HAL
open science

Baudelaire, Rimbaud et Verlaine

Cyril Labbé, Dominique Labbé

To cite this version:

Cyril Labbé, Dominique Labbé. Baudelaire, Rimbaud et Verlaine : Communication aux VIIIe Journées de l'ERLA. Banks David. Aspects linguistiques du texte poétique, L'Harmattan, pp.17-45, 2011. halshs-00273725

HAL Id: halshs-00273725

https://shs.hal.science/halshs-00273725

Submitted on 21 Apr 2008

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.

Université de Bretagne occidentale
Faculté des Lettres et Sciences sociales
Equipe de Recherche en Linguistique Appliquée

Nouvelles Journées de l'ERLA n° 8

Aspects linguistiques du texte poétique
(Brest 16-17 novembre 2007)

Baudelaire, Rimbaud et Verlaine

Cyril Labbé
Université Grenoble I
(cyril.labbe@imag.fr)

Dominique Labbé
Institut d'Etudes Politiques de Grenoble
(dominique.labbe@iep.grenoble.fr)

Résumé
Existe-t-il un genre poétique particulier ? Qu'est-ce qui différencie ce genre poétique du reste de la littérature ? On répond à ces deux questions en utilisant d'abord les œuvres de Baudelaire, Rimbaud et Verlaine. L'analyse est ensuite étendue à l'ensemble de la littérature de la seconde moitié du XIXe siècle. La poésie en vers se distingue de la prose par des densités d'emplois différentes des parties du discours. La poésie privilégie le groupe nominal (substantifs, adjectifs et déterminants) ; la prose utilise plus de verbes, d'adverbes et de pronoms. On présente enfin le vocabulaire et les thèmes propres à la poésie versifiée.

Mots clefs : France – Littérature – XIXe siècle – Poesie – Prose – Vers – Baudelaire – Rimbaud – Verlaine - Statistique lexicale.

Version provisoire
(en cours d'évaluation pour publication dans les actes de la conférence)

Cette année le choix des organisateurs des Journées de l'Erla amène à se poser une question préalable : existe-t-il réellement un langage propre à la poésie ? C'est seulement si l'on peut répondre par l'affirmative à cette première question qu'il devient pertinent de rechercher quels sont les aspects linguistiques particuliers des textes produits à l'aide de ce "langage".

La statistique permet d'envisager de manière originale cette question "préjudicielle". On s'aidera pour cela des œuvres de Baudelaire (1821-1867), Rimbaud (1854-1891) et Verlaine (1844-1896).

Après avoir présenté ce corpus et les opérations préalables nécessaires à son traitement informatique, nous montrerons qu'il existe une opposition claire entre la prose et les vers et que cette opposition se retrouve dans toute la littérature de la seconde moitié du XIXe siècle. Dans chacun de ces deux genres (prose vs poésie versifiée), les "parties du discours" ont des densités d'emploi très différentes. Ce sera enfin l'occasion d'évoquer succinctement le vocabulaire et les thèmes propres à la poésie versifiée.

1. Corpus et traitements préalables

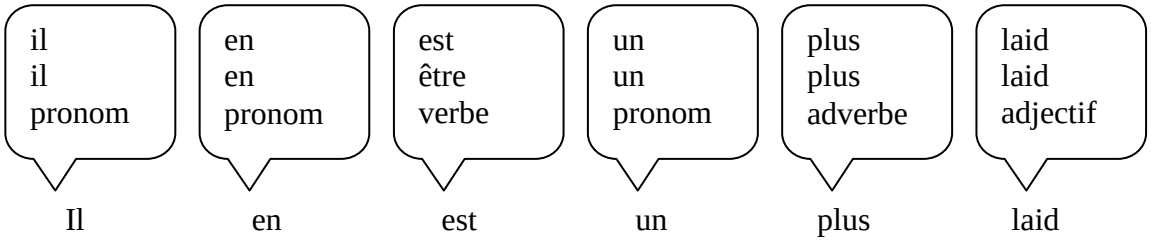
Baudelaire, Rimbaud et Verlaine : ces trois noms ont été si souvent rapprochés qu'il n'est sans doute pas nécessaire de justifier leur étude commune. La liste des œuvres et leurs principales caractéristiques statistiques sont présentées dans l'annexe 1 à la fin de cette communication. Chacun de ces textes a fait l'objet de trois traitements préalables (pour le détail de ces traitements : Labbé 1990).

En premier lieu, l'orthographe a été soigneusement corrigée et les graphies multiples ont été standardisées (événement et évènement ; puis et peux, etc.)… En effet, les textes que l'on trouve en ligne sont issus de scanners et comportent de nombreuses imperfections surtout sensibles dans la ponctuation, les majuscules, etc. Ainsi, la lettre "l" (minuscule) est souvent confondue avec le "i" majuscule ou avec le chiffre 1… Par exemple dans la version de *Une saison en enfer*, mise en ligne par la BNF ont trouve le mot « MongoI » (avec un "i" majuscule à la place du "l" minuscule terminal). Un nettoyage approfondi est donc nécessaire avant de constituer une base de données textuelles et avant toute statistique. Ces tâches sont partiellement effectuées par des automates, mais les interventions manuelles sont nécessairement nombreuses et suivent des règles bien précises.

En second lieu, des balises indiquent, au début, les sources du texte puis délimitent les poèmes, les chapitres, les titres, etc. On s'est limité ici à un balisage minimal : titres, date et lieu (quand ceux-ci sont présents). Beaucoup d'autres indications sont imaginables : numéros des vers dans la poésie, pages de l'édition de référence, typographies particulières, etc. Mais elles ne sont pas utiles pour nos objectifs.

Enfin, la lemmatisation attache à chacun des mots du texte une étiquette contenant la forme graphique normalisée et l'entrée sous laquelle le mot peut être retrouvé dans un dictionnaire. La figure 1 donne un exemple tiré de l'avertissement au lecteur placé en tête des *Fleurs du mal*.

Figure 1. Exemple d'étiquettes attachées à chacun des mots du texte



On remarque que, l'étiquette vient s'ajouter au texte sans l'altérer en rien. Par exemple, dans les *Fleurs du mal*, Baudelaire écrit assez systématiquement "Mal" avec une majuscule initiale. Cette majuscule initiale est conservée dans le texte mais elle est réduite en minuscule dans l'étiquette afin de signaler qu'il ne s'agit pas d'un "nom propre" (patronyme ou toponyme par exemple) mais d'un "substantif masculin" (à distinguer de l'adverbe de même graphie).

Cette levée des ambiguïtés est indispensable. Ainsi dans l'exemple ci-dessus, 5 des 6 mots sont ambigus :

- en : préposition ou pronom ?
- est : substantif (point cardinal) ou verbe "être" ?
- un : déterminant ou pronom ?
- plus : adverbe ou verbe "plaire" ?
- laid : adjectif ou substantif ?

Ces problèmes sont banaux : dans tout texte écrit en français, plus du tiers des mots peuvent être rattachés à plus d'une entrée de dictionnaire et il s'agit souvent des mots les plus fréquents.

L'étiquette comporte trois informations : la forme standardisée – majuscule initiale des mots communs ramenée en minuscule (comme pour "Il" premier mot du vers), réduction des formes multiples à une graphie standard, correction des fautes d'orthographe… - puis le **vocable** – c'est-à-dire l'entrée où se trouve la forme dans le dictionnaire et la catégorie grammaticale, telle qu'elle figure en seconde position dans cette entrée de dictionnaire. Ainsi, les conjugaisons d'un même verbe sont groupées sous son infinitif ou les pluriels du substantif sous le singulier ou encore les féminins et pluriels de l'adjectif sous le masculin singulier. Par exemple, “être v.” regroupe toutes les formes conjuguées de ce verbe, tandis que “est n. m.” ne se rencontre qu'avec le singulier et le pluriel.

La nomenclature des mots, apprise à l'ordinateur, est systématique (par exemple, en français, les substantifs se distinguent par le genre, donc tous les substantifs doivent se voir affecter le masculin ou le féminin), elle est exhaustive (tous les mots doivent y trouver leur place), elle est univoque (une seule entrée par mot) elle exclut tout double compte, elle ne comporte pas de catégorie ad hoc, ou fourre-tout, etc. Enfin la lemmatisation est réversible : on peut retrouver le texte original, sans altération, à partir du fichier étiqueté.

Les utilisations sont multiples. En premier lieu, le vocabulaire d'un corpus est aisément établi avec, sous les lemmes, les formes graphiques sous lesquelles les vocables sont attestés dans ce corpus. Par rapport aux traitements sur les formes graphiques brutes, la normalisation et la lemmatisation donnent une existence aux verbes (en rassemblant leurs flexions sous une étiquette commune), ce qui permet de retrouver certains mots - comme le point cardinal “est”, les substantifs “être”, “avoir”, “avions”… - dont les occurrences sont habituellement noyées dans l'océan des formes verbales homographes.

Pour bien comprendre les calculs présentés ci-dessous, il faut donc se souvenir qu'un texte est la succession d'un certain nombre de **mots** – dont le nombre total donne la **longueur** (voir en annexe 1 les caractéristiques des textes de Baudelaire, Rimbaud, Verlaine) – ces mots étant issus d'un **vocabulaire** nécessairement plus restreint puisque certains **vocables** (ou "mots différents") sont employés plusieurs fois dans le texte. Par exemple, "le", "les", "la", "l´" – et leurs équivalents avec une majuscule initiale - sont les différentes **formes graphiques** sous lesquelles l'article ou le pronom "le" apparaissent dans un texte. "le, article" et "le, pronom" sont des **vocables** (ou "entrées de dictionnaire"). Chacune des **occurrences** de ces deux vocables – sous les formes "le", "la", "les", "l´", "Le", "La", "Les", "L´" – constituent des mots du texte.

Un calcul de distance et des opérations de classification effectués sur ces textes permettent de vérifier que chaque auteur se différencie des deux autres mais aussi qu'il existe un genre "poésie versifiée" différent des autres genres littéraires.

2. Distances et classification

Il s'agit de mesurer la distance entre les textes comme on le fait entre des objets quelconques. L'unité de mesure est ici le vocable. La distance entre deux textes – nommée pour cela : "distance intertextuelle" – est le nombre de vocables différents que contiennent ces textes. Plus ce nombre est élevé, plus leur distance est grande (pour le détail des calculs : Labbé & Labbé 2001 ; Labbé & Labbé 2003).

Cette mesure comporte plusieurs limitations. En premier lieu, elle ne doit pas être appliquée à des textes trop courts où la présence de quelques vocables rares peut avoir une incidence exagérée. En tous cas, les textes doivent comporter plus de 1000 mots et, jusqu'à 4.000 mots environ, certaines distorsions sont possibles. Ceci interdit d'étudier les poèmes un à un et oblige à constituer des ensembles sur des bases raisonnées (qui sont discutées ci-dessous). En second lieu, les textes ne doivent pas avoir des longueurs trop différentes (le rapport entre la plus grande et la plus petite doit être inférieur à 1/10), ce qui oblige à découper les textes trop longs pour les rapprocher de la taille des autres.

La méthode sera d'abord illustrée avec les œuvres de Baudelaire.

Un exemple : C. Baudelaire

Le découpage des œuvres est le suivant (pour les longueurs, voir l'annexe 1).

Les *Fleurs du mal* ont été découpées en deux parties :

- *Les Fleurs du mal 1* : poèmes présents dans la première édition de 1857 ;

- *Les Fleurs du mal 2* : poèmes ajoutés dans l'édition de 1861 plus ceux apportés dans les *Epaves* (1866) et par l'édition posthume de 1868 ;

Les Paradis artificiels (1860) ont été découpés en trois parties de longueurs sensiblement égales en respectant les césures du texte : 1. "Le poème du haschisch" (publiée dans la *Revue contemporaine* en 1858) ; 2 et 3. "Un mangeur d'opium" (adaptation des *Confessions* de Quincey publiée dans la *Revue contemporaine* en 1860);

- *Le Spleen de Paris* comporte les 21 "Poèmes en prose" rassemblés sous ce titre dans les *Œuvres complètes*.

Le tableau 1 présente les résultats du calcul des distances appliqué à ces 6 textes. Ce tableau a une diagonale nulle car la distance d'un texte à lui-même est égale à 0. La distance intertextuelle étant symétrique, le tableau l'est également (par exemple, la deuxième case de la première ligne est égale à

la deuxième de la première colonne). Ces deux cases indiquent que les poèmes de la première édition des *Fleurs du mal* (Fleurs1) et ceux ajoutés par les éditions ultérieures (Fleurs2) sont séparés par une distance de 0,251, c'est-à-dire que, en moyenne, sur 4 mots, 3 sont communs à ces deux textes. La distance la plus faible (0,242) sépare les deux parties des *Paradis artificiels* qui sont toutes deux extraites du même texte original (*Un mangeur d'opium*). La plus grande distance (0,407) sépare Fleurs2 - poèmes ajoutés aux *Fleurs du mal* après la première édition - et le début des *Paradis artificiels*, c'est-à-dire des textes en vers (Fleurs2) et en prose (Paradis1).

Tableau 1. Distances intertextuelles séparant trois œuvres de C. Baudelaire.

	Fleurs1	Fleurs2	Paradis1	Paradis2	Paradis3	Spleen
Fleurs1	0,000	**0,251**	0,399	0,399	0,385	0,339
Fleurs2	**0,251**	0,000	**0,407**	0,397	0,383	0,341
Paradis1	0,399	**0,407**	0,000	0,271	0,269	0,289
Paradis2	0,399	0,397	0,271	0,000	**0,242**	0,299
Paradis3	0,385	0,383	0,269	**0,242**	0,000	0,294
Spleen	0,339	0,341	0,289	0,299	0,294	0,000

L'indice de la distance intertextuelle enregistre l'influence de 4 facteurs : l'auteur, le thème, le genre et l'époque. Le dernier facteur se comprend aisément : la langue est un organisme vivant dont le composant sémantique (le "lexique") évolue constamment. Il est donc nécessaire de comparer des textes contemporains afin de neutraliser l'influence de ce facteur "chronologique". L'influence du "genre" est surtout clair quand on compare l'oral à l'écrit et, secondairement, les vers et la prose (comme ici). En effet, de nombreuses expériences menées depuis une dizaine d'années ont abouti à l'étalonnage d'une échelle des distances (Labbé & Labbé 2001). Pour deux textes dont la longueur tourne autour de 10.000 mots, on observe que :

- les valeurs inférieures à 0,20 – moins d'un mot sur 5 est différent – sont rares. Elles ne se rencontrent que pour des textes contemporains écrits par un même auteur dans un même genre et portant sur un même thème ;

- les valeurs comprises entre 0,20 et 0,25 signalent normalement un auteur unique dans un même genre - mais écrivant à des époques un peu plus éloignées ou sur des thèmes légèrement différents. On en rencontre rarement chez deux auteurs différents. Il faut pour cela qu'ils aient écrit à la même époque, dans un même genre et sur un même thème. Dans ce cas, il est aussi très probable que l'un des deux auteurs s'est "inspiré" de l'autre ;

- les valeurs comprises entre 0,25 et 0,35 concernent un même auteur, écrivant dans un même genre mais, dans ce cas, il y a un changement important de thème et/ou d'époque. Si les auteurs sont différents, alors ils sont contemporains et traitent des sujets proches dans un même genre ;

- au-dessus de 0,35, des textes d'un même auteur appartiennent à des genres différents. Si le genre est le même, alors les auteurs sont différents.

Cette échelle s'applique bien aux données du tableau 1 :

- les poèmes publiés en 1857 et ceux qui ont été ensuite ajoutés dans les éditions ultérieures des *Fleurs du mal* sont bien du même auteur dans le même genre et sur des thèmes proches (distance 0,25) ;

- les deux extraits du *Mangeur d'opium* (Paradis1 et 2) possèdent également une forte unité (distance 0,24). En revanche, les textes intitulés *Poème du haschisch* et le *Mangeur d'opium* – réunis ensemble dans l'édition originale des *Paradis artificiels* - sont légèrement décalés ce qui signale quelques différences de thème et/ou de style) ;

- les *Fleurs du mal* et les *Paradis artificiels* sont séparés par des distances de l'ordre de 0,40 qui reflètent deux genres opposés (vers et prose) dont l'influence sera discutée plus bas ;

- Le *Spleen de Paris* se situe entre les deux, ce que justifie d'ailleurs bien le titre sous lesquels certains de ces textes ont parus d'abord ("Poèmes en prose"). Ce recueil est toutefois plus proche en moyenne des *Paradis* (prose) que des *Fleurs*, ce qui suggère que la versification, comme opération technique, pourrait avoir un impact important sur les textes poétiques.

La petite taille du tableau 1 permet une analyse directe de son contenu. Cependant, en face de tableaux plus vastes – celui issu du corpus Baudelaire-Rimbaud-Verlaine comporte 289 cases - le recours à des classifications est une nécessité. Il s'agit de rechercher - de manière automatique, afin de ne pas faire intervenir la subjectivité du chercheur - les meilleurs groupements possibles. Deux critères sont utilisés. D'une part, les distances entre les individus composant un même groupe doivent être les plus courtes possibles ; d'autre part, les distances séparant les différents groupes ainsi constitués, doivent être les plus grandes possibles (pour une présentation de la question : Sneath & Sokal 1973 et Benzecri 1980).

Une fois la classification opérée, il faut représenter graphiquement la population classée en restituant le mieux possible les positions respectives de chacun des individus par rapport à tous les autres. Nous utilisons la méthode dite de la "classification arborée" qui est classique en génétique (Felsenstein 2004a et 2004b ainsi que le site : http://evolution.genetics.washington.edu) ou en linguistique historique (Embleton 1986 et pour une revue récente : Holm 2007). Elle est le prolongement des méthodes multidimensionnelles utilisées en analyse des données (Lebart & Salem 1994).

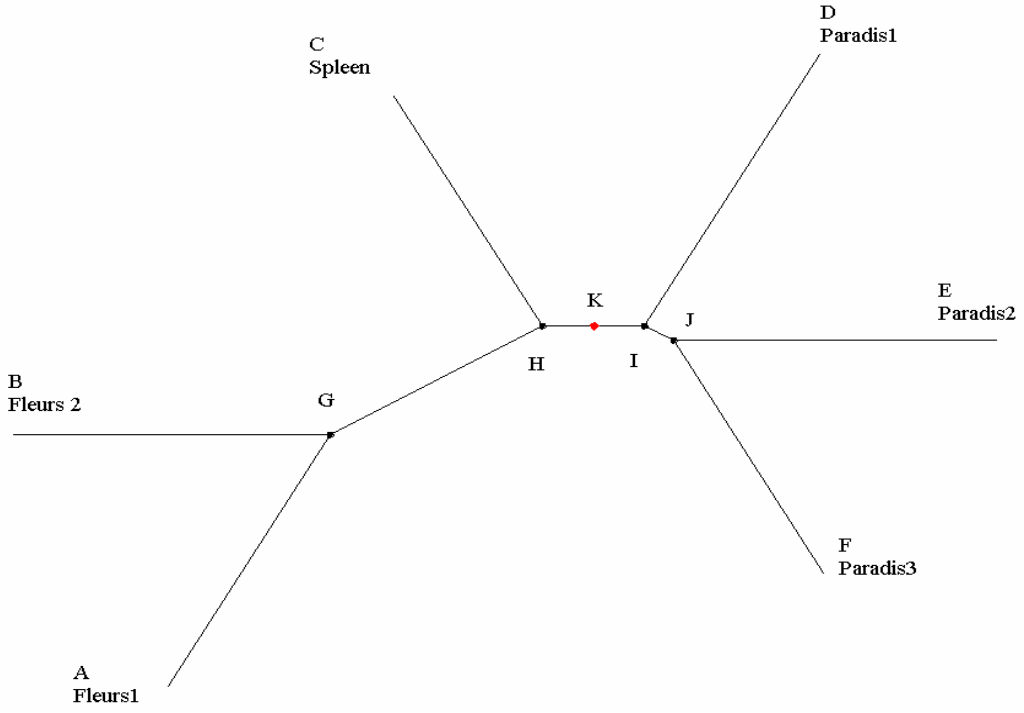
La classification arborée repose sur le théorème suivant : si tous les individus étudiés sont séparés par des distances, il existe un "arbre" qui représente exactement les positions respectives de ces individus les uns par rapport aux autres (pour la démonstration : Luong 1988). La construction d'un arbre "parfait" exige que toutes les combinaisons possibles soient examinées alors que leur nombre augmente exponentiellement en raison de l'effectif de la série. Divers algorithmes ont été imaginés pour éviter d'avoir à examiner toutes ces combinaisons. Nous utilisons l'algorithme mis au point par X. Luong[1]. Appliqué au tableau 1, cet algorithme a tracé l'arbre ci-dessous (figure 2).

Dans cet arbre, les points A, B, C, D, E, F (les titres des œuvres) sont les **feuilles** ; G, H, I et J les **nœuds** figurant les groupements successifs ; K est la **racine** de l'arbre. Les segments de droite joignant les feuilles et les noeuds sont les **arêtes**. La distance entre deux points est figurée par le **chemin** unissant ces points et la longueur de ce chemin est proportionnelle à la distance originelle correspondante (**arbre valué**). Pour "lire" cet arbre, il ne faut pas aller d'un point à un autre "à vol d'oiseau", mais suivre les chemins qui relient ces points. Cela permet de voir la proximité évidente des deux groupes de poèmes des *Fleurs du mal*, d'une part, et celle des deux extraits du *mangeur d'opium* (Paradis2 et 3) et leur très grand éloignement mutuel. Cela permet aussi de visualiser la position intermédiaire du *Spleen de Paris*, à mi-chemin entre la prose et les vers. Cependant, l'arête G-H est nettement plus longue que l'arête

[1] Code source dans Luong, 1988, les principes et les formules sont également présentées dans Luong 1994. Notre logiciel a été réalisé avec son aide et avec celle de M. Ruhlman (Ruhlman 2003).

H-I : la coupure essentielle oppose donc prose (à droite du graphe) et vers (à gauche), que les textes soient ou non à visée explicitement "poétique". Cela semble donc suggérer, au moins pour Baudelaire, l'existence d'un genre littéraire "poésie en vers" distinct des autres.

Figure 2. Classification arborée sur les œuvres de C. Baudelaire



Avant de discuter les caractéristiques de la prose et des vers, la même méthode de classification va être appliquée aux trois poètes étudiés afin de voir si l'on observe chez eux la même opposition entre vers et prose.

Baudelaire, Rimbaud et Verlaine.

Les poèmes de Rimbaud ont été découpés en deux ensembles selon la césure indiquée par tous les biographes comme étant le tournant décisif dans la vie du poète : sa rencontre avec Verlaine. Les dimensions de ces deux ensembles se prêtent bien à l'analyse (annexe 1). Dans l'œuvre de Verlaine, il existe trois recueils nettement plus courts que les autres : *Fêtes galantes* (1869), *La bonne chanson* (1870) et *Romances sans paroles* (1874). Cette petite taille (moins de 3000 mots) induit des risques de déformation dans la comparaison avec les autres. Mais leurs distances mutuelles montrent que ces trois recueils sont très proches entre eux (ce que laissait déjà prévoir la chronologie). C'est pourquoi, pour l'expérience ci-dessous, ils sont groupés ensemble afin de ne pas introduire de distorsion dans les calculs et de rendre les graphiques plus lisibles.

L'annexe 2 reproduit le tableau complet des distances. Le tableau 2 ci-dessous présente les résultats de l'expérience suivante : les textes d'un même auteur sont groupés en fonction de leur genre (vers ou prose) et les distances sont calculées entre ces ensembles. Le tableau se lit ainsi : les poésies en vers de Baudelaire sont séparés entre elles par une distance de 0,254, les textes en prose du même Baudelaire ont entre eux une distance moyenne de 0,277 ; la prose et les vers de 0,381. Ce qui correspond bien à l'échelle standard de la distance présentée ci-dessus.

Tableau 2. Distances intertextuelles entre vers et prose chez Baudelaire, Rimbaud et Verlaine

	Baudelaire poésie	Baudelaire prose	Rimbaud poésie	Rimbaud prose	Verlaine Poésie
Baudelaire poésie	0,254	0,381	0,336	0,383	0,334
Baudelaire prose	0,381	0,277	0,415	0,380	0,378
Rimbaud poésie	0,336	0,415	0,329	0,390	0,358
Rimbaud prose	0,383	0,380	0,390	0,000*	0,377
Verlaine poésie	0,334	0,378	0,358	0,377	0,305
Moyenne	0,338	0,366	0,366	0,383	0,350

* Pour Rimbaud, il n'y a qu'un texte en prose (la distance d'un texte à lui-même est nulle).

Les principales conclusions sont les suivantes :

- en dehors de la prose de Rimbaud – pour laquelle nous n'avons qu'un texte, les valeurs sur la diagonale sont toujours les plus faibles : les variables "auteur" et "genre" additionnent leurs effets pour minimiser la distance ;

- dans un même genre, les distances entre textes d'un même auteur sont inférieures à celles observées entre textes d'auteurs différents. Ainsi, les distances entre les poésies de Rimbaud (0,329) sont inférieures à celles qui les séparent de celles de Baudelaire (0,336) comme de celles de Verlaine (0,358). Autrement dit, dans un même genre, l'auteur est la variable la plus importante ;

- en revanche, chez un même auteur, les textes de genre différents sont séparés par des distances supérieures à celles observées entre textes d'un même genre par deux auteurs différents. Ainsi, les distances entre poésies en vers de Baudelaire et de Rimbaud (0,336) ou de Verlaine (0,334) sont inférieures aux distances entre les vers et la prose de Baudelaire (0,381). Autrement dit, les différences de genre l'emportent sur les différences d'auteur.

- la poésie en vers de Verlaine (du moins à partir des *Fêtes galantes*) se rapproche un peu plus de la prose que les poèmes des deux autres ;

- enfin Rimbaud est le plus décalé des trois tant pour la poésie que pour la prose. Il existe une coupure assez nette entre ses poèmes écrits avant sa rencontre avec Verlaine et ceux écrits ensuite mais, en moyenne, sa poésie est plus proche de celle de Baudelaire que de celle de Verlaine.

Le graphique 3 ci-dessous présente le résultat de la classification arborée opérée sur le tableau des distances en annexe 2 (la feuille correspondant aux trois recueils de Verlaine mentionnés ci-dessus est notée "VerlaineChFeRo").

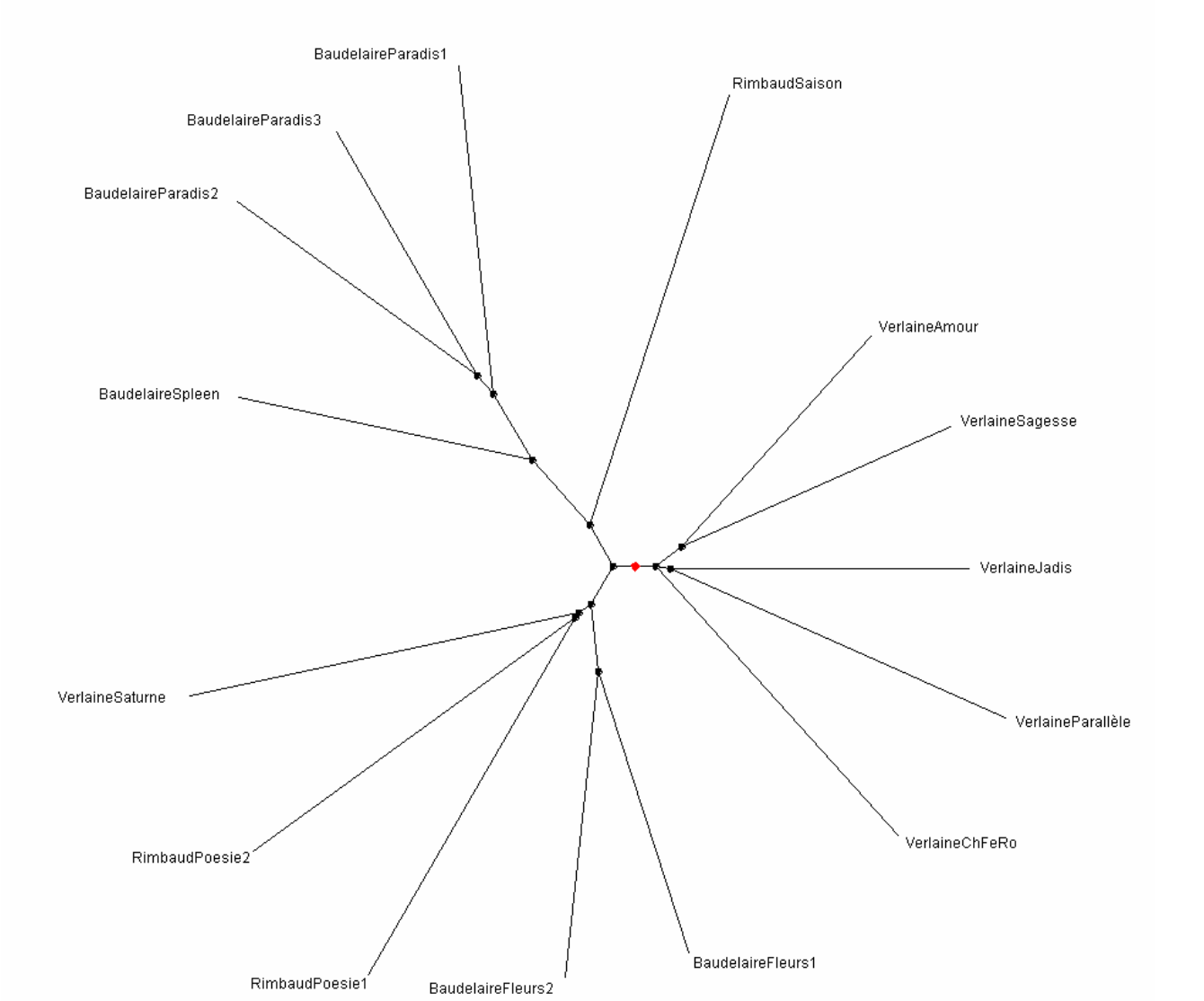
La classification isole deux groupes principaux de textes.

En haut, les textes en prose avec une position intermédiaire pour deux d'entre eux. Pour le *Spleen de Paris* (Baudelaire), cela s'explique par la nature "poésie en prose" déjà signalée ci-dessus. Pour *Une saison en enfer* (Rimbaud), la présence de quelques vers au milieu de la prose peut expliquer, au moins partiellement, cette position à "mi-chemin".

En bas, la poésie versifiée se sépare en deux sous-groupes. A gauche, on trouve les poésies de Rimbaud, groupées ensemble et quasiment à équidistance des *Fleurs du Mal* (Baudelaire) - également groupées ensemble - et du premier recueil de Verlaine (*Poèmes saturniens*, 1866). Toutes les œuvres postérieures de ce dernier (à partir des *Fêtes galantes*,

1869) sont groupées en bas à droite du graphique selon leur ordre chronologique. La coupure majeure dans l'œuvre de Verlaine serait donc antérieure à sa rencontre avec Rimbaud. Enfin, la longueur des branches signale l'aspect "décalé" de Rimbaud par rapport aux deux autres.

Graphique 3. Classification arborée sur les œuvres de Baudelaire, Rimbaud et Verlaine

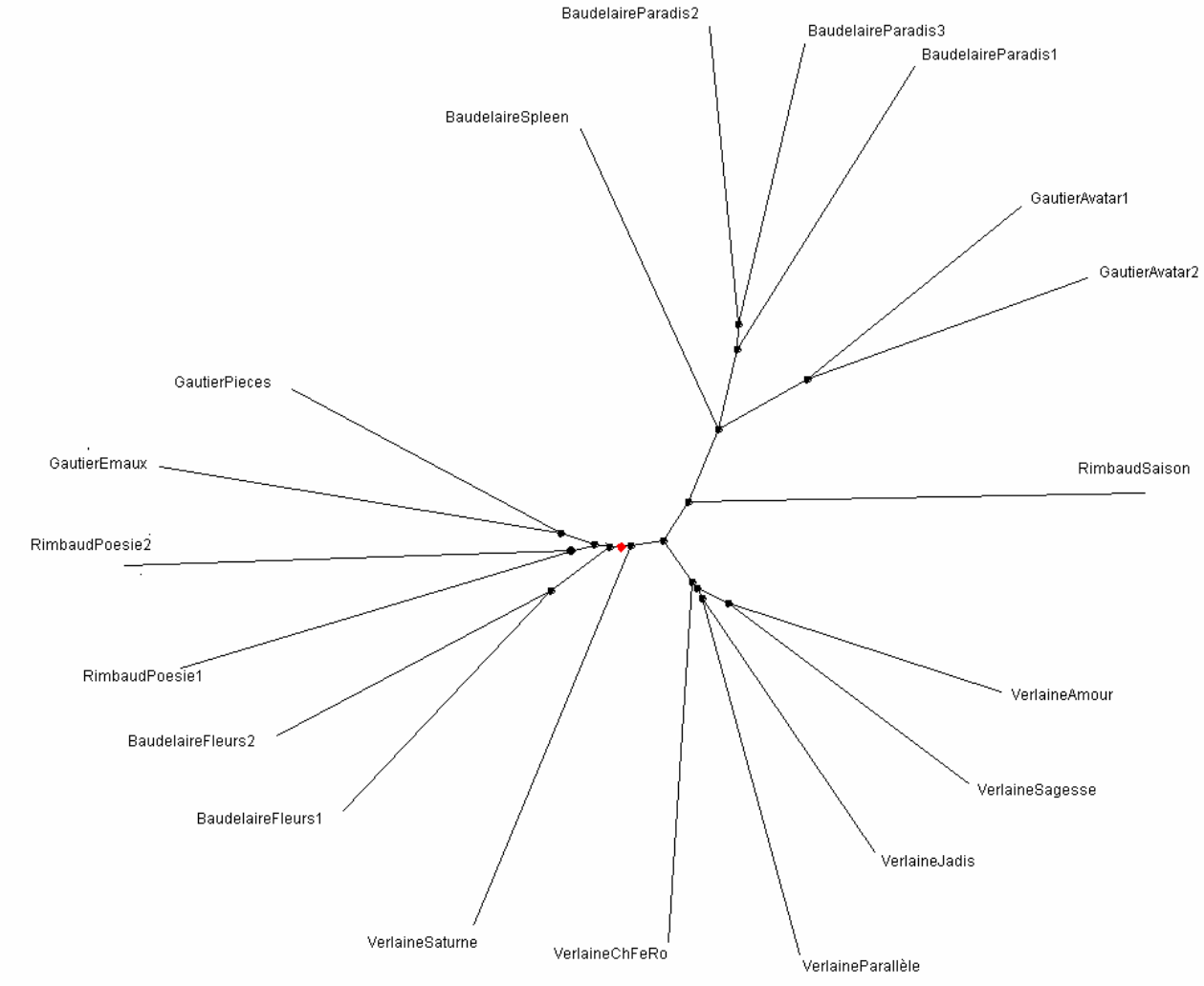


Cet arbre confirme donc l'opposition principale entre vers et prose déjà identifiée dans l'œuvre de Baudelaire. Nous proposons de vérifier cette opposition en élargissant le champ de la recherche.

Baudelaire, Gautier, Rimbaud et Verlaine.

Dans l'histoire littéraire française, plusieurs auteurs ont publié des vers et de la prose. V. Hugo est certainement le plus connu. Pour les contemporains des trois poètes que nous analysons, T. Gautier est l'un de ceux dont la proximité a souvent été soulignée, notamment avec Baudelaire. Nous avons donc répété l'expérience en introduisant les textes de Gautier présents dans nos corpus lemmatisés (annexe III) : deux recueils de poèmes (*Emaux et Camées* ; *Pièces diverses*) et un roman (*L'Avatar*) qui est découpé en deux parties égales pour le ramener à des dimensions comparables aux autres textes.

Graphique 4. Classification arborée sur les œuvres de Baudelaire, Gautier, Rimbaud et Verlaine



La classification antérieure se retrouve sans perturbations majeures et Gautier vient prendre place sur le graphique là où l'on pouvait l'attendre : sa poésie est très proche des *Fleurs du mal* et à peine plus éloignée de celle de Rimbaud et du premier Verlaine ; la prose – assez baroque - de l'*Avatar* est groupée avec les *Paradis* de Baudelaire et avec le *Spleen* et la *Saison en enfer*.

Cette expérience est proche des limites actuelles de la méthode. Bien qu'encore très fiable, l'arbre est d'une qualité légèrement moins bonne que celle des deux précédents et certains chemins doivent être interprétés avec prudence (sur les problèmes de qualité des arbres : Labbé & Labbé 2006). Par exemple, la proximité apparente de la poésie de Rimbaud avec celle de Gautier est légèrement exagérée ; de même que l'éloignement relatif des *Poèmes saturniens* par rapport aux autres œuvres de Verlaine, etc. Toutefois, les conclusions essentielles demeurent : il y a bien trois groupes de textes et une opposition principale entre la prose (en haut du graphique) et les vers (en bas). Au sein de ces groupes, les auteurs sont bien discriminés malgré certaines parentés évidentes.

Cette expérience confirme que les auteurs sont identifiables à condition d'analyser des textes contemporains et de même genre. Ainsi par exemple, il est impossible d'attribuer des vers anonymes, ou douteux, à un auteur connu si l'on n'a que de la prose de la main de cet

auteur. De même, il sera difficile d'attribuer un texte de "maturité" si l'on ne dispose que de textes de "jeunesse" de l'auteur putatif (comme le suggère le cas de Verlaine).

Cette expérience confirme donc la coupure essentielle, chez un même auteur, entre la poésie versifiée et la prose, ce qui permet de conclure à l'existence d'un genre particulier "la poésie versifiée". Quelles sont les caractéristiques propres à ces deux genres littéraires ?

3. Prose et Vers

Il reste à comprendre les raisons de cette opposition systématique et si nette entre prose et vers. Provient-elle de la contrainte du rythme et de la rime comme on le pense généralement ? Il faut remarquer que la distance intertextuelle ne mesure qu'indirectement ce facteur puisqu'elle porte sur le vocabulaire. On doit donc s'interroger sur les différences majeures existant à ce niveau-là.

Les fichiers lemmatisés offrent des éléments de réponse. Par exemple, l'étiquette contenant la catégorie grammaticale (appelée "partie du discours") permet de comparer les densités de ces catégories dans la prose et dans la poésie.

Comparaison des parties du discours dans les vers et la prose

Le tableau 3 présente cette comparaison entre *Spleen de Paris* et les *Paradis artificiel* d'une part (prose) et d'autre part les *Fleurs du mal* (vers). Ce tableau se limite aux effectifs relatifs de chaque catégorie grammaticale afin de permettre la comparaison entre corpus de longueurs différentes (les effectifs absolus peuvent être retrouvés grâce à l'annexe 1).

Le tableau 3 se lit de la manière suivante : dans ses textes en prose, Baudelaire utilise en moyenne 134,6 verbes pour 1000 mots (colonne A) contre 119,5‰ dans les *Fleurs* (colonne B) soit un recul moyen de 11,2% quand il passe de la prose aux vers (dernière colonne du tableau). Les lignes suivantes du tableau montrent que ce recul moyen est le produit de mouvements contradictoires : forte baisse du passé et de l'infinitif mais croissance considérable du participe présent.

Les lignes suivantes du tableau montrent que les pronoms ont des densités d'emploi reliées à celles des verbes et que leur poids relatif semble inverse à celui des substantifs, des adjectifs et des déterminants, ce qui suggère une opposition entre groupes nominal et verbal. Le premier groupe est constitué des pronoms, des verbes, des adverbes et des conjonctions de subordination ; le second est constitué des substantifs, adjectifs, déterminants, prépositions et conjonctions de coordination. Certes, le partage n'est pas exclusif : certains adverbes peuvent se glisser dans le groupe nominal, certaines prépositions dans le groupe verbal, certains pronoms relatifs s'utilisent dans les deux, etc. Enfin les locutions, les mots étrangers, etc. ne sont pas classables dans l'un ou l'autre des groupes.

Tableau 3. Poids des catégories grammaticales dans la prose et les vers de Baudelaire

Catégories	A. Prose (‰)	B. Vers (‰)	B-A (%)
Verbes	134,6	119,5	-11,2
Formes fléchies	*86,5*	*86,6*	*0,1*
Participes passés	*18,7*	*7,5*	*-59,6*
Participes présents	*4,8*	*8,4*	*+74,9*
Infinitifs	*24,7*	*17,0*	*-31,2*
Noms propres	8,8	8,2	-6,9
Noms communs	198,2	235,7	+19,0
Adjectifs	83,8	109,9	+31,1
Adj, participe passé	*13,2*	*17,2*	*+30,3*
Pronoms	103,6	86,5	-16,6
Pronoms personnels	*60,7*	*52,6*	*-13,4*
Pronoms démonstratifs	*8,6*	*4,6*	*-47,1*
Pronoms possessifs	*0,4*	*0,2*	*-44,7*
Pronoms indéfinis	*3,6*	*3,7*	*1,5*
Pronoms relatifs	*24,6*	*23,0*	*-6,4*
Déterminants	183,9	192,3	+4,6
Articles	*130,0*	*136,1*	*+4,7*
Nombres	*7,5*	*2,1*	*-71,9*
Possessifs	*22,1*	*39,9*	*+80,8*
Démonstratifs	*12,1*	*8,5*	*-29,6*
Indéfinis	*12,2*	*5,8*	*-52,7*
Adverbes	68,9	43,9	-36,2
Prépositions	149,5	133,0	-11,0
Conjonctions	65,1	64,9	-0,3
Conjonctions de coordination	*41,3*	*42,1*	*+1,8*
Conjonctions de subordination	*23,8*	*22,9*	*-3,8*
Mots étrangers	2,9	1,3	-56,0
Groupe nominal	665,5	721,2	+8,4
Groupe verbal	330,8	272,7	-17,6

En moyenne, dans les textes écrits en français depuis le XVIIe siècle, le groupe verbal couvre 38% de la surface des textes contre 62% pour le groupe nominal (en négligeant les locutions et mots étrangers). Par rapport à cette moyenne, C. Baudelaire privilégie donc manifestement le groupe nominal, spécialement l'adjectif. Cette tendance est encore accentuée quand il versifie. L'une des différences intéressantes concerne les pronoms (tableau 4).

La relative faiblesse des pronoms ne donne pas pour autant une poésie "impersonnelle". En effet, la troisième personne est une "non-personne" et les vrais pronoms personnels sont les première et deuxième personnes (Benveniste 1956). Or le sous-groupe, formé par *je*, *tu* et *nous*, augmente de 22 à 28‰ en passant de la prose aux vers. L'augmentation la plus frappante est celle du "tu". On en tire que chez Baudelaire, la poésie est un dialogue avec l'autre mais aussi une recherche (ou une nostalgie) de la fusion avec l'autre ("nous"). Le recul de la troisième personne (il, ils) ainsi que du démonstratif signale également une sorte de fermeture autour de cette relation avec l'autre et une tendance à l'exclusion des tiers, du récit, voire de la description. Nous n'entrons pas plus dans le détail de cette analyse qui fait appel à l'énonciation de la subjectivité dans le langage (Benveniste 1958 & 1970 ; Kerbrat-Orecchioni 1981).

Tableau 4. Densités des principaux pronoms dans la prose et les vers de Baudelaire

Pronoms	A. Prose (‰)	B. Vers (‰)	B-A (%)
je	15,7	14,2	-9,4
tu	0,8	7,2	+759,0
il	17,3	6,6	-62,1
on	2,2	1,9	-13,8
nous	2,6	3,6	+40,5
vous	3,3	3,1	-7,1
ils	2,5	2,1	-17,4
ce (pro)	5,7	3,5	-38,0
dont	2,0	2,5	+23,8
le (pro)	5,9	3,0	-49,4
que (pro)	4,3	5,7	+32,4
qui	11,2	11,5	+2,1

Ces caractéristiques sont-elles propres à Baudelaire ou bien existent-elles chez tous les auteurs qui utilisent les deux genres ? Le tableau 5 présente les résultats de l'expérience menée sur Rimbaud et Gautier (puisque de Verlaine, il n'y a que des vers dans le corpus).

Tableau 5. Comparaison du poids des catégories grammaticales dans la prose et les vers chez Rimbaud et Gautier.

Catégories	Rimbaud			Gautier		
	A. Prose (‰)	B. Vers (‰)	B-A (%)	A. Prose (‰)	B. Vers (‰)	B-A (%)
Verbes	159,9	122,2	-23,6	140,1	121,5	-13,3
Formes fléchies	*114,3*	*91,4*	*-20,0*	*92,4*	*86,0*	*-7,0*
Participes passés	*21,0*	*9,3*	*-55,8*	*17,8*	*7,0*	*-60,5*
Participes présents	*3,9*	*8,5*	*+119,8*	*6,1*	*9,7*	*+57,6*
Infinitifs	*20,8*	*13,0*	*-37,6*	*23,7*	*18,8*	*-20,8*
Noms propres	7,8	15,2	93,3	30,8	20,3	-33,9
Noms communs	217,3	241,5	+11,1	200,0	243,3	+21,7
Adjectifs	64,4	105,6	+63,9	80,8	91,5	+13,2
Adj, participe passé	*10,5*	*16,7*	*+58,7*	*17,9*	*18,4*	*+2,7*
Pronoms	138,6	81,0	-41,6	103,5	74,9	-27,6
Pronoms personnels	*102,7*	*50,5*	*-50,8*	*67,7*	*45,8*	*-32,4*
Pronoms démonstratifs	*9,9*	*7,2*	*-27,4*	*3,9*	*3,5*	*-9,6*
Pronoms possessifs	*0,3*	*0,1*	*-82,1*	*0,5*	*0,1*	*-83,6*
Pronoms indéfinis	*4,2*	*2,3*	*-45,7*	*3,0*	*2,9*	*-5,0*
Pronoms relatifs	*16,8*	*18,2*	*+8,1*	*22,6*	*19,4*	*-14,0*
Déterminants	171,1	189,1	+10,5	176,7	201,2	+13,9
Articles	*131,0*	*141,8*	*+8,3*	*128,9*	*145,8*	*+13,2*
Nombres	*2,2*	*5,4*	*+148,8*	*4,5*	*3,9*	*-13,3*
Possessifs	*22,4*	*30,4*	*+35,9*	*23,6*	*38,8*	*+64,3*
Démonstratifs	*5,7*	*5,4*	*-3,9*	*10,5*	*5,5*	*-47,6*
Indéfinis	*9,9*	*5,9*	*-40,0*	*9,3*	*7,2*	*-22,6*
Adverbes	69,4	44,6	-35,8	56,7	37,2	-34,4
Prépositions	120,6	141,3	+17,1	153,8	155,2	+0,9
Conjonctions	43,3	51,0	+17,7	55,1	51,3	-7,0
Conjonction coordination	*26,9*	*33,7*	*+25,4*	*33,2*	*34,8*	*+4,7*
Conjonction subordination	*16,5*	*17,3*	*+5,3*	*21,9*	*16,5*	*-24,8*
Mots étrangers	1,2	0,9	-20,4	1,7	2,1	+24,6
GN	608,1	726,2	+19,4	675,3	746,4	+10,5
GV	384,4	265,0	-31,1	322,1	250,0	-22,4

Rimbaud et Gautier partagent avec Baudelaire, la même tendance à privilégier le groupe nominal quand ils passent de la prose aux vers. Dans le détail, on remarque le même recul du passé et la forte montée du participe présent ainsi que le recul très important des pronoms dans les vers par rapport à la prose. Chez Rimbaud, ces mouvements sont de plus grande ampleur que chez les trois autres.

Enfin, chez Rimbaud et Gautier, comme chez Baudelaire, la poésie semble dominée par les relations interlocutives (je/tu/vous) et inclusive (nous).

Ces constats soulèvent au moins deux questions. Ces caractéristiques séparent-elles systématiquement la prose de la poésie quels que soient l'époque et l'auteur ? Existe-t-il un vocabulaire propre à la poésie différent de celui de la prose ?

Vers et prose dans la littérature française (seconde moitié du XIXe)

Nous proposons de répondre à ces questions en étudiant un plus grand nombre d'auteurs et d'œuvres contemporains de Baudelaire, Rimbaud et Verlaine[2].

De 1841-42 (premiers poèmes de Baudelaire) à 1896 (décès de Verlaine), il s'écoule un demi-siècle. Le début de la période correspond aux dernières publications de Chateaubriand (*La vie de Rancé*, 1844), aux premiers grands succès d'A. Dumas : *Le comte de Monte Cristo* paraît en feuilleton à partir de 1843. La fin de la période voit la première publication de M. Proust *Les plaisirs et les jours* en 1896.

Les deux corpus utilisés pour cette expérience sont constitués, pour le premier, de poésies françaises parues entre ces deux dates et, pour le second, d'œuvres de littérature générale en prose parues durant la même période - complètes ou sous forme d'extraits (liste en annexe 3). Ces textes ont été traités en suivant les méthodes présentées au début de cette communication.

Ces deux corpus ont été constitués au fil du temps pour des études particulières sur divers auteurs. Ils n'ont pas de prétention à l'exhaustivité ni à la "représentativité". Cependant, leur taille et leur diversité sont suffisantes pour pouvoir conclure que les caractéristiques mises au jour ne sont pas des accidents propres à l'un ou l'autre des auteurs mais des traits assez généraux (au moins pour la période considérée).

Le tableau 6 ci-dessous présente le résultat de cette expérience. Les mêmes tendances apparaissent clairement quand on passe de la prose aux vers : recul considérable du verbe et des autres constituants du groupe verbal et hausse du groupe nominal, spécialement de l'adjectif. L'augmentation des deux tiers du poids de l'adjectif suggère que la densité exceptionnelle de cette catégorie grammaticale semble être une caractéristique essentielle du "langage poétique".

Il resterait encore à déterminer le vocabulaire caractéristique de chacun de ces 2 ensembles. Il faudrait aussi analyser la longueur et la structure des phrases, etc. Les dimensions restreintes de cette communication ne permettent pas de présenter ces analyses qui ouvriront quelques aperçus nouveaux à la critique littéraire. Pour en suggérer l'intérêt, nous donnons ci-dessous deux exemples : le vocabulaire caractéristique et les principaux thèmes des *Fleurs du mal*.

[2] Nous avons déjà présenté une expérience de ce genre lors des journées de l'ERLA en 2003, pour caractériser le français oral (Labbé 2003).

Tableau 6. Comparaison du poids des catégories grammaticales dans la prose et les vers dans la seconde moitié du XIXe siècle

Catégories	A. Prose (‰)	B. Vers (‰)	B-A (%)
Verbes	160,8	125,2	-22,1
Formes fléchies	*109,7*	*92,0*	*-16,1*
Participes passés	*16,3*	*9,4*	*-42,0*
Participes présents	*8,5*	*7,8*	*-9,0*
Infinitifs	*26,3*	*16,0*	*-39,1*
Noms propres	21,9	18,3	-16,1
Noms communs	184,7	224,5	+21,6
Adjectifs	61,4	103,0	+67,9
Adj, participe passé	*13,1*	*16,4*	*+25,2*
Pronoms	135,7	89,1	-34,4
Pronoms personnels	*88,2*	*53,5*	*-39,3*
Pronoms démonstratifs	*8,2*	*6,1*	*-25,3*
Pronoms possessifs	*0,3*	*0,3*	*+5,4*
Pronoms indéfinis	*5,1*	*3,7*	*-27,5*
Pronoms relatifs	*26,0*	*21,8*	*-16,4*
Déterminants	165,1	186,4	+12,9
Articles	*116,6*	*131,6*	*+12,9*
Nombres	*8,8*	*4,7*	*-46,8*
Possessifs	*23,1*	*35,4*	*+53,2*
Démonstratifs	*7,6*	*7,4*	*-2,6*
Indéfinis	*9,0*	*7,3*	*-19,3*
Adverbes	67,3	48,3	-28,2
Prépositions	144,6	137,8	-4,7
Conjonctions	55,5	60,6	+9,1
Conjonctions coordination	*34,4*	*41,8*	*+21,5*
Conjonctions subordination	*21,1*	*18,8*	*-11,2*
Mots étrangers	1,1	1,3	+13,5
Groupe nominal	612,1	711,9	+16,3
Groupe verbal	385,0	281,3	-26,9

A la recherche du vocabulaire et des thèmes caractéristiques de la poésie

En quoi le vocabulaire et la thématique de la poésie versifiée se singularisent-ils par rapport à la prose ?

Le calcul suivant permet de répondre à cette question : un mot est *caractéristique* d'un texte, ou d'une partie d'un corpus, lorsque sa fréquence relative dans ce texte ou cette partie s'écarte *significativement* de celle observée dans l'ensemble du corpus (sur ce calcul : Labbé & Labbé 1997). Si fréquence relative dans le texte est supérieure à celle du corpus entier, cela signifie que ce mot est caractéristique, ou encore que l'auteur éprouve une attirance particulière pour lui ; dans le sens contraire, l'auteur éprouve une répugnance à employer ce terme (à propos des thèmes traités dans le texte considéré).

Ce calcul a été appliqué aux *Fleurs du mal* comparées à la littérature de la seconde moitié du XIXe (annexe 3). Les résultats sont présentés en annexe 4. Le lecteur familier de la poésie Baudelairienne ne sera pas surpris par la liste des substantifs et des adjectifs privilégiés dans ces poèmes. Voici la liste des substantifs les plus caractéristiques (on a moins de 1

chance sur 10.000 de se tromper en considérant qu'il y en a significativement plus dans les *Fleurs du mal* que chez la moyenne des littérateurs contemporains) :

coeur, oeil, ciel, âme, soleil, amour, beauté, nuit, soir, ange, esprit, corps, mer, parfum, mort, douleur, fleur, vin, fond, sang, désir, souvenir, plaisir, pitié, poète, sein, ténèbre, gouffre, race, enfer, horreur, baiser, pleur, remords, tombeau, misère, flamme, ennui, volupté, mort…

Une fois déterminé ce vocabulaire caractéristique, le logiciel relit le texte analysé – ici les *Fleurs du mal* - en recherchant les phrases les plus caractéristiques, c'est-à-dire celles qui contiennent le plus de vocables sur-employés dans cette œuvre et le moins de ceux qui sont sous-employés. A la rencontre d'un mot significativement sur-employé dans les *Fleurs du mal*, le score de la phrase est augmenté de 1 ; à l'inverse, ce score est diminué de 1 à la rencontre d'un mot significativement sous-employé dans ces textes. On obtient ainsi les "phrases canoniques" comme celles qui servent, dans un dictionnaire, à illustrer la définition du mot. On lira à la fin de l'annexe 4, les 20 phrases les plus caractéristiques des *Fleurs du mal* (le score donné entre parenthèses est le score absolu rapporté au nombre de mots de la phrase).

Pour rechercher les principaux thèmes traités, on relève les combinaisons de mots qui reviennent à l'identique à plusieurs reprises. La statistique textuelle a popularisé la notion de "segment répété" (Salem 1987) à laquelle la lemmatisation donne une portée supplémentaire en permettant l'élimination de tous les mots outils et le regroupement de toutes les flexions, notamment des verbes (Pibarot et Labbé, 1988). Par exemple, le segment le plus souvent répété dans les Fleurs du mal est *prend(re) pitié (de ma) longue misère* qui est le leitmotiv des *Litanies de Satan* (annexe 5). Tous les "syntagmes" énumérés dans ce tableau sont caractéristiques des *Fleurs du mal* et donnent un aperçu des images favorites de C. Baudelaire, du moins quand il écrivait des vers.

Conclusions

Le lecteur jugera sans doute que, dans ce qu'il vient de lire, beaucoup de choses lui étaient déjà connues. Même dans ce cas, l'intérêt des procédures automatisées n'est pas négligeable puisqu'elles font le travail mieux et plus vite qu'avec les procédures manuelles traditionnelles. Pourquoi continuer à relever à la main – avec le risque d'en laisser passer – des vocables ou des combinaisons de mots que l'ordinateur peut compter et restituer aisément, sans risque d'erreur ?

L'intérêt de la statistique appliquée au langage dépasse ce premier avantage évident. Par exemple, à notre connaissance, personne n'avait compté manuellement les verbes, les substantifs, les adjectifs, etc. de Baudelaire, Rimbaud et Verlaine, tout simplement parce que cette tache fastidieuse est hors de portée d'un observateur muni d'un crayon et d'une gomme.

Naturellement, la portée de l'opération dépasse le simple intérêt comptable. Il devient maintenant possible d'étudier les structures de phrases favorites d'un auteur, non plus avec quelques exemples choisis intuitivement, mais sur l'ensemble de ses textes…

Notre objectif était ailleurs. Il s'agissait de démontrer l'existence d'un genre particulier dans la littérature française - la poésie versifiée – puis, cette démonstration faite, de proposer quelques commencements de réponse à la question : quels sont les traits caractéristiques de ce genre particulier ? Notre recherche suggère une conclusion paradoxale. La principale caractéristique ne résiderait peut-être pas dans la rime ou dans le rythme du vers mais dans la densité particulière, dans ces poèmes, du groupe nominal et spécialement des adjectifs.

Par exemple, on a noté la propension de Baudelaire, Rimbaud et Gautier à sur-utiliser le participe présent dans leurs poésies. M. Cressot a signalé les caractéristiques particulières de ce participe qui serait la forme verbale la plus proche de l'adjectif : "Cette forme a pris au XIXe siècle un développement considérable, surtout à partir de Flaubert. Les écrivains qui attribuent aux choses une vie et une volonté secrètes, ont compris l'utilité de l'adjectif verbal pour leur expression dynamique du monde, et la possibilité d'atténuer, grâce à lui, la note trop éclatante des adjectifs en *–eur* et en *–teur*" (Cressot 1963, p. 150). Voilà effectivement un trait que partagent les poètes analysés.

Comment expliquer le poids considérable du groupe nominal dans la poésie versifiée ?

En 1950, P. Guiraud avait signalé que le nombre des substantifs et celui des verbes varient en sens inverse et que le substantif domine dans la "prose abstraite". En 2002, ici même, il a été montré comment, chez le même auteur, le passage de l'oral à l'écrit se traduit par une diminution très significative du poids du groupe verbal et une augmentation parallèle du groupe nominal (Labbé 2002). Autrement dit, l'expression spontanée privilégie le verbe, les pronoms, les adverbes. Le passage à l'écrit amène à remplacer un certain nombre de ces verbes par des substantifs, certains adverbes par des adjectifs, à réduire l'emploi du démonstratif, etc. A la suite de Guiraud, on peut donc penser que l'effort d'élaboration qu'implique l'écrit s'accompagne d'un mouvement d'abstraction au-delà de la perception ou de la visée immédiate de l'auteur. Nous savons maintenant que le passage de la prose à la versification entraîne un mouvement du même genre et de même sens, comme si l'auteur faisait un effort supplémentaire pour s'abstraire un peu plus du monde environnant. Si cette hypothèse est exacte, les contraintes de la versification (rimaison, longueur et rythme des vers) seraient alors des stimulants qui aident les poètes à réaliser cet effort d'abstraction – des manières de "passeurs" en quelque sorte - et non la cause de ces caractéristiques si particulières que semblent partager tous les poètes quand ils font des vers…

Références

Les versions électroniques des œuvres de Baudelaire, Rimbaud et Verlaine ont été déchargées sur divers sites notamment celui de la BNF (gallica) et de poesie.net. Les textes ont été corrigés en utilisant les éditions suivantes :

Compagnon Antoine (1993). *Charles Baudelaire. Les Fleurs du mal*. Paris : Le Seuil.

Le Dantec Y.-G & J. Borel (1962). *Paul Verlaine. Œuvres poétiques complètes*. Paris Gallimard, Pléiade.

Favre Y.-A. (1992). *Paul Verlaine. Œuvres poétiques complètes*. Paris : Robert Laffont, Bouquins.

Forestier L. (1998). *Arthur Rimbaud. Œuvres complètes*. Paris : Robert Laffont, Bouquins.

Pichois C. & Ziegler J. (1975). *Charles Baudelaire. Œuvres complètes*. Paris : Gallimard, Pléiade.

Statistique et traitement des données textuelles :

Benveniste E. (1956). "La nature des pronoms". Reproduit dans Benveniste 1966, p.251-265.

Benveniste E. (1958). "De la subjectivité dans le langage". Reproduit dans Benveniste 1966, p.258-266.

Benveniste E. (1970). "L'appareil formel de l'énonciation". *Langages*, 17, p. 12-18.

Benveniste E. (1966 & 1970). *Problèmes de linguistique générale*. Paris, Gallimard (rééd. 1980).

Benzecri J.-P. (1980). *L'analyse des données. 1. La taxinomie*. Paris, Dunod.

Cressot M. (1963). *Le style et ses techniques*. Paris : PUF (1ère édition : 1947).

Embleton S. (1986). *Statistics in Historical Linguistics*. Bochum, Brokmeyer.

Felsenstein J. (2004a). *Inferring Phylogenies*. Sunderland, Sinauer Ass.

Felsenstein J. (2004b). *Package of Programs for Inferring Phylogenies (PHYLIP).* Seattle : University of Washington.

Guiraud P. (1950). *Les caractères statistiques du vocabulaire*. Paris : PUF.

Guiraud P. (1960). *Problèmes et méthodes de la statistique linguistique*. Paris : PUF.

Holm H. J. (2007). "The New Arboretum of Indo-European "Trees". Can New Algorithms Reveal the Phylogeny and Even Prehistory of Indo-European ?". *Journal of Quantitative Linguistics*. 14-2, p. 167-214.

Jolivet R. (1982). *Descriptions quantifiées en syntaxe du français*. Genève-Paris : Slatkine-Champion.

Kerbrat-Orecchioni C. (1981). *L'énonciation de la subjectivité dans le langage*. Paris, A Colin.

Labbé C. & Labbé D. (1994). *Que mesure la spécificité du vocabulaire ?* Grenoble, CERAT. Repris dans : *Lexicometrica*, 3, 2001.

Labbé C. & Labbé D. (2001). "Inter-Textual Distance and Authorship Attribution Corneille and Molière". *Journal of Quantitative Linguistics*, 8-3, December 2001, p. 213-231.

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

Labbé C. & Labbé D. (2006). "A Tool for Literary Studies: Intertextual Distance and Tree Classification". *Literary and Linguistic Computing*. 21-3, 2006, p 311-326.

Labbé D. (1990). *Normes de saisie et de dépouillement des textes politiques, Cahier du CERAT n° 7*, Grenoble : CERAT-IEP.

Labbé D. (2002). "Le général de Gaulle en campagne". IIIe Journées de l'ERLA, *Aspects linguistiques du texte de propagande*, Brest, 15-16 novembre 2002. Reproduit dans : Banks D. (éd.). *Aspects linguistiques du texte de propagande*. Paris, L'Harmattan, 2005, p 213-233.

Labbé D. (2003). "Coordination et subordination en français oral". IVe journées de l'ERLA, *Coordination/subordination dans le texte de spécialité*. Brest 14-15 novembre 2003. Reproduit dans Banks D. (éd.). *La coordination et la subordination dans le texte de spécialité*. Paris, L'Harmattan, 2007, p. 161-182.

Labbé D. (2007). "Experiments on Authorship Attribution by Intertextual Distance in English". *Journal of Quantitative Linguistics*. April 2007, 14-1. p. 33-80.

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

Luong X. (1988). *Méthodes d'analyse arborée. Algorithmes, applications.* Thèse pour le doctorat ès sciences. Paris, Université de Paris V.

Luong X. (1994). "L'analyse arborée des données textuelles : mode d'emploi". *Travaux du cercle linguistique de Nice*. 16, p. 25-42.

Monière D. & Labbé D. (2006). "L'influence des plumes de l'ombre sur les discours des politiciens". In Condé C. & Viprey J.-M. *Actes des 8e Journées internationales d'Analyse des données textuelles*. Besançon, 19-21 avril 2006, II, p. 687-696.

Pibarot A. & Labbé D. (1998). "Les syntagmes répétés dans l'analyse des commentaires libres". in Mellet S. (ed*). 4e Journées d'analyse des données textuelles*. Nice, 1998, p. 507-516.

Salem A. (1987). *Pratique des segments répétés. Essai de statistique textuelle*. Paris, Klincksieck.

Sneath P. & Sokal R. (1973). *Numerical Taxonomy*. San Francisco, Freeman.

Ruhlman M. (2003), *Analyse arborée. Représentation par la méthode des groupements*. Grenoble, Polytech' – CERAT, août 2003.

Annexe I. Le corpus des œuvres de Baudelaire, Rimbaud et Verlaine.

Titre du fichier	Titre de l'oeuvre	Dates	Mots	Vocables
C. Baudelaire				
BaudelaireFleurs1	Les Fleurs du mal	1857	14 841	2 905
BaudelaireFleurs2	Les Fleurs du mal	1861-68	10 740	2 487
BaudelaireSpleen	Le Spleen de Paris	1862-64	11 307	2 537
BaudelaireParadis1	Les Paradis artificiels (Le poème du Haschisch)	1858	15 878	2 794
BaudelaireParadis2	Les Paradis artificiels (Un mangeur d'opium)	1860	15 972	2 868
BaudelaireParadis3	Les Paradis artificiels (Un mangeur d'opium)	1860	16 232	2 927
Total Baudelaire			**84 970**	**7 857**
A. Rimbaud				
RimbaudPoesie1	Poésies	1869-71	11 062	2 381
RimbaudPoesie2	Poésies	1871-72	10 640	2 658
RimbaudSaison	Une saison en enfer	1873	7 778	1 750
Total Rimbaud			**29 480**	**4 549**
P. Verlaine				
VerlaineSaturne	Poèmes saturniens	1866	9 048	2 306
VerlaineFetes	Fêtes galantes	1869	2 543	907
VerlaineChanson	La bonne chanson	1870	2 696	852
VerlaineRomances	Romances sans paroles	1874	2 960	922
VerlaineSagesse	Sagesse	1880	11 285	2 116
VerlaineJadis	Jadis et naguère	1884	17 441	3 218
VerlaineAmour	Amour	1888	14 448	2 726
VerlaineParallèle	Parallèlement	1889	7 696	1 996
Total Verlaine			**68 117**	**6 941**
Total			**182 567**	**10 277**

Annexe II. Distances intertextuelles Corpus Baudelaire, Rimbaud et Verlaine

	Baudel. Fleurs1	Baudel. Fleurs2	Baudel. Paradis 1	Baudel. Paradis 2	Baudel. Paradis 3	Baudel. Spleen	Rimbaud Poesie1	Rimbaud Poesie2	Rimbaud Saison	Verlaine Amour	Verlaine ChFeRo	Verlaine Parallèle	Verlaine Sagesse	Verlaine Saturne	Verlaine Jadis
BaudelaireFleurs1	-	0,254	0,399	0,399	0,385	0,339	0,341	0,347	0,384	0,325	0,328	0,342	0,314	0,324	0,334
BaudelaireFleurs2	0,254	-	0,407	0,397	0,383	0,341	0,327	0,331	0,382	0,340	0,345	0,355	0,340	0,318	0,340
BaudelaireParadis1	0,399	0,407	-	0,271	0,269	0,289	0,436	0,433	0,393	0,370	0,401	0,399	0,379	0,433	0,367
BaudelaireParadis2	0,399	0,397	0,271	-	0,243	0,299	0,424	0,429	0,390	0,367	0,400	0,391	0,388	0,423	0,361
BaudelaireParadis3	0,385	0,383	0,269	0,243	-	0,294	0,414	0,416	0,384	0,371	0,401	0,396	0,374	0,407	0,363
BaudelaireSpleen	0,339	0,341	0,289	0,299	0,294	-	0,386	0,386	0,354	0,333	0,358	0,352	0,332	0,385	0,330
RimbaudPoesie1	0,341	0,327	0,436	0,424	0,414	0,386	-	0,329	0,422	0,368	0,361	0,379	0,376	0,346	0,351
RimbaudPoesie2	0,347	0,331	0,433	0,429	0,416	0,386	0,329	-	0,359	0,355	0,361	0,368	0,341	0,334	0,360
RimbaudSaison	0,384	0,382	0,393	0,390	0,384	0,354	0,422	0,359	-	0,352	0,391	0,402	0,342	0,420	0,353
VerlaineAmour	0,325	0,340	0,370	0,367	0,371	0,333	0,368	0,355	0,352	-	0,287	0,284	0,243	0,348	0,265
VerlaineChFeRo	0,328	0,345	0,401	0,400	0,401	0,358	0,361	0,361	0,391	0,287	-	0,321	0,292	0,343	0,273
VerlaineParallèle	0,342	0,355	0,399	0,391	0,396	0,352	0,379	0,368	0,402	0,284	0,321	-	0,301	0,367	0,280
VerlaineSagesse	0,314	0,340	0,379	0,388	0,374	0,332	0,376	0,341	0,342	0,243	0,292	0,301	-	0,360	0,272
VerlaineSaturne	0,324	0,318	0,433	0,423	0,407	0,385	0,346	0,334	0,420	0,348	0,343	0,367	0,360	-	0,339
VerlaineJadis	0,334	0,340	0,367	0,361	0,363	0,330	0,351	0,360	0,353	0,265	0,273	0,280	0,272	0,339	-

Annexe III. Corpus comparatifs

1. Poésie en vers de la seconde moitié du XIXe siècle (251 466 mots ; extraits ou œuvres intégrales ; auteurs classés par ordre alphabétique) :

Théodore de Banville, *Dans la Fournaise*, Charles Baudelaire, *Les fleurs du mal* ; François Coppée, *Promenades et intérieurs*, *L'exilée* ; Théophile Gautier, *Emaux et Camées*, *Pièces diverses* ; Victor Hugo, *Contemplations* ; Charles Leconte de l'Isle, *Poèmes barbares* ; Stéphane Mallarmé, *Poésies* ; Rimbaud, *Poésies* ; Verlaine, *Poésies*.

2. Littérature générale en prose de la seconde moitié du XIXe siècle (1 951 592 mots ; extraits ou œuvres intégrales ; auteurs classés par ordre alphabétique) :

Jules Barbey d'Aurevilly, *Diaboliques* ; Charles Baudelaire, *Les paradis artificiels* ; François de Chateaubriand, *La vie de Rancé* ; François Coppée, *La bonne souffrance* ; Alexandre Dumas Père, *Monte Cristo*, *Balsamo*, *Vingt ans après* ; Alexandre Dumas Fils, *La Dame aux Camélias* ; Gustave Flaubert, *Madame Bovary*, *l'Education sentimentale*, *Salammbô*, *Bouvard et Pécuchet* ; Théophile Gautier, *Roman de la momie*, *Avatar* ; Edmond et Jules de Goncourt, *Germinie Lacerteux* ; V. Hugo, *Notre-Dame de Paris*, *Les Misérables* ; Joris-Karl Huysmans, *A rebours* ; Maupassant, *Une vie*, *Bel Ami*, *Mont Oriol*, *Un cœur simple*, *Pierre et Jean*, *Fort comme la mort* ; Gérard de Nerval, *Aurélia* ; Marcel Proust, *Les plaisirs et Les jours* ; George Sand, *La mare au diable*, *La petite Fadette* ; Jules Vernes, *Le tour de monde en 80 jours*, *De la terre à la Lune* ; Emile Zola, *Thérèse Raquin*, *Germinal*, *La bête humaine*.

Annexe IV. Le vocabulaire caractéristique des Fleurs du Mal

Vocables significativement sur-employés au seuil de 5%
(classement par catégorie grammaticale et indice décroissant)

Noms propres : Satan, Abel, Caïn

Verbes : dormir, chanter, contempler, rêver, enivrer, verser, traîner, rouler, plonger, remplir, danser, nager, haïr, endormir, caresser, noyer, ronger, planer, fleurir, pâmer, abreuver, bâtir, mêler, étaler, souvenir, tordre, déchirer, fuir, couler, jaillir, frémir, allumer, connaître, mordre, adorer, aimer, agiter, briller, chercher, rendre, voir

Substantifs : coeur, oeil, ciel, âme, soleil, amour, beauté, nuit, soir, ange, esprit, corps, mer, parfum, mort, douleur, fleur, vin, fond, sang, désir, souvenir, plaisir, pitié, poète, sein, ténèbre, gouffre, race, enfer, horreur, baiser, pleur, remords, tombeau, misère, dieu, flamme, ennui, volupté, mort, chat, gloire, poison, azur, secret, démon, monstre, haine, miroir, sommeil, rayon, douceur, paradis, métal, dent, clarté, aile, cerveau, charme, automne, cité, cadavre, reine, chant, jeunesse, squelette, splendeur, sanglot, jeu, vaisseau, univers, sorcière, astre, fantôme, destin, palais, plafond, péché, troupeau, serpent, penser, ennemi, humanité, prunelle, bijou, loisir, éternité, philtre, néant, muse, caveau, flambeau, climat, paresse, crâne, blasphème, ivrogne, ver, alcôve, boue, bourreau, encensoir, serment, spleen, vertige, flacon, majesté, appât, tambour, encens, voile, aurore, rêve, soulier, abîme, extase, regard, nature, faubourg, pavé, fosse, odeur, glaive, lecteur, ruisseau, spectre, or, fleuve, débris, globe, langueur, tableau, toile, front, musique, ardeur, lumière, travers, été, hiver, orgueil, lune, chanson, horizon, être, île, flot, peuple, air, fruit, bête, crime, santé, lueur, matin, espoir, vent, cri, fête, pied, larme, amant, mal, feu, lit, bord, chair, grâce, art, ombre

Adjectifs : plein, beau, vieux, noir, long, doux, profond, froid, éternel, lourd, charmant, étrange, pâle, divin, sombre, clair, immense, funèbre, vaste, amer, vivant, vert, ténébreux, semblable, pur, singulier, fatal, riche, inconnu, cruel, puissant, antique, maint, mystérieux, infernal, maigre, amoureux, vil, infâme, mystique, morne, maudit, parfumé, langoureux, chargé, vaincu, lumineux, savant, familier, avide, enfantin, fécond, mortel, sale, implacable, immonde, gonflé, frêle, brumeux, triomphant, plaintif, impur, moqueur, mélancolique, livide, subtil, stupide, digne, triste, sanglant, ridicule, noble, sinistre, superbe,

brisé, unique, parfait, lointain, nu, grand, aimable, obscur, affreux, fier, cher, rose, joyeux, infini, vide, lent, saint, terrible, fort, énorme, humain, pauvre

Pronoms : qui, tu, nous, dont, toi, nul, tout,

Adverbes : où, ainsi, parfois, hélas, jamais, jadis, loin, pourtant, lentement

Déterminants : le, mon, ton, leur, notre, votre, un, nul, ce, chaque

Conjonctions et prépositions : et, dans, comme, sans, ni, sous, vers, lorsque, parmi, ou, quand

Vocables significativement sous-employés au seuil de 5%

Verbes : songer, vouloir, falloir, pouvoir, entendre, mettre, appeler, apprendre, envoyer, lire, offrir, sortir, causer, sourire, reconnaître, asseoir, essayer, revoir, apporter, écrier, venir, sentir, déclarer, avoir, être, demander, parler, devoir, apercevoir, dire, reprendre, arriver, paraître, aller, arrêter, sembler, revenir, répondre, recevoir, entrer, passer, répéter, murmurer, demeurer, attendre, regarder, devenir, tenir, porter, comprendre, retourner, continuer, rencontrer, faire, disparaître, finir, rester, lever, remettre, croire, quitter, penser, présenter, écrire, oser, commencer, donner, dîner, trouver, manquer, laisser, écouter

Substantifs : voix, prêtre, rue, place, temps, peur, campagne, journal, affaire, partie, chaise, chapeau, femme, mademoiselle, feuille, peine, comtesse, garçon, manière, bout, monsieur, madame, heure, sorte, mot, figure, franc, porte, lettre, jour, moment, père, fois, suite, idée, coup, main, pensée, fille, argent, personne, mère, voiture, an, lendemain, mari, comte, ami, fait, salon, envie, parole, côté, chose, salle, fenêtre, maison, chambre, homme, ton, médecin, gens, docteur, reste, doute, besoin, pièce, question, cause, mois

Adjectifs : nouveau, sûr, gros, immobile, bon, petit, seul, jeune

Pronoms : il, lui, on, se, ils, celui, moi, lequel, quoi, quelqu'un, leur, le, y, en, ça, ce, vous, cela, autre, personne, rien, un, lui-même, je

Adverbes : même, trop, assez, surtout, seulement, debout, pas, puis, ne, peu, alors, oui, bien, beaucoup, maintenant, point, là, non, presque, d'abord, enfin, comment, tout, très, pourquoi, peut-être, si, fort, encore, vite, aussitôt, ensuite, plus, aussi, mieux, ailleurs

Déterminants : plusieurs, premier, son, deux, trois, quatre, autre, cent, même, quelque, dix, cinq, huit, aucun, vingt, certain, tout

Conjonctions et prépositions : près, parce que, cependant, jusque, avant, donc, tandis que, mais, chez, que, à, après, contre, depuis, entre, avec, en, devant, pendant, si, dès, pour, par, car

Phrases les plus caractéristiques en valeur relative
(les majuscules initiales des débuts de vers sont respectées)

Toi dont l'oeil clair connaît les profonds arsenaux Où dort enseveli le peuple des métaux, O Satan, prends pitié de ma longue misère ! (*Les litanies de Satan*, score 0.800)

O toi qui de la mort, ta vieille et forte amante, Engendras l'espérance, - une folle charmante ! (*Les litanies de Satan*, score 0.706)

De ce ciel bizarre et livide, Tourmenté comme ton destin, Quels pensers dans ton âme vide Descendent ? (*Horreur sympathique*, score 0,706)

Le ciel est triste et beau comme un grand reposoir (*Harmonie du soir*, score 0,700).

Nous imitons, horreur ! la toupie et la boule Dans leur valse et leurs bonds ; même dans nos sommeils La curiosité nous tourmente et nous roule, Comme un Ange cruel qui fouette des soleils (*Le voyage*, score 0,676).

De l'antique Vénus le superbe fantôme Au-dessus de tes mers plane comme un arôme, Et charge les esprits d'amour et de langueur (*Un voyage à Cythère*, score 0,667).

Sur ta chair le parfum rôde Comme autour d'un encensoir ; Tu charmes comme le soir, Nymphe ténébreuse et chaude (*Chanson d'après-midi*, score 0,650).

La gloire du soleil sur la mer violette, La gloire des cités dans le soleil couchant, Allumaient dans nos coeurs une ardeur inquiète De plonger dans un ciel au reflet alléchant (*Le voyage*, score 0,647).

Ainsi l'amant sur un corps adoré Du souvenir cueille la fleur exquise (*Le parfum*, score 0,643).

Toi qui poses ta marque, ô complice subtil, sur le front du Crésus impitoyable et vil, o Satan, prends pitié de ma longue misère ! (*Les litanies de Satan*, score 0,640)

Le violon frémit comme un coeur qu'on afflige, Un coeur tendre, qui hait le néant vaste et noir ! (*Harmonie du soir*, score 0,632)

Mon désir gonflé d'espérance Sur tes pleurs salés nagera Comme un vaisseau qui prend le large, Et dans mon coeur qu'ils soûleront Tes chers sanglots retentiront Comme un tambour qui bat la charge ! (*L'Héautontimorouménos*, score 0,629)

Toi qui mets dans les yeux et dans le coeur des filles le culte de la plaie et l'amour des guenilles, o Satan, prends pitié de ma longue misère ! (*Les litanies de Satan*, score 625)

Dans quel philtre, dans quel vin, dans quelle tisane, Noierons nous ce vieil ennemi, Destructeur et gourmand comme la courtisane, Patient comme la fourmi ? (*L'irréparable*, score 0,625)

Liqueur Qui me ronge, ô la vie et la mort de mon coeur ! (*Le Flacon*, score 0,615)

Et planait librement à l'entour des cordages ; Le navire roulait sous un ciel sans nuages, Comme un ange enivré d'un soleil radieux (*Un voyage à Cythère*, score 0,609).

Tu contiens dans ton oeil le couchant et l'aurore ; Tu répands des parfums comme un soir orageux ; Tes baisers sont un philtre et ta bouche une amphore Qui font le héros lâche et l'enfant courageux (*Hymne à la beauté*, score 0,605).

Tu ressembles parfois à ces beaux horizons Qu'allument les soleils des brumeuses saisons… (*Ciel brouillé*, score 0,600)

Comme de longs échos qui de loin se confondent Dans une ténébreuse et profonde unité, Vaste comme la nuit et comme la clarté, Les parfums, les couleurs et les sons se répondent (*Correspondances*, score 0,594).

Un coeur tendre, qui hait le néant vaste et noir, Du passé lumineux recueille tout vestige ! (*Harmonie du soir*, score 0,588)

Annexe V

Les syntagmes répétés les plus fréquents dans les *Fleurs du mal* (avec leur fréquence absolue)

prendre pitié (16) ; pitié long (15) ; long misère (15) ; être beau (10) ; ange plein (10) ; race Caïn (8) ; coeur être (6) ; être doux (6) ; avoir air (4) ; coeur plein (4) ; être plein (4) ; gouffre amer (4) ; grand œil (4) ; oeil feu (4) ; soleil couchant (4) ; tour tour (4) ; Victor Hugo (3) ; air être (3) ; calme volupté (3) ; ciel brouillé (3) ; ciel être (3) ; coeur gonflé (3), coeur mortel (3) ; dormir sommeil (3) ; doux secret (3) ; être affreux (3) ; être âme (3) ; être charmant (3) ; être heure (3) ; être œil (3) ; être ordre (3) ; être rêve (3) ; être semblable (3) ; être triste (3) ; faire rêver (3) ; grand cœur (3) ; mettre pied (3) ; monde être (3) ; nuit noir (3) ; oeil être (3) ; oeil plein (3) ; ordre beauté (3) ; suivre rythme (3) ; vieux faubourg (3)