Exploration morpho-sémantique en corpus. Les formes du genre de l'article scientifique en sciences humaines

Driss Ablali[1], Margareta Kastberg Sjöblom[2]

[1]Laseldi EA 02 281, Université de Franche-Comté, UFR SLHS, 30, rue Mégevand, 25 030 Besançon cedex, driss.ablali@univ-fcomte.fr

[2]ATST, Centre Jacques Petit, EA3187, Université de Franche-Comté, UFR SLHS, 30, rue Mégevand, 25 030 Besançon cedex, margareta.kastberg@univ-fcomte.fr

Abstract

The purpose of this paper is to carry out an exploratory study of a corpus in order to observe genre-related variations in the semantic field in the area of Corpus Linguistics. The corpus is extracted from several scientific articles in the Humanities, where seven different fields have been selected among the most representative of Social Sciences, (Linguistics, Literature, Philosophy, Sociology, Anthropology, History and Geography) and it contains over five million tokens. In order to observe the scientific article as a distinctive type of text and to obtain indications of divergence and of convergence between the different fields in this particular register, this analysis explores statistical techniques and applies lexical tools. The exploration of the Hyperbase tool (version 6.5) allows the extraction of lexical forms and tokens, statistically specific for this corpus, as well as isotopy correlations, networks and constellations, and makes it possible to observe, at a semantic and at a lexical level, the various associations between the different items of the corpus. This exploration makes it possible not only to observe and to distinguish the thematic fields of the different components of the corpus, but, furthermore, to, beyond the specific character of each field, characterize and describe a common characteristic, a real core, representative for the scientific article.

Résumé

Ce travail a pour ambition d'ouvrir un espace de confrontations sur la question de la sémantique des formes dans le domaine de la linguistique de corpus. Il s'agit d'un travail exploratoire sur sept domaines parmi les plus représentatifs des sciences humaines, *(linguistique, littérature, philosophie, sociologie, anthropologie, histoire et géographie)*, avec l'ambition de montrer que les concepts scientifiques peuvent être étudiés comme des thèmes. Et comme les formes ne prennent leur sens que dans un environnement contextuel en opérant de façon située dans un contexte donné, on s'attachera à voir leur distribution en tenant compte de leur entourage lexical et morphosyntaxique pour souligner combien est arbitraire la frontière du mot et combien s'avère indispensable une typologie des contextes dans le cadre d'une exploration sémantique du discours. L'accent sera mis sur les formes saillantes partagées par tous les discours à partir des réseaux isotopiques et des associations thématiques, réunissant les différents lexèmes les uns aux autres. Au moyen du logiciel Hyperbase, on cherchera à savoir s'il existe, derrière les formes partagées, un noyau dur commun aux différents sous-corpus, et qui serait spécifique aux sciences humaines, ou des variations d'acceptions fortement caractérisantes de chacun des discours.

Mots-clés : discours scientifique, genre, discours, morphologie, corrélations, textométrie.

1. Introduction

Dans le domaine de la recherche, l'article scientifique intéresse beaucoup la communauté des chercheurs, non pas dans une optique générique, mais en tant que texte qui fixe et met en débat la pensée de son signataire pour gagner la crédibilité des collègues. Et les études

génériques aussi bien en linguistique textuelle que dans les théories littéraires vont dans le même sillage : les études menées sur les genres textuels tendent à délaisser l'appréhension du texte scientifique comme genre spécifique. Il suffit de feuilleter une bibliographie exhaustive sur la généricité pour prendre la mesure du phénomène. En effet, les travaux portent davantage sur les textes littéraires, politiques ou journalistiques que sur les genres mineurs, comme les recettes de cuisine, les carnets de route ou le journal de bord. Mais quelques travaux récents nous paraissent pourtant plaider en faveur de la pertinence d'une exploration générique de l'article scientifique dans sa production textuelle.

Cet article explore, à travers la comparaison de sept domaines parmi les plus représentatifs des sciences humaines, *(linguistique, littérature, philosophie, sociologie, anthropologie, histoire et géographie)*, les indices de divergence et de convergence entre domaines que permettent d'observer les méthodes exploratoires de données textuelles pour le genre de l'article scientifique. L'objectif est de décrire, d'un point de vue lexical, les similitudes et les divergences entre sept corpus appartenant aux sciences humaines. Il s'agit donc d'un travail largement exploratoire qui cherche à détecter les constantes et les variations entre les domaines étudiés au sein du même genre, et à identifier les éléments fondamentaux de l'écriture pour ce dernier.

Les analyses menées sont fondées sur des analyses aux niveaux lexical et sémantique afin 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 domaines.

Notre propos ne vise pas tant à classer qu'à étudier ce qui peut rapprocher les discours sur le plan sémantique. Le parcours que présente cet article, on l'aura compris, n'aspire pas à l'objectivation : aucun des corpus utilisés ne peut prétendre représenter son discours d'appartenance. Il s'agit d'un travail dont l'ambition est d'ouvrir un espace de confrontations sur la question de la sémantique des formes dans le domaine de la linguistique du corpus.

2. Corpus d'étude

Le corpus est constitué de 700 articles, extraits de 21 revues francophones de sciences humaines, publiées entre 1990 et 2007. Il comprend exclusivement des articles intégraux et non des extraits. Il se répartit sur sept discours scientifiques distincts dont la réunion est justifiée par une proximité académique : leur appartenance au domaine des sciences humaines. Ils partagent également le même espace éditorial : ce sont tous des articles de recherche publiés dans des revues universitaires, et vu la taille de ce corpus, on pense que l'on peut le considérer comme un « échantillon » de la population de l'écriture universitaire, et plus précisément de l'article scientifique en sciences humaines. Les textes du corpus comptent 5 656 084 occurrences et 117 154 formes graphiques, repartis sur sept sous-corpus de taille relativement homogène.

Le tableau ci-dessous regroupe dans la colonne de gauche le domaine du discours ; la colonne du milieu comprend les revues dont sont extraits les articles, et celle de droite l'étendue de chaque domaine en nombre d'occurrences :

Discours	Revues	Occurrences
Géographie	*Cybergéo* *Les Cahiers de géographie du Québec*	714 057
Sociologie	*Criminologie* *Sociologie et société* *Recherches sociographiques* *Enfances, familles, générations*	841 499
Littérature	*Revue interdisciplinaire sur les textes modernes* *Etudes françaises* *Etudes littéraires* *Textes*	715 178
Ethnologie	*Anthropologie et société* *Etudes inuit studies*	931 034
Philosophie	*Methodos* *Philosophiques*	875 804
Histoire	*Revue d'histoire du XIX^{e} siècle* *Revue d'histoire de l'Amérique* *Les Cahiers d'histoire*	847 701
Linguistique	*Cahiers de praxématique* *Recherches linguistiques de Vincennes* *Cahiers de linguistique française* *Revue québécoise de linguistique*	730 811

Figure 1 : Le corpus

Ce corpus a été traité et exploré avec le logiciel *Hyperbase*[1], désormais bien connu, dans sa version 6.5 qui permet le traitement des formes graphiques et des lemmes en parallèle. En effet, grâce à une lemmatisation effectuée au préalable par l'analyseur *Cordial*, nous pouvons traiter non seulement les mots, mais aussi les lemmes, les codes grammaticaux, ou encore les enchaînements syntaxiques. L'exploration statistique du logiciel donne la possibilité d'analyses diverses, non seulement traditionnelles comme celles sur la richesse lexicale, l'accroissement lexical, la distance lexicale, la corrélation chronologique etc., mais plusieurs fonctions thématiques sont également disponibles. Le logiciel *Hyperbase* permet d'analyser les spécificités lexicales aussi bien d'un point de vue endogène qu'exogène, en donnant accès aux corrélats thématiques. Il recense aussi tous les termes situés dans l'environnement immédiat d'un mot donné afin d'en extraire le réseau isotopique.

3. Exploration lexicale

Quelles sont les caractéristiques thématiques de notre corpus ? Ce ne sont pas les thèmes à proprement dit qui nous intéressent, mais plutôt leur textualité. En d'autres termes, il s'agit de voir si derrière la spécificité terminologique de chaque domaine, il existe un noyau dur au

[1]Des informations détaillées sur le logiciel Hyperbase sont disponibles à l'adresse suivante : www.unice.fr/bcl.

niveau lexical qui maintient un équilibre global en présence de déséquilibres partiels générés par chaque discipline. Nous avons cherché les items lexicaux spécifiques de ce discours avec le recours au logiciel *Hyperbase* qui, de façon très précise, permet d'analyser les spécificités des différents textes.

L'analyse des spécificités est une démarche classique, que le logiciel accomplit en s'appuyant sur *Frantext*, et plus précisément sur le corpus du XX^{e} siècle ; elle permet, au niveau exogène, de mettre en relief les spécificités lexicales de notre corpus. La liste ci-dessous donne à voir pour chaque item répertorié, de gauche à droite, l'écart, mesurant la spécificité, le nombre d'occurrences dans le corpus de référence, et sous la rubrique « texte », le nombre d'occurrences dans le corpus faisant l'objet de notre recherche. Il est aisé de constater dans cette liste hiérarchique la présence de mots caractérisant[2] le discours scientifique en tête de liste :

Ecart	Corpus	Texte	Mot	Ecart	Corpus	Texte	Mot
626.30	40921	50558	)	173.00	1371	2455	recherche
585.39	96	2034	processus	172.72	187	859	dimension
579.24	40245	46765	(	167.45	185	829	catégories
450.12	120	1754	canada	165.08	517	1397	construction
356.72	513	2920	analyse	162.96	303	1043	interprétation
347.26	142	1478	activités	161.11	2281	3041	discours
328.82	69	972	sociologie	160.99	2326	3072	rapport
296.44	121	1166	sartre	159.39	274	969	définition
276.11	660	2594	social	158.01	100	571	utilisation
262.60	469	2071	relation	157.17	1049	1945	permet
261.08	248	1484	structure	155.28	108	584	analyses
257.04	339	1716	pratiques	154.14	777	1626	niveau
237.33	389	1705	sociales	154.02	119	609	contraintes
237.11	166	1100	sociaux	152.25	399	1129	production
222.33	866	2430	sociale	151.48	308	981	poincaré
203.18	73	622	facteurs	150.17	206	789	information
201.09	138	852	représentation	149.10	810	1612	fonction
200.38	538	1715	théorie	148.98	813	1614	cadre
187.87	528	1597	notamment	148.14	179	724	économiques
186.63	535	1598	données	146.74	224	806	philosophique
184.30	979	2173	culture	146.24	115	569	médicaments
183.93	323745	83832	des	142.96	474	1165	textes
182.05	108	682	caractéristique	142.64	131	594	mathématique
181.00	646	1714	développement	141.75	140	611	international
180.45	256	1054	sociétés	141.07	550	1245	population
178.91	433	1374	identité	140.70	1279768	241744	de
178.73	432	1371	sciences	140.52	1306	1978	système
177.87	505	1481	université	138.88	4038	3694	selon
177.43	1227	2367	texte	138.54	2273	2656	espace
175.47	390	1277	économique	131.71	153	596	constituent
173.00	1371	2455	recherche	131.26	459	1057	individus
175.47	390	1277	économique	131.21	770	1394	notion

Figure 2 : Vocabulaire spécifique du corpus

Nous trouvons d'un côté les mots qui structurent le travail, l'analyse et la recherche scientifiques, avec des items comme *analyse*, *activité*, *structure*, *pratiques*, *théorie*, *développement*, *sciences*, *université*, *recherche* etc., et de l'autre les termes désignant l'objet du travail scientifique comme *sociologie*, *social*, *société*, *texte*, *économique*, *discours* etc. Au niveau de la ponctuation, des signes comme les parenthèses permettent de caractériser ce genre d'écriture. Cette construction phrastique particulière est due à la structure

[2] Pour éviter certaines ambiguïtés nous nous basons dans cette analyse sur la forme graphique et non sur le lemme.

argumentative et démonstrative de l'article scientifique, dont la thèse constitue le parangon. Cette économie de pensée est un principe épistémologique résultant de processus cognitifs aussi importants que l'objectivation et la rationalisation. Et comme le dit J.-M. Berthelot (2003 : 28) « La science réduit les dimensions du réel, résumé les expériences, épure son vocabulaire ».

On peut constater que cette liste ne contient pas de verbe, ni même d'adjectif ou d'adverbe. Il n'y a aucune articulation du discours : la phrase semble constituée de substantifs juxtaposés. Il s'agit ici évidemment d'un effet de genre textuel, reflété par la grande présence d'un discours intellectuel, faisant appel au substantif. Cet intérêt prononcé pour le substantif n'est pas fortuit, il détermine le recours aux candidats concepts des différents domaines du corpus, comme il renvoie aux méthodologies scientifiques à l'œuvre. On peut aussi noter que, comme l'auteur n'est pas supposé se mettre en valeur de manière explicite, les adjectifs et les adverbes qui renforcent et valorisent la première personne sont moins présents dans la phrase scientifique.

Or, dans la liste ci-dessous nous trouvons les spécificités négatives, c'est-à-dire les mots statistiquement sous-employés dans le corpus. Ici, nous pouvons constater encore le même effet du genre avec un déficit important de pronoms personnels, bien plus représentés dans le corpus de référence que dans ce corpus. L'écriture scientifique emploie davantage de pronoms personnels *nous* et *on* au détriment de la première personne comme *je, me*, *moi* et *mon*. Le chercheur en effet n'écrit pas en tant que sujet de la vie quotidienne, mais en tant que figure appartenant à un domaine d'activité. On note aussi la même chose concernant les auxiliaires à la première personne du singulier, *ai, suis, nous*, référant à la personne privilégiée par l'auteur pour actualiser son discours et développer ses hypothèses de recherche.

Notons au passage aussi le déficit des différents signes du dialogue ainsi que du point, reflétant une autre caractéristique de ce discours : une phrase longue et énumérative. Le déficit du point est compensé par l'excédent de la virgule, ce qui permet également de confirmer la longueur de la phrase scientifique. Ce discours est en effet caractérisé par des phrases énumératives qui procèdent par accumulation, donnant au discours un caractère parfois répétitif, à cause du peu de variété syntaxique et de l'itération des mêmes structures. Ainsi, nous constatons que la phrase scientifique ne change pas beaucoup d'un corpus à l'autre, qu'il n'y a aucune tendance disciplinaire. La longueur de la phrase dépend, en réalité, surtout du genre. Ici, pour le rappeler, c'est le même genre qui est en question, mais dans des disciplines différentes. En d'autres termes, la discipline n'a pas d'influence considérable sur la ponctuation de la phrase, dès lors que le genre est le même.

Quant à l'énumération, elle est liée aux objets étudiés, aux exemples cités ainsi qu'aux ouvrages mentionnés. Cette énumération développe une stratégie textuelle qui assure une gestion scientifique pour la lisibilité et pour l'aspect démonstratif de l'article. Elle correspond à des normes de nature prescriptives, qui réguleraient les pratiques d'écriture de l'article.

Ecart	Corpus	Texte	Mot	Ecart	Corpus	Texte	Mot
-226.94	396110	7169	je	-88.70	72487	2107	dit
-178.98	576485	35782	il	-86.93	205106	15894	n'
-159.10	171164	1652	vous	-82.97	55831	1196	ma
-140.55	157060	3171	j'	-74.19	51679	1558	rien
-139.96	141768	2010	me	-72.94	62241	2622	là
-132.04	235834	11634	elle	-71.82	431159	45949	que
-127.43	148707	4270	était	-71.78	41150	842	suis
-122.28	124605	2877	avait	-70.63	51596	1836	quand
-116.64	89602	704	tu	-67.54	115651	8691	sa
-115.12	305816	21951	pas	-56.10	35823	1458	jamais
-113.62	147626	6044	lui	-65.02	103601	7661	si
-113.27	92788	1311	m'	-62.50	143578	12492	son
-108.69	267396	19002	ne	-62.08	157506	14175	mais
-106.68	92692	2010	ai	-61.29	30000	614	yeux
-106.35	300597	23078	qu'	-61.18	31842	775	mes
-102.92	82020	1519	moi	-60.22	446799	50650	un
-100.29	79885	1607	mon	-58.28	97700	7760	bien
-99.09	1652388	194888	.	-58.06	36888	1425	puis
-89.65	136791	8170	tout	-57.14	90585	7086	ses
-88.95	58150	883	ça	-56.38	202774	20479	se

Figure 3 : Vocabulaire spécifique négative du corpus

4. De l'autonomie du lexique

Le logiciel permet également l'observation du vocabulaire spécifique de chacun des sous-corpus, c'est-à-dire une comparaison endogène. Cette spécificité est déterminée par le calcul de l'écart réduit pour chaque forme dans chaque partie du corpus. Les textes sont comparés, les uns après les autres, avec le corpus dans son ensemble. Ces comparaisons internes se justifient facilement, puisque le corpus est expressément conçu pour mettre en valeur les différences qui opposent les textes dans ce même ensemble. S'il est homogène, le calcul relèvera, comme ici, toujours des écarts intéressants.

Littérature	*Linguistique*	*Philosophie*	*Histoire*	*Sociologie*	*Ethnologie*	*Géographie*
personnage	énoncé	Poincaré	historien	sociologie	médicament	eau
autobiographie	linguistique	mathématique	révolution	goût	anthropologie	quartier
tournant	corpus	Kant	histoire	couple	maladie	échelle
narrateur	verbe	physique	siècle	social	autochtone	spatial
autobiographique	locuteur	géométrie	Canada	artiste	culture	mer
autofiction	dénomination	mathématique	Montréal	Bourdieu	communauté	géographie
récit	lexical	texte	Québec	parental	anthropologue	géographique
écrire	sémantique	Descartes	français	enfant	musulman	pôle
roman	syntaxique	science	madawaska	galerie	islam	port
écriture	exemple	Locke	républicain	famille	tuberculose	bungalow

Figure 4 : Spécificités lexicales des sous-corpus

Les résultats sont très nets, les dix premiers mots reflètent parfaitement le profil caractéristique de chaque domaine. Si l'on jette un coup d'œil sur les spécificités lexicales de la « géographie », par exemple, on a des lexèmes tels que : *territoire*, *sud*, *eau*, *géographie,*

urbain, *silicium*, *spatiale* et *quartier*. Le corpus « linguistique » est caractérisé par des mots comme *linguistique*, *corpus, langues*, *sémantique*, *dénomination*, *verbe*, *syntaxique* et *phonologie*, ce qui n'est pas très étonnant. En revanche, on se rend compte facilement que la forme *langue*, objet pourtant intuitivement premier de la linguistique, est détrônée par d'autres formes, comme *énoncé*, *corpus*, *verbe*, et surtout *exemple*, qui montre clairement que l'exemple, comme données attestées, reste l'objet de prédilection des linguistes. La linguistique, qui vise à l'objectivation, s'intéresse en effet davantage à ses observables qu'à d'éventuelles thématiques. Un autre fait, non des moindres, à souligner concernant la philosophie, qui compte parmi ses dix premiers mots, et contre toute attente, un lexème comme *texte*, auquel on s'attendait en linguistique. Mais cela ne fait que confirmer que le texte a encore du mal à s'imposer dans le giron de la linguistique, qui fait de la phrase son principal cheval de bataille.

Avant d'aller plus loin dans notre exploration, regardons d'abord ce qui lie et divise les sept domaines en fonction de leur lexique ; c'est donc par l'analyse de la distance – ou connexion – lexicale, que nous nous proposons de commencer cette partie de notre analyse.

L'analyse arborée[3] de la distance lexicale de notre corpus fait apparaître immédiatement les spécificités du lexique scientifique.

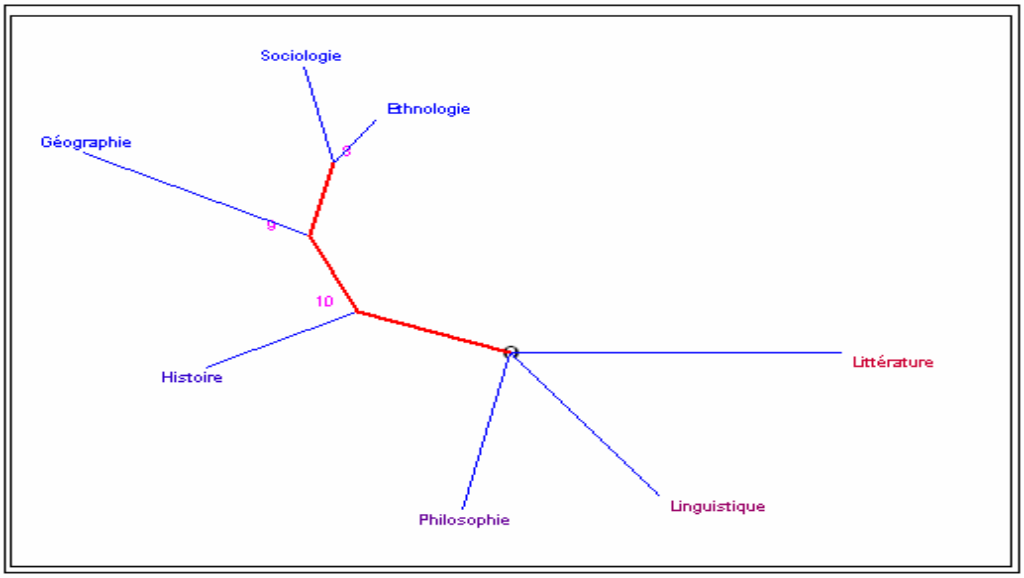

Figure 5 : Analyse arborée de la distance lexicale des sous-corpus

Le calcul de la distance entre les vocabulaires des sept corpus vise en effet à répondre à la question suivante : quels sont les textes les plus proches et les plus éloignés du point de vue de leur contenu lexical et thématique ? Les branches de l'arbre permettent de constater

[3] La technique de l'analyse arborée élaborée par Xuan Luong permet de représenter les résultats du calcul de la distance lexicale d'une façon différente. L'algorithme produit des graphes qui rendent compte de la proximité, ou de l'éloignement des textes étudiés en une seule représentation graphique, sous forme radiale. Le modèle de l'arbre est un graphe connexe et sans circuit, et il est caractérisé par l'ensemble des distances entre ses éléments, la longueur des branches représentant fidèlement la distance entre les textes. Par ailleurs, sa structure est aussi importante, faisant apparaître l'ordre et la force des regroupements ou des oppositions entre les différents éléments. L'avantage de cette technique par rapport à l'analyse factorielle, est qu'on n'a plus à distinguer et à croiser des facteurs, dont chacun n'explique qu'une partie de la variance. L'analyse arborée permet également, grâce à la représentation par branches, d'éviter les inconvénients du saut minimal, la technique employée dans les dendrogrammes.

plusieurs regroupements de textes et rend compte de la spécificité et la proximité, ou de l'éloignement thématique, des textes. En bas de l'arbre, et sur la même branche, se trouvent ensemble « philosophie », « linguistique » et « littérature ». Cette réunion est justifiée par une proximité domaniale. En effet, les trois disciplines partagent en grande partie le même univers lexical : l'étude du sens et de la signification, de la subjectivité et du sujet, de l'homme et des textes. Il s'agit d'un savoir totalisant une réflexion visant une interprétation globale du monde, du langage et des œuvres. La proximité s'observe ainsi sur fond d'un lexique partagé qui témoigne d'emprunts réciproques entre les trois corpus. En haut de l'arbre, l'ethnologie et la sociologie sont attachées également à une même branche : la connexion thématique entre les deux corpus s'explique par la nature de l'objet d'étude qui concerne l'ensemble des caractères sociaux et culturels des groupes humains. A la gauche de l'arbre sur une autre branche se trouve, à l'écart des autres, la géographie. Le discours du géographe, il faut le rappeler, est le seul qui ne puisse se passer des graphiques. Ce qui signifie que l'étude de l'espace des sociétés, ou de la dimension spatiale du social, c'est-à-dire la façon dont les sociétés établissent les distances qui séparent leur composants (individus, entreprises, États, ressources, etc.), a une autre assise, autre que lexicale. Ce qui pourrait expliquer sa grande distance par rapport au lexique des autres corpus. Quant au discours de l'historien dont l'objet est l'étude des faits et des événements du passé, il développe également ses thématiques dans un lexique et un vocabulaire à l'écart des autres. Les différents domaines semblent en effet sensibles au lexique, qui les divise assez nettement, tout comme les structures encadrantes semblent les réunir.

5. Noyau dur lexico-sémantique en sciences humaines

Si nous avons obtenu les spécificités lexicales des sept sous-corpus, nous ne disposons pas de leurs intersections lexicales et conceptuelles, essentielles pourtant à une entreprise d'exploration thématique. Notre objectif maintenant est d'extraire la liste des formes lemmatisées partagées par les sept sous-corpus pour examiner leur(s) intersection(s). Ce qui nous permettra de sélectionner les formes dont nous observerons le fonctionnement en corpus, à partir de l'extraction de cooccurrences et de réseaux isotopiques, associant les différents lexèmes les uns aux autres.

L'analyse des intersections entre les formes lexicales les plus fréquentes dans les sept discours laisse toutefois entrevoir la présence d'un univers conceptuel commun aux sciences humaines.

Regardons de près les lexèmes qui semblent caractéristiques de ce discours et qui l'identifient au niveau de la typologie. Certains éléments lexicaux forment en effet un cadre général dans le discours de l'article scientifique, comme des pierres angulaires du discours, servant à tenir l'argumentation intellectuelle. Parmi les mots les plus spécifiques de notre corpus figurent des items comme *analyse*, *donnée*, *méthode*, *cadre*, *résultat*, *étude*, *question*.

Les sept discours observés se partagent d'abord l'objet *analyse*, 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 *donnée*, *méthode*, *cadre*, *résultat*, *étude*, *question*. Les intersections obtenues confirment la plus grande proximité des sept disciplines qui partagent un nombre sensiblement plus important de formes. Car malgré le fait que les sept disciplines pensent des phénomènes complexes avec des outils différents, nous sommes en présence d'un paradigme lexical qui s'impose non sans flottements sémantiques.

Le premier mot *analyse*, le plus spécifique de ce corpus, pourrait effectivement être qualifié comme étant la clé de voûte qui régule l'activité de recherche de l'article scientifique au sein de la chapelle des sciences humaines. Pour analyser son environnement thématique, les éléments qui l'entourent, nous avons procédé à l'extraction thématique de cet item lexical. L'extraction automatique du contexte d'un item lexical par le logiciel *Hyperbase*, ici le paragraphe, permet la création d'un sous-corpus qui est soumis à un calcul de spécificité particulier, puisqu'on recherche une relation privilégiée entre les lexèmes eux-mêmes à l'intérieur de ce sous-corpus. Cette procédure tient compte de l'ensemble indéfini de tous les mots qui peuvent se trouver dans l'entourage du pôle a*nalyse*. En confrontant *analyse* au sous-corpus constitué par les mots qui gravitent autour du pôle, nous pouvons extraire l'environnement thématique suivant (ordre hiérarchique, début de liste) :

Ecart	Corpus	Texte	Mot	Ecart	Corpus	Texte	Mot
37.58	2920	2917	analyse	8.27	77	18	approfondi
33.09	115529	2118	l'	8.20	105	20	comparative
20.71	83832	2118	des	8.16	1945	86	permet
16.31	241744	4948	de	8.16	1074	60	résultats
15.44	54402	1352	une	8.11	931	55	montre
14.73	71707	1677	d'	8.07	597	43	propose
13.73	2823	157	notre	7.62	1614	73	cadre
12.23	1598	104	données	7.55	149	21	morphologique
11.72	23007	613	sur	7.31	425	33	section
11.58	3041	145	discours	7.21	30	11	multicritère
11.40	102	29	détaillée	7.17	2017	81	avons
10.90	13230	387	nous	7.02	710	42	méthode
10.62	14887	419	cette	6.89	418	31	variables
9.82	452	45	spatiale	6.89	131	18	pertinent
9.71	37	17	conversation	6.88	1288	59	questions
9.13	16	12	factorielle	6.80	351	28	méthodes
9.09	808	56	corpus	6.79	1699	69	critique
8.95	281	33	classification	6.70	220	22	propose
8.77	399	38	auto	6.60	622	37	facteurs
8.58	104	21	hiérachique	6.52	82	14	qualitative

Figure 6 : Environnement thématique de l'item analyse

Ce tableau reflète bien que les items qui entourent le pôle *analyse* ne sont pas les objets de l'analyse scientifique que décrit l'article, mais qu'il s'agit bien de l'articulation de ce discours essentiellement nominal, faisant appel aux substantifs et aux adjectifs. Le premier pronom au pluriel domine encore l'article scientifique qui, comme nous l'avons déjà dit, emploie très peu de verbes. En effet, il semble que l'articulation même de ce discours, les lexèmes qui organisent sa structure font appel à un réseau lexical bien spécifique pour l'article scientifique.

Nous avons cherché aussi, non seulement l'univers lexical qui entoure un mot-pôle, mais aussi les cooccurents les plus proches d'un mot-pôle spécifique pour ce discours, *problème.* On aborde ici les isotopies spécifiques, non seulement à partir des fréquences, mais par une autre approche : celle de l'étude des séquences. Cette analyse permet de considérer les associations de mots dans leur environnement immédiat, en ignorant la partition des textes.

L'extraction des corrélats par le logiciel *Hyperbase* regroupe dans cette analyse les substantifs. Ensuite, une analyse d'associations[4] reprend ce tableau des substantifs, calcule les distances et trie le détail des associations deux à deux. Le graphique ci-dessous montre les résultats de l'analyse des associations du mot *problème*.

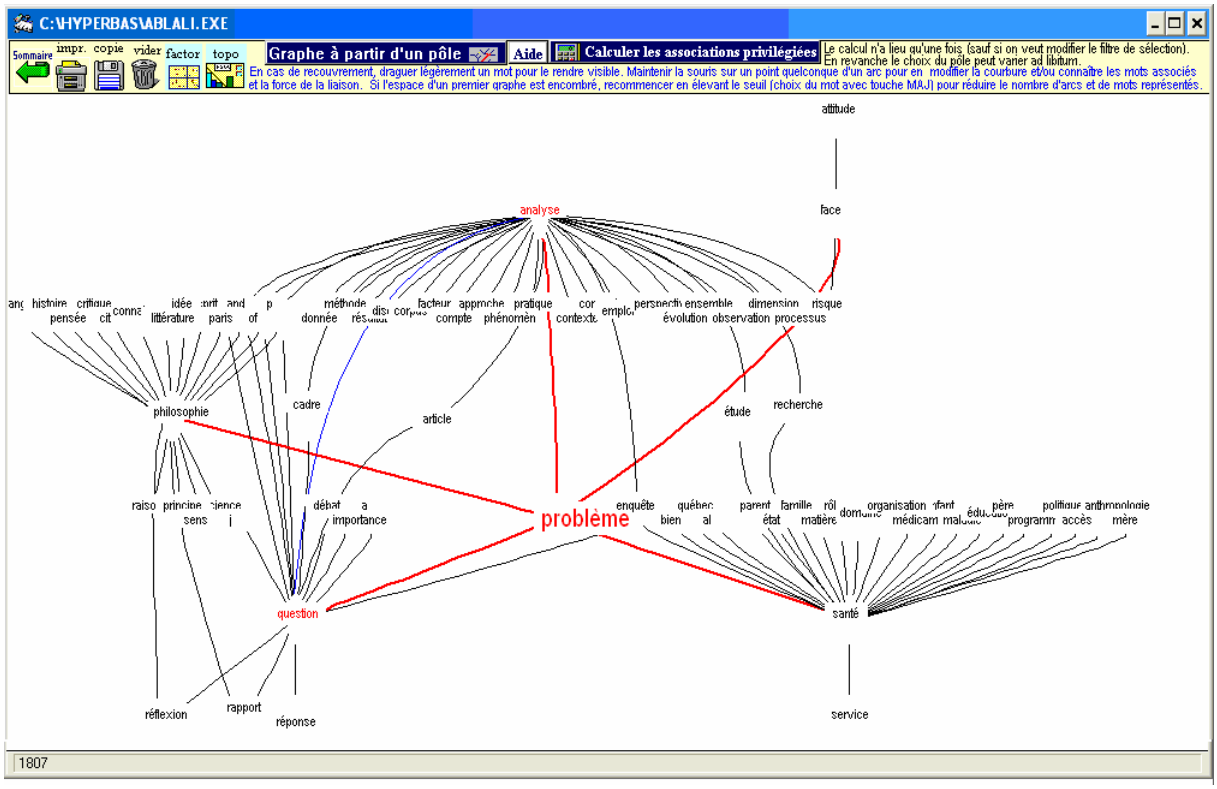

Figure 7 : Associations privilégiés du mot problème

Les mots en gras correspondent ici à des nœuds de forte fréquentation, et les mots en petite police correspondent à des nœuds moins fréquentés, n'ayant pas de fréquentation directe avec le mot-pôle. Les traits gras correspondent aux cooccurrences directes avec le pôle, et les traits fins aux cooccurrences indirectes, c'est-à-dire en cooccurrence avec les cooccurrents.

[4] Le calcul du graphe arborescent, des nœuds et des arcs est assuré par le logiciel libre GRAPHVIZ. Les données sont fournies à ce programme selon les spécifications du langage DOT et les résultats bruts sont repris par Hyperbase dans une représentation graphique qui tient compte non seulement des positions mais aussi des pondérations. La mesure de la cooccurrence est empruntée au Rapport de Vraisemblance, proposé par Dunning en 1993. Cet indice s'appuie sur quatre paramètres

- a : nombre de cooccurrences des deux mots dans le champ exploré (ici le paragraphe)

- b : nombre d'occurrences du premier mot en l'absence du second

- c : nombre d'occurrences du second mot en l'absence du premier

- d : nombre d'occurrences des autres mots RV = -21ogL = 2(sl-s2)

pour s 1 = a log a + b log b + c log c + d log d + (a+b+c+d)log(a+b+c+d)

s2 = (a+c)log(a+c) + (b+d)log(b+d) + (a+b)log(a+b) + (c+d)log(c+d)

A partir de la valeur 4, l'indice de Dunning est considéré comme échappant au hasard, au seuil de 5%.

Il est aisé de constater que ce pôle, *problème*, central dans le discours de l'article scientifique, fait appel en premier lieu à d'autres éléments qui structurent le discours et non pas comme on aurait pu s'y attendre aux différents thèmes de nos sept domaines des sciences humaines.

En effet, les cooccurrents les plus forts de *problème* sont *question* et *analyse*. Autour du mot *analyse* gravitent les termes qui articulent les méthodes de recherche, autour de *question* les mots argumentatifs, et ce n'est que dans le deuxième ou le troisième cercle que nous trouvons réellement des thèmes qui sont l'objet et le but de la problématique scientifique avec *santé*, entouré des termes reflétant l'organisation médicale, et de *philosophie* faisant appel aux termes plus abstraits de ce domaine plus intellectuel.

6. Conclusion

L'objectif de cette étude n'était pas de dresser un parangon de l'article scientifique en sciences humaines. Elle n'avait pas non plus la prétention de traiter tous les aspects sous lesquels on pourrait définir le discours scientifique. Un genre peut être appréhendé sur différents niveaux, ce qui signifie que son identité est toujours relative.

En revanche, cette approche contrastive des sept corpus nous a permis d'observer une plus grande intersection entre les sept domaines, intersection que l'on retrouve aux niveaux morphosyntaxique et thématique. Les différences thématiques s'observent ainsi sur fond d'un lexique partagé qui témoigne d'un noyau dur éclipsé derrière les spécificités lexicales de chaque domaine. Il ressort en effet de notre analyse qu'au-delà d'un lexique divergent selon les différents domaines disciplinaires, il se tisse un lien lexical commun qui n'est pas seulement un jargon terminologique, mais un vocabulaire qui véritablement structure l'article scientifique. En outre, le profil morphosyntaxique qui émerge de nos différentes analyses est celui d'une écriture qui privilégie la catégorie du substantif, la virgule, la parenthèse, avec la juxtaposition des éléments nominaux. L'article scientifique obéit bien à une volonté de forme précise très prononcée, voire homogénéisée, au niveau morphosyntaxique aussi bien qu'au niveau sémantique. Il conviendra naturellement d'approfondir et de préciser les observations mises à jour, en mettant l'accent sur des phénomènes énonciatifs en relation avec la question du sujet, sur les temps verbaux en rapport avec les pronoms personnels.

Références

Ablali D. (2006). Ecrire en critique. Exploration morphosyntaxique sur corpus. In Rastier F., Ballabriga M. (dir.), *Corpus en Lettres et Sciences sociales : des documents numériques à l'interprétation*, pp. 207-214.

Ablali D. (in press). Contribution de la lexicométrie à l'approche sémantique des corpus. La forme *Texte* dans un corpus des études littéraires. In Williams G. (éd.), *Les 4es journées de la Linguistique du corpus.* Rennes, Presses universitaires de Rennes.

Ablali D. (in press). La fabrique de l'article scientifique en sciences humaines. Exploration sur corpus. In Williams G. (éd.), *Les 5es journées de la Linguistique du corpus*. Rennes, Presses universitaires de Rennes.

Adam J.-M. (2005). *Les textes types et prototypes : Récit, description, argumentation, explication et dialogue*. Paris, Arman Colin, collection Fac. Linguistique.

Berthelot J.-M. (2003). *Figures du texte scientifique*. Paris, PUF.

Biber D. (1993). Using register-diversified corpora for general language studies. In *Computational Linguistics*, 19(2), pp. 243-258.

Kastberg Sjöblom M. (2006). Peut-on refuser les genres littéraires ? Etude quantitative d'un corpus informatisé. In Olsen M. et Swiatek E. (éd.), *XVI^e Congrès des Romanistes Scandinaves.* Roskilde Universitetcenter, http://www.ruc.dk/cuid/publikationer/publikationer/XVI-SRK-Pub/.

Kastberg Sjöblom M (2006). La sémantique lexicale et les genres : analyse systématique d'un corpus québécois. In Williams G. (éd.), *Les 4^es journées de la Linguistique du corpus*. Rennes, Presses universitaires de Rennes.

Loiseau S., Poudat C., Ablali D. (2006). Exploration contrastive de trois corpus de sciences humaines. In *JADT*, Besançon, Les cahiers de la MSH Ledoux, pp. 631-642.

Rastier F. (1987). *Sens et textualité*. Paris, Hachette.

Reppen R, Fitzmaurice S, Biber D. (éd.). 2002. *Using Corpora to Explore Linguistic Variation.* Amsterdam, John Benjamins.