Comparer des AFC de cooccurrence généralisée

Jean-Marie Viprey

ELLIADD-MSHE-UBFC – France

Abstract

In statistic-based discourse analysis, general cooccurrence (Qocc) expresses into a vocable/vocable array dedicated to Correspondence Analysis, which enhances to reticular structure of vocabulary following an explicit set of parameters and a selection of vocables to consider. Many scholars want to compare 2 or n textual sets following this approach. We propose a method based upon *angulation*. Two points in a cloud get an angle by center which is comprised between 0° and 180°, independent of the occurrence of the concerned vocables though representative of the system of gaps from equidistribution. Once that angle is determined for each pair of points in the text A, then in B, from a common list of vocables, one have at disposal $n^2/2$ differences of angles, the means of which for each vocable attests its relative migration inside the cloud, consequently a possible change of its harrissian distribution, *i.e.* of its signification. We propose an application to a Balzac-Stendhal-Flaubert corpus, global then focalized upon the lemma HOMME.

Résumé

En analyse de discours à consistance statistique la cooccurrence généralisée(Qocc) s'exprime par une matrice vocable/vocable destinée à l'AFC, qui met en évidence la structure réticulaire du vocabulaire d'après un paramétrage explicite et un choix des vocables à considérer. Beaucoup de chercheurs souhaitent comparer 2 ou n ensembles textuels selon cette optique. Nous proposons une méthode fondée sur l'angulation. Deux points d'un nuage forment par le centre un angle compris entre 0° et 180°, indépendant de l'effectif des vocables considérés tout en reflétant le système des écarts à l'équidistribution. Une fois déterminé cet angle pour toutes les paires de points dans un ensemble textuel A, puis dans une autre, B sur une liste commune de n vocables, on dispose $n^2/2$ différences d'angles, dont la moyenne pour un vocable témoigne de sa migration relative dans le nuage, donc d'un éventuel changement de sa distribution harrissienne, *i.e.* de sa signifiance. On propose une application à un corpus Balzac-Stendhal-Flaubert, globale puis focalisée sur le lemme HOMME.

Mots-clés : cooccurrence, projection géodésique, distance, distribution, vocable.

1. Introduction

En analyse textuelle de discours à consistance statistique on envisage deux types de distributions : distribution du vocabulaire parmi les partitions du corpus, désormais *macro-distribution* ; distribution harrissienne, au sens de cooccurrence, focalisée ou généralisée – *Qocc* (Massonie 1986, Viprey 2006), désormais *micro-distribution*.

La *Q-occurrence* s'exprime en une matrice vocable/vocable destinée à une analyse multidimensionnelle, ici l'AFC, qui met en évidence la structure réticulaire du vocabulaire d'un ensemble textuel d'après un paramétrage explicite (*empan* de cooccurrence) et un choix des vocables à considérer.

Dans divers champs disciplinaires, on est amené à souhaiter comparer 2 (ou N par paires) ensembles textuels selon cette optique, c'est-à-dire à comparer 2 nuages de points issus de l'analyse de 2 matrices de *Qocc* par l'AFC. On peut vouloir mettre en évidence par ce moyen le contraste entre 2 auteurs littéraires, 2 orateurs, 2 classes de locuteurs, à partir de l'étude micro-distributionnelle d'une liste commune d'items lexicaux fortement occurrents dans

chaque sous-corpus. Cette confrontation de nuages de points sur un plan factoriel ou en géodésique (Viprey 2006) risque de ne conduire qu'à des estimations impressionnistes ou à des jugements très grossiers. De manière générale, un certain nombre d'utilisateurs de l'AFC sont amenés à regretter l'absence d'une méthode synthétique pour une telle comparaison[1]. Le propos de cette contribution est modeste. Il s'agit de proposer une telle fonctionnalité aux chercheurs en analyse de discours qui utilisent de façon privilégiée l'AFC et la *Qocc*, avec *Hyperbase, TXM, Iramuteq* ou tout autre environnement concerné.

Il ne s'agit pas ici de discuter au fond des diverses méthodes permettant de comparer 2 matrices de distribution de même format, ni des vertus et défauts respectifs de ces méthodes liés notamment aux tailles et différences de taille des sous-corpus.

Nous proposons une méthode de comparaison fondée sur l'angulation dans le nuage de points étudié. Dans la géode comme sur le plan factoriel, deux points d'un nuage forment par le centre un angle compris entre 0° et 180°[2]. Contrairement au khi-2, cet angle est indépendant de l'effectif des vocables considérés tout en reflétant fidèlement le système des écarts à l'équidistribution. Une fois déterminé cet angle pour toutes les paires de points dans une projection A (ensemble textuel A), puis dans une projection B (ensemble textuel B) à partir d'une liste commune de vocables, on dispose d'une différence d'angles, qui témoigne pour chaque paire de son changement de statut lexico-thématique de A à B. La moyenne des différences de tous les appariements d'un vocable témoigne de sa migration relative dans le nuage, donc du degré de changement de sa distribution lexicale (au sens harrissien), donc de signifiance.

2. Rappels sur la projection géodésique

Cette proposition a été faite aux JADT 2006 (Viprey 2006). Elle consiste à concevoir le nuage de points résultant des coordonnées sur les 3 premiers axes, qui résument l'essentiel de l'information, dans l'espace d'une géode. On peut projeter chacun de ces points à la surface d'une sphère de même centre, et déployer cette sphère en planisphère selon une méthode de type Mollweide. Des niveaux de gris et des tailles de police permettent de signaler l'éloignement du centre et donc le caractère plus ou moins significatif de la position. Le planisphère doit être lu comme tel, c'est-à-dire que ses bords droit et gauche se rejoignent.

Cette projection présente le double avantage de remédier à l'enchevêtrement du nuage sur le plan de 2 facteurs, et surtout d'augmenter significativement la part d'information représentée. Les zooms régionaux respectent strictement la géométrie de la géode. Contrairement à toutes les tentatives antérieures connues de prendre en compte 3 facteurs (de type *MacSpin*), celle-ci ne présente de risques de mésinterprétation qu'aisément corrigibles. On peut ainsi visualiser sur un écran de 24 pouces un nuage de plusieurs centaines de points, aisément disponible à l'interactivité (repérages, colorations, zooms, clics de navigation…).

3. Principe de l'angulation

Dans l'espace géodésique, deux points forment par le centre un angle compris entre 0° et 180°. C'est cet angle, et le système des $n^2/2$ angles ainsi formés, qui constituent la base du

[1] Voir encore récemment Guaresi 2015 pour un contraste entre candidats député-e-s hommes/femmes.

[2] Valeur éventuellement réductible à un indice compris entre 0 et 1

concept d'AFC, entre conjonction, quadrature[3] et opposition (Cibois 1994). Contraint qu'il est par ses limites définitionnelles, il constitue un indice de (dis-)similarité de profil distributionnel indépendant des variables qui affectent ordinairement les autres indices, comme le Khi-2. Ainsi ne présente-t-il pas de corrélation avec l'effectif des vocables concernés, tout en présentant une corrélation positive avec le Khi-2, qui témoigne que l'essentiel de l'information du tableau de données a bien été retenue (bien au-delà du pourcentage de l'inertie calculé par l'algorithme).

Certes, la (dis-)similarité exprimée est d'autant plus significative que les points sont éloignés de l'origine des axes[4]. Ainsi le produit scalaire des deux vecteurs concernés peut-il apporter une information corrective utile à la seule considération de l'angle géodésique. Le produit scalaire est d'autant plus élevé que l'angle est fermé, mais aussi que les vecteurs sont longs. Cependant, les produits scalaires s'avèrent impossibles à comparer d'une matrice représentée à l'autre, ce qui est l'objet de cette étude.

Autour d'un point projeté sur la sphère, on peut visualiser une zone angulaire plus ou moins étendue qui est une matérialisation de l'entour *isotropique* (Viprey 2005) du vocable considéré. Cet entour est formé des vocables présentant les profils micro-distributionnels les plus similaires dans les limites de l'espace des 3 premiers axes.

4. Objectif de la comparaison

L'objectif est de comparer la structure cooccurrentielle du vocabulaire de partitions de corpus, prises d'abord par paires. Dans cette étude, il s'agit d'un corpus représentatif de trois auteurs narratifs français du XIXème siècle : 13 romans de Balzac(7), Stendhal(3) et Flaubert(3)[5].

Les textes ont été lemmatisés au moyen du logiciel *DiaTag*, de manière semi-automatique afin d'en assurer le contrôle rigoureux. La lemmatisation a paru utile ici pour les raisons de principe désormais largement admises, mais aussi pour des raisons méthodologiques (il était nécessaire de coder les mots composés afin d'éviter de prendre en compte la cooccurrence factice de leurs composants, comme *femme* et *chambre* dans *femme de chambre*) et statistiques (le regroupement des données favorise la constitution d'effectifs critiques en plus grand nombre). Ont été lemmatisées toutes les formes occurrentes, quel que soit leur effectif, dès lors qu'elles étaient susceptibles de se rattacher à un lemme lexical (substantif, adjectif, verbe, certains adverbes) fortement occurrent (fig.1).

	BALZAC	FLAUBERT	STENDHAL	TOTAL
N	777931	253325	585377	1616833
V formes	34952	21928	23353	47916
V lemmes	17342	11427	12021	22425
V non lemmatisé	1411	859	735	2020

Fig.1. Structure quantitative des 3 sous-corpus

Nous avons retenu après la lemmatisation contrôlée 289 lemmes lexicaux[6] dépassant le seuil de 250 occurrences au total, de 50 chez Balzac et de 35 chez chacun des deux autres auteurs,

[3] En géodésique un point est en quadrature avec toute une circonférence.

[4] Nous négligeons ici la distance entre cette dernière et le barycentre (Lebart & Salem 1994 : 89)

[5] De Balzac : *Béatrix, César Birotteau, Les Chouans, Le Père Goriot, Le Lys dans la vallée, Un Ménage de garçon, Les Parents pauvres* ; de Stendhal : *La Chartreuse de Parme, Lucien Leuwen, Le Rouge et le noir* ; de Flaubert : *Madame Bovary, Salammbô, La Tentation de Saint Antoine*. Textes de la base BASILE/ARTFL.

et de 6 dans chacun des romans du corpus. La création de cette liste commune à toutes les explorations suivantes est rendue nécessaire par l'objectif de disposer de matrices de formes rigoureusement semblables, permettant les calculs et mesures de comparaison.

Nous avons comptabilisé en matrices carrées symétriques pour chacun des sous-corpus et pour le corpus entier l'ensemble des cooccurrences de ces 289 vocables, dans un empan de 15 mots à gauche et à droite limité à la phrase.

Ces matrices ont été soumises à l'AFC, dont les résultats ont été projetés selon la méthode géodésique. Les visualisations qui en résultent sont lisibles séparément sans difficulté aussi bien comme synthèses que sous l'angle particulier d'un zoom « régional ». Il est également aisé de se rendre compte d'une stabilité grossière des résultats d'un sous-corpus à l'autre et d'un sous-corpus au corpus entier. Cette stabilité a été mainte fois montrée et expliquée (Brunet 1981), dès lors que l'on étudie dans un genre, une langue et une époque relativement circonscrits.

Néanmoins, bien que la liste des vocables soit identique, et même si (et/ou parce que) cette stabilité s'exprime, il est impossible et il serait imprudent de comparer visuellement, sans autre artefact, les représentations même 2 par 2.

Or les différences micro-distributionnelles, au-delà des contrastes purement quantitatifs en « sacs de mots » traditionnellement mis en exergue, donnent un accès privilégié à la structure lexicale qui caractérisent les univers de langage spécifiques qui motivent et constituent les œuvres de ces auteurs, leurs thèmes. Elles permettent de caractériser finement la signification singulière que l'auteur investit et propose dans tel vocable, dans tel autre, dans telle association. La présente étude, par la modestie de son corpus et de son étendue, n'est que la préfiguration d'une exploration globale, sous cet angle notamment, de la littérature narrative du XIX^{ème} siècle. Elle a pour but premier de présenter et valider une méthode.

5. Méthode de la comparaison

Chaque paire de vocables considérée forme dans chaque représentation géodésique un angle au centre compris entre 0° et 180°. Dans Balzac 15GDP0[7], ENFANT[8] et MORT(sf) forment un angle très fermé de 7°, signe de profils cooccurrentiels très similaires. Dans S15GDP0, ils forment un angle presque identique (8°). La *distance*[9] est de 1°, elle est parmi les plus faibles possibles. Les fig.2 et 3 montrent les deux zones isotropiques déterminées par cette paire de vocables proches, zones qui présentent des points communs et des variations significatives, sur lesquelles nous reviendrons plus loin.

---

[6] A l'exclusion notamment des noms propres (sauf *France, Paris* et *Dieu*) et des désignateurs de fonctions pouvant renvoyer à des personnages trop localisés (*comtesse, maréchal, reine, roi, abbé,…*).

[7] 15 pour empan de 15 mots, GD pour gauche Et droit, P pour limite de phrase, 0 pour aucune exclusion des cooccurrents les plus proches. Désormais on codera également le nom de l'auteur, sur le modèle B15GDP0, F pour Flaubert, S pour Stendhal.

[8] Les lemmes sont présentés en petites capitales.

[9] Nous choisissons *distance* pour nous rattacher à la problématique que recouvre ce terme en textométrie, c'est-à-dire l'étude des (dis-)similarités entre textes, ou entre sous-ensembles textuels, continus ou discontinus. Ainsi de la *distance intertextuelle*.

AVANCER et TUER présentent une *distance* du même ordre, tout en occupant dans les deux représentations des positions quasiment antipodiques (resp. 175° et 174°). A l'inverse, NOUVEAU et TOUJOURS, qui forment un angle de 14° dans S15GDP0, sont antipodiques dans B15GDP0 (176°), soit une *distance* de 162°. Ainsi peut-on classer les $n^2/2$ couplages de vocables les plus différents, par leur angle, entre les deux sous-corpus.

Néanmoins, l'intérêt descriptif majeur en vue d'une exploration rationnelle, ordonnée, est d'établir la moyenne de ces *distances* pour chacun des 289 vocables, identifiant ainsi les vocables conservant le plus, ou modifiant le plus leur distribution lexicale entre les deux sous-corpus. En l'occurrence les résultats vont de 21° (ARRIVER, AIMER, HOMME, MOURIR, etc) à 60° (CHERCHER), 61° (FOND), 71° (MONTRER) et 77° (NOUVEAU).

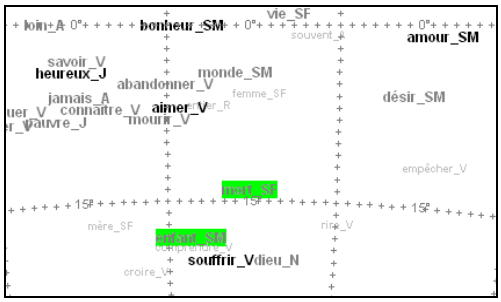

Fig.2 Zoom ENFANT,MORT - Balzac

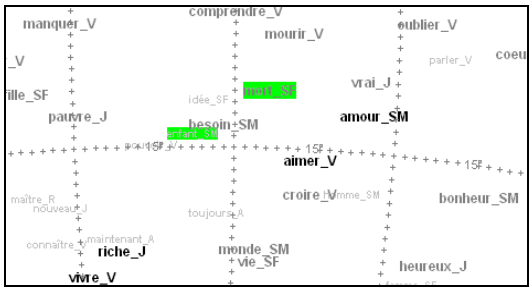

Fig.3 Zoom ENFANT,MORT – Stendhal

La *distance* moyenne s'établit à 30.3°. Entre B15GDP0 et F15GDP0, cette *distance* moyenne monte à 33.7°. Les vocables les plus similaires sont MATIN (22°), SENTIR et CONDUIRE (23°), les plus différents sont SIGNE (69°), RAPPELER (70°), HOMME (73°) et NOUVEAU (75°).

Entre Stendhal 15GDP0 et Flaubert 15GDP0, la *distance* est de 32.9°. Les vocables les plus similaires sont MAISON, BOIRE, LEVRE et CAMPAGNE (22°), les plus différents sont DECOUVRIR (73°), HOMME (74°) et SIGNE (76°).

En réalité, s'il paraît rationnel d'observer individuellement les indices forts, les *distances* faibles sont à observer plus globalement, tant elles sont finement graduées. La fig.4 montre la gamme de ces distances[10].

	B/F 15GDP0	B/S 15GDP0	F/S 15GDP0
g<25	24	96	35
25<g<39	113	86	113
30<g<35	70	45	59
35<g40	26	23	32
49<g<50	33	27	31
50<g<60	13	8	10
g>60	10	4	9
distance	**33.7**	**30.3**	**32.9**

Fig.4 Gammes des distances, 3 sous-corpus 2 à 2 [11]

[10] A partir de la fig.4, sauf mention contraire expresse, les indices figurant dans les tableaux sont des mesures d'angles en degrés.

[11] Cette figure montre en outre deux gammes très semblables, celles qui ont pour élément commun Flaubert, et une singulière, qui exclut cet auteur. Cela est évidemment en rapport avec la différence des *distances*.

C'est donc aux indices forts que nous consacrerons notre attention dans le bref compte rendu d'enquête exploratoire qui suivra. Il nous faut cependant d'abord présenter quelques caractéristiques des *distances* ainsi mesurées, qu'elles concernent des vocables individuels ou les ensembles textuels où agissent ces vocables.

Elles ne prétendent en aucun cas au statut d'éléments de *preuve*, même déguisé, notamment en matière d'auctorialité, et cela bien qu'elles mesurent d'une façon extrêmement précise des éléments de la plus fine granularité rationnellement descriptible dans les textes, et surtout que, contrairement à d'autres méthodes prenant en compte la cooccurrence, elles soient insensibles aux effets de taille, que ce soit l'effectif des vocables ou le volume des corpus. Au contraire par exemple de la comparaison massive des effectifs de couples de cooccurrents selon les méthodes par ailleurs traditionnelles de *distance lexicale*. En tout état de cause, l'objectif de notre démarche n'est pas d'établir de telles distances massives et de catégoriser les textes dans une optique probatoire ou documentaire, mais de donner accès au fonctionnement différentiel de vocables et de groupes significatifs de vocables (l'*isotropie*) par confrontation des vues les plus fines pouvant être prises et interprétées sur l'organisation micro-distributionnelle. Cependant, la *distance* moyenne qui est indiquée en dernière ligne de la fig.4 indique sans le moindre doute une parenté objective entre les sous-corpus Balzac et Stendhal *dans le cadre strict* de la triple comparaison avec Flaubert, du corpus choisi, de la lemmatisation, et des paramètres d'empan de cooccurrence.

Nous avons également vérifié que la *distance* entre vocables est indépendante de la longueur des vecteurs concernés dans la géode. On aurait pu craindre que de grands écarts factices soient favorisés par la proximité du centre des géodes, zone où les aléas distributionnels sont plus influents[12]. Ce n'est pas le cas. On prendra soin cependant de mettre ces données en regard des résultats, afin de ne pas donner la même importance aux points proches des centres et aux points très excentrés, dans l'interprétation.

Par ailleurs, les variations d'empan ne peuvent qu'influencer les résultats. Mais la fig.5 qui met en regard les résultats globaux dans l'empan 15GDP0 et dans les empans 8GDP0 15GDP5, ainsi que pour 15GD0[13], montre bien une régularité intéressante. Flaubert induit une *distance* moyenne systématiquement plus grande que celle du « couple » Balzac/Stendhal.

nb voc.	empan	BF	BS	FS
289	15GDP0	33.7	30.3	32.9
289	8GDP0	31.8	28.9	33.2
289	15GDP5	37,25	32,5	36,9
146	15GDP0	34,87	29,88	32,6

Fig.5 Distances moyennes, 3 sous-corpus 2 à 2

6. Un cas d'étude : le vocable HOMME

Nous observerons d'abord le résultat de la comparaison des sous-corpus Balzac et Flaubert sous l'aspect 15GDP0, qui nous semble une approximation correcte des standards de ce type d'étude. Les vocables présentant les plus grandes distances sont sur la fig.6.

[12] Avec un « basculement » de part et d'autre de l'origine des axes, pouvant générer des angles à 180° non significatifs. Ici le produit scalaire est un correctif qui s'impose à l'évidence.

[13] Cette comparaison est faite sur les 146 vocables les plus fréquents seulement, en raison de la chute d'effectifs que représentent les empans courts, chute qui rendrait les calculs moins fiables.

La fig.7 indique les résultats de la comparaison Stendhal/Flaubert sous le même aspect. On repère la présence commune, aux tous premiers rangs, de HOMME[14]. La fig.8 livre les mêmes informations à propos de la comparaison Balzac/Stendhal. On peut relever des points communs respectivement avec les deux têtes de liste précédentes, mais aucun vocable n'est commun aux 3 listes. En revanche, HOMME apparaît cette fois parmi les plus faibles distances (fig.9). Certes les écarts entre lignes successives sont très serrés, comme expliqué *supra*, mais le cas se différencie nettement de SIGNE, qui présente ici un indice assez moyen de 33.31°. C'est pourquoi nous choisissons d'explorer plus précisément les emplois de HOMME dans les 3 sous-corpus à la lumière de ces indices.

nouveau_J	75,08		coup_SM	64,46		embrasser_V	55,13		commencer_V	51,88
homme_SM	72,87		entier_R	61,64		route_SF	53,27		étonner_V	51,32
rappeler_V	69,23		jeune_homme_SM	61,14		lire_V	53,24		chercher_V	50,59
signe_SM	68,84		retourner_V	60,42		oser_V	53,06		ancien_J	50,27
essayer_V	66,72		montrer_V	59,75		partir_V	52,63		rentrer_V	50,07
aussitôt_A	64,58		revenir_V	55,65		offrir_V	52,19		ordre_SM	49,98

signe_SM	76,24		agiter_V	63,54		chercher_V	57,04		domestique_R	48,82
homme_SM	73,57		retourner_V	61,96		brûler_V	55,91		rentrer_V	48,53
découvrir_V	73,24		vieux_J	60,24		dame_SF	55		partir_V	48,21
crier_V	68,43		fond_SM	59,25		essayer_V	53,25		charger_V	46,93
battre_V	66,05		beau_R	58,93		force_SF	52,94		empêcher_V	46,73
aussitôt_A	64,12		revenir_V	57,9		route_SF	51,63		rencontrer_V	46,35

nouveau_J	76,88		découvrir_V	54,58		prêtre_SM	49,85		continuer_V	46,81
montrer_V	71,02		quelquefois_A	52,51		beau_R	48,93		cour_SF	46,71
fond_SM	61,54		saint_R	51,97		paraître_V	48,84		agiter_V	46,66
chercher_V	60,19		présenter_V	51,37		vieux_J	48,28		garçon_SM	46,37
rappeler_V	57,34		voix_SF	50,91		cacher_V	47,82		corps_SM	46,29
commencer_V	56,47		dame_SF	50,27		jeune_homme_SM	47,58		mot_SM	45,65

arriver_V	20,77		pauvre_J	21,58		premier_J	21,86		avancer_V	22,02
aimer_V	21,2		monde_SM	21,62		sourire_V	21,87		conduire_V	22,1
homme_SM	21,26		jamais_A	21,64		toujours_A	21,94		tuer_V	22,12
mourir_V	21,34		monter_V	21,68		amour_SM	21,95		rendre_V	22,2
vrai_J	21,48		mari_SM	21,8		vivre_V	21,97		vie_SF	22,21
mort_SF	21,54		fille_SF	21,86		asseoir_V	21,98		comprendre_V	22,24

Fig.6 à 9. Distances fortes B/F, S/F, B/S et faibles B/S

Dans Balzac (B15GD0P), homme se trouve dans une zone isotropique que présente la fig.10.

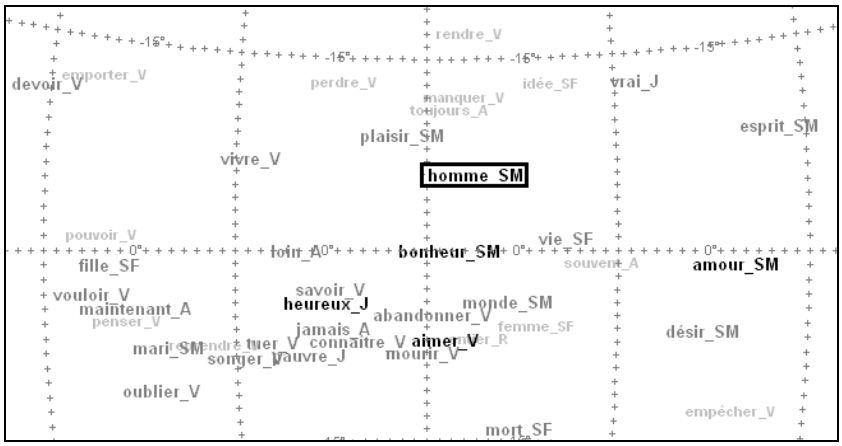

Fig.10. Zone isotropique de HOMME dans Balzac B15GD0

[14] Nous laissons de côté les outils possibles d'une confrontation plus systématique des résultats. On notera la présence commune également de SIGNE, RETOURNER, REVENIR, ESSAYER, ROUTE, RENTRER.

La fig.11 montre la position de HOMME dans Stendhal (S15GD0P). On y repère de nombreux vocables communs, ce qu'explicite plus systématiquement la fig.12 où il s'agit des angles formés respectivement avec HOMME par les vocables les plus proches dans les 2 sous-corpus. On lit dans ce tableau[15] une grande similitude de distribution, et on y repère ce qui pourra ensuite être observé plus finement sur les zooms.

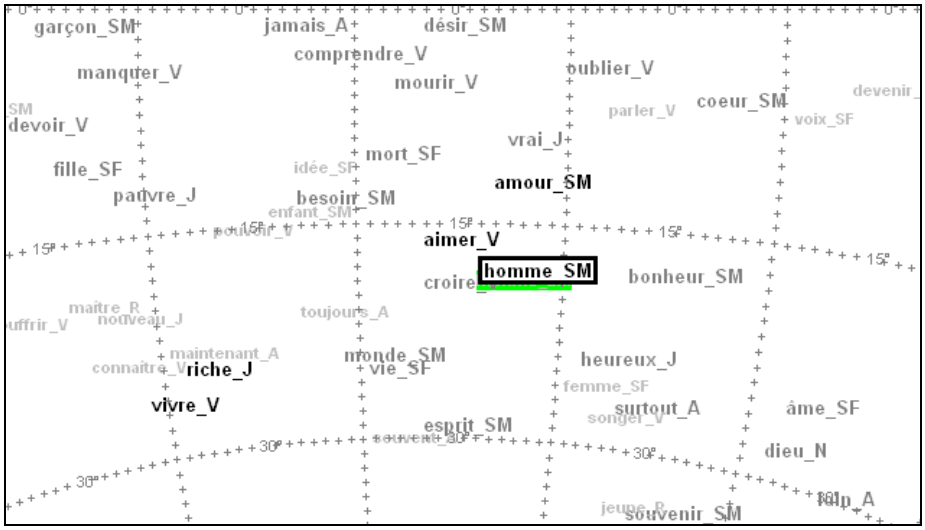

Fig.11. Zone isotropique de HOMME dans Stendhal S15GD0

VOCABLE	DIFF °	° BLZ	° ST		VOCABLE	DIFF °	° BLZ	° ST		VOCABLE	DIFF °	° BLZ	° ST
toujours_A	7	5	12		souvent_A	1	12	13		connaître_V	11	16	27
manquer_V	25	6	31		femme_SF	3	13	10		vivre_V	7	17	24
plaisir_SM	33	7	40		entier_R	21	13	34		pauvre_J	7	19	26
bonheur_SM	4	7	10		aimer_V	8	13	5		tuer_V	14	20	34
vie_SF	0	9	10		mourir_V	2	14	15		mort_SF	9	20	12
idée_SF	5	10	15		savoir_V	32	14	47		amour_SM	14	21	7
monde_SM	0	10	11		loin_A	13	14	28		désir_SM	3	22	18
rendre_V	11	11	22		vrai_J	5	15	10		songer_V	11	23	12
abandonner_V	46	12	58		heureux_J	7	16	9		esprit_SM	13	24	11
perdre_V	19	12	32		jamais_A	7	16	23		enfant_SM	9	24	15

Fig.12. Comparaison Balzac/Stendhal

La situation est très différente dans Flaubert (fig.13 et 14).,

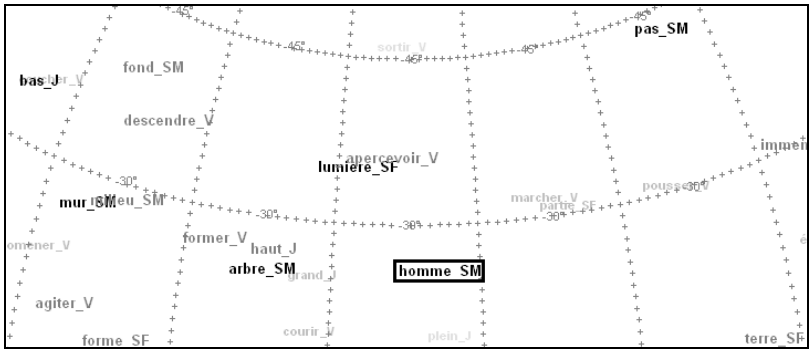

Fig.13. Zone isotropique de HOMME dans Flaubert F15GD0

[15] Les angles respectifs de HOMME et des vocables en ligne sont indiqués dans les colonnes BLZ pour Balzac et ST pour Stendhal, la différence en col.2..

HOMME échoit dans une zone très peu dense du nuage, avec des vocables dont la signification commune est intuitivement inattendue. La fig.13 et le tableau en fig.14 (comparaison Balzac/Flaubert) sont éloquents. Le vocable a migré, *grosso modo* à 90°. Comment l'expliquer ?

VOCABLE	DIFF °	° BLZ	° FL		VOCABLE	DIFF °	° BLZ	° FL		VOCABLE	DIFF °	° BLZ	° FL
plein_J	101	107	6		arbre_SM	140	156	16		fermer_V	123	153	30
grand_J	118	129	11		eau_SF	104	124	20		suivre_V	129	160	31
disparaître_V	164	175	11		sortir_V	137	158	20		terre_SF	50	81	31
marcher_V	130	141	11		former_V	125	146	21		arrêter_V	125	156	31
apercevoir_V	149	161	12		pousser_V	125	147	22		fond_SM	117	148	31
lumière_SF	114	126	12		découvrir_V	84	109	25		mur_SM	123	155	32
courir_V	132	144	13		descendre_V	132	161	29		nuit_SF	99	131	32
partie_SF	101	114	13		milieu_SM	145	174	29		brûler_V	85	118	33
ciel_SM	70	85	14		pas_SM	121	151	30		immense_J	83	116	34

Fig.14. Comparaison Balzac/Flaubert

Le vocable a une fréquence proche de 0.2% dans les 3 sous-corpus, nous y reviendrons *infra*. Mais il est employé dans des contextes très différents, ce que montre bien la fig.15, plan des 2 1ers facteurs de l'AFC de la distribution des cooccurrents de HOMME.

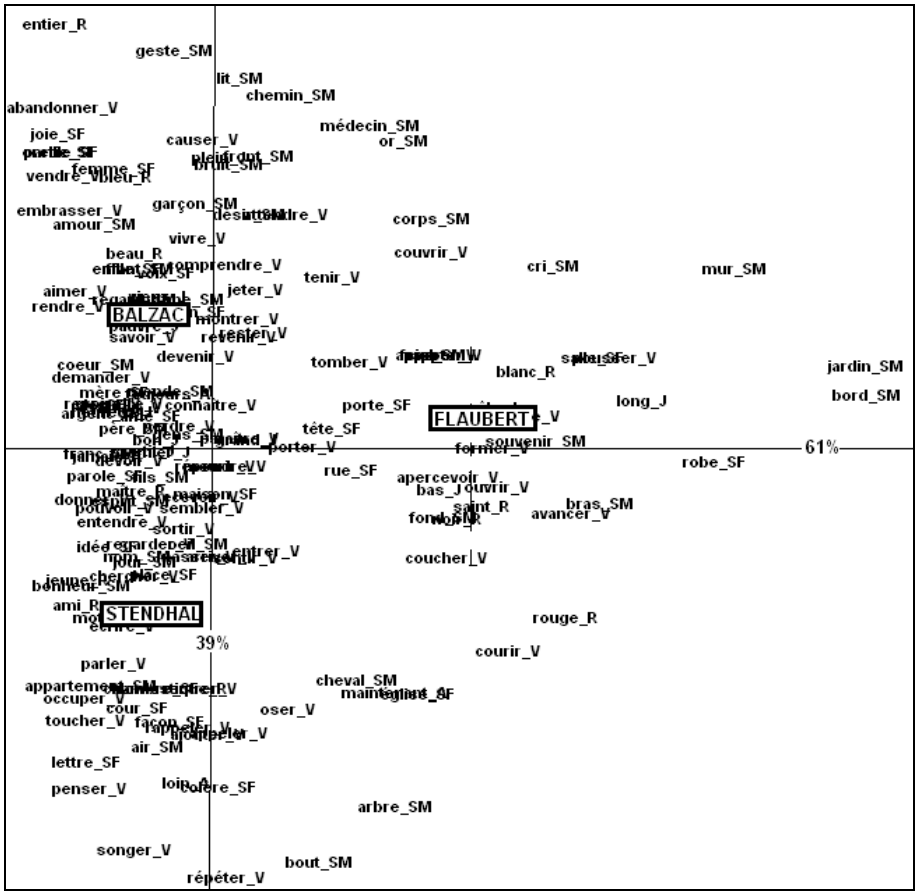

Fig.15. Distribution de vocables fréquents dans les cotextes de HOMME pour tout le corpus

Il est patent que Flaubert n'emploie pas HOMME comme le font en commun Balzac et Stendhal, même si ceux-ci se différencient sur l'axe 2. En fait, le dépouillement du vocabulaire cooccurrent de HOMME chez Flaubert révèle d'abord une beaucoup plus grande *dispersion*, autrement dit une bien moins active fonction de *polarisateur*. Parmi les 20 plus fréquents des cooccurrents de HOMME chez les 3 auteurs confondus (en effectifs), un seul, CONNAITRE, est mieux représenté comme cooccurrent chez Flaubert (en écart-réduit). Mieux

encore : parmi les 56 vocables d'effectif>1000 sur tout le corpus, 5 seulement sont mieux représentés comme cooccurrents chez Flaubert que chez Balzac et Stendhal confondus : REPONDRE (n°16), TETE (n°31), TENIR (n°38), CONNAITRE (n°43) et DEVENIR(n°48).

Ce « détour » par l'approche « sophistiquée » de la cooccurrence nous permet de ne pas en rester, aveuglés, à la différence brute que présentent les effectifs du lemme HOMME entre les 3 auteurs. C'est Stendhal qui l'emploie le plus, avec une fréquence de 2.25‰, devant Balzac (2.02‰) et Flaubert loin derrière (1,7‰). Mais la différence majeure, qui oppose à nouveau Flaubert aux 2 autres, est la proportion des singuliers et des pluriels. Alors que le singulier prédomine nettement chez Stendhal (79%) et chez Balzac (72%), il ne représente que 49% chez Flaubert, où le pluriel prédomine légèrement. Ainsi ces données, intéressantes par elles-mêmes, peuvent-elles ici venir en appui interprétatif aux données distributionnelles, plutôt que l'inverse.

C'est en outre, chez Balzac plus encore que chez Stendhal, le cotexte HOMME/FEMME (H/F) qui donne le ton. Compte tenu des fréquences respectives des 2 vocables, Balzac présente une proportion double de cooccurrences H/F avec Stendhal, respectivement en nombre brut 189 et 36. Cette proportion est 3 fois moindre chez Flaubert que chez Stendhal, avec 14 cooccurrences seulement[16]. Les cotextes des 33 triples cooccurrences HOMME/FEMME/AMOUR OU AIMER sont caractéristiques de l'emploi par ailleurs varié, même dans ce cadre, de HOMME par Balzac.

7. Conclusion, perspectives

C'est bien sûr l'observation des cooccurrents lexicaux en cotexte, dans les environnements appropriés, qui permettra de donner à ces calculs leur pleine vertu heuristique. Mais une saisie synthétique est nécessaire pour ordonner et rationaliser cette observation. Cette méthode devrait aider tous les chercheurs qui travaillent sur des corpus contrastifs, et tiennent à l'heuristique de l'AFC. Nous explorerons par la suite les possibilités de faire servir cette approche à la discussion générale des méthodes de *distance* lexicale entre corpus.

Références

Brunet E. (1981). *Le Vocabulaire français de 1789 à nos jours*. Slatkine-Champion.

Cibois P. (1994). *L'Analyse factorielle*. PUF.

Guaresi M. (2015). « Les thèmes dans le discours électoral de candidature à la députation sous la Cinquième République. Perspective de genre (1958-2007) » in *Mots* 108.

Harris Z. S. (1969) « Analyse du discours » in *Langages* n°13 J.Dubois & J.Sumpf éds. Larousse.

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

Massonie, J.-P. (1986). " Q-occurrences libres ", in Brunet, *Méthodes quantitatives et informatiques dans l'étude des textes*. 611-623. Slatkine-Champion.

Viprey J-M. (2005). « Structure non séquentielle du texte » in *Langages* n° 161, *Unité(s) du texte* 65-82. Larousse.

Viprey J-M. (2006). « Ergonomiser la visualisation AFC dans un environnement d'exploration textuelle : une projection 'géodésique' ». In *JADT 2006*. PUFC.

[16] La proportion est 3 fois moindre parce que le nombre « brut » de cooccurrences est 3 fois moindre alors que le nombre « théorique » d'après les paramètres des sous-corpus est identique.