Sémantique générique et statistique

Véronique Magri-Mourgues

▶ **To cite this version:**

Véronique Magri-Mourgues. Sémantique générique et statistique. JADT 2008 - 9e Journées d'Analyse statistique des Données Textuelles, Mar 2008, Lyon, France. pp.753-764. hal-00596471

HAL Id: hal-00596471
https://hal.science/hal-00596471

Submitted on 27 May 2011

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Sémantique générique et statistique

Véronique Magri-Mourgues[1]

[1] Laboratoire BCL, Université de Nice Sophia-Antipolis, CNRS
MSH de Nice, 98 bd E. Herriot, 06 200 Nice.

Abstract

The purpose of this paper is to complete a preceding work whose aim was to characterise the "travel narrative manner" compared with "novels." This time, by studying a corpus based on two lemmatized under-corpora by way of Cordial, travel narrative texts and novels, the analysis investigates lexical and semantic contrasts between the two generic groups. The efficiency of the methods of "lexical correlations and lexical associations" as they are included in Hyperbase, is assessed. In order to visualize contrasts operating between the two generic groups, we use factorial correspondence analysis or graphics around a word-pole.

Résumé

Cette contribution se donne pour objectif de compléter le travail de caractérisation générique entrepris lors d'une précédente recherche. Cette fois, à partir d'un corpus fondé sur deux sous-corpus lemmatisés par l'étiqueteur Cordial, l'un constitué de récits viatiques, l'autre de textes fictionnels, l'analyse explore les contrastes lexicaux et sémantiques entre les deux genres textuels. Elle utilise pour cela les outils des associations lexicales et des corrélats lexicaux intégrés dans le logiciel Hyperbase. Les présentations sous forme d'analyses factorielles des correspondances ou de graphes à partir d'un mot-pôle illustrent les contrastes ainsi observés.

Mots-clés : sémantique, linguistique, cooccurrence, corrélat, stylométrie, typologie textuelle, corpus textuel, genre.

1. Introduction

Lors de mon intervention aux JADT 2006, mon objectif avait été de caractériser par contraste la poétique du récit de voyage. Mon corpus était constitué de binômes d'œuvres d'un même écrivain, un récit de voyage d'une part, un roman d'autre part. Douze écrivains du XIXe siècle avaient ainsi été sollicités comme échantillon représentatif des genres mis en présence. Le corpus avait été étudié sur la base des répartitions d'étiquettes grammaticales au moyen des logiciels Cordial et Hyperbase. L'analyse contrastive permet de pointer les singularités de l'un ou l'autre sous-corpus par rapport à l'ensemble qui sert de norme de référence.

Les tests statistiques auxquels j'avais soumis mon corpus avaient fait ressortir des divergences et des particularités d'ordre essentiellement morpho-syntaxique. J'avais montré que le récit de voyage privilégie la classe nominale tandis que le texte fictionnel a des affinités avec celle du verbe. Dans le premier ensemble de textes, les phrases longues s'associent aux marqueurs de la coordination pour esquisser un discours plus fluide. Enfin, les tests effectués avaient montré la primauté du genre masculin, du pluriel et de l'article défini dans le récit de voyage, ce qui semblait plaider en faveur d'une double orientation, vers la concrétisation et vers la généralisation.

L'enjeu de la communication présente est de compléter les analyses en conservant le même objectif de caractérisation générique. La perspective était essentiellement d'ordre grammatical et s'appuyait sur l'observation des parties de discours ; aujourd'hui, j'entends éclairer davantage des contrastes d'ordre sémantique, en testant des outils d'analyse encore au stade expérimental intégrés dans le logiciel d'analyse hypertextuelle, Hyperbase, comme les fonctions qui concernent les associations lexicales privilégiées, les corrélats lexicaux et qui permettent d'esquisser des réseaux lexicaux. Il s'agit de mettre à l'épreuve ces outils en vue de la caractérisation générique. L'approche sémantique par le biais des cooccurrences lexicales fait-elle ressortir des contrastes exploitables en termes de genre textuel ?

2. La distance intertextuelle

Lors de ma précédente communication, comme je m'intéressais moins aux variations lexicales qu'aux particularités syntaxiques potentielles, j'avais soumis mon corpus au calcul préalable de la distance intertextuelle, fondé sur les codes grammaticaux et sur les structures syntaxiques. Le codage proposé par Cordial permet en effet de travailler soit sur les graphies, soit sur les lemmes, soit sur les codes grammaticaux, soit encore sur les structures syntaxiques, entendues comme un groupe de codes. Cette étude s'intéresse cette fois-ci à la structure thématique des œuvres ; pour cette raison, c'est la distance intertextuelle calculée sur les lemmes qui sera mise à l'épreuve pour tester la mise en série des textes du corpus ; le dépouillement du texte sous forme de lemmes permet le regroupement des formes qui se rattachent à la même entrée du dictionnaire en neutralisant les variations grammaticales. Pour l'étude de la connexion entre deux textes, deux protocoles d'analyse sont proposés par le logiciel Hyperbase. L'un repose sur la méthode dite Jaccard qui se fonde sur le critère de la présence ou de l'absence d'un mot donné dans les deux textes considérés sans se préoccuper de sa fréquence. Si un mot est commun aux deux textes, il tend à les rapprocher ; la distance entre les deux textes augmente au contraire si le mot ne se rencontre que dans un seul. Tous les mots et tous les appariements des textes deux à deux sont envisagés. La version d'Hyperbase utilisée rapproche encore les textes par l'exclusion du même vocabulaire : les mots du corpus total qui ne se trouvent dans aucun des deux textes confrontés sont également pris en compte. La connexion intertextuelle se fonde par conséquent à la fois sur les mots utilisés par les deux textes et sur les mots également rejetés. L'autre méthode est celle de Labbé[1] qui repose sur un algorithme évaluant la distribution réelle des fréquences dans les textes confrontés.

La représentation graphique proposée est celle de l'analyse développée par Xuan Luong, qui permet non seulement de représenter les distances entre les textes de manière exactement proportionnelle, mais encore de mettre en évidence les nœuds de regroupement des textes au sein d'une structure arborée[2] particulièrement fiable et stable. Elle permet de présenter une structuration classificatoire des textes en ajoutant la contrainte de la proximité entre les classes.

[1] Voir C. Labbé et D. Labbé (2003).

[2] La disposition dans l'espace du graphe, l'orientation ou les directions des branches importent peu pour l'évaluation de la distance intertextuelle ; seule compte la distance physique entre les points du graphique, ici les titres des œuvres, ainsi que les ramifications qui regroupent les points en bouquets ou au contraire les isolent ; les diverses bipartitions possibles de l'arbre manifestent la hiérarchie classificatoire.

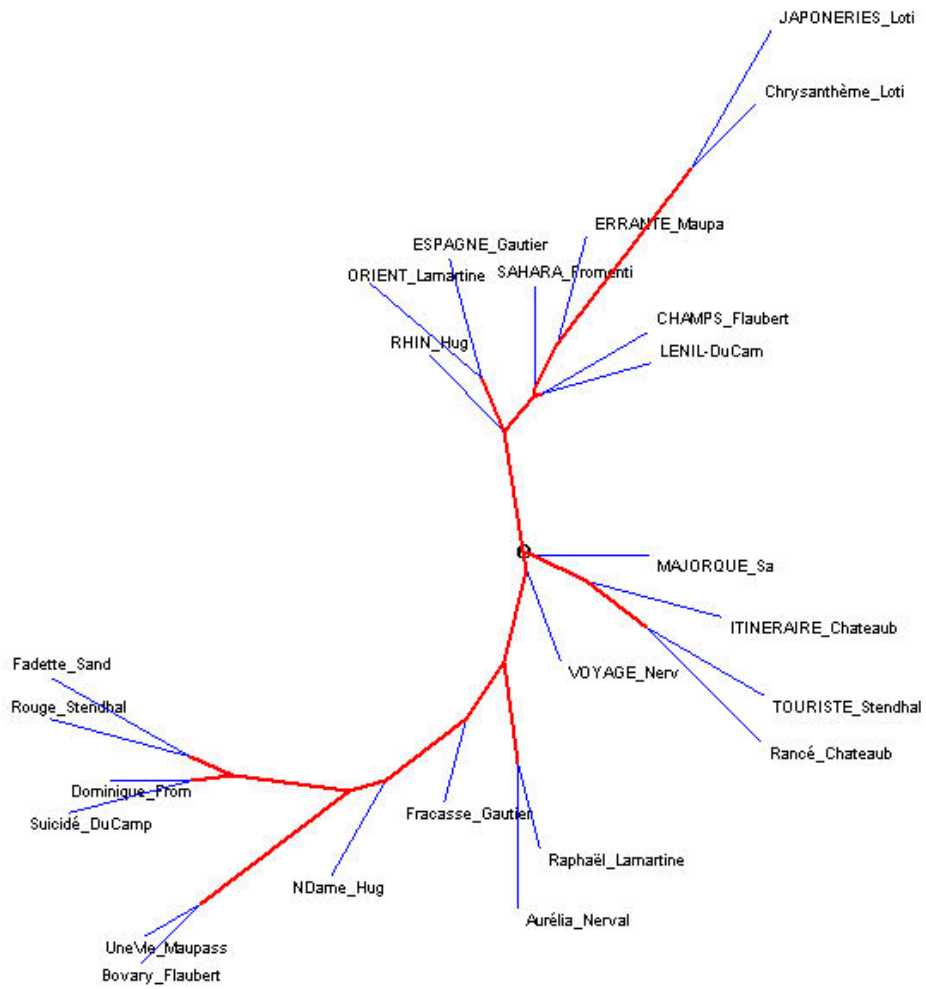

Figure 1 - La distance intertextuelle sur les lemmes. Calcul fondé sur la fréquence (méthode Labbé).

On observe sur ce graphique la formation de trois ensembles : un groupe bien structuré rassemble la plupart des récits de voyage en haut du graphique, un autre un peu plus discordant réunit les récits fictionnels en bas du graphique. Certains « romans » se dissocient de l'harmonie d'ensemble : *Raphaël, Aurélia, Le Capitaine Fracasse* et *Notre-Dame de Paris* s'écartent en effet des deux ramifications principales du groupe fictionnel. Un autre groupe intermédiaire de quatre récits de voyage auquel s'associe *La Vie de Rancé* se détache au centre du graphe.

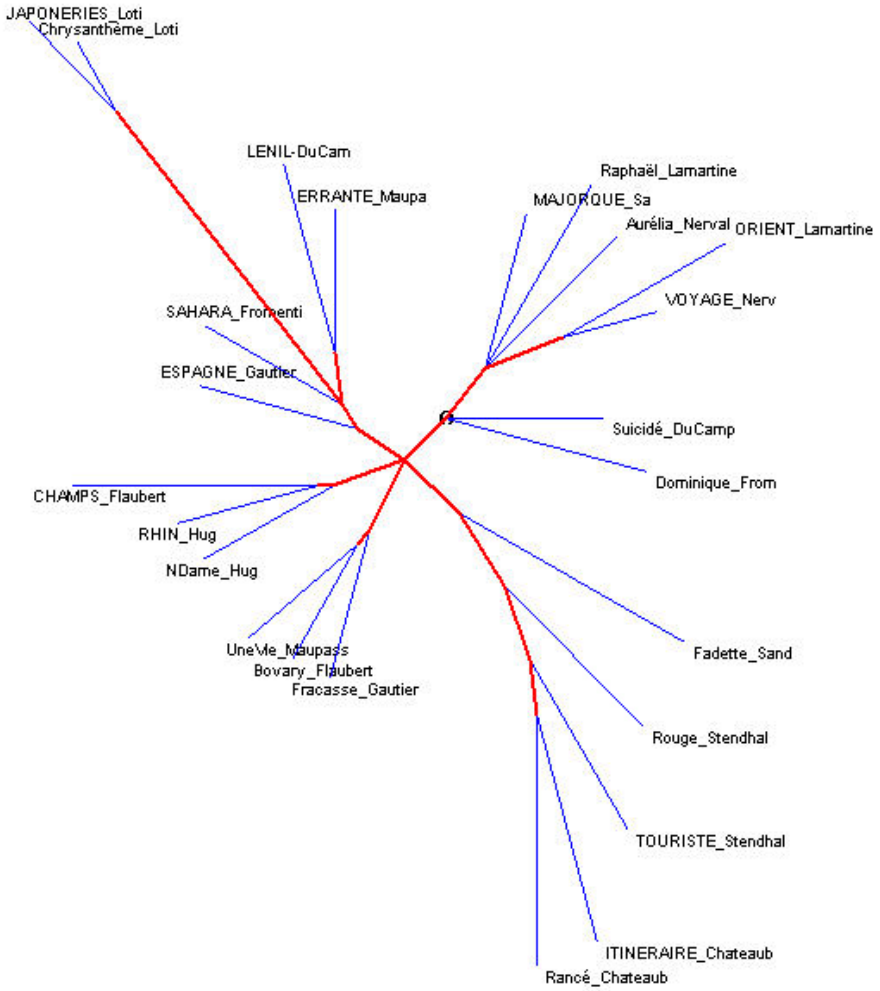

Figure 2 - La distance intertextuelle sur les lemmes. Calcul fondé sur la présence ou l'absence de la variable (méthode Jaccard)

Ce graphique est plus éclaté que le précédent et le partage par genres se dessine moins nettement. On observe encore des bouquets génériques constants d'un graphique à l'autre : un ensemble viatique avec les titres des œuvres situées sur la moitié gauche de l'arbre (*Le Nil*, *La Vie errante, Un Été dans le Sahara, Voyage en Espagne, Par les champs et par les grèves, Le Rhin*) et un ensemble fictionnel plus restreint qui regroupe trois œuvres (*Une Vie, Madame Bovary et Le capitaine Fracasse*). Des solidarités binaires sont encore stables, quel que soit le calcul utilisé, qui établissent une connexion forte entre *La petite Fadette* et *Le Rouge et le Noir*, *Dominique* et *Mémoires d'un suicidé*, *Itinéraire de Paris à Jérusalem* et *La Vie de Rancé*, du même auteur. Les variations d'un graphe à l'autre manifestent le rapprochement des œuvres par auteur et par-delà l'hétérogénéité générique. Cela se vérifie pour trois écrivains : Lamartine (*Voyage en Orient* et *Raphaël*), Nerval (*Voyage en Orient* et *Aurélia*), Stendhal (*Le Rouge et le Noir* et *Mémoires d'un touriste*). Avec cette méthode, ce sont les aspérités du lexique qui sont valorisées autrement dit les mots rares, notamment les hapax, au détriment des mots de fréquence élevée comme les mots grammaticaux forcément présents dans tous les textes. Cette méthode fait cependant ressortir des thématiques communes et, en l'occurrence, des thématiques privilégiées par les écrivains, quel que soit le genre dans lequel ceux-ci s'inscrivent. Elle permet une approche des constantes lexicales d'un auteur, à même d'amorcer une caractérisation de son univers imaginaire, qui paraît devoir établir une connexion entre deux textes plus forte que les déterminations génériques. On peut reprocher à cette méthode de trop mettre l'accent sur l'hétérogénéité du vocabulaire. D'autre part, si les

mêmes lemmes sont employés, cela n'implique en rien qu'ils le soient avec les mêmes acceptions pour chaque occurrence. On doit franchir alors une autre étape qui mène à l'appréciation non plus seulement lexicale mais sémantique des sous-corpus.

3. Des cooccurrences aux corrélats lexicaux

3.1. Méthodologie

Pour tester les nouvelles fonctions intégrées à Hyperbase, toujours avec l'objectif d'une caractérisation contrastive mais cette fois d'ordre sémantique, il a fallu que je reconsidère la constitution de mon corpus d'étude. Alors que j'avais pu travailler sur une base unique constituée des vingt-quatre textes pour ma précédente communication, j'ai dû cette fois segmenter la base en deux partitions et travailler sur chacune d'elles séparément. La confrontation se fera alors par graphiques juxtaposés. J'ai d'autre part exclu de mon corpus d'étude Loti dont les œuvres avaient été constamment rapprochées par les calculs statistiques effectués – d'une part à cause de l'ambiguïté générique qui caractérise celles-ci, d'autre part à cause de leur proximité thématique qui peut être établie *a priori*. Je travaille donc cette fois sur deux sous-bases de onze textes chacune. La particularité enfin de cette étude thématique est de ne pas pouvoir distinguer les sous-corpus constitués par les œuvres de chaque écrivain. Chaque ensemble est traité comme un tout, comme un exemple de tel ou tel genre ; ceci est particulièrement pertinent dans une perspective générique puisque les particularités individuelles éventuelles sont gommées au profit des tendances d'ensemble.

Dans le logiciel Hyperbase, les fonctions nouvelles qui peuvent orienter l'interprétation vers le sémantique reposent sur une observation des séquences et non plus seulement sur les comparaisons des fréquences. Autrement dit, un mot est évalué par rapport à son cotexte, à son environnement proche.

La problématique de l'interprétation est de pouvoir passer du niveau lexical au niveau sémantique. Les cooccurrences doivent être interprétées comme des corrélats sémantiques pour pouvoir « être considérées comme des lexicalisations partielles d'un thème » (F. Rastier, 2001). Avec les fonctions des associations privilégiées et des corrélats lexicaux, on considère les mots dans leur environnement immédiat, sans tenir compte de la partition en textes.

3.2. Le test des associations privilégiées

Le premier test auquel j'ai soumis mon corpus est l'analyse de la distribution des « associations privilégiées » dans le corpus[3]. L'analyse repose sur le choix préalable, effectué par le logiciel, des 400 substantifs les plus fréquents tout en évitant ceux dont la fréquence est trop élevée en raison de leur sens très usuel ; ils sont extraits du dictionnaire, sous la forme des lemmes. Le texte est exploré paragraphe par paragraphe et utilise le programme mis au point par L. Lebart (LX3AFC.EXE) intégré dans Hyperbase. Le seuil choisi pour que la cooccurrence soit retenue est de 3 ; le calcul des associations privilégiées repose sur un tableau général des cooccurrences fondé sur le calcul de la distance entre les mots de la liste, pris deux à deux. Les résultats se présentent au travers d'une analyse factorielle des correspondances, où les items se regroupent par le facteur de collocation ; la projection planaire propose la distribution des plus fortes corrélations lexicales[4], sous la forme d'items

[3] Voir Brunet (2006).

[4] La mesure de la cooccurrence repose sur le rapport de vraisemblance de Dunning (1993).

regroupés par nuages de points. L'attraction qui s'exerce entre deux points est inversement proportionnelle à la distance qui les sépare sur le graphique.

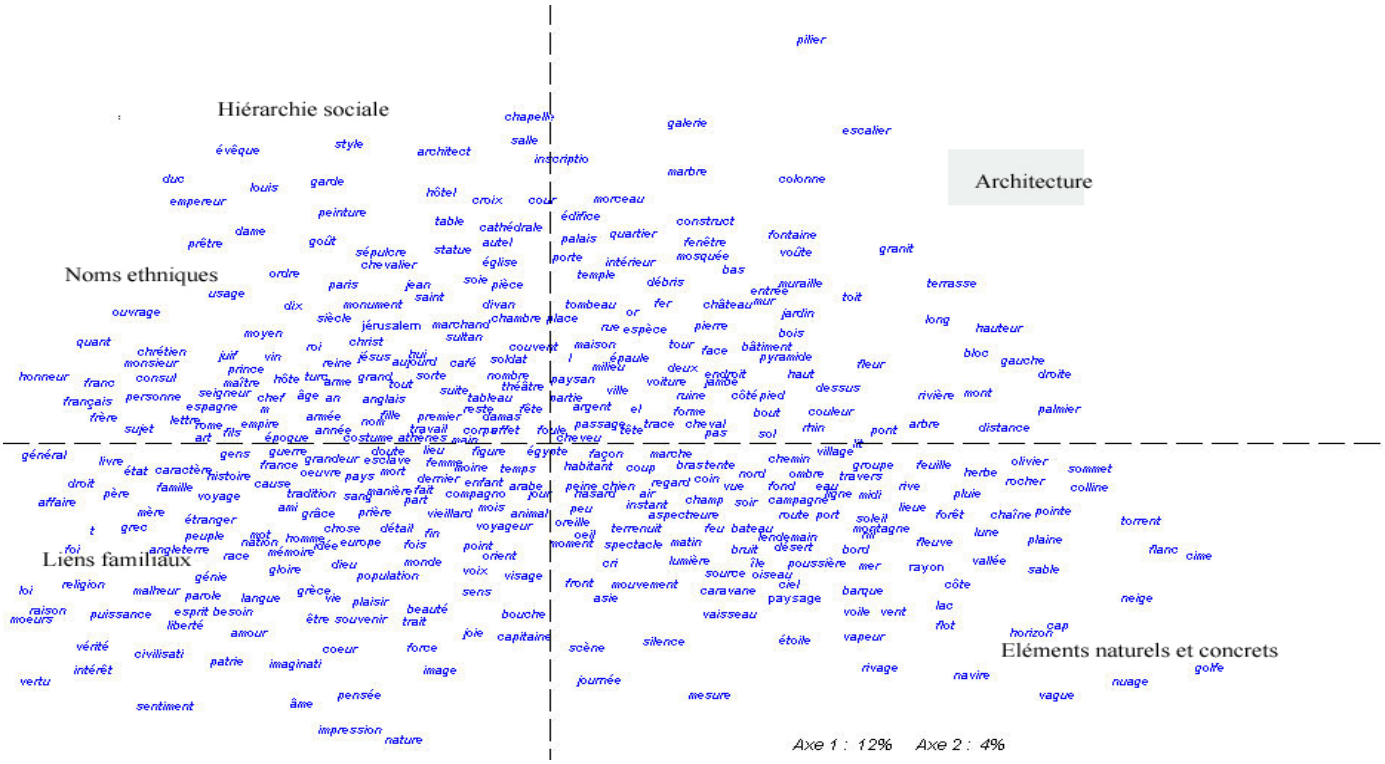

Figure 3 - Analyse factorielle des associations privilégiées pour le récit de voyage

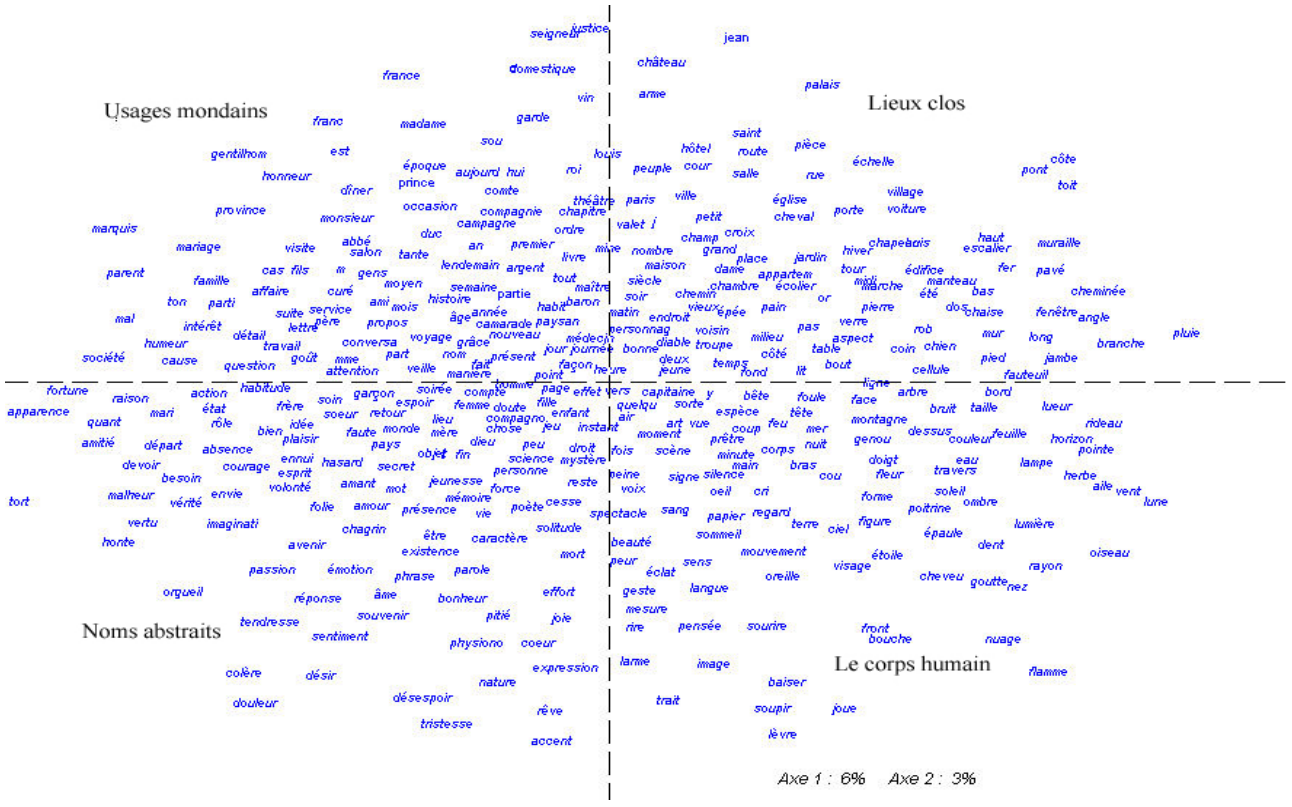

Figure 4 - Analyse factorielle des associations privilégiées pour le texte fictionnel

La distribution des items sur le graphique établit des corrélations entre mots du lexique qui se trouvent rapprochés dans l'espace du texte. Si on observe la répartition par quadrants, il est possible d'interpréter ce graphique en se fondant sur les sèmes communs qui peuvent justifier les cooccurrences. La proximité sémantique expliquerait le développement de réseaux lexicaux dans un texte. En comparant les deux graphiques, on observe que la distribution des items se fait de manière similaire pour les deux genres, dans le sens où dans les quadrants gauches se trouvent les termes qui ont trait aux relations sociales pour le quadrant haut, et aux sentiments pour le quadrant bas, tandis que les quadrants droits rassemblent d'une part les constructions architecturales et d'autre part les éléments de la nature et les parties du corps humain. Cependant, c'est dans l'observation plus précise des items présents dans chaque sphère lexicale que des différences se révèlent entre les genres. La sphère du social est pour le roman celle des usages mondains avec des lemmes comme « mariage, dîner, salon, visite » tandis que le texte factuel privilégie le réseau de la hiérarchie sociale, que les liens soient politiques ou religieux, avec des termes comme « évêque, prêtre, empereur, prince, seigneur, roi » et fait apparaître les noms ethniques qui tissent les liens entre les nations « turc, juif, chrétien, Espagne, français ». Les liens ethniques débordent dans le quadrant gauche inférieur avec des noms collectifs comme « peuple, nation » et des lemmes caractéristiques du voyage « étranger, race ». Dans le quadrant inférieur gauche, le texte fictionnel rassemble un réseau développé de termes abstraits qui transcrivent des sentiments « amitié, passion, honte, orgueil, tendresse, passion, colère, douleur, désir, désespoir, tristesse, pitié, amour » par exemple tandis que le récit de voyage développe beaucoup moins ce réseau, délaisse les nuances et privilégie des abstractions plus usuelles et de sens plus général comme « raison, imagination, impression ». À ce réseau lexical répond, dans le quadrant droit correspondant, l'inventaire des éléments naturels qui entrent dans les séquences descriptives et qui se développe au gré de quasi-synonymes quelquefois au sein d'une même isotopie « route, port, fleuve, côte, lac, rivage, rive, mer, golfe, montagne, forêt, désert ». Alors que les seules parties du corps mentionnées sont ici « l'oreille » et « l'œil », le texte fictionnel renverse l'équilibre, réservant la part congrue aux termes concrets de la nature et faisant la part belle à l'énumération des parties du corps humain. Enfin, les quadrants nord-est s'opposent par le réseau de termes architecturaux et de termes plus exotiques pour le récit de voyage (« marbre, édifice, palais, voûte, colonne, muraille, mosquée, temple, pyramide ») et par le réseau de termes qui dénotent des lieux clos pour le roman (« hôtel, salle, cour, palais, pièce, église, jardin »).

3.3. Le test des corrélats lexicaux. Les mots-pôles

C'est à partir du même tableau que celui établi pour les calculs précédents des associations privilégiées qu'on peut obtenir les graphes qui suivent. Le programme analyse et trie le détail des associations deux à deux et propose une représentation sous forme de graphes des liens préférentiels qui tissent un réseau lexical autour d'un mot choisi pour pôle. Un parcours interprétatif, qui repose sur l'hypothèse initiale que « le contexte proche est structuré par des isotopies »[5], conduit de l'analyse lexicale à l'analyse thématique, des cooccurrents aux corrélats. Cette étude se veut purement expérimentale et les trois mots choisis comme mots-pôles le sont en raison de caractéristiques inhérentes et hypothétiques et non pas comme mots-vedettes ou comme lexicalisation synthétique d'un thème défini *a priori* comme objet d'étude : il s'agit de deux substantifs dont la polysémie est prometteuse « langue » et « nature » et un autre, « homme » *a priori* sensible à l'interprétation sémantique.

[5] Rastier (1996).

Sur les graphes, les conventions suivantes sont adoptées : les mots encadrés ou les mots en rouge désignent les coocurrents directs du mot-pôle, reliés par un trait rouge, qui dessinent le premier cercle des cooccurrents. Les tracés en bleu matérialisent les relations des mots du premier cercle entre eux ; les tracés en noir dessinent le deuxième cercle c'est-à-dire les relations des mots du premier cercle avec d'autres mots qui n'ont pas de contact avec le mot-pôle. L'épaisseur du trait – trait en pointillés, trait maigre ou gras – correspond à la force de la liaison.

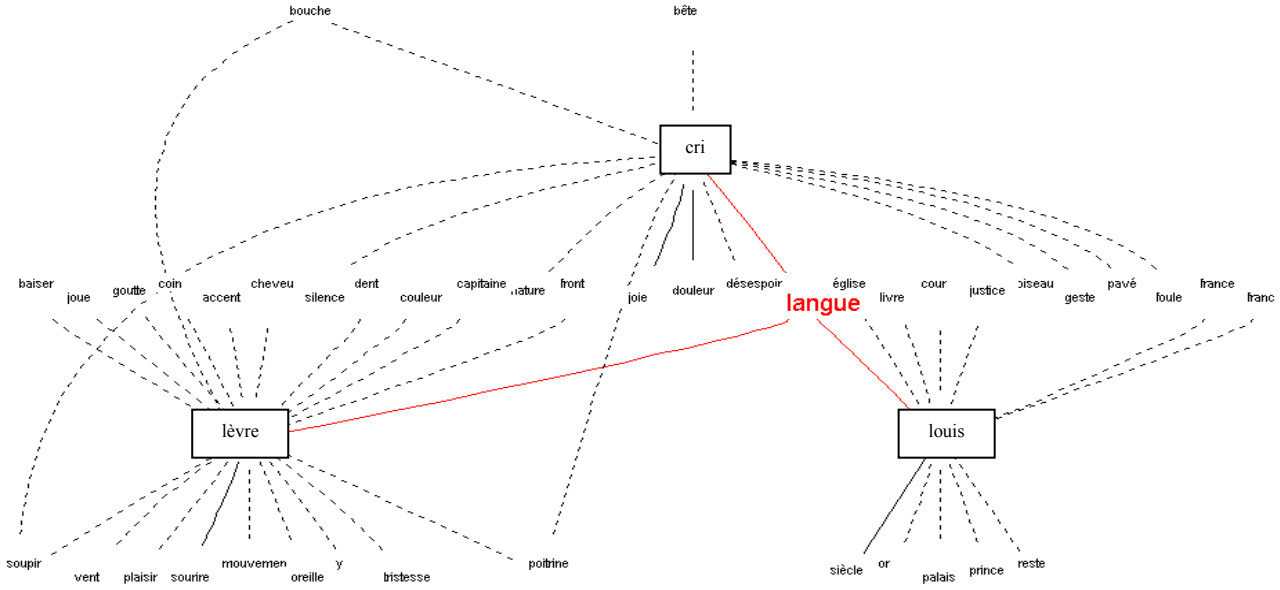

Figure 5 - Mot-pôle « langue » dans le corpus fictionnel

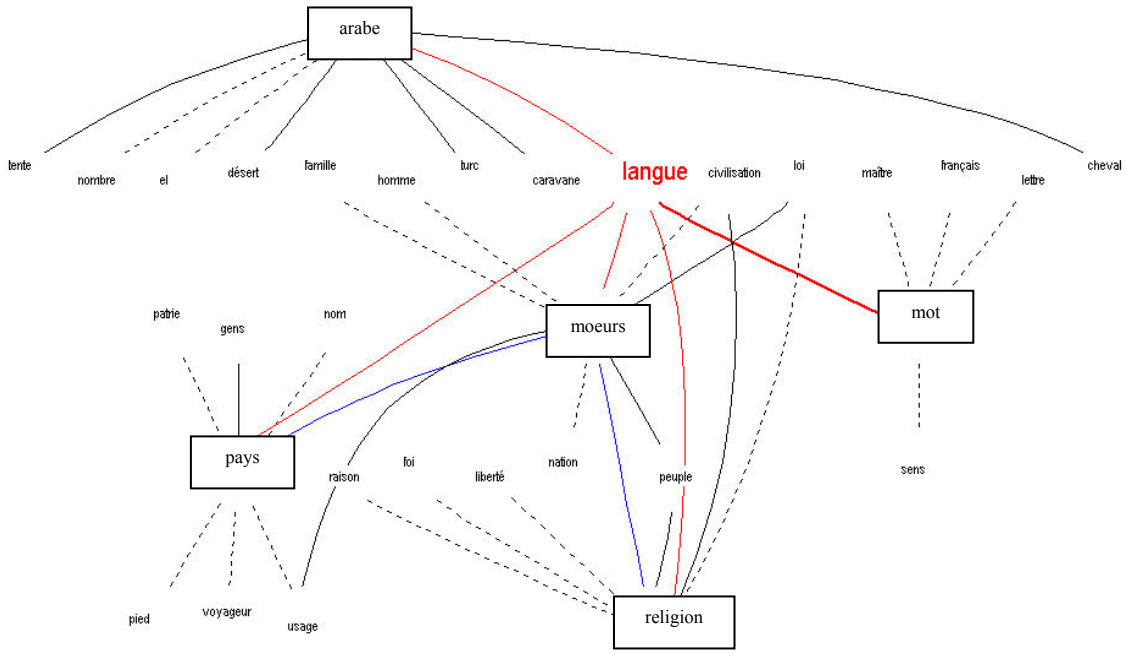

Figure 6 - Mot-pôle « langue » dans le corpus factuel

Pour ce mot, la répartition est assez tranchée selon les genres et aisée à commenter. Le texte fictionnel privilégie l'acception de la langue comme organe et comme organe de la parole ; le mot est associé à des termes concrets comme « lèvre » ou à des mots expressifs comme

« parole » ou « cri ». En revanche, c'est le sens de la langue comme système de signes caractéristique d'un pays que le récit de voyage exploite.

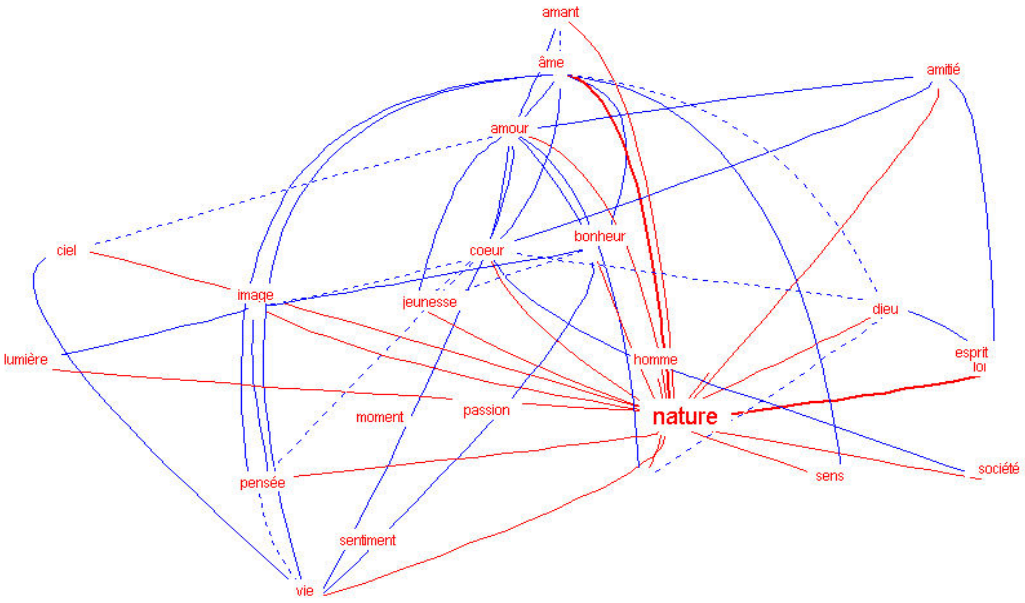

Figure 7 - Mot-pôle « nature » dans le corpus fictionnel

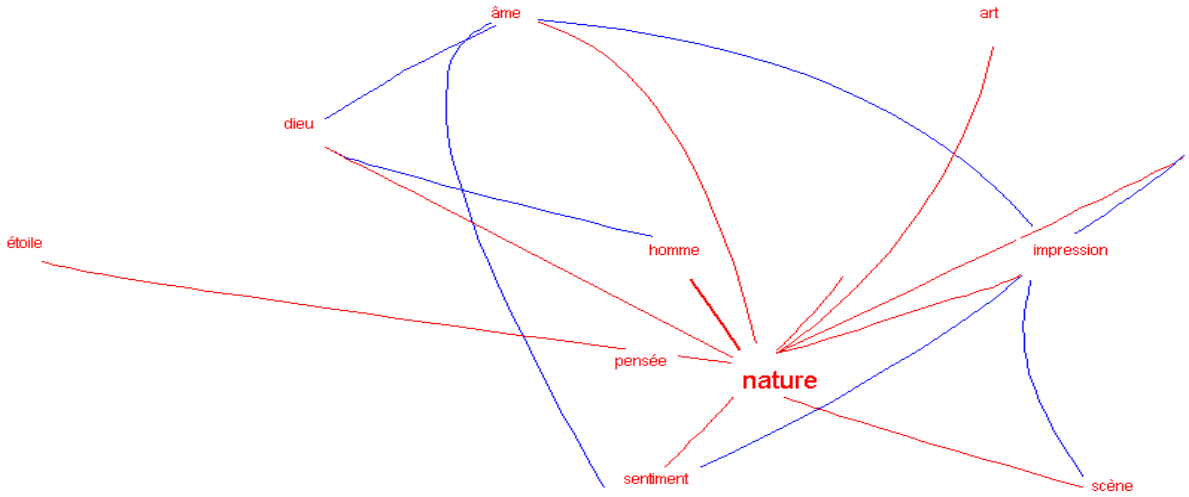

Figure 8 - Mot-pôle « nature » dans le corpus factuel

Pour ces deux figures, une variante toute récente du logiciel a été expérimentée : la possibilité de supprimer les liaisons annexes qui s'établissent entre les « mots-amis » du mot-pôle et d'autres mots qui n'ont cependant aucun lien avec le mot-pôle. Les graphes s'en trouvent allégés et plus clairs, étant entendu que sont supprimés seulement les liens qui n'affectent pas directement le mot-pôle, autrement dit les liens collatéraux. On garde à l'esprit que les liens directs peuvent être eux-mêmes perturbés par les artefacts que sont les expressions phraséologiques, comme « loi de nature », qui ne traduisent qu'un usage de langue sans relever du choix du locuteur. Cependant, le chercheur a toujours la possibilité de rétablir la totalité des liens et d'évaluer le profit qu'il peut en tirer. Les résultats sont là encore intéressants. Pour le texte fictionnel, les liens les plus forts sont avec les mots « loi et âme ». Autrement, l'opposition entre « nature » et « société » est matérialisée par l'apparition de ce dernier terme dans l'entourage immédiat du mot-pôle et il semble que ce soit l'acception du terme « nature » comme « nature humaine » qui soit privilégiée ici au vu du réseau lexical

développé du sentiment qui compte des termes comme « amour, amant, amitié, bonheur, cœur, passion, pensée, sentiment ». Si on observe à présent le graphe du mot-pôle « nature » pour le corpus factuel, on constate un lien privilégié avec le mot « homme » ; ce couple « homme – nature » paraît bien pouvoir fonctionner comme armature du récit de voyage dont l'enjeu est cette confrontation du voyageur – individu singulier et exemplaire représentatif – et des décors naturels traversés. De même, on trouve dans l'environnement proche du mot « nature » des substantifs comme « impression, scène, pensée, sentiment » qui traduisent l'effet produit par le paysage naturel par exemple ; la relation établie avec le mot « art » conforte cette hypothèse d'un regard plus critique porté par le voyageur sur la nature et d'une vision à tendance plus esthétisante.

Enfin, des recherches similaires ont été menées pour le mot-pôle « homme », extrait du tableau des cooccurrences. Cependant, une autre option de présentation a cette fois été choisie : celle qui propose conjointement deux histogrammes. Cette présentation permet de comparer l'environnement d'un mot dans les deux bases parallèlement. Une liste de mots dite littéraire sert de point de comparaison[6]. Sur l'axe des ordonnées figurent les écarts réduits affectés à chaque mot selon le calcul hypergéométrique des cooccurrences.

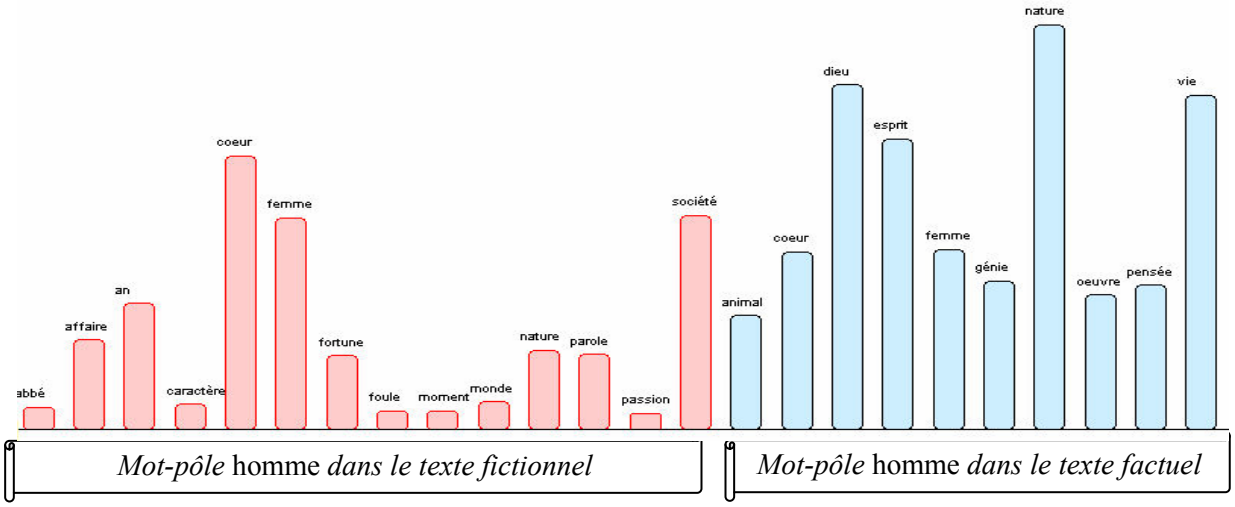

Figure 9 – Histogrammes comparés du mot-pôle « homme » dans les deux corpus

Dans le corpus fictionnel, le mot « homme » entre dans plusieurs réseaux lexicaux : un réseau affectif en premier lieu avec les mots « cœur » et « femme » et, dans une moindre mesure « passion » ; un réseau social et économique ensuite dessiné par les mots « société, affaire, fortune, monde, foule, parole » ; un réseau religieux enfin à peine esquissé toutefois par le terme « abbé ». Le texte factuel retrouve, quant à lui, le lien très fort entre l'homme et la nature et dessine davantage l'opposition entre le règne humain et animal ; il affiche par ailleurs une visée plus didactique et philosophique avec des substantifs abstraits comme « esprit, génie, pensée, vie » et ébauche une réflexion métaphysique par le lien fort établi avec le mot « dieu », absent de l'environnement pour le corpus fictionnel.

[6] Voir le logiciel *Hyperbase*. La liste littéraire de 400 mots est fondée sur un choix de romans présents dans la base dénommée *Exemplem*. Le roman est, dans cette perspective, envisagé comme le prototype de l'œuvre fictionnelle.

3. Conclusion et perspectives

Les tests que j'avais effectués, lors de ma précédente contribution, sur la distribution des catégories morpho-syntaxiques dans le but de caractériser un ensemble générique s'étaient révélés pertinents. On avait ainsi pu expliquer les contrastes observés par l'opposition généralisée entre le mode descriptif - qui privilégie la classe nominale - pour le récit de voyage et le mode narratif - dynamique qui manifeste une prépondérance pour la classe du verbe - pour le texte fictionnel et romanesque en particulier, et proposer ainsi l'hypothèse d'une différenciation générique sur le plan grammatical.

L'exploration lexicale et sémantique des corpus paraît également ouvrir de nouvelles pistes de recherche stimulantes : la distribution des isotopies sur les AFC révèle déjà un traitement spécifique à chaque ensemble ; si la répartition des thématiques générales reste stable d'un genre à l'autre, c'est dans le détail des items regroupés que résident les contrastes lexicaux susceptibles d'interprétations sémantiques. L'inscription générique se manifeste aussi dans le traitement spécifique de thèmes communs ; les réseaux lexicaux qui se dessinent autour d'un mot-pôle décrivent des choix qui incombent moins à l'auteur qu'au genre où ils prennent place. Les perspectives ouvertes par ces nouvelles fontionnalités sont prometteuses et devraient pouvoir être encore davantage exploitées par la multiplication des tests. L'hypothèse d'une différenciation générique sur le plan lexical et sémantique devra être confirmée par d'autres expérimentations.

Références

Bourion E. (2001). *L'aide à l'interprétation des textes électroniques*. Thèse, Université de Nancy II. Ed. pdf. électroniques. *Texto !* mars 2003 [en ligne]. Disponible sur : http://www.revue-texto.net/Corpus/Publications/Bourion/Bourion_Aide.html Thèse présentée et soutenue le 14 décembre 2001 à l'Université de Nancy II en vue de l'obtention du titre de Docteur en sciences du langage. Jury : Étienne Brunet, Bernard Combettes, Nicole Mozet, François Rastier (directeur), Laurent Romary(co-directeur).

Brunet E. (2006). Navigation dans les rafales. Disponible sur : http://www.cavi.univ-paris3.fr/lexicometrica/jadt/JADT2006-PLENIERE/JADT2006_EB.pdf.

Brunet E. (2007). Logiciel *Hyperbase*.

Labbé C. et Labbé D. (2003). La distance intertextuelle. *Corpus*, 2, p.95-118.

Rastier F. (1996). La sémantique des thèmes - ou le voyage sentimental. *Texto* [en ligne]. Disponible sur : http://www.revue-texto.net/Inedits/Rastier/Rastier_Themes.html.

Rastier F. (2001). *Arts et sciences du texte*. Paris, PUF.

Rastier F. et Pincemin B. (1999). Des genres à l'intertexte. *Cahiers de praxématique* 33, p.83-111.

Le corpus

Auteur	Texte factuel	Occurrences	Lemmes	Texte fictionnel	Occurrences	Lemmes
Chateaubriand	*Itinéraire de Paris à Jérusalem,* 1812	247351	19524	*La Vie de Rancé,* 1844	9288	2443
Du Camp	*Le Nil, Egypte et Nubie,* 1854	62886	7544	*Mémoires d'un suicidé,* 1853	214922	12325
Flaubert	*Par les champs et par les grèves,* 1848	271775	16123	*Madame Bovary,* 1857	71011	7920
Fromentin	*Un Eté dans le Sahara*, 1857	245665	19910	*Dominique,* 1869	214934	16475
Gautier	*Voyage en Espagne,* 1843	61178	8929	*Le capitaine Fracasse,* 1863	71091	5662
Hugo	*Le Rhin,* 1842	274976	19553	*Notre-Dame de Paris,* 1832	24959	4659
Lamartine	*Voyage en Orient,* 1835	148033	15567	*Raphaël,* 1849	226317	20284
Maupassant	*La vie errante,* 1890	43086	7225	*Une vie,* 1883	145032	13977
Nerval	*Voyage en Orient, 1851*	100015	11566	*Aurélia*, 1855	102877	11347
Sand	*Un hiver à Majorque,* 1842	9013	2281	*La petite Fadette,*1849	90879	9891
Stendhal	*Mémoires d'un touriste,* 1838	52993	7952	*Le Rouge et le Noir,* 1830	91583	9769
Total		1516971	50506		1262893	44637