Exploration contrastive de trois corpus de sciences humaines

Sylvain Loiseau[1], Céline Poudat[2], Driss Ablali[3]

[1]UMR 7114 Modyco, UFR LLPhi, Université Paris X Nanterre, 200 avenue de la République, 92001 Nanterre Cedex

[2]CORAL, Université d'Orléans, UFR Lettres Langues et Sciences Humaines, 10 rue de Tours, 45072 Orléans Cedex 02

[3]Laseldi, Université de Franche-Comté, UFR SLHS, 30 rue Mégevand, 25030 Besançon cedex

Abstract

The present study investigates the discursive differences between three humanities corpora (linguistics, literary criticism and philosophy). We use different exploratory techniques to contrast the three discourses and focus on the morphosyntactic and lexical levels of description. Morphosyntactic variables are significantly discriminating, whereas lexical descriptors enable us to explore the corpora thematic and to reveal behind apparently shared concepts, different variations of sense that are strongly characteristic of each of the observed discourses.

Résumé

Cet article explore, à travers la comparaison de trois corpus de sciences humaines (linguistique, critique littéraire, et philosophie), les indices de divergence et de convergence entre discours que permettent d'observer les méthodes exploratoires de données textuelles. Les analyses menées sont fondées sur les niveaux morphosyntaxiques et lexicaux : si les descripteurs morphosyntaxiques suffisent à différencier les trois corpus de manière significative, les variables lexicales permettent d'explorer la thématique du corpus et de mettre à jour, derrière des concepts apparemment partagés, des variations d'acceptions fortement caractérisantes de chacun des discours.

Mots-clés : analyse exploratoire de données textuelle, statistique textuelle, linguistique de corpus, analyse contrastive de corpus, discours des sciences humaines

1. Introduction

L'objectif de cet article est de décrire les similitudes et les points de divergence entre trois corpus appartenant aux sciences humaines : un corpus de linguistique, un corpus de philosophie et un corpus de critique littéraire. L'enjeu est de s'interroger sur les possibilités offertes par les méthodes quantitatives et les perspectives typologiques élaborées en linguistique de corpus pour caractériser des discours, c'est-à-dire des unités extrêmement vastes de la langue dont les définitions restent pour l'instant indépendantes de caractérisations issues de corpus. Il s'agit donc d'un travail largement exploratoire.

La démarche contrastive adoptée ici peut être un moyen de pallier l'absence de représentativité des corpus de discours : aucun des corpus utilisés ne peut prétendre représenter son discours d'appartenance. Les points de contraste offrent donc des indices précieux pour les caractériser et les opposer.

La démarche contrastive a un second versant : la mise à jour de points de proximité entre les corpus, qui pourraient être des indices d'un éventuel noyau dur des sciences humaines. Là encore, la réunion des trois corpus ne peut prétendre représenter les sciences humaines – c'est au contraire une certaine proximité au sein de celle-ci, qui garantit la pertinence de la méthode

contrastive. Seul le rattachement de la linguistique aux sciences humaines ne pose d'ailleurs pas de difficulté.

La méthode adoptée consistera à évaluer en premier lieu les intersections des corpus en se fondant sur une Analyse en Composantes Principales (ACP) associée à une Classification Ascendante Hiérarchique (CAH) fondée sur le niveau de description morphosyntaxique. On cherchera à valider si les trois discours observés sont significativement distincts au moyen d'ellipses de confiance.

Après avoir obtenu les spécificités morphosyntaxiques et lexicales des trois discours au moyen d'un tri systématique de signification, on observera leurs intersections lexicales en relevant les concepts les plus représentés. Ces informations serviront enfin de fondement à une entreprise d'exploration thématique qui distinguera les cooccurrents de concepts apparemment partagés pour décrire, derrière ces concepts, les variations sémantiques de leur emploi dans chacun des discours.

2. Corpus d'étude

Les trois corpus ont été constitués dans le cadre de projets indépendants : ils n'adoptent donc pas la même stratégie en ce qui concerne la représentativité. Leur réunion est justifiée par une proximité intuitive : la linguistique et la critique littéraire partagent un objectif de description précise de faits langagiers. La philosophie et la critique partagent des méthodes d'interprétation éloignées de l'empirisme. Tous trois partagent un objet partiellement commun, comme nous le verrons, la description du sens.

Les trois corpus sont de taille comparable : 1 469 439 mots pour la linguistique, 1 901 904 mots pour la critique et 1 288 477 mots pour la philosophie.

- Le corpus « linguistique » est constitué de 224 articles extraits de 32 numéros de revues (soit 11 revues) francophones de sciences du langage, publiées entre 1995 et 2003. Il est supposé « représentatif » des sciences du langage et différents domaines spécifiques à la linguistique française y sont représentés (syntaxe, sémantique, phonologie, sociolinguistique, etc.). Soulignons que la variation stylistique est neutralisée dans le corpus, dans la mesure où 226 auteurs y sont représentés.

- Le corpus « critique » comprend des articles tirés de revues et de sites français ou francophones universitaires relevant du domaine de la critique littéraire. Il contient 249 articles extraits de 11 revues, publiées entre 1970 et 2004.

- Le corpus « philosophie » est constitué de 14 livres publiés par Deleuze de 1953 à 1994. Il comprend 5 essais[1] et 9 commentaires[2]. Ce corpus ne vise donc pas une représentativité du discours philosophique, mais il essaye de représenter l'empan maximal de la production d'un philosophe, depuis les thèses publiées (*Différence et répétition*, *Spinoza et le problème de l'expression*), les commentaires académiques (sur Hume ou Kant), les essais novateurs représentatifs des préoccupations de la philosophie contemporaine (*Logique du sens*, *Anti-Œdipe*, *Mille-Plateaux*), et jusqu'aux textes où le philosophe élargit son domaine de compétence, notamment à la littérature (*Proust et les signes*, *Critique et clinique*).

[1] *Logique du sens, Différence et répétition, L'Anti-Œdipe, Mille-Plateaux Qu'est-ce que la philosophie ?,* et *Critique et clinique.*

[2] *Empirisme et Subjectivité, La Philosophie critique de Kant, Proust et les signes, Spinoza et le problème de l'expression, Spinoza philosophie pratique, Nietzsche et la philosophie, Le Pli et Foucault.*

3. Contrastes et spécificités des discours

L'ensemble des expérimentations de la présente section a été mené avec le logiciel DTM développé par Ludovic Lebart[3].

3.1. Intersections et (non-)recouvrements des trois discours

Afin de déterminer les intersections, et les (non-) recouvrements éventuels des trois discours observés, nous avons mené sur le corpus global une ACP associée à une CAH à partir d'un ensemble de 88 descripteurs morphosyntaxiques étiquetés par Cordial® (Synapse Développement). Le niveau d'annotation morphosyntaxique présente en effet l'intérêt d'avoir déjà démontré son efficacité en matière de classification textuelle (e.g. Malrieu et Rastier, 2001 ; Poudat, 2003), et d'être suffisamment développé pour être automatisable.

On observe un pallier dans la décroissance des valeurs propres au niveau des trois premiers facteurs, qui extraient 26.22% du nuage de points – ce qui est significatif dans la mesure où le tableau de départ ne contient que 88 variables. On note donc que le nuage est essentiellement concentré dans un espace à trois dimensions.

Nb	Valeur propre	% d' inertie	% cumulé	
1	9.44	10.85	10.85	**********
2	7.63	4.62	19.63	**********
3	5.473	6.59	26.22	**********
4	3.46	3.98	30.20	**********
5	3.09	3.56	33.76	**********

Figure 1 : Diagramme des 5 premières valeurs propres

La figure 2 propose une représentation des descripteurs morphosyntaxiques et des variables supplémentaires « discours » sur les deux premiers axes factoriels : ces derniers opposent ainsi le discours linguistique aux discours philosophique et critique. La linguistique semble ainsi se rattacher à un pôle plus « scientifique », caractérisé par l'usage intercorrélé de parenthèses, de cardinaux et de propositions indépendantes, tandis que le discours critique est associé à une longueur plus importante des phrases et des paragraphes, et contiendrait visiblement plus de tirets, liés à la présence de dialogue dans les textes littéraires. La philosophie semble enfin discriminée par une présence plus importante d'adverbes et de négations :

[3] DTM étant la version universitaire de SPAD-T http://www.enst.fr/egsh/lebart/

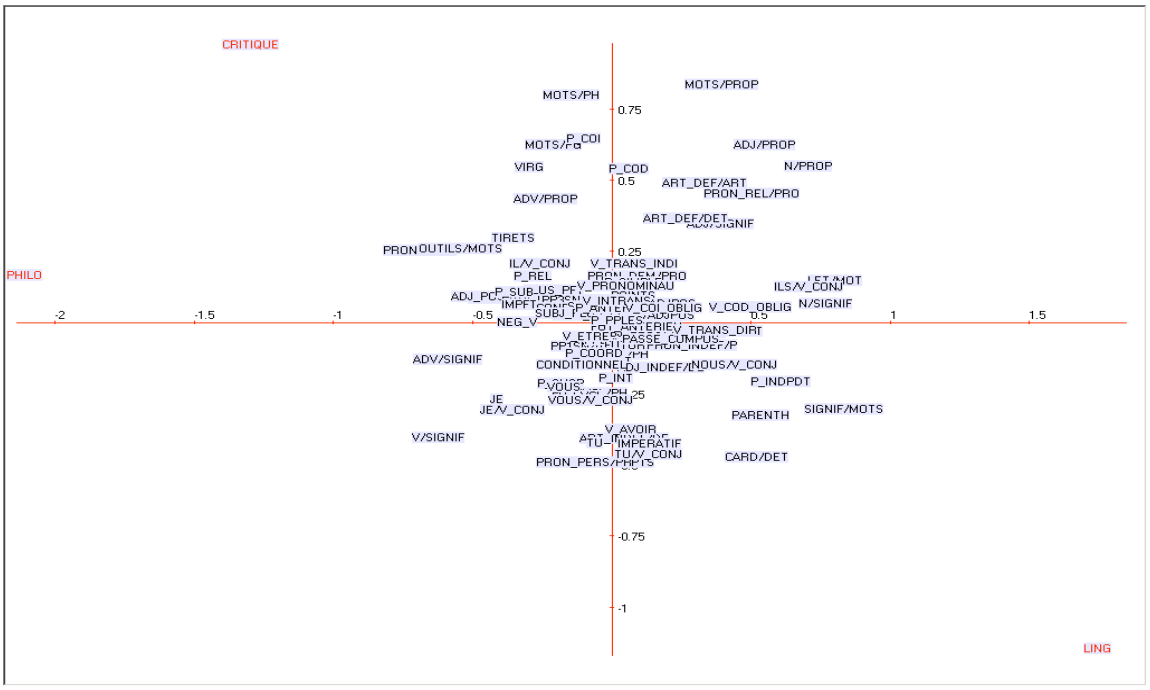

Figure 2 : Positionnement des 88 descripteurs morphosyntaxiques et des trois variables supplémentaires sur les deux premiers axes factoriels

Nous n'approfondirons pas le détail des éléments associés aux trois discours étudiés sur les différents axes factoriels, dans la mesure où nous avons entrepris *infra* une entreprise de validation externe (3.2.).

L'objectif de cette section est davantage d'apprécier les recouvrements éventuels entre discours. On note que les trois discours sont significativement distincts sur le plan morphosyntaxique, comme le montrent les ellipses de confiance suivantes, qui valident ce non recouvrement :

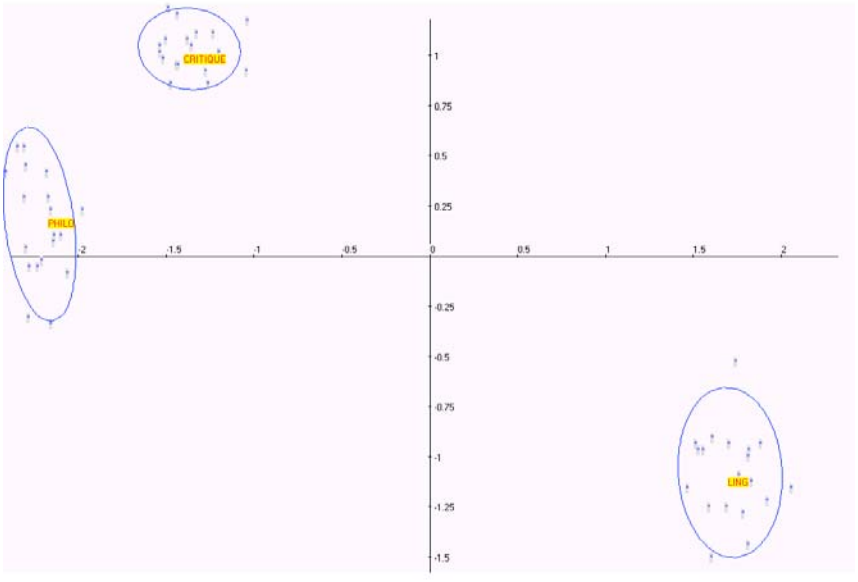

Figure 3 : Ellipses de confiance autour des trois discours observés sur le premier plan factoriel

On peut d'ailleurs observer que les ellipses sont relativement petites ce qui renforce la significativité des positions des trois catégories.

On retrouve cette opposition lorsque l'on examine les 12 partitions textuelles obtenues grâce à la CAH :

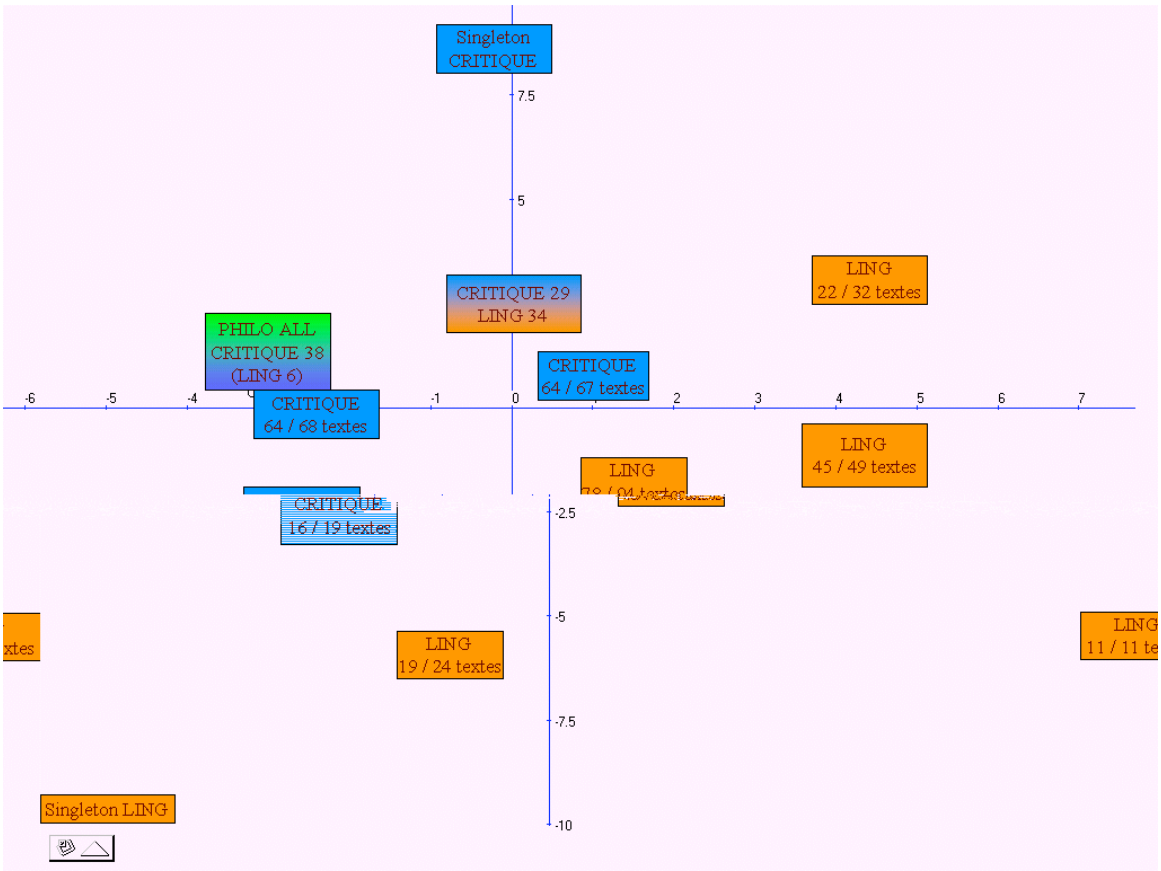

Figure 4 : Positionnement des douze classes sur les deux premiers axes factoriels

3.2. Validation externe : spécificités morphosyntaxiques et lexicales des trois discours

Afin d'obtenir les spécificités des trois discours observés, l'ensemble du corpus a été confronté avec la variable supplémentaire « discours », qui a été positionnée *a posteriori* sur les axes principaux au moyen d'une *valeur-test* qui « exprime la signification statistique de la coordonnée de la variable sur l'axe » (Lebart, 2004 : 712). On appellera la méthode utilisée *Tri Systématique de Signification* (TSS).

Nous avons ainsi obtenu les spécificités morphosyntaxiques et lexicales des trois corpus.

3.2.1. Caractéristiques morphosyntaxiques

On remarque d'abord que les discours linguistique et critique sont les premiers à s'opposer, dans la mesure où l'ensemble des descripteurs caractérisant positivement le discours linguistique caractérise négativement le discours critique.

Le discours linguistique se situe bien sur un pôle plus scientifique : il est ainsi significativement caractérisé par un usage plus important de parenthèses (v. test de 13.44), de cardinaux (9.41) et de ponctuations deux points (5.66). On observe également une plus grande

longueur des mots (10.79), de même qu'une proportion plus conséquente de mots signifiants (8.87), liée aux énumérations d'exemples lexicaux, fréquente dans les articles linguistiques. À l'inverse, le discours critique est caractérisé par un usage plus important de mots-outils (8.10).

On relève au contraire un usage moins important de tirets en linguistique (-16.46), premier descripteur caractérisant positivement le plus digressif discours critique (17.66), qui privilégie le tiret au numéro pour construire ses listes. On observe également que les textes de critique contiennent des paragraphes de longueur plus importante, à l'inverse des textes de linguistique.

Au niveau des temps verbaux, le discours linguistique se caractérise par un usage plus important du présent et du futur, et moins important des temps narratifs du passé simple et de l'imparfait (et de leurs homologues composés), tandis qu'on note que le pronom NOUS est clairement privilégié au pronom JE.

Les valeurs-tests associées au discours philosophique sont plus faibles, eu égard au plus petit nombre de textes pris en compte : le discours philosophique est significativement associé aux négations (6.67), aux virgules (5.38) et aux adverbes (4.51). Il serait ainsi plus polémique. On relève également une absence de temporalité avec une prédominance nette du présent parmi les temps observés.

3.2.2. Caractéristiques lexicales

Les spécificités lexicales des trois discours obtenues (figure 5) confirment pour partie certaines des remarques déjà effectuées : le discours linguistique se caractérise ainsi par l'usage de chiffres, et plus particulièrement des numéraux [1, …, 4], de même que par des marques de formalisation (*x* ou *b* par exemple). On relève un intérêt particulier des textes pour le *verbe* (v. test de 54.67 pour *verbe* et 50.84 pour *verbes*) et la *sémantique* (49.80), tandis que le discours critique est significativement associé à la *fiction* (54.29), au *récit* (54.08) et au *roman* (53.09), de même qu'au *lecteur* (49) et à l'*écriture* (48.39). On observe également une forte présence de Sartre (45.75), liée à certaines revues – la plupart des numéros de la revue RITM de Nanterre contient en effet des textes consacrés à cet auteur.

Le discours philosophique est quant à lui plus orienté vers les concepts de *puissance* (55.46), de *corps* (53.64) et d'*essence* (49.10). Notons qu'il est significativement associé à *Dieu* (48.66).

Ces spécificités lexicales participent déjà à l'exploration thématique des trois discours analysés.

4. Exploration thématique

Si nous avons obtenu les spécificités des mots des trois corpus, nous ne disposons pas de leurs intersections lexicales et conceptuelles, essentielles pourtant à une entreprise d'exploration thématique. Parmi les variables lexicales envisageables, ce sont les substantifs les plus fréquents que nous avons sélectionnés. En effet, les noms sont des parties du discours non vides susceptibles de pointer sur des concepts scientifiques, contrairement aux adverbes, verbes ou adjectifs. Ils sont donc potentiellement plus discriminants et présentent l'avantage d'être facilement extractibles. Le poids des substantifs au singulier et au pluriel (dans la mesure où ils sont susceptibles de renvoyer à des concepts différents, e.g. "la langue" en linguistique ne renvoie pas à la même notion que "les langues") a également été pris en compte.

Nous avons ainsi extrait la liste des 50 premiers substantifs singulier et pluriel et examiné leur(s) intersection(s) (4.1.), ce qui nous permettra de sélectionner les concepts dont nous observerons le fonctionnement en corpus, à partir d'une méthodologie originale d'ACP sur co-occurrents (4.2.).

4.1. Intersections conceptuelles des trois discours observés

Les 50 premiers concepts employés correspondent globalement aux spécificités obtenues avec DTM : la critique se caractérise en effet par un usage important des concepts relatifs à son matériel d'étude (*roman, fiction, lecteur, récit, œuvre, histoire,* etc.) et à ses thèmes de prédilection (la *mère*, la *femme*, la *voix* ou la *vie*), tandis que le discours philosophique observé s'intéresse davantage aux concepts de *désir*, *nature, terre, ordre, puissance* ou *mouvement*. La linguistique se concentre davantage sur ses objets, et plus spécifiquement au *verbe,* particulièrement représenté dans le corpus. On relève également de très nombreuses occurrences de *phrase, énoncé, mot, locuteur* ou *interprétation*, qui semble un concept plus linguistique que critique, la critique privilégiant la *lecture* à l'*interprétation*.

L'analyse de ces 50 premiers concepts illustre ainsi les différences d'objet des trois disciplines : si la linguistique, qui vise l'objectivation, se focalise sur ses données observables (e.g. *phrase*, *mot*), la critique s'intéresse à son objet *texte*, tout en se concentrant sur différents thèmes manifestement récurrents, alors que la philosophie aborde frontalement le conceptuel.

L'analyse des intersections entre les 50 substantifs les plus fréquents dans les trois discours laisse toutefois entrevoir la présence d'un univers conceptuel commun aux sciences humaines.

Les trois discours observés se partagent d'abord l'objet *sens*, qu'ils tentent d'élucider selon leurs méthodologies d'analyse. On note la présence d'un fond métalinguistique commun à travers l'usage partagé de termes comme *objet, sujet, forme* ou *langage.*

Les intersections obtenues confirment la plus grande proximité des disciplines philosophique et critique, qui partagent un nombre sensiblement plus important de concepts (8 concepts au singulier partagés, *vs.* 6 entre critique et linguistique, et 4 entre linguistique et philosophie, et 9 au pluriel, *vs.* 5 entre critique et linguistique, et 6 entre linguistique et philosophie).

Les concepts au pluriel relevés confirment les tendances et les orientations des disciplines linguistique et critique, avec des substantifs de haute fréquence comme *verbes, énoncés* ou *noms* pour la première, et *romans, poèmes* ou *femmes* pour la seconde. Les entrées philosophiques au pluriel semblent davantage idiolectales – et donc deleuziennes : si les substantifs au singulier les plus représentés renvoyaient à un ensemble de concepts traditionnellement objets de la philosophie, les substantifs au pluriel sont plus spécifiques à la philosophie deleuzienne qu'au discours philosophique : *machines, concepts, différences*, etc.

Le corpus « Philosophie » contenant un nombre plus restreint de texte, et donc d'objet, les concepts relevés ont des fréquences significativement plus basses que dans les deux autres corpus : le premier concept linguistique est ainsi *sens* avec 2136 occurrences, tandis que la critique mobilise le *texte*, avec 3219 occurrences, alors qu'on ne recense que 1553 occ. de *puissance*, premier concept philosophique relevé.

4.2 Exploration thématique

Si, jusqu'à présent, on a utilisé les intersections de leur lexique pour caractériser les proximités et spécificités des trois corpus entre eux, on peut adopter à l'inverse comme poste d'observation de la variation entre les corpus la variation des acceptions entre concepts employés.

Pour cela on a mis en œuvre une méthode utilisant les espaces vectoriels introduits par les travaux de Salton en Recherche d'information, et fondée sur la représentation des distances entre les cooccurrents d'un concept dans un espace multidimensionnel. Les cooccurrents d'un concept sont rapprochés par les méthodes de classifications factorielles et permettent de mettre à jour une facette du sens de ce concept, voire une de ses acceptions. On a donc cherché à déterminer (i) si les concepts étaient caractérisés par les mêmes co-occurrents au sein de leurs corpus d'appartenance ; (ii) si les pôles de sens entre lesquels ces cooccurrents se distribuaient étaient similaires, voire si l'empan de l'hétérogénéité du sens entre les corpus attestait de fonctionnements sémantiques différents (plus ou moins grande « polysémie » ou plus ou moins grande technicité dans l'emploi du même concept entre les différents corpus).

L'expérience a consisté à extraire les cooccurrents d'une sélection de concepts dans la fenêtre de la phrase, telle que produite par l'analyseur Cordial. Cette fenêtre s'est imposée pour des questions de taille : les matrices générées par le paragraphe étaient trop lourdes à manipuler, tandis que les empans fondés sur une fenêtre graphique (N mots à droite et à gauche) sont moins adaptés aux explorations thématiques (Grefenstette, 1996). On n'a gardé que les cooccurrents d'une fréquence absolue supérieure à 50 et inférieure à 300 dans le corpus considéré. De très nombreux cooccurrents sont donc sélectionnés, la plupart n'étant réalisés que dans très peu d'observations (phrases). L'un des paramètres décisifs dans l'emploi de telles méthodes est donc le critère utilisé pour sélectionner les cooccurrents entre lesquels seront calculées les proximités. Différentes stratégies ont été explorées dans la littérature existante, comme la sélection des variables en fonction de leur fréquence ou de leur corrélation au mot pôle (Schütze, 1998).

Dans cette expérience, nous avons développé une méthode originale (Loiseau, 2003) : les cooccurrents sélectionnés comme variables sont bien ceux qui sont le mieux corrélés au mot-pôle, mais on a également manipulé la matrice soumise à l'analyse factorielle en remplaçant les phrases d'origine par autant de sous-corpus regroupant les phrases qui contiennent l'une des variables sélectionnées. Cette méthode permet de « densifier » les données de départ et de renforcer la représentativité des variables. On ne classe non plus des phrases, mais des sous-corpus réalisant chacun des cooccurrents.

Les mots pôles observés selon cette méthode ont donc été sélectionnés par leur égale importance dans les trois corpus. On a d'abord procédé à une analyse de *sens*, le premier concept partagé par les trois discours, avant d'observer les premiers concepts à l'intersection des trois paires de corpus : *système*, le premier concept partagé par la philosophie et la critique, *texte*, le premier concept partagé par la linguistique et la critique, et *monde*, le premier concept partagé par la critique et la philosophie. Ces trois concepts sont d'ailleurs déjà emblématiques des proximités deux à deux des corpus.

4.2.1. Sens

Le mot pôle lui-même (*sens*) est retiré afin de ne pas influencer la disposition de ses cooccurrents.

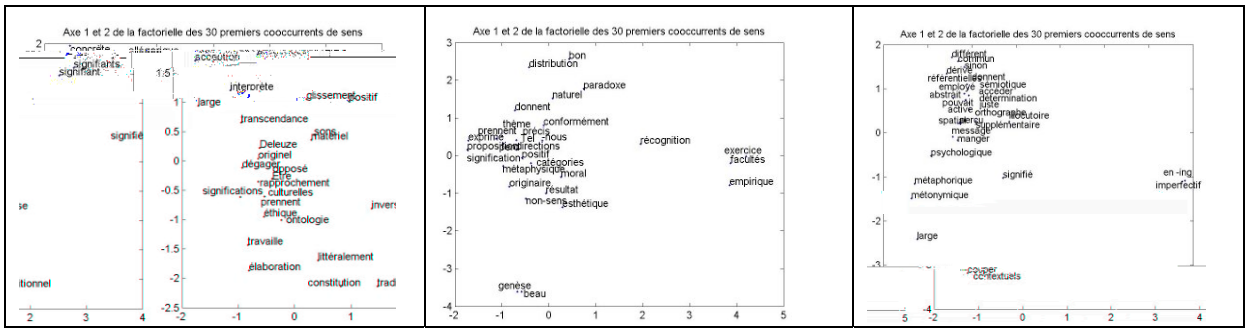

Figure 5 : Cooccurrents de sens *dans la critique, la philosophie et la linguistique. Les donnés sont calculées sur 2257 phrases (critique), 1992 phrases (philosophie) et 2141 phrases (linguistique).*

La comparaison des profils des cooccurrents permet d'observer des différences significatives entre les trois discours. La linguistique se caractérise par une bien plus grande technicité du concept : les cooccurrents illustrent en premier lieux des collocations du vocabulaire de la discipline, comme « sens métaphorique », « sens métonymique », « sens abstrait », « sens contextuel ». Un quasi-synonyme comme *signifié* est centré dans le plan, indiquant une faible différenciation des deux termes. Les différents aspects de *sens*, et les différents sous-domaines qui leur correspondent, sont identifiables : *sémiotique, illocutoire, psychologique* – bien que ces cooccurrents ne s'organisent pas en pôles sur le plan, ce qui indique une stabilité du sens de *sens* à travers les domaines. Il est remarquable que tous les concepts négativement corrélés à *sens* sont des marques de formalisation : nombres, symboles mathématiques (<, >, =, +) ou (*, [,]).

La critique illustre une plus grande différenciation des facettes de sens. *Signifiant* et *signifié* se détachent, témoignant de l'importation de l'acception linguistique du concept. Sur le second axe factoriel, on trouve d'un côté des cooccurrents du domaine de l'interprétation littéraire (*allégorique, interprète*), de l'autre des cooccurrents témoignant visiblement du point de vue de la description des oeuvres, et de leur genèse : *travail, élaboration* et *constitution*. On note enfin au centre du plan un vocabulaire philosophique qui indique une tendance spéculative : *Deleuze, Être, ontologie, éthique*. Par opposition aux cooccurrents de la linguistique, on observe une plus grande hétérogénéité, et ainsi une moindre technicité du terme.

La philosophie étend enfin cette hétérogénéité : du pôle esthétique (*beau, genèse, esthétique*), à celui de la logique (*paradoxe*), tandis que le premier axe met en valeur un pôle empirique et une recherche de technicité (*empirique, exemple*). On relève aussi un emprunt au vocabulaire technique du traitement du sens en linguistique (*distribution*).

4.2.2 Texte dans les corpus linguistique et critique

Sélection des cooccurrents > 100

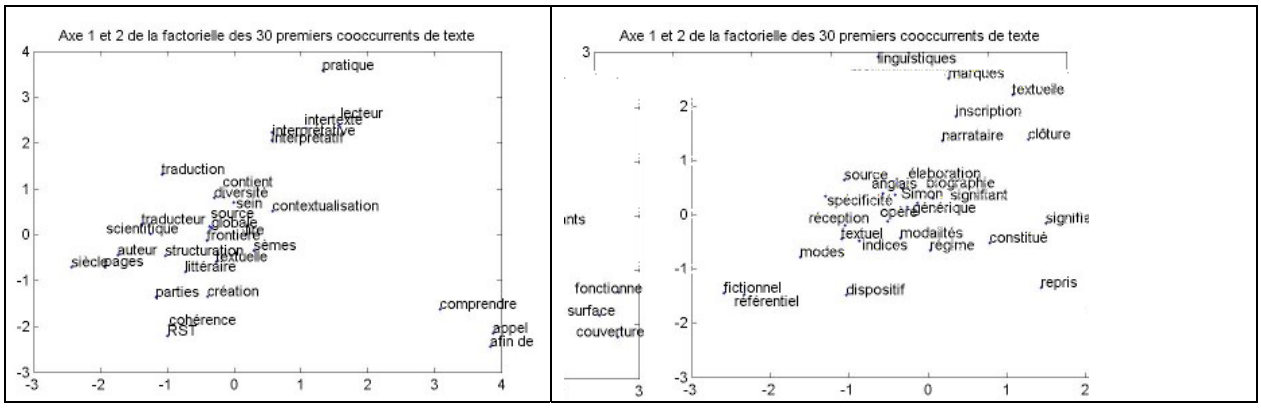

Figure 6 : Les cooccurrents de texte *dans les corpus linguistique (à gauche) et critique (à droite)*

La comparaison des deux profils illustre l'absence de recouvrement entre les cooccurrents du concept dans les deux corpus, à l'exception du dérivé *textuel*. On observe donc bien, derrière un concept partagé, une différence forte d'acception. La linguistique se caractérise comme précédemment par une concentration des cooccurrents dans l'espace de description et une faible interprétativité des deux premiers axes factoriels. Un pôle littéraire se dégage (en bas), avec *littéraire*, *création*, *siècle*, *auteur*, tandis qu'un pôle interprétatif se dégage en haut (*intertexte*, *interprétation*, *interprétatif*). Au centre du plan, on relève de nombreux cooccurrents qui ne s'organisent pas en groupes de contextes : *scientifique*, *traducteur*, *sources*, *contextualisation*…

Dans le corpus critique, on observe un pôle influencé par la linguistique (et qui se positionne explicitement par rapport à elle) : *linguistique*, *marque*, et des concepts à la frontière de la critique et la linguistique comme *narrataire*. On peut lui opposer, en bas, un registre plus orienté vers la description et l'interprétation que par la modélisation qui revient à travers *fonctionne*, *indices*, *couverture*, *surface*, auquel se mêle le registre littéraire (*réception*, *autobiographie*, *fictionnel* et *référentiel*).

4.2.3 Monde dans les corpus philosophie et critique

Sélection des cooccurrents > 50

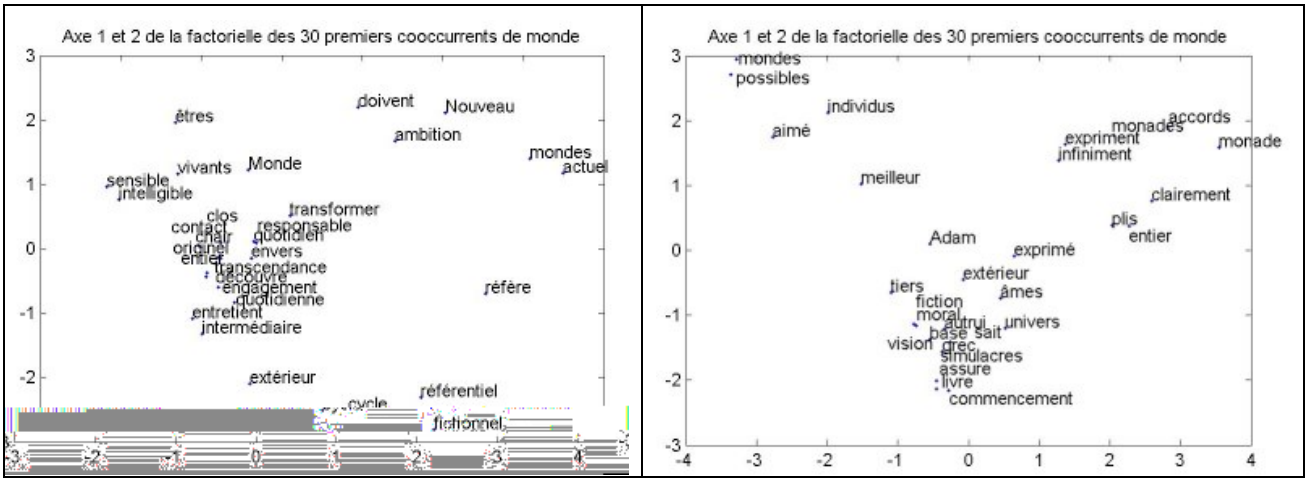

Figure 7 : Les cooccurrents de monde *dans les corpus critique (à gauche) et philosophie (à droite)*

On relève ici une plus grande proximité entre les discours, malgré l'absence de lexique partagé. Tandis que la critique témoigne d'un intérêt pour un pôle philosophique inscrit sur le premier axe factoriel (à gauche), on remarque un décalage par rapport au vocabulaire effectivement employé dans le corpus philosophie : il s'agit de concepts généraux ou philosophiquement anciens, qui ne sont plus spécifiques à la philosophie : *sensible*, *intelligible*, *transcendance*. La thématique sartrienne de l'engagement illustre également ces zones conceptuelles charnières. Enfin, le pôle plus spécifiquement littéraire semble réduit à son étiage, si l'on excepte *référentiel* et *fictionnel*.

Du côté de la philosophie, les regroupements en contextes homogènes et donc en facettes de sens sont plus manifestes : on peut opposer un pôle leibnizien à droite (*monades*, « mondes possibles ») à une modalité littéraire (*fiction*, *livre*), qui prolonge la thématique de l'opposition du sensible et de l'intelligible dans le corpus critique (*simulacre*, *vision*).

4.2.4 Système dans les corpus philosophie et critique

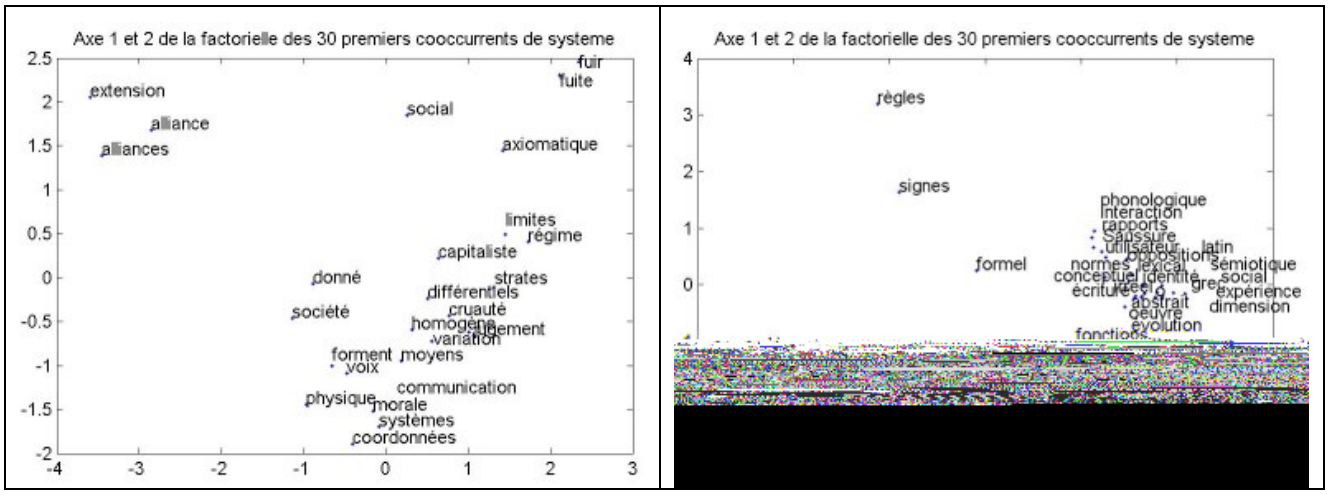

Figure 8 : Les cooccurrents de système
dans les corpus philosophie (à gauche) et linguistique (à droite)

La technicité du concept *système*, partagé par la philosophie et la linguistique, se décline différemment entre les deux corpus. Du côté de la linguistique, les cooccurrents ne s'organisent absolument pas en groupes de contextes. Un lexique technique mais général est mobilisé, à travers des concepts communs à l'ensemble de la discipline (système, signe, règle, fonctions, Saussure) et, corollairement, les marques de leur déclinaison dans des sous domaines (*social*, *sémiotique*, *phonologique*). Du côté de la philosophie, on relève au contraire derrière ce concept technique une très forte hétérogénéité sémantique : les domaines sélectionnés sont nombreux (*physique*, *société*, *capitalisme*, *morale*), et se mélangent avec d'autres concepts du même ordre de technicité apparente : *variation*, *homogène*, *différentiels*, *axiomatique*). Les concepts d'apparence méthodologiques et formels cohabitent donc avec la généralité conceptuelle et l'hétérogénéité sémantique.

5. Conclusion(s)

La comparaison des trois corpus nous a permis d'observer une plus grande intersection entre la philosophie et la critique, que l'on retrouve aux niveaux morphosyntaxique et thématique.

Si philosophie et critique étaient morphosyntaxiquement plus proches, ils attestent d'un fonctionnement thématique proche ; l'exploration thématique a en effet montré que les concepts étaient difficilement discriminés en linguistique, tandis que l'on observait des pôles pertinents et beaucoup plus interprétables sur les corpus critique et philosophique. Les différences thématiques s'observent ainsi sur fond d'un lexique partagé qui témoigne d'emprunts réciproques entre les deux discours.

La linguistique, qui vise à l'objectivation, s'intéresse en effet davantage à ses observables qu'à d'éventuelles thématiques ; on touche ici à une relative inadaptation de méthodes thématiques exploratoires pour des corpus de langue plus spécialisée comme la linguistique.

Il conviendra naturellement d'approfondir et de préciser les observations mises à jour, en augmentant le nombre de concepts sélectionnés et la taille et la diversité discursive du corpus d'étude (e.g. histoire et sociologie).

Références

Ablali D. (à paraître). Contribution de la lexicométrie à l'approche sémantique des corpus. La forme « texte » dans un corpus de critique universitaire. In Williams G., editors, *Les Journées Internationales de Linguistique de Corpus*, Presses universitaires de Rennes.

Grefenstette G. (1996). Evaluation Techniques for Automatic Semantic Extraction : Comparing Syntactic and Window Based Approaches. In Boguraev B. and Pustejovsky J., editors, *Corpus Processing for Lexical Acquisition*, The MIT Press.

Lebart L. (2004). Validité des visualisations de données textuelles. In *Actes des 7emes Journées internationales d'Analyse statistique des Données Textuelles* : 708-715.

Loiseau S. (2005). Thématique et sémantique contextuelle d'un concept philosophique. In Williams G. editors, *La linguistique de corpus*, Presse Universitaires de Rennes.

Loiseau S. (2003). *CorpusReader, un logiciel d'extraction de traits sur corpus richement annotés*. http://panini.u-paris10.fr/~sloiseau/CR.

Malrieu, D., Rastier, F. (2001). Genres et variations morpho-syntaxiques. In Daille B., Romary R., *Traitement automatique des langues : linguistique de corpus*, vol. 42 n°2, Atala/ Hermès : 547-577.

Poudat, C. (2003). Characterization of French linguistic research papers with morphosyntactic variables. In Fløttum K. & Rastier F., editors, *Academic discourses — Multidisciplinary Approaches*. Novus.

Poudat C., Loiseau S., (sous presse). Authorial presence in academic genres. In Tognini Bonelli E., editors, *Strategies in Academic Discours*, John Benjamins.

Rastier, F. (2001). *Arts et Sciences du texte*. Presses Universitaires de France.

Schütze H. (1998). Automatic Word Sense Discrimination. *Computational Linguistics*, vol. 4, n° 1.