L'exploitation de la BDTS (Banque de Données Textuelles de Sherbrooke) au moyen d'HYPERBASE

Étienne Brunet

Etienne Brunet

L'exploration de la BDTS *(Banque de Données Textuelles de Sherbrooke)* au moyen d' Hyperbase

Parmi les grands projets de recherche que suscite l'intérêt passionné du Québec pour sa langue, les uns sont d'ordre terminologique[1], d'autres envisagent l'aspect diachronique[2], et certains étudient la langue, telle qu'elle est *hic et nunc*, sans s'interdire de montrer ce qu'elle devrait être. La base que nous nous proposons d'explorer est de cette dernière espèce[3]. Ses deux promoteurs, Pierre Martel[4] et Hélène Cajolet-Laganière, visent à établir, à partir de l'observation, le français standard du Québec, c'est-à-dire « la variété de français socialement valorisée que la majorité des Québécois francophones tendent à utiliser dans les situations de communication formelle. » Le but ultime est la réalisation d'un dictionnaire du français au Québec, qui serait au Québec, mutatis mutandis, ce que le TLF est à la France. La méthode est en effet semblable et empruntée aux sciences de l'observation. On se propose, non pas de partir de dictionnaires existants, surtout s'ils viennent de France[5], mais de constituer un corpus représentatif et d'y observer les usages, qu'ils correspondent ou non à ceux de l'hexagone. Ainsi avait fait Paul Imbs il y a quarante ans quand il fondait les bases du TLF, premier dictionnaire établi sur un corpus informatisé. Mais le projet québécois diffère dans la composition du corpus. Le TLF avait une assise principalement littéraire, même si quelques textes techniques (20 % du total) avaient complété la documentation, pour couvrir certaines zones de la terminologie scientifique, peu familières aux écrivains. Le rapport est inversé dans le corpus québécois où la part dévolue aux écrivains ne dépasse pas 20%. La cible n'est pas le français soutenu, non plus que le parler populaire, mais une norme standard dont la littérature, mais aussi les journaux, l'administration, les essais techniques portent témoignage.

---

[1] Le nom d'André Clas est lié à cette entreprise.

[2] Claude Poirier, responsable de la base *QuébéText*, a la charge du dictionnaire historique.

[3] C'est l'un des onze corpus lexicaux que propose le réseau québécois, à l'adresse : http://www.spl.gouv.qc.ca/corpus/index.html

[4] Pierre Martel, du temps qu'il était à la tête du Conseil de la langue Française au Québec, recommandait dès 1990 la création d'un corpus original québécois. Libéré de sa charge administrative, il a mis en œuvre ce projet avec quelques collègues de Sherbrooke et de Laval.

[5] On peut citer, parmi d'autres exemples, l'édition québécoise du *Petit Robert*. Ces tentatives ont reçu un accueil mitigé. Le temps de la colonisation linguistique est révolu. À l'image du Brésil pour le portugais et des États-Unis pour l'anglais, le Québec affirme son indépendance en matière de langue.

Le dosage des sources documentaires, retenues pour la BDTS, est le suivant : oral 10%, ouvrages didactiques 20%, textes littéraires 20%, articles de journaux 20%, textes spécialisés 30%. À l'heure actuelle les textes dépouillés constituent une masse de 37 millions de mots. D'ores et déjà, une partie de ce corpus est exploitable et disponible sur Internet. On trouvera en annexe la composition de ce corpus provisoire, qui est gros de 2 millions d'occurrences et qui répartit en huit parts égales l'oral, la littérature, les journaux, les textes sociopolitiques, administratifs, environnementaux, technologiques et scientifiques[6]. Ce corpus a été indexé et on peut l'interroger sur Internet, en lui proposant n'importe quelle entrée, comme le mot « yeule »[7] dans l'exemple ci-dessous (figure 1).

Figure 1. Exemple de consultation de la base BDTS sur Internet

[6] Comme le corpus littéraire n'est pas détaillé dans cette annexe, précisons qu'on y trouve un roman (*L'ange exterminé*, de Gérald Bodin), des contes et nouvelles (*Le cassé et autres nouvelles*, de Jacques Renaud, et *Contes sur la pointe des pieds*, de Gilles Vigneault), une pièce de théâtre (*La Dalle-des-Morts*, de F.A. Savard) et un essai de critique littéraire (*Écrire de la fiction*, de Noël Aubert). L'exemple de *yeule* qui fait l'objet de la figure 1 est emprunté au corpus littéraire, d'où le « joual » n'est pas absent. Mais si le joual peut à l'occasion se rencontrer dans la documentation, il n'entre pas dans la nomenclature du québécois standard.

[7] En cinq mots l'orthographe pittoresque et suggestive de la citation évoque le parler du pays, qui n'est pas sans rappeler celui des campagnes dans l'Ouest de la France, où l'on observe pareillement l'allongement et l'épaississement du *a* précédé de *r (fâme)* et le relâchement articulatoire qui transforme en yod une occlusive initiale : *gueule* > *yeule*, *dieu* > *yeu*, *queue* > *yeue*. Ajoutons que *pis* pour *puis* s'entend encore en France en milieu rural, et dans de nombreuses variétés sociolinguistiques.

Cependant dans le prototype mis en place la base se borne à égrener des contextes, un par écran[8], sans donner lieu à quelque synthèse, non plus qu'à une exploitation lexicométrique. Avec l'accord des auteurs[9], les mêmes données ont été soumises à notre logiciel HYPERBASE pour constituer une base hypertextuelle et statistique.

On trouvera ci-dessous (figure 2) le menu principal. Comme on se propose de permettre le téléchargement de cette base – si les réalisateurs de la BDTS le souhaitent et si le copyright ne s'y oppose -, notre ambition est moins d'exploiter et de commenter les résultats auxquels elles conduisent que d'expliquer leurs fonctions et leur mode d'emploi, afin que les chercheurs du Québec en fassent un usage mieux qualifié.

Figure 2. Le menu principal de la base Québec

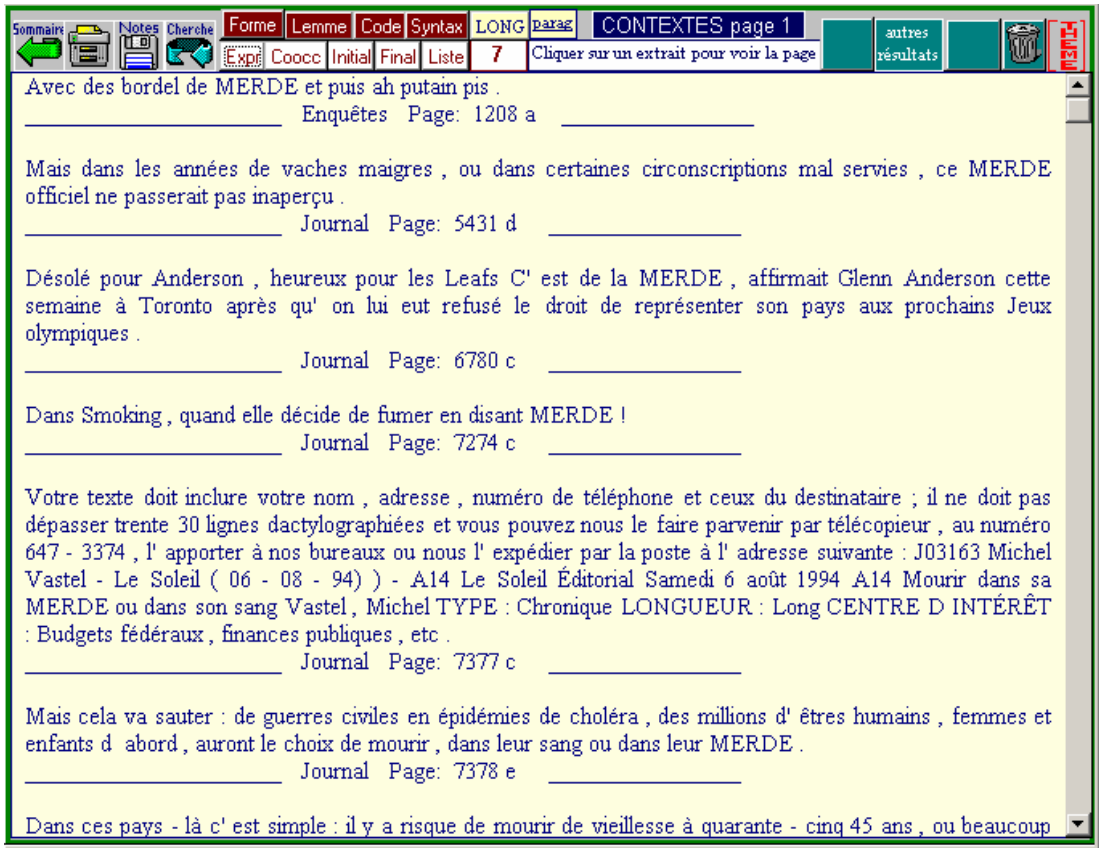
On distinguera deux séries de fonctions : les unes, documentaires, sont groupées horizontalement au haut de l'écran, les autres, vouées à la statistique, occupent la marge droite. On n'insistera guère sur les premières, car leur utilité

[8] Cela ne va pas sans quelque lourdeur, lorsqu'un mot a une certaine fréquence et qu'on ne dispose pas d'un haut débit. Le défilement des contextes se faisant pas à pas et exigeant à chaque citation l'intervention de l'utilisateur, le dialogue s'en trouve ralenti.

[9] Loin d'être hostile aux méthodes lexicométriques, Pierre Martel a publié en 1992 un *Dictionnaire de fréquence des mots du français parlé au Québec* (avec Normand Beauchemin et Michel Théoret).

s'accorde avec leur facilité. Il serait oiseux de s'appesantir sur les programmes de concordance ou de recherche de contextes. Un exemple, relatif au mot de Cambronne, suffit à illustrer cette fonctionnalité (figure 3). Les Québécois ne sont pas bégueules plus que d'autres, mais soucieux des genres et des convenances, ils n'emploient ce mot que dans les accès d'humeur, ce qui n'est guère autorisé qu'au théâtre, dans la conversation ou dans la presse (en réalité, presque tous les exemples sont relevés dans le même journal).

Figure 3. Les contextes du mot de Cambronne

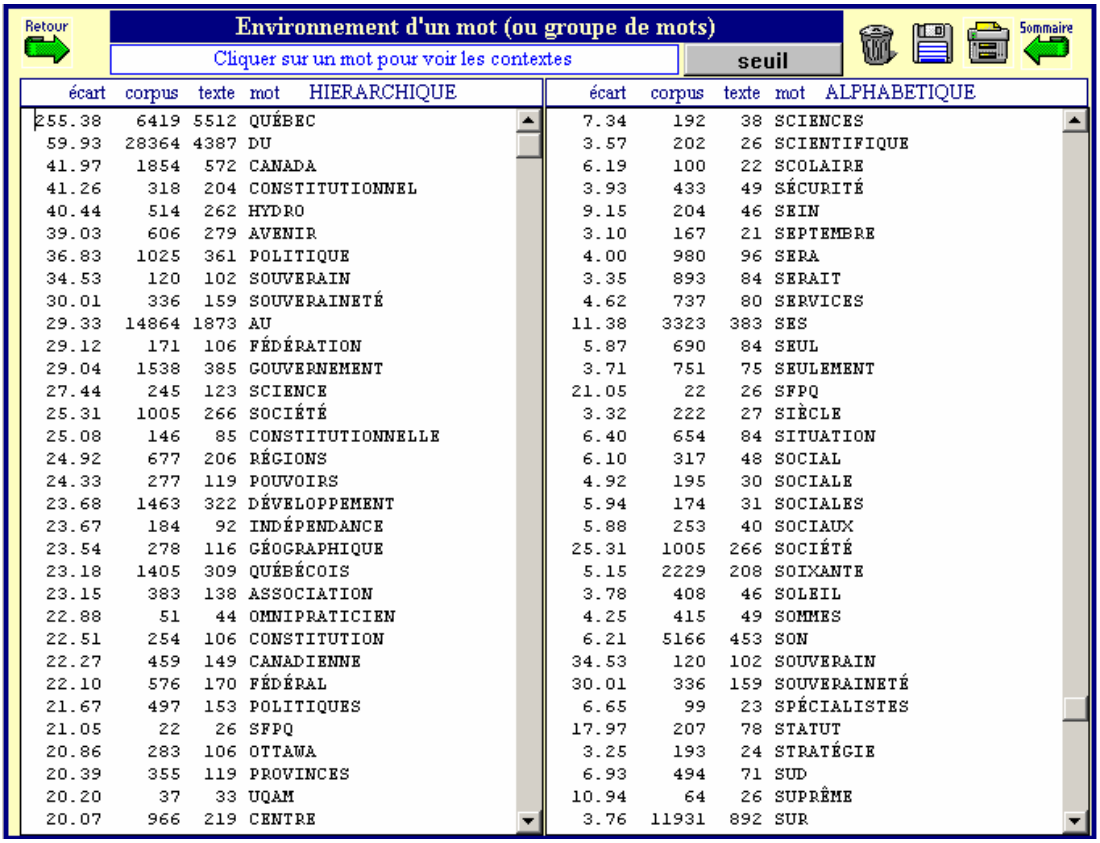


Mais même dans cette fonction documentaire très traditionnelle la statistique pointe son nez. Quand un mot (ou un ensemble de mots) produit une moisson suffisante de contextes (par exemple plus de 5000 pour le mot *Québec*), la fonction THEME observe tous les mots présents dans l'entourage immédiat du mot choisi pour pôle et compare la fréquence de ces corrélats dans ce sous-ensemble à celle qui est la leur dans le corpus entier. Dans la liste qui en résulte (figure 4) on découvre une constellation thématique qui circonscrit ce pays (ou la ville qui porte le même nom), avec une coloration éminemment politique : tous les corrélats sont liés au problème constitutionnel de la fédération et le *Canada* est de loin le terme le plus souvent associé au *Québec*[10], avec 572

[10] Les relations sémantiques ne sont pas seules en cause. La syntaxe explique la présence de certains éléments, comme *du* et *au* qu'impose l'accord avec le mot-pôle.

mentions sur 1854 (la *France* en regard fait piètre figure, avec 23 citations sur 190 et un écart réduit qui dépasse à peine le seuil).

Figure 4. L'environnement lexical du mot Québec (5476 contextes relevés)

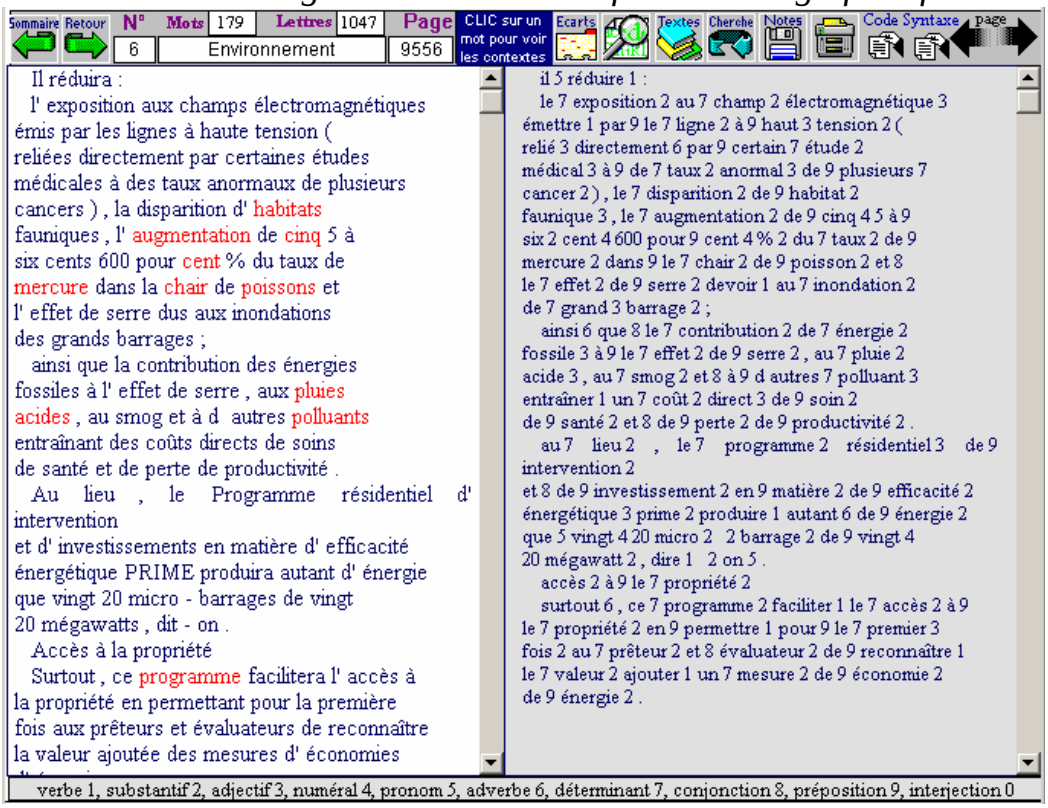


écart	corpus	texte	mot	HIERARCHIQUE
255.38	6419	5512	QUÉBEC	
59.93	28364	4387	DU	
41.97	1854	572	CANADA	
41.26	318	204	CONSTITUTIONNEL	
40.44	514	262	HYDRO	
39.03	606	279	AVENIR	
36.83	1025	361	POLITIQUE	
34.53	120	102	SOUVERAIN	
30.01	336	159	SOUVERAINETÉ	
29.33	14864	1873	AU	
29.12	171	106	FÉDÉRATION	
29.04	1538	385	GOUVERNEMENT	
27.44	245	123	SCIENCE	
25.31	1005	266	SOCIÉTÉ	
25.08	146	85	CONSTITUTIONNELLE	
24.92	677	206	RÉGIONS	
24.33	277	119	POUVOIRS	
23.68	1463	322	DÉVELOPPEMENT	
23.67	184	92	INDÉPENDANCE	
23.54	278	116	GÉOGRAPHIQUE	
23.18	1405	309	QUÉBÉCOIS	
23.15	383	138	ASSOCIATION	
22.88	51	44	OMNIPRATICIEN	
22.51	254	106	CONSTITUTION	
22.27	459	149	CANADIENNE	
22.10	576	170	FÉDÉRAL	
21.67	497	153	POLITIQUES	
21.05	22	26	SFPQ	
20.86	283	106	OTTAWA	
20.39	355	119	PROVINCES	
20.20	37	33	UQAM	
20.07	966	219	CENTRE	

écart	corpus	texte	mot	ALPHABETIQUE
7.34	192	38	SCIENCES	
3.57	202	26	SCIENTIFIQUE	
6.19	100	22	SCOLAIRE	
3.93	433	49	SÉCURITÉ	
9.15	204	46	SEIN	
3.10	167	21	SEPTEMBRE	
4.00	980	96	SERA	
3.35	893	84	SERAIT	
4.62	737	80	SERVICES	
11.38	3323	383	SES	
5.87	690	84	SEUL	
3.71	751	75	SEULEMENT	
21.05	22	26	SFPQ	
3.32	222	27	SIÈCLE	
6.40	654	84	SITUATION	
6.10	317	48	SOCIAL	
4.92	195	30	SOCIALE	
5.94	174	31	SOCIALES	
5.88	253	40	SOCIAUX	
25.31	1005	266	SOCIÉTÉ	
5.15	2229	208	SOIXANTE	
3.78	408	46	SOLEIL	
4.25	415	49	SOMMES	
6.21	5166	453	SON	
34.53	120	102	SOUVERAIN	
30.01	336	159	SOUVERAINETÉ	
6.65	99	23	SPÉCIALISTES	
17.97	207	78	STATUT	
3.25	193	24	STRATÉGIE	
6.93	494	71	SUD	
10.94	64	26	SUPRÊME	
3.76	11931	892	SUR	

Même la lecture est assistée par la statistique. La page courante où on lit le texte met en relief (grâce à la couleur) les mots qui sont caractéristiques de ce texte, comme les termes *habitat*, *pluies acides*, *mercure*, *polluants* dans l'exemple de la figure 5, emprunté à un document sur l'environnement. Dans d'autres pages du même texte la couleur désigne à l'attention les mots qui marquent ce souci de protéger notre planète : *hydrocarbures*, *pollution*, *concentration*, *assainissement*, *toxique*, *contamination*, *phosphore*, *eau*, *lac*, *rivière*, etc. Les mots ainsi soulignés le sont parce qu'ils sont reconnus comme appartenant au vocabulaire spécifique du texte considéré, qu'on peut atteindre directement en faisant appel à la fonction SPÉCIFICITÉS. Le calcul (fondé sur la loi hypergéométrique) peut s'exercer sur les graphies et sur les lemmes, mais aussi sur les codes grammaticaux et les structures syntaxiques.

Figure 5. Mise en relief des mots significatifs



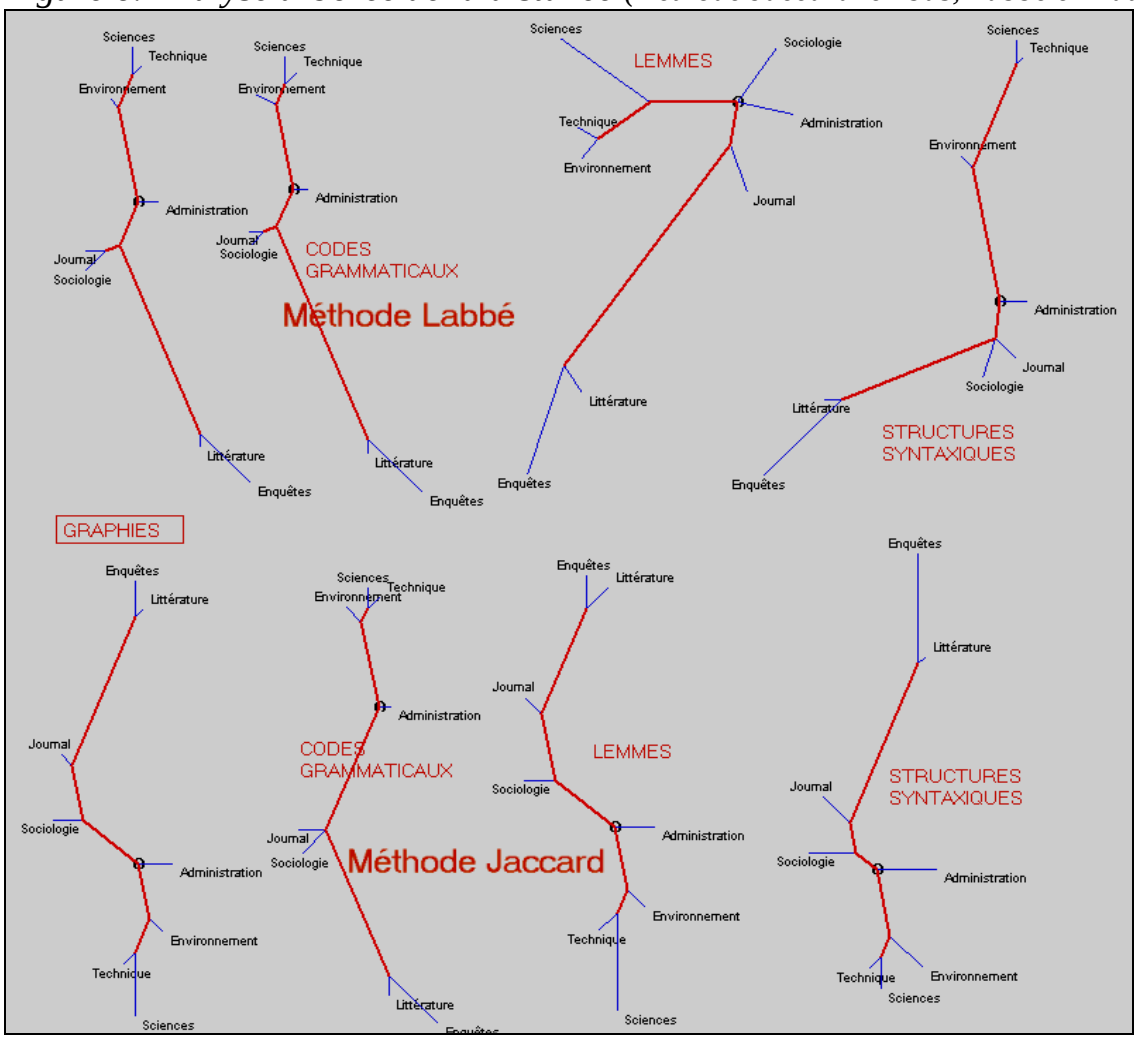
On remarquera en effet que le corpus a été lemmatisé et catégorisé[11], le texte de la figure 5 apparaissant dans deux séquences alignées, à gauche les formes, à droite les lemmes. En réalité, deux autres champs, pareillement alignés, contiennent les codes grammaticaux et les structures syntaxiques. On peut ainsi choisir une forme et repérer immédiatement tous les passages où elle apparaît, mais le choix peut se faire aussi sur les trois autres objets. Ainsi en cliquant sur le code *Vmif3sv* (V=verbe, m=principal, i=indicatif, f=futur, 3 = troisième personne, s=singulier), qui correspond au premier verbe (*réduira*) du précédent extrait, on peut faire défiler les 4054 contextes où un futur est pareillement employé à la troisième personne du singulier. Mais leur nombre est tel qu'on préfèrera les compter, observer leur distribution dans les textes du corpus (ils se concentrent dans la prose journalistique ou sociopolitique) ou faire la comparaison avec d'autres catégories. Dans la phase de préparation et d'indexation, le logiciel s'emploie à de tels décomptes, en relevant dans le corpus 2875412 occurrences (mots ou ponctuations), 67690 formes différentes et 43326 lemmes. Naturellement ces relevés sont faits aussi pour chaque texte.

En prenant appui sur les lemmes (le calcul peut se faire aussi sur les graphies, sur les codes et sur les structures), il est possible de calculer la distance

[11] Cette lemmatisation, due au logiciel *Cordial*, est indépendante de celle dont certains textes du corpus ont pu bénéficier à l'Université de Sherbrooke, comme indiqué en annexe. Satisfaisante dans son ensemble, la catégorisation de *Cordial* présente néanmoins quelques faiblesses, par exemple ici *polluants* est indiqué comme adjectif.

qui sépare un texte de tous les autres respectivement. Pour chaque couple de textes, on prend en compte tous les mots rencontrés et leur répartition, partagée ou exclusive, dans les deux textes. En réalité le calcul peut se faire en tenant compte ou non de la fréquence. Si l'on travaille sur V, on utilisera la formule de Jaccard, améliorée par nos soins, et l'on se bornera à enregistrer la présence ou l'absence des mots (graphies ou lemmes). En prenant en considération la fréquence et donc en travaillant sur N, on pourra utiliser la formule de Labbé[12]. La figure 6[13] dessine la carte typologique qui résume l'ensemble de ces mesures de proximité ou d'éloignement. Ainsi pourrait-on représenter la carte géographique d'un pays quand on connaît le tableau des distances de ville à ville.

Figure 6. Analyse arborée de la distance (méthode Jaccard en bas, Labbé en haut)

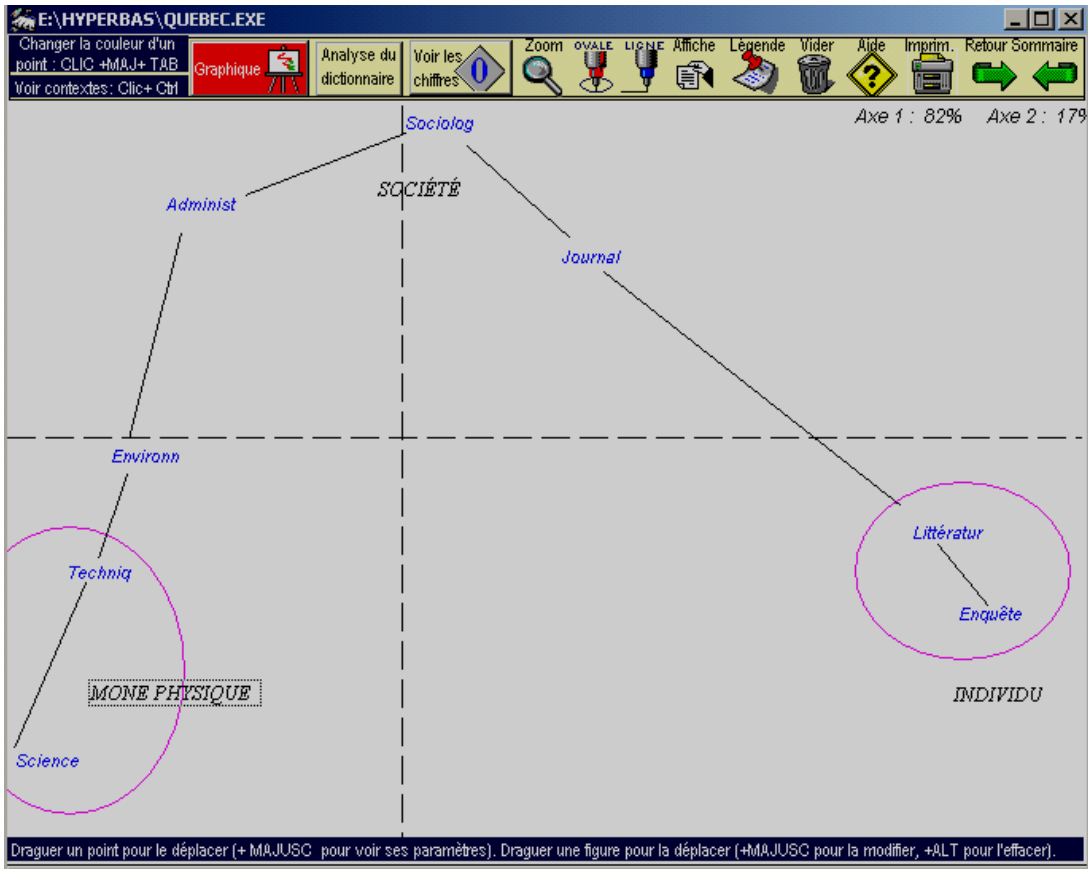


[12] Le coefficient de proximité qu'on obtient et qui évolue entre 0 et 1 n'a pas la vertu différentielle que son auteur lui alloue. C'est une mesure globale qui ne distingue pas, dans la distance observée, ce qui tient à l'écrivain, au genre, à l'époque et au sujet et qui ne permet en aucune façon de dire si Molière et Corneille sont un seul et même auteur.

[13] La synthèse et la représentation du tableau des distances sont dues à Luong et Barthélémy, inventeurs de l'analyse arborée, que nous avons incorporée au logiciel HYPERBASE.

Observons toutefois que seule compte la longueur des segments qu'il faut suivre pour aller d'un point à un autre ; les angles, les directions et l'orientation sont arbitraires et indifférents. Cela ne gêne en rien la lisibilité et la stabilité du résultat, les huit graphes obtenus étant superposables, quelle que soit la méthode utilisée et quel que soit l'objet analysé. Aux deux bouts opposés du graphe, on retrouve toujours les deux mêmes paires : enquêtes orales et littérature d'une part, textes scientifiques et techniques d'autre part. À ce dernier couple se rattachent étroitement les textes sur l'environnement. Deux types de texte sont toujours associés, dans une position médiane : les textes sociopolitiques et les journaux. Restent les textes administratifs : ils accompagnent, assez mollement, ce couple, tout en se rapprochant des textes techniques. Tout se passe comme si la typologie des discours était polarisée par l'opposition monde humain vs monde physique. De l'individu (oral et littérature), on passe à la société (presse et sociologie), puis à la gestion de l'État et de la planète (l'homme y tient encore un peu de place), et enfin aux objets physiques de la technique et de la science. En s'appuyant sur le même tableau de distances, l'analyse factorielle (de correspondance) propose une typologie tout aussi claire (figure 7).

Figure 7. Analyse factorielle de la distance intertextuelle

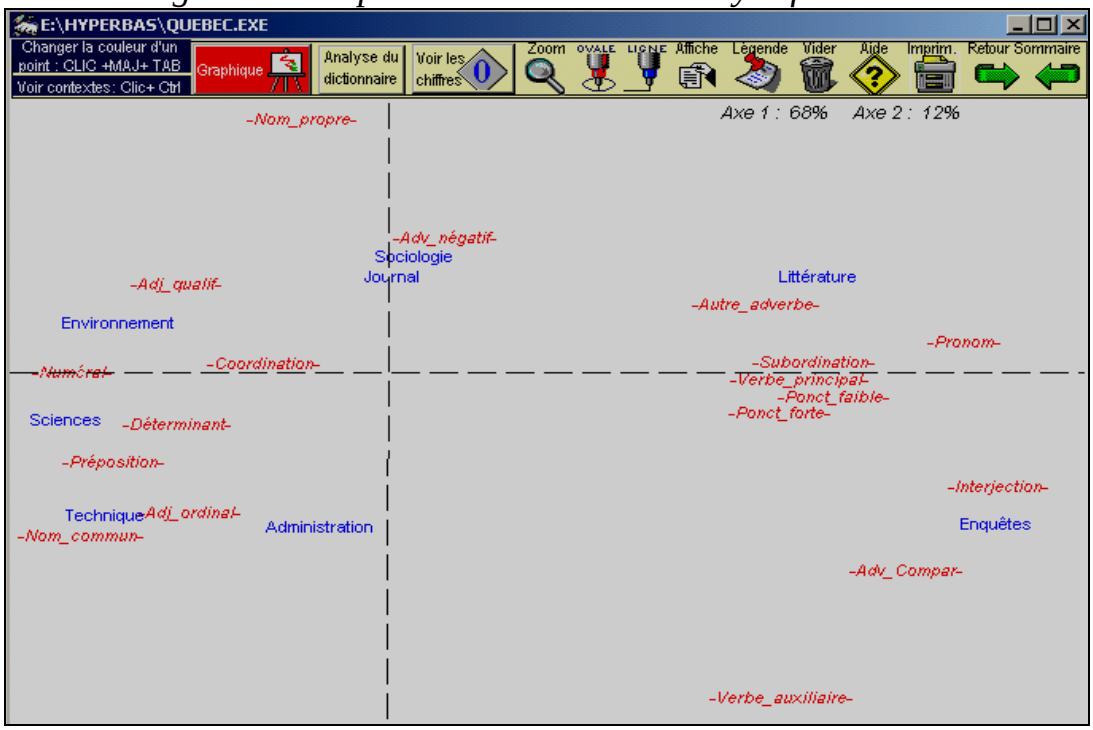


Comme les autres graphes se recoupent, seul est reproduit celui qui concerne les graphies selon le calcul Jaccard. Cette fois l'espace est orienté et les points cardinaux ont une signification : le premier facteur, qui concentre la plus grande

part de la variance (82%) et qui oppose la droite et la gauche, met en relief la tension humain/non humain. Les points s'ordonnent en formant un croissant, comme il arrive souvent dans les données sérielles, et en suivant l'ordre proposé par l'analyse arborée.

D'aucuns penseront que le choix des sujets gouverne partiellement les alliances et les oppositions. Il est peu probable en effet que les conversations et les contes traitent des mêmes questions que les traités scientifiques. Tout ne se réduit pas pourtant au thème, puisque l'analyse des codes et des structures syntaxiques – d'où le sujet est radicalement absent – reproduit la même typologie. Il reste cependant une incertitude : les graphes qui précèdent montrent les positions, mais sans les expliquer. Or il est un moyen, au moins dans l'analyse des codes grammaticaux, de percer le secret. Constituons un tableau à deux dimensions, dont chaque case indiquera l'effectif d'une partie du discours dans un texte particulier. On élimine ainsi l'influence du thème en évacuant le sens des mots et en ne retenant du texte que l'aspect grammatical. Cette fois l'analyse factorielle peut représenter en même temps les lignes (les parties du discours) et les colonnes (les textes), ce qui permet de mettre en rapport les unes et les autres et de transformer l'observation en explication.

Figure 8. Les parties du discours. Analyse factorielle.



Considérons en effet les éléments qui figurent en rouge sur le graphe et qui représentent les lignes du tableau. Comment ne pas voir qu'ils s'ordonnent pareillement en deux camps, dont la rivalité a été maintes fois constatée dans d'autres corpus. Le verbe campe solidement à droite, en compagnie de ses acolytes habituels : pronoms, adverbes et subordonnants. Le substantif règne à

gauche, qu'il s'agisse du nom propre ou du nom commun. Les déterminants l'accompagnent, et aussi les adjectifs, les numéraux et les prépositions[14]. Or dans cet univers grammatical bipolaire, les types de discours s'ordonnent de la même façon que dans les analyses précédentes : l'oral et le littéraire à droite, dans la zone d'influence du verbe, le technique et le scientifique à gauche, parmi les catégories nominales. On retrouve ici comme ailleurs les lignes de force qui structurent le discours et qui sans doute ne sont pas propres au français, qu'il soit québécois ou hexagonal. Nous les avons observées dans le grand corpus de *Frantext*, qui recouvre cinq siècles de notre histoire, et aussi dans un corpus plus récent que nous avons constitué, sous le nom de *Francil*, avec des données extraites des pays francophones – et qui contenait certains textes du présent corpus. Nous renvoyons là-dessus à notre étude parue dans la *Nouvelle histoire de la langue française*[15]. Nous observions que la dichotomie traditionnelle qui a longtemps opposé l'oral et l'écrit a perdu de sa vigueur, et que s'y substitue une opposition grandissante entre le littéraire et l'utilitaire. Le français utilitaire, celui de l'information et de l'exposé scientifique, utilise systématiquement le substantif et ne recourt plus guère au verbe, simple copule de transfert ou d'égalité, analogue au signe = dans les développements mathématiques. Par opposition à ce troisième larron[16], venu troubler la partie, l'oral et le littéraire, qui tous les deux exploitent les ressources expressives du verbe, ne se trouvent plus face à face, mais côte à côte. C'est exactement leur position dans le graphique 8.

D'autres faits linguistiques se lisent dans la même figure, qui influencent surtout le second facteur et dont la cause est aisée à découvrir. Passons sur les interjections, qui ne sauraient trouver refuge ailleurs qu'au voisinage de l'oral et de la littérature. Un scientifique qui se respecte ne saurait lâcher un « tabernacle[17]» au cours de son exposé. Passons sur les ponctuations : qu'elles soient des pauses provisoires, comme la virgule, ou des haltes dans le discours, comme les points, elles produisent une segmentation beaucoup plus serrée dans le discours littéraire et plus encore à l'oral. La phrase de l'exposé technique est plus longue, alors même que les verbes y sont plus rares. Reste à expliquer l'antinomie marquée sur l'axe vertical entre les noms propres et les noms communs. Le discours scientifique ne connaît guère que ces derniers. Les rares

[14] La syntaxe explique la liaison forte qui s'établit entre le substantif et les déterminants. Elle ne justifie qu'en partie la relation substantif-préposition, car la préposition peut introduire aussi bien un pronom et un infinitif. De plus beaucoup de prépositions entrent dans la composition des subordonnants et annoncent une proposition subséquente, et donc un verbe.

[15] Sous la direction de Jacques Chaurand, Le Seuil, 1999, « Ce que disent les chiffres », 673-727.

[16] En réalité ce troisième larron était déjà dans la place, aux temps anciens. Il régnait dans le sabir des gens de loi, d'église ou de médecine. Mais il est vrai qu'il parlait latin et que l'honnête homme s'en tenait éloigné.

[17] Il y a 40 occurrences de « tabernacle » dans le corpus, toutes observées à l'oral, sauf six exemples relevés dans la presse et une citation – allusive – dans un texte sociologique.

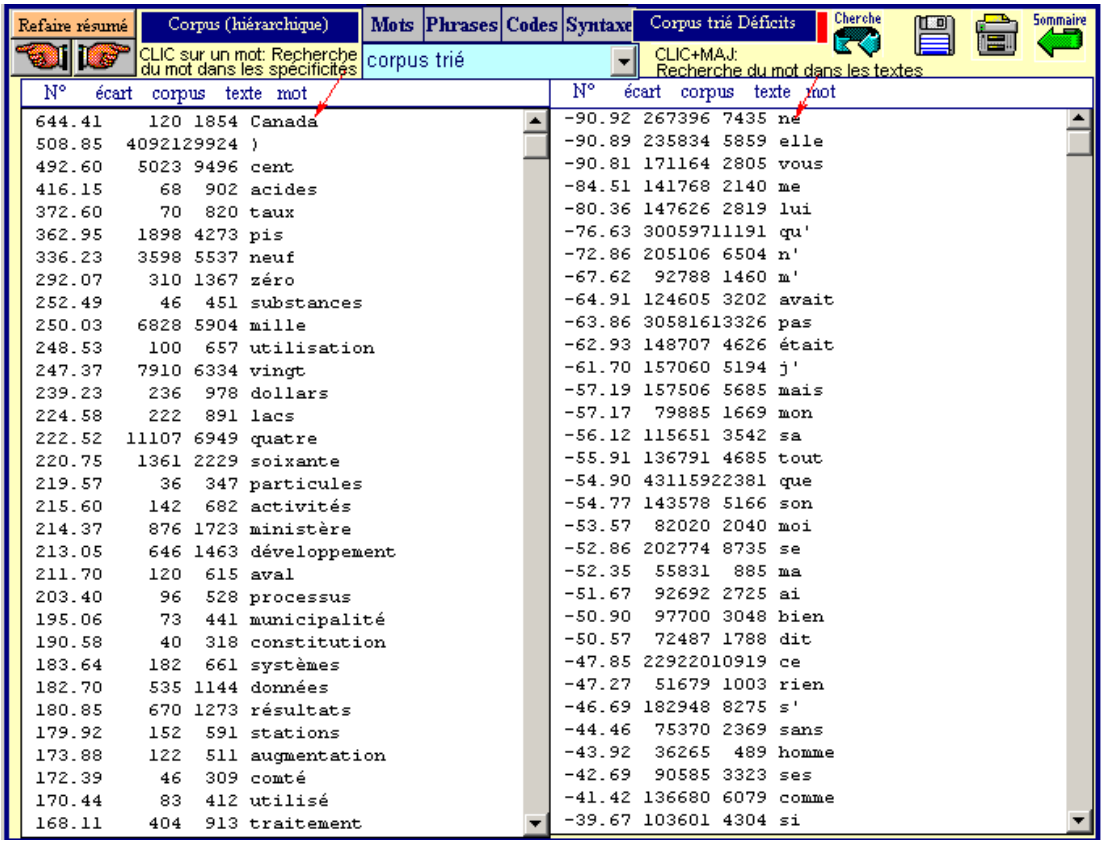
noms propres qu'il accepte sont ceux des savants qui ont donné leur nom à une loi, à un théorème, à une mesure[18] ou des marques qui revendiquent un brevet ou un produit ou ceux des collègues dont on approuve ou dont on conteste les dires. Dans les sciences sociales ou politiques au contraire, le discours est localisé et personnalisé, il est inscrit dans un cadre historique dont les références sont des lieux, des dates (les dates sont des sortes de noms propres dans le temps) et des personnages publics.

On vient de constater que les tendances aperçues dans la base *Frantext* trouvaient un écho dans la présente base. N'en concluons pas trop vite que les deux corpus se recouvrent. Nous avons vu que leur composition diffère. Dans le temps *Frantext* déborde la *BDTS* puisque cinq siècles y sont entassés, alors que le corpus *BDTS* est une coupe synchronique de faible épaisseur. Dans l'espace des champs disciplinaires, c'est l'inverse : *Frantext* paraît étroitement littéraire auprès de la *BDTS*, qui englobe une gamme étendue de discours diversifiés. Si on rapproche les deux corpus l'un de l'autre grâce à un calcul de spécificités (c'est le plus gros corpus, donc *Frantext*, qui sert de norme au plus petit), l'irrédentisme du Québec apparaît de façon irréductible (figure 9). Non que les coutumes locales soient mises en évidence : il n'y a guère que le premier mot de la liste (*Canada*) qui soit la marque d'un territoire et quelques indices (*pis, comté, dollars*), au reste peu significatifs[19]. Vu de *Frantext*, le corpus québécois semble un univers de comptables et d'ingénieurs, où les chiffres sont la préoccupation majeure (*cent, taux, neuf, zéro, mille, vingt, quatre, soixante*) et où l'activité, utilitaire et matérielle, évolue entre la recherche scientifique (*systèmes, données, particules, résultats*), la production industrielle (*acides, substances, traitement, processus, utilisation, développement*) et la gestion publique (*ministère, municipalité, constitution*). La liste est, bien sûr, beaucoup plus longue que le court extrait que nous présentons, mais les substantifs continuent à y exercer une écrasante majorité et leur tonalité ne varie pas. On risque donc de ne pas obtenir exactement ce que l'on recherchait, à savoir l'image contrastée des discours tenus en France et au Québec, mais le résultat trop prévisible de choix différents dans la constitution des corpus. On a souvent reproché à *Frantext* de ne pas être la référence extérieure, universelle et neutre dont on a besoin. C'est trop demander : un corpus, si vaste soit-il, même diversifié, même étalonné, ne sera jamais la référence absolue, comme le niveau de la mer. Cela est vrai pour *Frantext*, et, pour les mêmes raisons, de la *BDTS*.

[18] Mais dans ce cas la minuscule est vite adoptée, ce qu'on observe avec joule, ampère, watt, ohm, berquerel, bel et décibel, etc.

[19] Curieusement c'est la liste des spécificités négatives (colonne de droite du tableau 8), qui fournit l'indice le plus clair. Quand on connaît le sous-emploi au Québec de la négation *ne*, trouver ce mot en tête des déficits est un signe non trompeur.

Figure 9. Les spécificités de la BDTS *par rapport à* Frantext



En revanche les comparaisons internes se justifient aisément, puisqu'un corpus est expressément conçu pour mettre en valeur les différences qui opposent les textes que l'on réunit dans le même ensemble. Si le corpus est hétérogène, l'intérêt est faible et les spécificités auront un aspect trivial. S'il est homogène, le calcul relèvera toujours des écarts et des nuances que la conscience linguistique peut n'avoir pas sentis au premier abord. Afin de rester dans une relative homogénéité, nous ne présentons dans le tableau 10 que quatre sous-ensembles « humains », en rejetant à l'extérieur les textes qui s'occupent d'autre chose : de science, de technique et d'environnement. Ce qui frappe d'abord c'est le parallélisme des deux premières colonnes, vouées aux enquêtes et à la littérature. Certes les expressions trop populaires qu'on relève dans certaines interviews (*pis, ben, pron, faque*) ne sont pas de mise chez les écrivains. Mais le discours fait appel aux mêmes embrayeurs, particulièrement aux pronoms des deux premières personnes, aux démonstratifs (*ça* ou *ce*), aux verbes simples (*être, faire, dire, penser*), à la conjonction *que*, aux ponctuations fortes (surtout le point d'interrogation). On pourrait songer à attribuer au théâtre ces points de convergence, si le théâtre était dominant dans le sous-ensemble littéraire. Or on n'y trouve qu'une pièce, de faible étendue. Une meilleure explication serait à rechercher dans le style des romans ou contes qui ont du succès au Québec et qui font la part belle au dialogue. Ceux qui ont été choisis

dans le corpus exploitent en effet cette veine populaire. Mais on aime à croire aussi que les mêmes ressources du langage sont communes à l'expression orale et à l'expression littéraire, comme on vient de le voir.

Tableau 10. Le vocabulaire spécifique de quatre sous-ensembles (extrait)

Enquêtes (formes)

N°	Ecart	Corpus	Texte	Mot
1	32.82	7059	5428	ça
1	32.82	5394	4012	là
1	32.82	4273	4075	pis
1	32.82	1375	1118	ben
1	32.82	1341	1341	pron
1	32.82	1010	760	sais
1	32.82	559	413	étais
1	32.82	398	398	faque
1	32.82	213	201	cabane
1	32.81	13326	3655	pas
1	32.81	12624	3994	on
1	32.81	10297	4880	c'
1	32.81	7626	3350	je
1	32.81	6671	2183	y
1	32.81	6307	2063	?
1	32.81	5194	3117	j'
1	32.81	4626	2269	était
1	32.81	4320	1218	fait
1	32.81	4282	1541	ils
1	32.81	3535	2496	tu
1	32.81	3202	1538	avait
1	32.81	2725	1554	ai
1	32.81	2471	1248	quand
1	32.81	2140	899	me
1	32.81	2040	1305	moi
1	32.81	1669	745	mon
1	32.81	1460	619	m'
1	32.81	1357	569	va
1	32.81	970	441	eu
1	32.81	948	510	parce que
1	32.81	944	515	suis
1	32.81	889	485	oui
1	32.81	713	483	vraiment
1	32.81	673	425	pense
1	32.81	638	382	as
1	32.81	619	415	avais
1	32.81	619	408	faisait
1	32.81	584	344	hiver
1	32.81	509	348	ah
1	32.81	404	277	gars
1	32.80	11482	322632	.
1	32.80	28436	5271	est
1	32.80	22381	3566	que
1	32.80	19893	3995	il
1	32.80	16552	2837	a

Littérature(formes)

N°	Ecart	Corpus	Texte	Mot
2	32.82	345	345	Didace
2	32.82	303	303	Phonsine
2	32.82	273	248	ti
2	32.81	7626	2785	je
2	32.81	5859	2604	elle
2	32.81	5724	3668	!
2	32.81	2819	1165	lui
2	32.81	2805	1064	vous
2	32.81	2140	944	me
2	32.81	1669	721	mon
2	32.81	973	499	puis
2	32.81	816	532	t'
2	32.81	447	342	yeux
2	32.80	19893	4547	il
2	32.80	6307	2078	?
2	32.80	6079	1679	comme
2	32.80	3542	1124	sa
2	32.7914547526708	,		
2	32.7911482321653	.		
2	32.55	195	195	Acayenne
2	32.32	1460	625	m'
2	32.21	195	194	Philomène
2	31.31	5166	1418	son
2	31.11	5194	1419	j'
2	30.70	8735	2072	se
2	30.42	1788	681	dit
2	30.41	10297	2342	c'
2	29.08	4685	1274	tout
2	28.01	13326	2785	pas
2	27.81	564	320	tête
2	27.81	448	281	coeur
2	27.55	489	293	homme
2	27.53	213	184	silence
2	27.26	1003	442	rien
2	26.91	400	257	toi
2	26.83	11191	2385	qu'
2	26.80	885	405	ma
2	26.26	313	220	rue
2	25.52	3323	933	ses
2	25.33	8275	1839	s'
2	25.32	140	135	demanda

JOURNAL (formes)

N°	Ecart	Corpus	Texte	Mot
4	32.81	966	571	centre
4	31.34	336	279	illustration
4	30.50	978	517	dollars
4	30.23	10163	2676	:
4	29.65	258	229	doc
4	24.24	48818	9650	-
4	23.39	444	266	parti
4	22.76	972	426	Montréal
4	22.65	191	158	94
4	22.45	145	133	gène
4	22.33	245	181	science
4	22.20	9496	2283	cent
4	21.10	5904	1525	mille
4	20.93	813	360	recherche
4	20.84	782	350	etc
4	20.76	656	312	intérêt
4	20.25	1665	575	ans
4	19.91	203	149	photo
4	19.88	133	116	actualité
4	19.77	470	245	1994
4	19.57	278	177	géographique
4	17.94	85	82	Festival
4	17.80	408	209	Soleil
4	17.48	579	258	longueur
4	17.29	307	172	presse
4	16.79	245	147	Devoir
4	16.59	229	140	milliards
4	16.53	299	164	explique
4	16.37	469	216	université
4	16.29	178	119	Claude
4	16.04	192	123	sciences
4	15.97	6334	1476	vingt
4	15.31	115	88	libéral
4	15.16	70	65	photographe
4	14.92	66	62	éditorial
4	14.90	99	79	gènes
4	14.63	203	119	Robert
4	14.54	6504	1470	n'
4	14.42	173	107	août
4	14.29	109	81	encadré
4	14.26	295	147	équipe
4	14.07	58	55	embryon

ADMINISTRATION (formes)

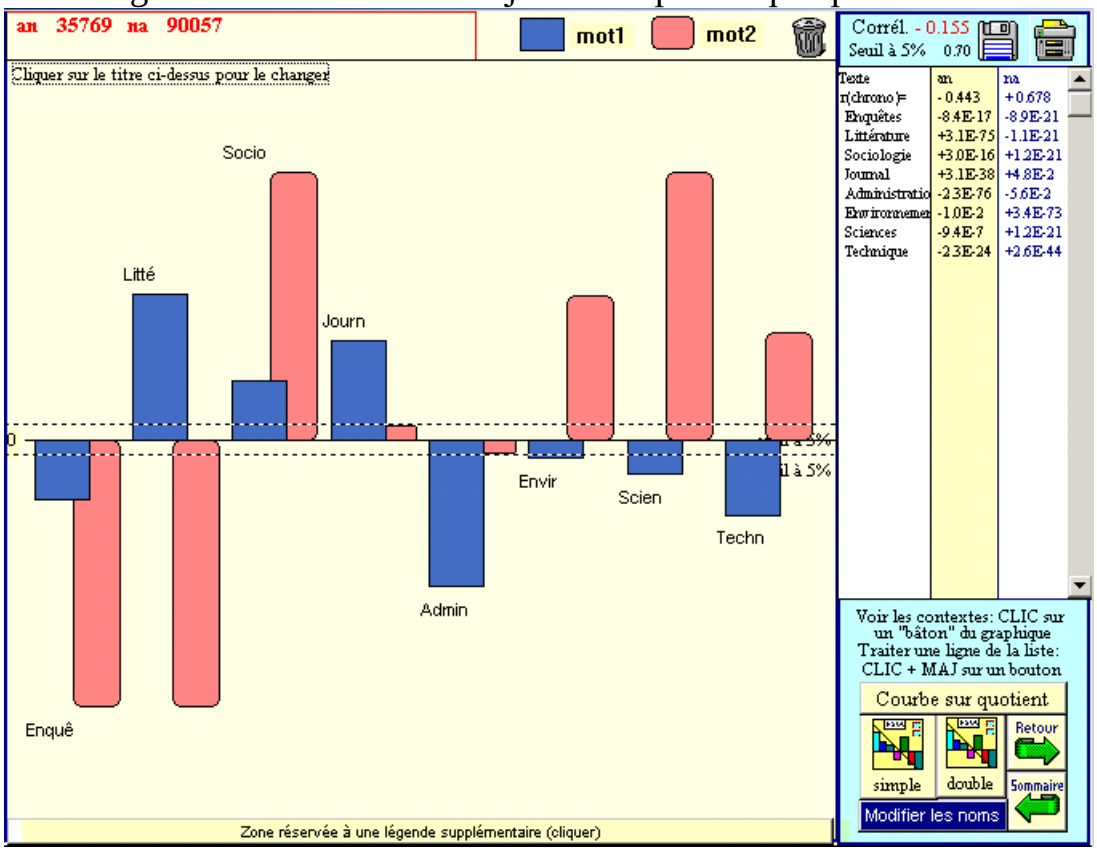
N°	Ecart	Corpus	Texte	Mot
5	32.82	963	774	article
5	32.82	955	766	président
5	32.82	948	889	employé
5	32.82	393	347	règlement
5	32.82	317	296	employeur
5	32.82	312	284	syndicat
5	32.82	301	272	employée
5	32.82	274	255	leader
5	32.82	238	234	indemnité
5	32.82	225	224	motion
5	32.82	201	201	alinéa
5	32.81	1723	1108	ministère
5	32.81	1426	619	travail
5	32.81	1210	882	loi
5	32.81	1179	851	ministre
5	32.81	1176	648	projet
5	32.81	947	499	commission
5	32.81	773	534	emploi
5	32.81	572	436	avis
5	32.81	392	301	assemblée
5	32.80	9235	2868	ou
5	32.71	224	211	travailleur
5	31.83	255	223	dispositions
5	31.64	291	239	vertu
5	31.60	16552	3252	a
5	31.21	339	257	date
5	30.96	215	198	congé
5	29.50	423	277	revenu
5	29.32	441	282	municipalité
5	29.27	881	415	droit
5	28.74	383	257	employés
5	28.26	50642	8062	l'
5	27.46	1538	555	gouvernement
5	27.44	283	210	régie
5	27.41	214	180	délai
5	27.16	913	401	traitement
5	26.50	285	205	prévu
5	26.38	369	235	décision
5	26.03	133	131	n°
5	25.93	169	151	convention
5	25.73	871	375	jours
5	24.54	51597	7965	le
5	24.51	323	206	comité
5	24.30	194	155	prévue
5	23.32	156	133	copie

Nous laisserons au lecteur le soin de voir ce qui peut rapprocher ou opposer les deux autres colonnes, dévolues aux textes journalistiques et administratifs. Les textes sociopolitiques, si l'espace ne nous était pas mesuré, auraient trouvé place entre les deux. Ces trois sous-ensembles appartiennent à la langue générale, plus diversifiée et presque désordonnée dans la presse, plus spécialisée et concentrée dans l'administration.

On laissera pareillement au lecteur le soin d'interroger la base et d'y exploiter beaucoup de fonctions qu'on ne peut détailler ici. L'accès aux codes grammaticaux et aux structures syntaxiques permet des investigations jusque là

impraticables. Un exemple relativement simple de ces structures est montré dans la figure 11 : il s'agit des séquences adjectif+substantif et substantif+adjectif, autrement dit, de l'anté- ou post-position de l'adjectif. Certains adjectifs admettent les deux constructions, parfois en changeant de sens (*un grand homme* vs *un homme grand*), mais la plupart ont une préférence pour l'une ou pour l'autre. Le graphique indique que l'antéposition est le fait de ceux qui surveillent leur plume, écrivains ou journalistes, alors que la postposition est en faveur lorsqu'on veut seulement transmettre une information. Toutes les combinaisons qu'elles soient à deux, trois ou n éléments sont recensées et indexées par le logiciel, ce qui donne accès à des recherches documentaires ou statistiques fort complexes.

Figure 11. Courbes de l'adjectif antéposé et postposé

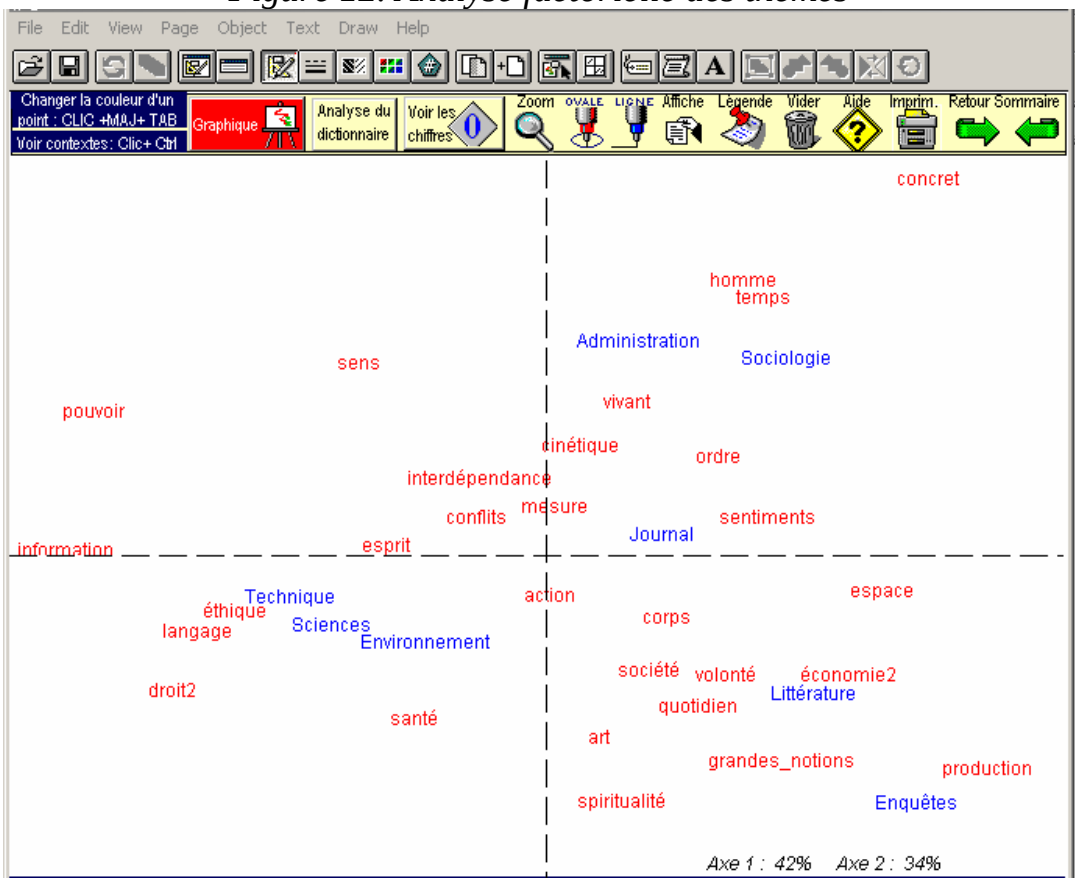


La phraséologie, le rythme du discours, voire les sonorités, tout est sujet à mesure, et parfois à découverte, y compris même la thématique. Il ne s'agit pas seulement de circonscrire une constellation de corrélats autour d'un mot, comme indiqué dans la figure 4. Ni de caractériser un texte par un ensemble d'extraits spécifiques[20]. L'ambition, permise par un traitement sémantique de *Cordial*, vise à rendre compte des idées, des sentiments, des actions, bref des thèmes exprimés dans un texte. En réalité *Cordial* fait appel à un thésaurus de référence, où sont cataloguées les disciplines, les concepts et les connaissances. Tout un jeu

[20] Le calcul des spécificités est appliqué non seulement aux formes, aux lemmes, aux codes grammaticaux et aux structures syntaxiques mais aussi aux phrases caractéristiques.

d'étiquettes hiérarchisées est mis en place, parmi lesquelles chaque mot du texte doit faire son choix. Sans doute ces étiquettes sont-elle parfois trop proches des représentations modernes, et s'appliquent-elles malaisément à certains corpus, sans compter les bévues auxquelles l'homographie et même la polysémie peuvent donner lieu. Derrière « cinétique » on peut comprendre « mouvement »; mais que recouvrent les termes d' « interdépendance », de « production » et de « grandes-notions »?

Figure 12. Analyse factorielle des thèmes



Pourtant, malgré les faiblesses et les incertitudes du codage sémantique, les résultats auxquels il conduit ne sont pas dénués d'intérêt. On les a reproduits dans la figure 12. On y découvre que la même aimantation des textes, déjà observée au niveau lexical et syntaxique, se retrouve au niveau thématique.

En conclusion, on pourrait s'étonner que l'on puisse obtenir la même image, alors que les objets considérés sont étrangers les uns aux autres. Certes entre les lemmes et les graphies, il y a une part commune. Les premiers sont un regroupement des secondes. De même un pont relie les codes aux structures, ces dernières étant des combinaisons de codes. Mais quel lien nécessaire existe-t-il entre les lemmes et les codes grammaticaux? Ou entre les graphies et les structures ? Sans parler des codes sémantiques, dont la délimitation est incertaine. On pourrait imaginer que les sous-ensembles du corpus s'orientent différemment, selon le point de vue mis en œuvre, de même que la carte d'un

pays est susceptible de configurations variables, selon qu'on envisage les élections, les revenus, les convictions religieuses, l'espérance de vie ou la fécondité. Et pourtant les sociologues savent que les variables que l'on croit indépendantes sont parfois liées par des accords secrets ou par une commune soumission à une influence cachée. Ainsi en est-il de la surdétermination du langage.

Annexe : Composition du corpus BDTS

Source : Centre d'analyse et de traitement informatique du français québécois (CATIFQ).

Le présent corpus comprend quelques deux millions d'occurrences (61 843 formes) tirées de 1054 textes différents. Il constitue un sous-ensemble de la BDTS (qui contient plus de 37 millions de mots à l'heure actuelle). Ce corpus est composé de huit sous-ensembles d'environ 250 000 mots chacun et traités selon une norme commune, ce qui rend leurs données comparables; plusieurs d'entre eux sont en outre lemmatisés. Ils sont représentatifs de divers domaines, types de discours et niveaux de langue suivants :

1. Textes techniques : langue spécialisée

Corpus constitué par Normand Maillet. Corpus lemmatisé. 8384 vocables. 12 513 unités complexes ou syntagmes.
Composition : 100 textes extraits de rapports, guides, manuels de formation, normes, procédures, etc.
Domaines : aluminium, environnement, mines, pâtes et papier, télécommunication, transport, hydro-électricité, informatique, et autres.

2. Textes scientifiques : langue spécialisée

Corpus constitué par Linda Pépin. Corpus lemmatisé. 8653 vocables. 2257 unités complexes ou syntagmes
Composition : 100 textes extraits de mémoires, de thèses, d'articles scientifiques et de rapports de recherche.
Domaines : biologie, chimie, physique, génie chimique et génie mécanique.

3. Textes sociopolitiques : langue générale

Corpus constitué par Nadine Vincent. Corpus lemmatisé. 8355 vocables
Composition : 100 textes de mémoires sélectionnés à partir des 583 mémoires présentés à la Commission.
Domaines : mémoires ou extraits de mémoires de la Commission Bélanger-Campeau.

4. Textes administratifs : langue générale. Corpus non lemmatisé

Composition : 54 textes
Domaines : débats de l'Assemblée nationale, conventions collectives, textes juridiques, textes du Bureau d'audiences publiques sur l'environnement (BAPE), rapports du vérificateur général du Québec, etc.

5. Textes journalistiques : langue générale. Corpus non lemmatisé

Composition : 52 textes tirés de *L'Actualité*. 55 textes tirés du *Devoir*. 17 textes tirés d'*Interface*. 63 textes tirés de *La Presse*. 57 textes tirés de *Québec Science*. 75 textes tirés du *Soleil*

6. Textes littéraires : langue générale

Corpus non lemmatisé
Composition : 25 textes
Domaines : romans, chansons, essais, textes de poésie, pièces de théâtre, etc.

7. Textes environnementaux : langue générale

Corpus constitué par Steeve Tremblay. Corpus semi-lemmatisé
Composition : 107 textes
Domaines : textes divers du ministère de l'Environnement du Québec, journaux, périodiques et magazines spécialisés dans le domaine, etc.

8. Textes oraux : langue générale

Corpus constitué par Gérald Charland. Corpus lemmatisé
Composition : 20 enquêtes
Domaines : enquêtes sociolinguistiques effectuées dans la région des Bois-Francs