à un facteur contextuel s'accompagne d'une diversification des cooccurrents.

3.2. Les attractions cooccurrentielles se renforcent-elles ?

En parallèle de la diversification des formes coocccurrentes, on relève pour certains un renforcement de l'attraction avec le mot-cible, marqué par une hausse de l'indice de spécificité. En triant les cooccurrents atteignant au moins une fois sur la période un indice de spécificité très élevé (entre 30 et 50), on obtient, pour chaque journal, une petite série de formes dont on peut observer l'évolution chronologique.

LE FIGARO	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012
l	-	-	-	-	42	32	**51**	45	30	**51**	27	18	35	39	12
Haut					13	32	**44**	19	21	7	21	21	19	18	
modèle					8	7	8	**51**	6	7		7	6	6	
réussie					10	6	10	**31**		23	7				
immigration					22	16	17	23	20	**30**	6		18	26	10

Tableau 1 : Evolution de l'attraction avec le mot-cible *intégration* des formes très spécifiques (*Le Figaro*, indices de spécificité). En grisé, les périodes de pic de fréquence du mot-cible

On observe dans l'ensemble un renforcement des attractions pendant les périodes de forte fréquence (en gris), excepté pour *immigration*. Un retour au texte permet d'identifier les séquences discursives qui sont en grande partie responsables de cette attraction (*ministère/ministre de l'immigration, de l'intégration et de l'identité nationale, Haut Conseil à l'Intégration, modèle français d'intégration, intégration réussie*).

LIBERATION	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012
l	13	14	8	15	22	25	35	18	29	20	16	23	21	**	32
d	10			8	**35**	8	7	18	13	29	12	8	7	6	12
immigration	6				13	6	7	11	21	**44**	7	10	18	20	20
contrat					**				14	10	7		7		

Tableau 2 : Evolution de l'attraction avec le mot-cible des formes très spécifiques (*Libération*, indices de spécificité)

Dans LIB, les attractions les plus fortes sont également relevées pendant les périodes de forte fréquence. En dehors des déterminants, les attractions sont plus irrégulières que dans FIG. On remarque pour le déterminant défini des indices moins élevés avant la période de haute fréquence (entre 8 et 15), mais les indices postérieurs à la première phase de fréquence élevée gardent en quelque sorte la mémoire de la hausse et demeurent élevés. Les concordances sur ces formes révèlent deux types d'unités : des figements qui viennent du contexte, c'est le cas des noms d'institutions ou de mesures politiques, qui sont un indice du sentiment de la saillance du mot pour les locuteurs dans le contexte politique ; et des figements « discursifs », qui viennent de la manière dont on parle du fait social, dont on le construit par le discours (*intégration réussie, modèle français d'intégration*). Les éléments mis au jour pour répondre aux deux premières questions laissent penser qu'en période de pic de fréquence, la saillance se manifeste par la diversification des cooccurrents (plus de cooccurrents différents) et, dans un mouvement inverse, par le développement d'une « phraséologie *ad hoc* », d'une routinisation de l'expression concernant le fait social. Ces observations, menées sur un corpus chronologique, rejoignent celles faites par Née (2012) à propos de *insécurité* et à partir des segments répétés. Les cooccurrents semblent régis, à l'instar du texte, par un principe de reprise et par un principe de nouveauté.

3.3. Les cooccurrents changent-ils de nature pendant les périodes de forte fréquence ?

Les deux journaux partagent de nombreux cooccurrents communs et on assiste tout au long de la période à une politisation du thème de l'intégration, assortie de la création d'institutions et de mesures politiques d'une part et à l'émergence de discours évaluatifs qui servent de base à l'action publique/politique. Ces discours émergent en 2002, comme le montrent des cooccurrents tels que *ratées, panne* (LIB) ou *réussie, ratée, échec, problème* (FIG). Une différence importante existe cependant dans l'environnement cooccurrentiel de la forme *intégration*. Elle est entourée de cooccurrents dès le début de la période dans LIB, et ces cooccurrents actualisent le thème de l'immigration (ex : *modèle, islam, Immigration, jeunes, Haut + Conseil*). Dans FIG, la forme n'a pas de cooccurrents en 1998 et 1999 (avec les seuils sélectionnés). Le seul cooccurrent qui filtre en 2000 est *handicapés*. Ce n'est qu'en 2001 que l'environnement cooccurrentiel actualise le thème, avec *jeunes, issus, ethniques, échec*. Le phénomène de saillance est donc plus complet dans FIG puisque la hausse et le rattachement au thème sont plus progressifs. Bien sûr, le niveau des seuils est arbitraire. Toutefois, une absence ou un faible nombre de cooccurrents indique que ces cooccurrents ne sont pas significatifs compte tenu du seuil posé.

Pour avoir une vue d'ensemble des cooccurrents les plus proches de la forme *intégration*, nous aurons recours, dans un premier temps, aux segments répétés (Salem, 1987). A la différence des cooccurrents, le rapport établi entre les formes dans le calcul de segment répété est moins polysémique du fait du lien très fort établi entre les unités, ce qui est un avantage dans notre entreprise. Le tableau 3 ne montre que les SR dont la fréquence est supérieure à 6. La comparaison entre 2001 et 2002 révèle une nette augmentation du nombre de SR en parallèle avec la hausse de fréquence. En 2001, le SR ne comporte que l'amorce du complément prépositionnel tandis qu'en 2002 apparaissent des groupes lexicaux au milieu de séquences grammaticales : *intégration + des étrangers, des immigrés, européenne* et *réussie*. On en voit dans ce dernier SR la confirmation de l'importance de l'évaluation de l'intégration. Le pic du nombre de SR, en 2004, correspond à une année de très forte fréquence de *intégration* (519 occurrences). En 2010, les SR sont en nombre beaucoup plus limité et constituent quasi-exclusivement en des structures grammaticales à compléter. Toutefois la relation fréquence-nombre de SR n'est pas mécanique. 2005 est l'année de la plus forte fréquence de *intégration* (558) et le nombre de SR est faible (4).

2001	intégration des	7	**2004**	intégration des	47
2002	intégration des	41		intégration de	45
	intégration de	20		intégration et	23
	intégration européenne	19		une intégration	21
	une intégration	15		intégration à	21
	intégration et	15		intégration de la	19
	leur intégration	11		son intégration	18
	intégration de la	11		intégration des immigrés	13
	intégration des immigrés	10		intégration et de	11
	intégration en	8		intégration à la	10
	intégration est	8	**2010**	intégration des	20
	intégration dans	8		intégration et	11
	intégration ne	7		intégration de	10
	intégration et de	6		intégration des immigrés	8
	intégration réussie	6			
	intégration des étrangers	6			

Tableau 3 : Segments répétés comportant *intégration* (*Le Figaro*) en 2001, 2002, 2004, 2010 (fq>6)

La routinisation du discours autour de l'intégration se stabilise progressivement dans FIG, ce que l'on voit en mettant au jour les cooccurrents repris d'une année sur l'autre. Ce sont des cooccurrents qui seront moins dépendants des fluctuations dues aux évènements. La présence d'une forme dans une colonne implique qu'il s'agit d'un cooccurrent déjà présent l'année précédente : « Troyes », présent en 2003, était déjà un cooccurrent en 2002, mais pas en 2001, ce qui explique qu'il n'apparaisse pas dans la colonne « 2002 ».

2002		**2003**		**2004**		**2005**	
l	42	modèle	7	modèle	8	modèle	51
d	26	politique	1	politique	8	Observatoire	6
		l	32	l	51	notre	8
		réussie	6	réussie	10	politique	11
		Haut	32	Haut	44	l	45
		Troyes	8	arrivants	7	réussie	31
		contrat	25	issues	11	société	6
		immigration	16	discriminations	12	Haut	19
		immigrés	14	accueil	20	Kriegel	6
		échec	6	HCI	20	accueil	6
		d	11	contrat	27	HCI	11
		Conseil	13	immigration	17	contrat	10
				échec	13	immigration	23
				Blandine	9	populations	7
				d	20	CAI	9
				Kriegel	9	échec	7
						Blandine	6
						d	29

Tableau 4 (1) : FIGARO. Cooccurrents repris d'une année sur l'autre 2002→ 2005 (forme, indice de spécificité).

Entre 2002 et 2005, un nombre croissant de cooccurrents sont repris, notamment entre 2002 et 2003. En 2002, la fréquence de *intégration* augmente considérablement. Elle passe de 63 en 2001 à 379. Jusqu'à cette date, on l'a vu plus haut, le nombre de cooccurrents est faible (6 en 2001). Il n'est que de 24 en 2002, tandis qu'il sera de 37 en 2003 (pour 345 occurrences de *intégration*, soit moins qu'en 2002). Dans cette phase de consolidation des cooccurrents, le

nombre de cooccurrents repris d'une année sur l'autre augmente. On observe une tendance inverse à la fin de la période, alors que la fréquence baisse globalement (malgré une remontée en 2007).

2006		2007		2008		2009		2010		2011		2012	
modèle	6	politique	6	immigration	6	discriminations	6	accueil	6	immigration	26	immigration	10
politique	11	loi	10	d	13	d	8	d	6	modèle	6	l	12
l	30	immigration	30	Conseil	12	Conseil	11	Conseil	6	l	39	Office	11
société	6	des	6	l	27	l	18	modèle	6	Haut	18		
Haut	21	d	29	réussie	7	Haut	21	l	35	Office	19		
difficultés	7	Conseil	7	Haut	21	immigrés	10	Haut	19	immigrés	7		
Kriegel	6	modèle	7	contrat	21	discriminations	6	HCI	7				
HCI	6	l	51	immigrés	10			immigrés	12				
contrat	8	Haut	7										
immigration	20	contrat	19										
multiculturel	6												
d	24												
faillite	8												
Conseil	11												

Tableau 4 (2) : FIGARO. Cooccurrents repris d'une année sur l'autre 2006 → 2012 (forme, indice de spécificité).

On observe que l'émergence d'un stock de cooccurrents réguliers se fait progressivement à travers une reprise de certains cooccurrents d'une année sur l'autre. Nous souhaitons revenir sur un fait évoqué plus haut : *intégration* a 24 cooccurrents en 2002 (379 occurrences de la forme, partie de 880 000 occs.) pour 37 en 2003 (fréquence de 345, partie de 700 000 occs.) : cette hausse dans le nombre de cooccurrents, dans des parties de taille similaire, confirme, de notre point de vue, la saillance, qui est partiellement décorrélée de la fréquence.

Bilan et perspectives

Les indicateurs de saillance pris individuellement ne fournissent pas des preuves suffisantes, mais ils suggèrent tout de même une accumulation de résultats qui vont dans le sens d'une augmentation de la densité du réseau cooccurrentiel en parallèle de l'augmentation de la fréquence : on observe à la fois une diversification des cooccurrents et la routinisation du discours. L'illustration en annexe propose une visualisation partielle de la phase de croissance de la saillance qui permet de croiser les résultats issus des différents traitements. La taille inégale des parties est une véritable question, cela d'autant plus que l'objet même de la recherche, la saillance dans le discours médiatique, a de grandes chances d'être à l'origine de parties inégales si le corpus est thématique, la saillance suscitant des articles. Ceci dit, on compare des parties de taille similaire, on relève tout de même des phénomènes qui montrent que la problématique peut retenir l'attention. Nous en donnons un exemple plus haut, en 3.3. Question de la taille des parties Deux pistes s'ouvrent à l'issue de cette recherche. La première est d'approfondir la modification apportée à l'usage d'un mot par la saillance née du contexte. Les cooccurrents ne rendent compte que d'associations binaires qui ne donnent qu'un accès partiel au discours. Un calcul de polycooccurrents (Martinez 2012) permettrait de voir si la routinisation du discours affecte des segments plus longs. La seconde piste serait de tenter de calculer l'augmentation du nombre de cooccurrents en rapport avec la fréquence. Cela pourrait se faire au moyen de méthodes statistiques telles que la régression.

Références bibliographiques

Fiala Pierre (1987). « Pour une approche discursive de la phraséologie - Remarques en vrac sur la locutionalité et quelques points de vue qui s'y rapportent, sans doute », Langage et société, Volume 42, Numéro 1, pp. 27-44.

Fleury S. et Zimina M. (2014). « Trameur: A Framework for Annotated Text Corpora Exploration. » *Proc. of COLING 2014 (25th International Conference on Computational Linguistics)*: System Demonstrations, August 2014, Dublin, Ireland, pages 57-61.

Habert B. (1985). « L'analyse des formes « spécifiques » [bilan critique et proposition d'utilisation], *Mots*, 11, p 127-154.

MacMurray Erin (2012). Discours de presse et veille stratégique d'évènements. Approche textométrique et extraction d'informations pour la fouille de textes. Thèse de l'Université Sorbonne nouvelle, Paris 3. En ligne : https://tel.archives-ouvertes.fr/tel-00740601

Martinez W. (2012). « Au-delà de la cooccurrence binaire… Poly-cooccurrences et trames de cooccurrence », *Corpus*, 11, p 191-216.

Mayaffre D. et Viprey J.-M. (dirs.) (2012). « La cooccurrence, du fait statistique au fait textuel », *Corpus*, 11.

Mayaffre D. (2008). « De l'occurrence à l'isotopie. Les co-occurrences en lexicométrie. Syntaxe & Sémantique, 9, pp.53-72.

Née E. (2012). *L'Insécurité en campagne électorale*, Paris, Honoré Champion.

Salem A. (1987). Pratique des segments répétés, essai de statistique textuelle, Paris, Publications de l'INaLF, Klincksieck, Col. "Saint-Cloud".

Schröter M. et Veniard M. (2016, à paraître). « Contrastive Analysis of Keywords in Discourses. *Intégration* and *Integration* in French and German discourses about migration », *Journal of Language and Culture*.

Tournier M. (1985). « Texte « propagandiste » et coocurrences. Hypothèses et méthodes pour l'étude de la sloganisation, Mots, 11, pp.155-187

Veniard M. (2015). « Comparaison de discours institutionnels européens et de discours médiatiques sur la politique d'intégration des immigrants. Le cas de la France, avec référence à la Grande-Bretagne », Colloque Discours sur et de l'Europe. Débats et controverses, Bruxelles, Université LIbre de Bruxelles, 17-18/12/2015.

Veniard M. (2013). *La nomination des évènements dans la presse. Essai de sémantique discursive*, Besançon, Presses Universitaires de Franche-Comté.

Annexes

Annexe 1. Profils quantitatifs des mots choisis sur la chronologie du corpus :

LIBERATION	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012
immigration	507	376	574	515	897	514	605	841	1325	1398	675	567	547	680	741
intégration	4	2	4	4	15	4	19	12	17	52	22	20	13	5	6
diversité	15	20	17	24	30	15	40	54	76	160	106	148	48	46	59

Tableau 5 : Evolution chronologique (en fréquence absolue) dans *Libération*

LE FIGARO	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012
immigration	70	22	24	252	1325	1000	1396	2307	1657	1809	866	500	692	786	646
intégration	15	7	10	63	379	345	519	558	264	300	148	117	157	139	91
diversité	0	0	3	15	42	72	145	122	122	130	121	153	65	39	39

Tableau 6 : Evolution chronologique (en fréquence absolue) dans *Le Figaro*

Annexe 2. Synthèse des résultats sur LIBE

Le graphique qui suit donne à voir une synthèse quantitative du fonctionnement de *intégration* dans LIBE du point de vue de leur évolution quantitative (fréquence et spécificité) en parallèle avec celles de leurs cooccurrents et des segments répétés et sur le début de la chronologie étudiée.