Existe-t-il un genre épistolaire ? Hugo, Flaubert et Maupassant

Cyril Labbé, Dominique Labbé

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

Nouvelles Journées de l'ERLA n° 10

Aspects linguistiques du texte épistolaire

(Brest 20-21 novembre 2009)

Existe-t-il un genre épistolaire ?

Hugo, Flaubert et Maupassant

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 épistolaire particulier ? Qu'est-ce qui différencie ce genre du reste de la littérature? On répond à ces deux questions en utilisant les correspondances de Victor Hugo, de Flaubert et Maupassant comparées au reste de leurs œuvres (romans, théâtre, poésie). Les lettres utilisent plus le vocabulaire usuel. Elles sont centrées sur la relation je-vous. Elles sont ancrées dans le temps, dans l'espace géographique et social. Elles privilégient le groupe verbal (verbes, pronoms et adverbes) au détriment du groupe nominal. Elles sont un substitut à la conversation.

Mots clefs : France – Littérature – XIXe siècle – correspondance – Hugo – Flaubert - Maupassant - Statistique lexicale.

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 correspondance ? C'est seulement si l'on peut répondre affirmativement à cette première question qu'il devient pertinent de rechercher quels sont les aspects linguistiques particuliers de ce « genre épistolaire ».

La statistique permet d'envisager de manière originale cette question "préjudicielle" trop peu explorée jusqu'à maintenant. A notre connaissance, il y a une seule recherche, assez ancien, en statistique lexicale sur la correspondance commerciale (Lyne 1985).

Il faudrait disposer - au moins pour quelques individus - d'une vaste collection de lettres et d'un étalon de comparaison : des textes appartenant à d'autres genres émis par les mêmes personnes, si possible à la même époque. Nous avons choisi d'illustrer la méthode en utilisant trois écrivains français : Victor Hugo (1802-1885), Gustave Flaubert (1821-1880) et Guy de Maupassant (1850-1893).

Après avoir présenté le corpus et les conventions de dépouillement, on comparera leurs lettres avec le reste de leurs œuvres.

I. Corpus et conventions de dépouillement

Un corpus est un ensemble de textes rassemblés en vue d'une étude particulière. Ces textes subissent un certain nombre de traitements préalables.

Les corpus

Il reste un grand nombre de lettres d'Hugo. Leur édition a été faite avec soin et elle est accessible en ligne, ce qui épargne une nouvelle saisie. Surtout, Hugo s'est essayé à tous les genres : romans, poésie, théâtre…. Il est donc possible de comparer sa correspondance à la plupart des autres genres. Nous avons sélectionné les lettres des années 1849 à 1870 afin de couvrir la totalité de son exil (décembre 1852 à septembre 1870). Les années 1849-1852 doivent permettre de répondre à la question de savoir si l'exil a pu modifier les pratiques épistolaires d'Hugo. On a regroupé les lettres par années civiles en scindant trois années (1852, 1867, 1869) pour lesquelles il y a une quantité trop importante de lettres par rapport aux autres (voir tableau en annexe). Au total, ce corpus comporte 1 126 lettres, pour une longueur totale de 292 173 mots. La longueur moyenne des lettres (259 mots), varie assez peu selon les années.

A titre d'éléments de comparaison, nous avons utilisé deux romans - la totalité de *Notre-Dame de Paris* (1831) et le premier tome des *Misérables* (1862) - un recueil de poésie – les *Contemplations* (dont la composition s'étend entre 1830 et 1852) - et deux pièces de théâtre : *Hernani* (1830) et *Ruy Blas* (1838). Pour les romans et la poésie, le découpage suit les césures par « livres » effectuées par Hugo lui-même. Deux œuvres ont été écrites partiellement (*Contemplations*) ou totalement en exil (les *Misérables*) et sont donc contemporaines de la correspondance. Au total, ce corpus de comparaison comporte 441 881 mots. Ces 734 046 mots ne sont qu'une faible partie de l'œuvre d'Hugo, mais étant répartis entre les différents genres possibles (poésie, théâtre, roman et correspondance), ils sont suffisant pour examiner la question posée : existe-t-il, chez Hugo, un genre "épistolaire" différent des autres genres littéraires.

Nous avons également retenu la correspondance entre Flaubert et Maupassant. Entre 1875 et la mort de Flaubert (1880), les deux hommes ont échangé une correspondance régulière. Maupassant était fils d'une amie d'enfance de Flaubert et celui-ci s'était pris d'un

sentiment quasi-paternel pour le jeune homme qu'il a guidé lors de ses débuts littéraires. Ces lettres se répondent les unes aux autres, commentent les mêmes événements et portent sur les mêmes thèmes. Nous les avons comparées à leurs romans et nouvelles (tableaux 3 et 4 en annexe). Ce corpus – qui représente un total de 1 241 664 mots - a déjà fait l'objet d'une analyse présentée dans Labbé et Labbé (2007b). Il servira d'élément de comparaison pour juger de la singularité de la correspondance des deux hommes.

Conventions de dépouillement

Ces conventions sont exposées dans Labbé (1990). En synthèse, elles consistent à :

1. Isoler le « para-texte » du texte.

En l'occurrence, la correspondance a été découpée en années selon le découpage choisi par l'éditeur. Chaque lettre a été dotée d'un entête avec la date et le destinataire. La signature est également isolée du texte, car la conserver aurait eu comme conséquence inévitable que les deux noms propres les plus fréquents seraient « Victor » et « Hugo »…

2. Homogénéiser les graphies :

Les graphies multiples ont été standardisées (événement et évènement ; puis et peux, etc.) et l'orthographe a été soigneusement corrigée. En effet, les textes que l'on trouve en ligne sont issus de scanners et comportent des imperfections surtout sensibles dans la ponctuation, les majuscules, etc. Par exemple, la lettre "l" (minuscule) est parfois confondue avec le "i" majuscule ou avec le chiffre 1…

De plus, les lettres contiennent de nombreuses abréviations que Hugo, Flaubert ou Maupassant ne se permettent pas dans leurs livres. Par exemple, lorsqu'il date ses lettres, Hugo utilise indifféremment *dim* ou *dimanche*, *8bre*, *9bre*, *10bre* ou *octobre*, *novembre* et *décembre*. Les unités monétaires sont « francs » ou *f.*, *fr*, *frs* ; sa résidence est tantôt *Hauteville-House* ou *Hauteville house, H-H*, etc… Hugo commence parfois ses phrases par une minuscule (spécialement quand la première lettre est accentuée). Il met une minuscule à certains noms propres (*édouard*, *émile*…). D'autres noms propres sont en abrégé, spécialement ses enfants (*Adèle* est le plus souvent *A* ou *Ad*). On trouve indifféremment : *Monsieur* (M, M.), *Messieurs* (MM ou Mm), *Madame* (Mme), *Mademoiselle* (Mlle). Etc.

En définitive, on se doute que la caractéristique la plus évidente du « style » épistolaire est un certain relâchement dans la graphie par rapport à l'imprimé. Le lecteur verra « dim » et « dimanche », comme un même mot, mais l'ordinateur n'a pas cette intelligence. Si on ne lui dit pas qu'il doit rattacher ces deux formes sous la même entrée (« dimanche, nom masculin »), il les comptera pour deux mots différents. Et le premier résultat des traitements sera : « par rapport aux autres écrits d'Hugo, Flaubert et Maupassant, leurs lettres sont caractérisées par le très grand nombre de graphies hétérodoxes ».

La règle est donc « un mot = une seule graphie ».

3. Lemmatiser les textes

La lemmatisation attache une étiquette à chaque mot du texte. Par exemple, « pouvoir, substantif masculin » ou « pouvoir, verbe à l'infinitif » ; « est, nom masculin » ou verbe « être » à la troisième personne de l'indicatif présent. Ces problèmes d'homographie 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, réduction des formes multiples à une graphie standard… - puis le **vocable** – c'est-à-dire l'entrée de dictionnaire et la catégorie grammaticale. 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.

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. 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” ou les “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** (les tableaux en annexe donnent ces caractéristiques pour les différents textes utilisés dans cette expérience) – 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 constitue un **mot** du texte.

II. Existe-t-il un genre «épistolaire » ?

La question posée revient à se demander si les lettres – ainsi standardisées et lemmatisées – partagent des traits particuliers qui les distinguent des autres textes des mêmes auteurs. Un calcul de distance et des opérations de classification apportent la réponse.

La distance intertextuelle

Il s'agit de mesurer la distance entre les textes comme on le fait entre des objets. L'unité de mesure est ici le vocable associé à sa fréquence. La distance entre deux textes – nommée pour cela : "distance intertextuelle" – est le nombre de vocables différents que contiennent ces textes. Pour le détail des calculs : Labbé & Labbé 2001 ; Labbé & Labbé 2003 ; Labbé & Labbé 2006.

Cette mesure comporte des 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 1.000 mots et, jusqu'à 4.000 mots environ, certaines distorsions sont possibles. Ceci interdit d'étudier les lettres ou les poèmes un à un et oblige à constituer des ensembles sur des bases raisonnées (ici les césures adoptées par l’auteur ou l’année civile pour la correspondance). En second lieu, les textes ne seront pas de longueurs trop différentes (le rapport entre le plus grand et le plus petit doit être inférieur à 1/10), ce qui oblige à découper les textes trop longs pour les rapprocher de la taille des autres.

La distance intertextuelle enregistre l'influence de 4 facteurs : le genre, l'auteur, le thème 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. De nombreuses expériences menées depuis une douzaine d'années ont abouti à l'étalonnage d'une échelle des distances (Labbé & Labbé 2001), qui comprend quatre zones principales.

Pour deux textes dont la longueur tourne autour de 10.000 mots – comme c’est le cas ici -, on peut distinguer quatre zones principales dans le continuum des distances

1. Une valeur inférieure à 2.000 mots différents pour 10.000 signifie que les deux textes partagent plus des 4/5e de leur surface. Cette situation assez rare se rencontre quand les deux textes sont écrits par un même auteur dans un même genre, sur un même thème et à la même époque. Pour illustrer cette idée, le tableau 1 donne les 10 distances les plus faibles enregistrées dans le corpus Hugo. Ces valeurs sont obtenues dans un même genre (surtout la correspondance) et à des dates voisines. Chez Hugo la correspondance semble la catégorie la plus homogène.

Tableau 1. Les distances les plus faibles au sein du corpus Hugo
(nombre de mots différents pour 10 000)

Rang	Texte 1	Texte 2	Distance
1	Correspondance 1852 A	Correspondance 1852 B	1 846
2	Correspondance 1867 B	Correspondance1868	1 852
3	Correspondance 1865	Correspondance 1866	1 857
4	Correspondance 1866	Correspondance 1867 A	1 879
5	Correspondance 1855	Correspondance 1862	1 890
6	Correspondance 1862	Correspondance1866	1 972
7	Correspondance 1866	Correspondance 1867 B	1 980
8	Correspondance 1867 B	Correspondance 1869 B	1 981
9	Contemplations 3	Contemplations 6	1 981
10	Correspondance 1868	Correspondance 1869B	1 987

2. Les valeurs comprises entre 2 000 et 2 500 signalent un auteur unique travaillant dans un même genre - mais les textes ont été écrits à des époques un peu plus éloignées ou ils portent sur des thèmes différents. Par exemple, les deux premiers livres des *Misérables* ont entre eux 2 371 mots différents (pour 10.000) alors que les livres 2 et 6 ne sont séparés que par une distance de 2 190 (les distances de ce deuxième livre avec les quatrième et cinquième livres sont également très faibles). Cela suggère que la rédaction n'a pas suivi le plan du livre ou que le deuxième livre a subi des révisions assez lourdes après la rédaction des derniers livres du premier tome.

3. Les valeurs comprises entre 2 500 et 3 500 peuvent concerner les textes d'un même auteur. Si les deux textes sont écrits dans un même genre, il y a un changement d'époque et de thèmes. Si le genre, le thème et l'époque sont identiques, il y a d'autant plus de chances d'être en présence de deux auteurs différents que les distances sont plus élevées. A titre d'exemple, le tableau 2 donne les distances observées pour les lettres échangées entre Flaubert et Maupassant.

Tableau 2. Distances entre les lettres de Flaubert (GF) et de Maupassant(GM)

	GF à GM (1874-78)	GF à GM (1879-80)	GM à GF (1875-78)	GM à GF (1879-80)
GF à GM (1874-78)	0	**2 534**	2 886	2 716
GF à GM (1879-80)	**2 534**	0	2 972	*2 659*
GM à GF (1875-78)	2 886	2 972	0	**2 322**
GM à GF (1879-80)	2 716	2 659	**2 322**	0

Les distances pour un même auteur (en gras sur le tableau) – Flaubert : 2 534 et Maupassant 2 322 – sont inférieures aux distances entre auteurs. Etant donné que thèmes,

époques et genre sont semblables, cette différence ne peut être imputée qu'à l'auteur. On vérifie également que, lorsque des textes sont produits par deux auteurs différents travaillant à la même époque, sur un même thème et dans un même genre, ces textes sont normalement séparés par une distance supérieure à 2 500 mots.

4. Au dessus de 3 500, pour un auteur unique, la différence de genre est certaine et plus on s'élève au-dessus de ce seuil, plus il est probable que des différences de thèmes et de dates de composition s'ajoutent à la différence de genre. Chez Hugo, Flaubert et Maupassant, la plupart des distances les plus fortes séparent les lettres et d'autres écrits, spécialement les romans. A titre d'exemple, le tableau 3 donne les 10 distances les plus élevées enregistrées dans le corpus Hugo.

Tableau 3. Les distances les plus élevées dans le corpus des œuvres d'Hugo.

Rang	Texte 1	Texte 2	Distance
1026	Notre-Dame 3	Correspondance 1867A	4 996
1027	Notre-Dame 3	Ruy Blas	5 072
1028	Notre-Dame 3	Correspondance 1870	5 073
1029	Notre-Dame 3	Correspondance 1869A	5 084
1030	Notre-Dame 3	Correspondance 1867B	5 125
1031	Notre-Dame 3	Correspondance 1852B	5 133
1032	Notre-Dame 3	Correspondance 1852A	5 151
1033	Notre-Dame 3	Correspondance 1869B	5 196
1034	Notre-Dame 3	Correspondance 1868	5 246
1035	Notre-Dame 3	Hernani	5 385

Ce dernier tableau révèle surtout la position singulière du troisième livre de Notre-Dame de Paris. Ce livre ne comporte que deux longs chapitres sur l'histoire, l'urbanisme et l'architecture. Le premier est consacré aux vicissitudes de la cathédrale au cours du temps, le second à une vue de Paris « à vol d'oiseau » depuis les tours de la cathédrale. On a souvent dit que, pour ces deux chapitres, Hugo s'est « inspiré » de plusieurs ouvrages d'histoire et d'antiquité, spécialement celui de Henri Sauval (1724). Les fortes distances avec le reste de l'œuvre signalent sinon un plagiat, du moins l'existence d'un « corps étranger » dans l'oeuvre. Pour conclure avec certitude, il faudrait disposer, en fichier électronique, des textes de Sauval pour mesurer les distances entre les passages concernés et ce livre 3.

Les tableaux complets des distances sont impossibles à reproduire et à analyser (celui du corpus Hugo comporte 1 378 cases différentes non nulles). Les regroupements sont donc une nécessité. Une première démarche consiste à tenter des regroupements *a priori* selon les hypothèses de départ. Le tableau 4 ci-dessous donne, chez Hugo, les distances moyennes au sein de chaque genre et entre ces genres.

La première case du tableau indique que la distance moyenne entre les 6 livres des *Contemplations* est égale à 2 540, etc. La diagonale indique la distance « intra-genre ». La correspondance présente l'homogénéité la plus forte (seule distance intra-groupe inférieure à 2 500) malgré le fait que sa rédaction s'étale sur plus de 20 ans. L'hétérogénéité la plus forte s'observe entre *Hernani* (1830) et *Ruy Blas* (1838), alors que ces deux pièces – en vers - ont été composées à moins de 10 ans d'intervalle, ce qui montre une évolution importante entre ces deux dates (et la rupture que *Ruy Blas* apporte dans l'œuvre d'Hugo).

Tableau 4. Distances moyennes entre les textes d'Hugo classés par genre.

	Poésie	Romans	Théâtre	Correspondance
Poésie	2 540	3 875	3 907	4 355
Roman	3 875	3 001	4 010	4 060
Théâtre	3 907	4 010	3 099	3 521
Correspondance	4 355	4 060	3 521	2 435

Les autres cases indiquent les distances moyennes inter-genres. Pour s'en tenir à la correspondance (dernière colonne ou dernière ligne) : elle est très éloignée de la poésie (distance moyenne : 4 355), un peu moins du roman (moyenne : 4 060) et un peu plus proche du théâtre (3 521). A priori, ce dernier résultat peut surprendre puisque *Ruy Blas* et *Hernani* sont toutes deux en vers et composées bien avant l'exil d'Hugo alors que la correspondance est en prose...

Les textes appartenant à un même genre sont donc plus proches entre eux et plus éloignés des ceux écrits dans les autres.

De plus, en dehors de la diagonale du tableau, toutes les distances moyennes sont supérieures à 3 500, ce qui, en fonction de l'échelle présentée ci-dessus, conduit à supposer l'existence de quatre genres distincts. Mais s'agit-il d'une loi générale ? Ou simplement d'une tendance qui admettrait des exceptions ? Le recours à la classification permet de répondre à ces questions.

La classification

Il s'agit de rechercher les meilleurs groupements possibles sans faire intervenir la subjectivité du chercheur. 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 2004) ou en linguistique historique (Embleton 1986, Holm 2007).

La classification arborée repose sur la propriété suivante : si tous les individus étudiés sont séparés par de véritables distances, il existe un "arbre" qui représente exactement les positions respectives de ces individus les uns par rapport aux autres (Luong 1988). La construction d'un arbre parfait exigerait 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 celui de Luong[1]. Appliqué aux 1 378 distances différentes du corpus Hugo, cet algorithme a tracé l'arbre ci-dessous (figure 1).

Dans cet arbre, les points terminaux (les titres des œuvres) sont les **feuilles** situées sur des **branches** reliées au **tronc** central par des **nœuds** qui figurent les groupements successifs. La distance entre deux textes est figurée par le **chemin** unissant les feuilles correspondantes et la **longueur** de ce chemin est proportionnelle à la **distance** originelle. Pour "lire" cet arbre, il ne faut pas aller d'un point à un autre "à vol d'oiseau", mais suivre les chemins reliant ces points.

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

Figure 1. Classification arborée sur les œuvres d'Hugo

Deux blocs principaux s'opposent : la correspondance (en haut) et le roman (en bas).

En bas, les extraits des *Misérables* sont clairement séparés de ceux de *Notre-Dame de Paris*. A la différence de thèmes – le moyen-âge pour *Notre-Dame* ; l'époque contemporaines pour Les *Misérables* – s'ajoutent les dates éloignées de création. Pratiquement trente ans séparent l'écriture de ces deux romans. Le graphe confirme aussi la position singulière du troisième livre de *Notre-Dame de Paris* déjà signalée.

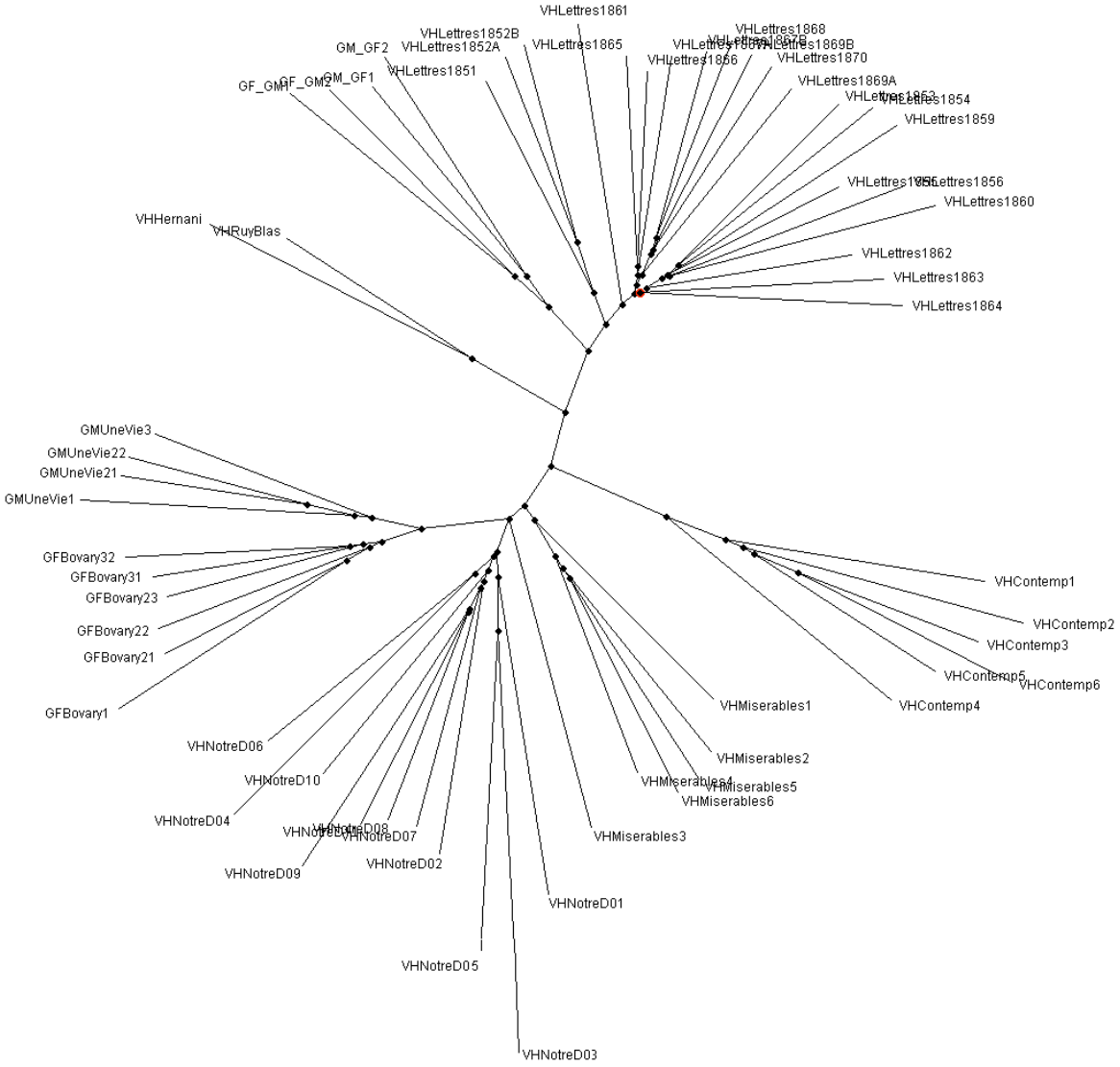
En haut, les lettres sont toutes groupées sans aucune exception. Les branches qui les relient entre elles et au tronc, sont les plus courtes. A gauche, les lettres des années 1851 et 1852 (avant l'exil) sont à part et semblent un peu plus proches de l'écrit. Puis un nœud regroupe la plupart des années suivantes (1853, 54, 55, 56, 59, 60), enfin les autres années sont les plus éloignées des autres genres, avec quelques sous-groupes là encore à dominante chronologique. Cela fournit une réponse à l'une des questions posées au début de cette communication : l'exil accentue bien le caractère singulier de la correspondance.

Deux autres groupes viennent se greffer sur le tronc qui relie les romans à la correspondance. Les *Contemplations*, c'est-à-dire la poésie en vers, sont proches des romans. Les différentes feuilles sont relativement écartées les unes des autres ce qui confirme les indications d'Hugo concernant les dates de création différentes des poèmes rassemblés dans ce recueil. A l'opposé, les deux pièces de théâtre sont groupées et clairement détachées du rest des œuvres imprimés. Le nœud qui les relie au tronc central se situe pratiquement à mi-chemin entre la correspondance et les romans.

Une conclusion s'impose : chez Hugo, il y a bien quatre genres distincts : la prose romanesque, les vers poétiques, le théâtre en vers et la correspondance.

L'expérience a été rééditée en ajoutant au corpus Hugo (VH) les lettres échangées entre Flaubert (GF) et Maupassant (GM) ainsi que les deux romans les plus proches : *Madame Bovary* et *Une vie* (voir Labbé 2007b). Ces textes étant découpés suivant le même principe que les romans d'Hugo. La figure 2 ci-dessous présente les résultats de cette expérience.

Figure 2. Classification arborée de la correspondance d'Hugo (VH), Flaubert (GF) et Maupassant (GM) avec le reste de leurs œuvres



Par parenthèse, cette figure montre la principale difficulté de l'expérience : un graphique contenant plus d'une soixantaine d'étiquettes est difficile à déchiffrer (c'est la

raison pour laquelle toutes les œuvres de Flaubert et Maupassant n'ont pu être introduites). Cette figure montre également la robustesse de la méthode : la classification des œuvres d'Hugo réapparaît semblable à celle de la figure 1. Les textes de Flaubert et de Maupassant s'y intègrent de la manière suivante.

En bas à gauche, *Madame Bovary* (Flaubert) et *Une vie* (Maupassant) forment deux ensembles distincts, mais proches, qui se rattachent aux romans d'Hugo.

En haut à gauche, les 4 fichiers de la correspondance sont groupés ensemble. Là encore, les deux auteurs sont correctement discriminés. Ces 4 branches sont rattachées au tronc qui relie le théâtre et la correspondance d'Hugo. Autrement, dit les échanges entre Flaubert et Maupassant étaient, en moyenne, un peu plus littéraires que les lettres d'Hugo mais ils appartiennent au même ensemble.

Il existe donc un genre épistolaire. Ce genre est nettement à part et il s'oppose à tous les autres. Cependant, il est plus proche du théâtre en vers que de la poésie et du roman. Attention : ces conclusions ne sont pas évidentes. Jusqu'à maintenant deux oppositions principales partageaient les corpus : l'oral et l'écrit (Labbé 2003), la prose et les vers (Labbé & Labbé 2007a). En fonction de ces oppositions, on s'attendrait à ce que le théâtre et la poésie soient regroupés - puisque tous deux écrits en vers - et que le roman soit classé avec la correspondance puisqu'ils sont tous deux écrits en prose.

Il reste maintenant à donner un contenu lexical à ce genre épistolaire et à comprendre les raisons de sa proximité avec le théâtre.

III Le vocabulaire du genre épistolaire

On isole la correspondance et on regroupe le reste (« autres » dans les tableaux ci-dessous). Les vocables sont classés par catégories grammaticales et par fréquence, ce qui permet de voir quel est le vocabulaire caractéristique de la correspondance.

Comparaison des vocables les plus fréquents

Les tableaux 5 à 9 ci-dessous présentent les fréquences comparées des vocables les plus utilisés dans la correspondance et les autres œuvres (les noms propres sont discutés plus bas). Pour neutraliser la différence de longueur entre les deux sous-corpus, les fréquences absolues sont converties en fréquences relatives exprimées en pour mille mots (‰). Les vocables en italiques se trouvent dans les deux parties du tableau et peuvent être considérés comme « communs » sous réserve des différences de fréquence qui sont discutées plus bas.

Dans le tableau 5, en dehors des verbes « écrire » et « envoyer » qui, pour des raisons évidentes, sont propres à la correspondance, on trouve les mêmes verbes usuels des deux côtés du tableau. Il faut cependant se souvenir que, dans tout texte écrit en français, les verbes *être*, *avoir* et *faire* occupent les trois premières places et dans cet ordre. Ce n'est que dans la suite du classement qu'apparaissent les différences entre auteurs et entre genres. Ainsi, chez Hugo, la présence du verbe « dire » en quatrième position mérite d'être soulignée, particulièrement dans la correspondance, puisque justement, on écrit aux personnes à qui on ne peut parler. En fait, la « conversation » occupe une grande place dans les œuvres d'Hugo comme cela sera souligné plus bas.

Tableau 5. Les verbes les plus employés

Autres (sous-corpus A)				Correspondance (sous-corpus B)		
Rang	Verbe	F absolue	F relative(‰)	Verbe	F absolue	F relative(‰)
1	*être*	10 852	25.6	*être*	8 816	30.2
2	*avoir*	6 890	16.2	*avoir*	6 131	21.0
3	*faire*	2 531	6.0	*faire*	2 004	6.9
4	*dire*	2 402	5.7	*dire*	1 376	4.7
5	*voir*	1634	3.9	écrire	891	3.1
6	*aller*	1224	2.9	*pouvoir*	859	2.9
7	venir	878	2.1	*vouloir*	808	2.8
8	*pouvoir*	806	1.9	envoyer	789	2.7
9	savoir	744	1.8	*voir*	703	2.4
10	*vouloir*	696	1.6	*aller*	689	2.4
Total			67.5			79.0

Une question vient naturellement à l'esprit : est-ce que les différences entre les deux parties des tableaux peuvent être considérées comme significatives ? Cette question revient à se demander si B peut être considéré comme un échantillon aléatoire tiré dans A. Il y a 292 173 mots dans B (N_b). Pour la première ligne, avec un risque d'erreur $\alpha = 5\%$, peut-on considérer que la densité des verbes dans B (PV_b = 30,2 ‰) ne diffère pas significativement de celle observée dans A (PV_a = 25,6 ‰) ? En utilisant l'approximation normale, il faudrait que PV_a soit compris dans l'intervalle suivant :

$$PV_b - 1{,}96\sqrt{\frac{PV_b * (1000 - PV_b)}{N_b}} \leq PV_a \leq PV_b + 1{,}96\sqrt{\frac{PV_b * (1000 - PV_b)}{N_b}}$$

Ce qui donne un intervalle {29,6 ‰ – 30,8 ‰} nettement supérieur à la fréquence du verbe *être* dans le reste de l'œuvre (colonne A : 25,6 ‰). On peut donc considérer, avec une chance infinitésimale de se tromper que Hugo sur-utilise le verbe *être* dans sa correspondance par rapport à ce qu'il fait dans ses autres écrits. Le même calcul appliqué aux autres lignes du tableau 5 donne les résultats suivants (tableau 8). Le signe + indique un sur-emploi caractéristique et le signe – un sous-emploi également caractéristique.

Tableau 6. Ecarts entre les fréquences des verbes usuels dans la correspondance comparée aux autres textes du corpus Hugo

Verbe	Ecart
être	+
avoir	+
faire	+
dire	-
écrire	+
pouvoir	+
vouloir	+
envoyer	+
voir	-
aller	-

Aucun des écarts de fréquence ne peut être dû au hasard. Autrement dit, le genre imprime sa marque même sur les mots les plus usuels et apparemment les plus banaux. De plus, les 7 sur-emplois sont plus nombreux que les 3 sous-emplois : la majorité des fréquences

sont plus élevées dans la correspondance (partie droite du tableau 5). La dernière ligne de ce tableau indique que, dans la correspondance, les 10 verbes les plus usuels ont une densité totale 17% plus élevée que dans le reste de l'oeuvre. Autrement dit, quand il écrit une lettre, Hugo a tendance à utiliser plus les verbes usuels que lorsqu'il compose un texte destiné à la publication.

La même tendance peut être observée à propos des noms (tableau 7).

Tableau 7. Les substantifs les plus fréquents

Autres				Correspondance		
Rang	Nom	F absolue	F relative(‰)	Nom	F absolue	F relative(‰)
1	homme	1215	2.9	*monsieur*	1320	4.5
2	oeil	776	1.8	lettre	1239	4.2
3	*monsieur*	729	1.7	livre (n m)	706	2.4
4	dieu	642	1.5	madame	516	1.8
5	jour	612	1.4	franc	485	1.7
6	chose	573	1.4	jour	480	1.6
7	enfant	568	1.3	*main*	462	1.6
8	tête	548	1.3	coeur	429	1.5
9	âme	545	1.3	chose	426	1.5
10	*main*	534	1.2	ami	413	1.4
Total			15.9			22.2

Tous les écarts sont significatifs. Deux noms seulement sont communs : *monsieur* et *main*, avec un classement assez différent. *Monsieur* et *Madame* sont épistolaires, les lettres portent d'abord sur les *livres* (ceux dont il accuse réception, ceux qu'il réclame et surtout ceux qu'il a en chantier). Le tableau précise qu'il s'agit du nom masculin car dans les lettres d'Hugo, il est fait référence 19 fois à la *livre sterling*. La place, en cinquième position du *franc* - c'est-à-dire la monnaie et non l'adjectif -, signale l'importance des problèmes financiers dans la correspondance d'Hugo. *Cœur* et *ami* indiquent l'aspect affectif des lettres. Enfin, on retrouve le même sur-emploi qu'avec les verbes (dernière ligne du tableau).

Cette caractéristique se retrouve chez les adjectifs (tableau 8).

Tableau 8. Les adjectifs les plus fréquents

Autres				Correspondance		
Rang	Adjectifs	F absolue	F relative(‰)	Nom	F absolue	F relative(‰)
1	*grand*	827	2.0	cher	1029	3.5
2	*bon*	536	1.3	*grand*	746	2.6
3	*petit*	524	1.2	*bon*	718	2.5
4	*beau*	512	1.2	charmant	403	1.4
5	noir	390	0.9	*petit*	381	1.3
6	vieux	353	0.8	*beau*	370	1.3
7	seul	343	0.8	doux	254	0.9
8	jeune	332	0.8	excellent	254	0.9
9	pauvre	308	0.7	noble	204	0.7
10	sombre	301	0.7	vrai	201	0.7
Total			10.4			15.6

Dans les textes français, les adjectifs *grand* et *petit* sont très souvent les adjectifs les plus employés. En revanche, les rangs de *bon* et de *beau* sont propres à Hugo. La première place de « cher » ne surprend pas : accolé au prénom du destinataire ou à *Monsieur*, *Madame*, *ami*… il ouvre, et parfois conclut les missives. En dehors de cela les adjectifs comme : *noir*,

vieux, *seul*, *pauvre*, *sombre* sont très utilisés dans la poésie, les romans, le théâtre, car ces œuvres sont pessimistes. A leur place, dans la correspondance, on trouve : *charmant*, *doux*, *excellent*, *noble*, *vrai*… Les lettres véhiculent des valeurs nettement plus optimistes ! Pourtant, l'exil et les difficultés matérielles auraient dû avoir la conséquence inverse. Enfin, la dernière ligne confirme le phénomène noté précédemment : un usage nettement plus intensif des adjectifs usuels (+50% entre la partie gauche et la partie droite du tableau)…

Les mêmes constats peuvent être faits à propos de toutes les autres catégories grammaticales - pronoms, adverbes, déterminants, prépositions et conjonctions – la quasi-totalité des écarts sont significatifs et Hugo a tendance à sur-employer les plus usuels.

Le tableau 9 donne le résultat pour les principaux pronoms.

Tableau 9. Les pronoms les plus fréquents (‰)

Pronom	A (Autres)	B (Correspondance)	(B-A)/A (%)
il	22,6	12,4	-45,0
je	14,7	43,5	+196,6
se	11,2	3,2	-71,3
qui	10,7	6,6	-38,2
ce (pro)	9,3	9,6	+3,0
que (pro)	7,6	8,1	+7,7
vous	7,1	27,5	+287,1
le (pro)	6,9	8,0	+15,6
on	5,3	3,1	-40,9
lui	3,9	3,1	-21,2
nous	3,6	4,5	+25,8
tu	3,3	4,8	+42,6
ils	2,4	1,3	-45,2
moi	1,9	4,8	+151,7
cela (pro)	1,6	2,3	+45,8
dont	1,2	0,7	-36,7
celui (pro)	1,0	0,8	-17,0
lequel	0,7	0,5	-22,8
toi	0,6	0,8	+27,1
Total	115,4	145,7	+27,0

En dehors de « ce que », tous les pronoms varient significativement en plus ou en moins. Les pronoms personnels sont affectés des plus amples mouvements. On peut résumer ces mouvements de la manière suivante :

Pronoms de la :	A Autres (‰)	B Correspondance (‰)	(B-A)/A (%)
Troisièmes personnes	45,4	23,2	-49,0
Premières et deuxièmes personnes	31,3	85,9	+175,0

- tous les pronoms la troisième personne reculent : *il* et *ils* (-45%), *on* (-41%), *se* (-71%) et *lui* (-21%). Chez Flaubert et Maupassant, on assiste au même phénomène : *il* (-53%) ; *ils* (-71%) ; *se* (-72%) ; *lui* (-30%). Selon Benveniste, la troisième personne est la « non personne », parce qu'elle désigne un tiers exclu de l'interlocution, ou parce qu'elle est employée pour désigner quelqu'un ou quelque chose dont on a parlé avant, soit parce qu'il s'agit d'un « impersonnel » (comme dans « il faut »). Dans la correspondance, ces trois cas passent au second plan.

- la fréquence des autres personnes augmente considérablement : triplement de fréquence du *je* et quasi-quadruplement du *vous* (multiplication par 3,9). Même phénomène

chez Flaubert et Maupassant : *je* +300% ; *vous* : +320%. Les lettres sont donc le lieu d'une interlocution (*je*/*vous*), voire d'une interpellation du *vous* par le *je*. Chez Hugo, le *tu* est réservé à sa femme et à ses enfants ; pas de tutoiement entre Flaubert et Maupassant. Le sens le plus fréquent du *nous* est « *moi* et *vous* destinataire de cette lettre », mais, chez Hugo, il désigne aussi souvent le cercle qui l'entoure dans son exil.

- dans la correspondance d'Hugo comme dans celle de Maupassant et Flaubert, le système des pronoms relatifs est simplifié au profit de *que* - le plus simple et le plus polyvalent – et au détriment des formes dont l'emploi est plus complexe (*lequel*, *dont*, *qui*). En cela, la correspondance ressemble à l'oral où se produit un phénomène identique (Labbé 2003).

Enfin, on note que, lorsque Hugo écrit une lettre, le poids des pronoms usuels augmente de 27% par rapport à ses autres écrits. L'augmentation chez Flaubert et Maupassant est du même ordre.

Ces listes, mêmes brèves, sont très instructives. Elles soulèvent aussi des questions. On peut notamment se demander si les mouvements considérables que l'on vient de décrire affectent tout le vocabulaire et quels sont les vocables caractéristiques de la correspondance (au-delà des plus fréquents qui viennent d'être décrits).

Vocables caractéristiques de la correspondance

Il est proposé de généraliser le raisonnement présenté ci-dessus à propos du verbe *être*. Un mot est *caractéristique* d'un texte, ou d'une partie d'un corpus, lorsque sa fréquence relative, dans ce texte ou cette partie, est *significativement* supérieure à celle observée dans l'ensemble du corpus (sur ce calcul : Labbé & Labbé 1994). Cela signifie que l'auteur a éprouvé une attirance particulière pour ce mot quand il rédigeait le texte considéré. En sens contraire, on s'intéressera aussi aux sous-emplois caractéristiques : les mots envers lesquels l'auteur a ressenti une répugnance particulière (à propos des thèmes traités dans le texte).

Voici un échantillon des substantifs les plus caractéristiques (avec moins de 1 chance sur 10.000 de se tromper en considérant qu'il y en a significativement plus dans les lettres d'Hugo que dans ses œuvres) :

> monsieur, lettre, livre, madame, franc, main, coeur, ami, temps, esprit, page, mot, journal, poète, succès, reste, vers, volume, oeuvre, mois, raison, article, décembre, épreuve, avril, point, droit, liberté, travail, mer, mai, dimanche, exil, merci, mars, ligne, juin, question, janvier, amie, fils, juillet, affaire, réponse, cas, avenir, avis, fait, suite, besoin, novembre, fin, poste, août, théâtre, publication, feuille, talent, édition, titre, février, envoi, drame, amitié, détail, exemplaire, manuscrit, devoir, éditeur, octobre, adresse, septembre, confrère, progrès, proscrit, honneur, famille, poésie, conscience, courrier, semaine, sujet, pli, préface

Cette courte liste montre que le vocabulaire de la correspondance comprend d'abord les termes dédiés aux échanges épistolaires : datation (mois, jours), *lettre*, *réponse*, *poste*, *envoi*, *adresse*, *courrier*, *pli*… Outre les problèmes d'argent (*franc*), l'objet essentiel de la correspondance d'Hugo en exil portait sur ses *livres*, son *travail*, les *envois* de *journaux*, les *volumes*, les *épreuves*, l'*édition*, *manuscrit*, *éditeur*, *préface*…. Un autre sujet est suggéré par la présence de *droit*, *liberté*, *exil*, *devoir*, *progrès* et *proscrit*. Mais il s'agit surtout d'évoquer son éloignement pour justifier telle ou telle difficulté, retard ou sollicitation. Après les premiers mois, les questions proprement politiques sont rarement abordées.

Même ainsi reconstitué, le vocabulaire caractéristique manque de chair. Il est nécessaire de l'illustrer par des exemples.

Phrases caractéristiques de la correspondance

Les phrases caractéristiques sont détectées de la manière suivante. Une fois déterminé le vocabulaire caractéristique, le logiciel relit les textes analysés – ici l'ensemble de la correspondance - en recherchant les phrases qui contiennent le plus de vocables sur-employés dans les lettres et le plus petit nombre de vocables sous-employés. A la rencontre d'un mot significativement sur-employé dans la correspondance, le score de la phrase est augmenté de 1. A l'inverse, ce score est diminué de 1 à la rencontre d'un mot significativement sous-employé. On obtient ainsi les "phrases canoniques" comme celles qui servent, dans un dictionnaire, à illustrer la définition du mot. Voici les 2 phrases les plus caractéristiques de la correspondance d'Hugo (entre parenthèse le score absolu rapporté à la longueur de la phrase).

> « Maintenant fais attention : les 250 francs pour Meurice, les 50 francs pour Charles et les 25 francs pour Georges ayant été payés directement par moi je m'en rembourse et tu les retiendras sur les 375 francs italiens pour les appliquer comme suit : tes 50 francs prélevés qui élèveront ton mois à 250 francs ; il reste 325 francs : 1 ton mois : 200 (qui sera en effet 250) » (A François-Victor Hugo, 25 juin1868, score 58%).

> « Je vous répète ce que je vous ai dit déjà à plusieurs reprises, que je désire conserver ma liberté et vous laisser la vôtre, que cela ne m'empêchera en aucune manière d'écouter, et, j'espère, d'accueillir vos propositions, si vous jugez à propos de m'en faire, quand j'aurai un ouvrage prêt à paraître, et que, dans tous les cas, il me semblerait bien difficile, sinon impossible, de faire, avec quelque éditeur que ce soit, un traité d'ensemble avant d'avoir terminé le livre 93 » (A Albert Lacroix (éditeur), 3 décembre 1867, score 42%).

Elles illustrent les deux thèmes dominants dans ce sous-corpus. En premier lieu les problèmes d'argent, puis la création littéraire et les relations avec les éditeurs. Dans les quarante phrases les plus caractéristiques, aucune ne parle explicitement politique.

IV. Le style épistolaire

La statistique appliquée au langage permet également de répondre à la question de savoir quel est le style de la correspondance.

La stylistique s'est détournée des recensements statistiques qui est le "ventre mou de la stylistique française", selon l'expression de G. Molinié (1986 , p. 54). Pourtant, une stylistique quantitative serait possible, comme le montre le recensement des indices possibles effectué par Molinié lui-même (1986 : 146-156). Trois dimensions sont proposées ici : la richesse du vocabulaire, les parties du discours, la longueur et la structure des phrases.

Richesse du vocabulaire

Quand un critique dit d'un auteur qu'il a un vocabulaire « riche », cela signifie généralement que ce critique a trouvé, dans le dernier livre de cet auteur, un certain nombre de mots « rares » (selon son jugement personnel). Ce genre de jugement ne peut évidemment pas faire l'objet d'une mesure objective…

En fait, cette richesse du vocabulaire recouvre deux choses assez différentes (voir Hubert & Labbé, 1994 et Labbé, 1998).

D'une part, la *diversité du vocabulaire* mesure le soin mis à varier ses expressions et à éviter des répétitions à intervalles rapprochés. Cette caractéristique est mesurée grâce à l'indice de la diversité du vocabulaire : nombre de vocables différents employés dans des portions de texte de même longueur afin de permettre des comparaisons entre oeuvres et entre auteurs (par exemple, 10 000 mots, du moins quand les textes comparés ont tous une longueur supérieure à cette dimension). Par exemple, un indice de 1 500 signifiera que, dans le texte ou

le corpus considéré, on rencontre en moyenne 1 500 vocables différents dans tous les extraits possibles de 10 000 mots.

D'autre part, la *spécialisation* du vocabulaire mesure la capacité d'un auteur à réserver l'emploi de certains vocables au traitement d'un thème particulier. Cet indice pourra être exprimé en « pour 1000 » ou pour « 10000 » selon la longueur des corpus et la précision souhaitée. Un indice de 100 ‰ signifie que, sur 1000 mots pris au hasard dans le corpus considéré, 100 (ou un dixième) ne sont employés que dans un passage donné, pour traiter un thème particulier.

Le tableau 10 ci-dessous donne les résultats obtenus sur les quatre sous-corpus des œuvres d'Hugo, sur ceux de Flaubert et Maupassant, puis sur quelques poètes et romanciers contemporains.

Tableau 10. Diversité et spécialisation du vocabulaire chez Hugo, Flaubert, Maupassant et quelques contemporains.

	Diversité (pour 10 000 mots)	Spécialisation (‰)
Victor Hugo		
Poésie *Contemplations* (vers)	2 179	0
Théâtre *Hernani*, *Ruy Blas* (vers)	1 488	297
Roman *Notre-Dame de Paris* (prose, 1832)	2 426	0
Roman *Misérables* (prose, 1860)	2 170	0
Correspondance (prose 1849-1870)	1 567	190
Flaubert et Maupassant		
Lettres de Flaubert à Maupassant (1875-1880)	1 711	317
Lettres de Maupassant à Flaubert (1875-1880)	1 615	56
Flaubert (tous les romans)	2 548	0
Maupassant (tous les romans)	1 916	108
Autres auteurs		
Banville *Fournaise* (poésie vers)	2 240	9
Baudelaire (toutes les poésies vers et prose)	2 320	104
Coppée *Bonne souffrance* (poésie vers)	2 250	65
Dumas *Monte Cristo* (roman)	1 499	230
Gautier (2 recueils de Poésie en vers)	2 542	0
Gautier *Avatar* (Roman)	2 520	0
Huysmans *A rebours* (roman)	2 855	282
Leconte de l'Ile *Poèmes antiques* (vers)	2 023	33
Rimbaud (toutes les poésies)	2 249	215
Sand (2 romans)	1 923	0
Verlaine (toutes poésies vers)	2 196	115
Zola (4 romans)	2 005	112
Moyenne poésie	2 260	77
Moyenne romans	2 181	105

Le tableau montre trois choses. Premièrement, la diversité et la spécialisation du vocabulaire ne sont pas des caractéristiques intrinsèques à un auteur (Hugo couvre presque tout l'éventail des combinaisons possibles). Deuxièmement, les différences ne tiennent pas à l'opposition entre vers et prose. En effet, les *Contemplations* et le théâtre d'Hugo sont tous deux en vers. Troisièmement, les différences ne tiennent pas au genre : dans *Notre-Dame de Paris*, la diversité du vocabulaire est très élevée, dans les *Misérables* elle est beaucoup plus réduite. De même, chez T. Gautier, il n'y a pas de différence entre les deux recueils de

poèmes (*Emaux et camées* et *Poésies diverses*) et le roman *l'Avatar* : même spécialisation nulle et même diversité élevée. Tout au plus, peut-on noter que la correspondance a une diversité plus faible que les autres genres et que les deux dernières lignes du tableau 12 indiquent que les poètes ont une légère tendance à privilégier un peu plus la diversité, les romanciers la spécialisation.

Il s'agit donc de choix stylistiques contingents. D'une part, l'auteur choisit entre la simplicité du propos (comme Dumas voire Sand et Zola) – au risque de paraître simpliste – ou, à l'inverse, comme Gautier ou Huysmans : la diversité au risque de paraître maniéré. D'autre part, il choisit d'utiliser les mêmes mots quels que soit le sujet (comme Hugo dans ses romans ou sa poésie) – au risque de la monotonie - ou de rechercher l'expression précise ne s'appliquant qu'à l'objet et au thème dont il traite (comme Hugo dans son théâtre ou dans sa correspondance) mais avec le risque d'un excès de couleur locale et d'exotisme (comme le théâtre d'Hugo), de préciosité - comme Huysmans dans *A rebours* -, voire d'obscurité (comme Rimbaud). D'autres auteurs comme Zola ont une grande régularité, injectant dans leurs romans une dose à peu près constante de vocabulaire propre au milieu qu'ils décrivent : les chemins de fer dans *la Bête humaine*, les mines dans *Germinal* ou la finance dans *l'Argent*.

Evidemment, nous n'avons aucune transcription de propos spontanés tenus par les écrivains du XIXe. Les expériences faites au XXe montrent que, chez une même personne, l'oral se différencie de l'écrit par une diversité plus faible – en tous cas inférieure à 1 500 mots différents pour 10000 mots – et par une spécialisation assez souvent plus élevée. En sens inverse plus le degré d'élaboration est important plus la diversité augmente et plus la spécialisation a tendance à diminuer (Labbé & Monière, 2008, p 24-29). Le tableau 5 ci-dessus suggérerait donc que la correspondance est, avec le théâtre, le genre le plus proche de l'oral, alors que la poésie en serait la plus éloignée. Au fond, quoi de plus normal ? Depuis Corneille, l'idéal du théâtre est de faire parler les personnages de manière réaliste et non « comme des auteurs ». A l'opposé, l'idéal de la poésie est de faire pénétrer le lecteur dans les pensées du poète. La fiction y ajoute un élément supplémentaire : le récit.

Les catégories grammaticales

Le poids donné aux différentes catégories grammaticales est le second élément utilisé pour étudier le style. En effet, le genre a une influence importante sur la densité de la plupart des catégories grammaticales et la correspondance n'échappe pas à cette règle. Le tableau 11 résume la comparaison entre les lettres d'Hugo et le reste de son œuvre.

Ce tableau se lit de la manière suivante. Dans le reste de l'œuvre – théâtre, roman, poésie (colonne A) – on rencontre en moyenne 158,4 verbes pour 1000 mots. Dans la correspondance (colonne B), cette densité passe à 164.1, soit une augmentation de 3,6 % (dernière colonne). Les quatre lignes suivantes détaillent ces densités et ces fluctuations pour les différents types de verbes (fléchis, participes passé et présents et infinitifs).

La légère augmentation du nombre de verbes s'explique surtout par le poids du participe passé : le récit occupe une place particulière dans la correspondance. La même chose s'observe chez Flaubert et Maupassant : verbes (+8%), passé (+77%).On écrit des lettres d'abord pour raconter.

Les pronoms suivent la même tendance mais en l'amplifiant (+24.8). Il s'agit essentiellement des pronoms personnels et des possessifs (le *vôtre, mien, tien, nôtre*) qui renforcent encore la tension interlocutive discutée ci-dessus. En revanche, le recul des relatifs illustre encore une fois la simplification de la syntaxe dans le courrier par rapport aux autres écrits.

Tableau 11. Comparaison des parties du discours dans la correspondance et dans l'œuvre de V. Hugo

Catégories	A-B Autres ‰	B Correspondance ‰	(B-A)/A %
Verbes	158.43	164.13	+3.6
Formes fléchies	*115.32*	*115.88*	*+0.5*
Participes passés	*15.70*	*19.66*	*+25.2*
Participes présents	*6.12*	*3.09*	*-49.6*
Infinitifs	*21.29*	*25.51*	*+19.8*
Noms propres	24.24	43.37	+78.9
Noms communs	200.79	166.06	-17.3
Adjectifs	61.86	54.86	-11.3
Adj. participe passé	*10.65*	*6.03*	*-43.4*
Pronoms	129.92	162.16	+24.8
Pronoms personnels	*77.28*	*109.73*	*+42.0*
Pronoms démonstratifs	*12.03*	*13.38*	*+11.1*
Pronoms possessifs	*0.25*	*0.87*	*+248.0*
Pronoms indéfinis	*5.26*	*5.11*	*-2.8*
Pronoms relatifs	*27.23*	*24.34*	*-10.6*
Déterminants	172.58	160.48	-7.0
Articles	*121.47*	*87.39*	*-28.1*
Nombres	*9.62*	*25.99*	*+170.3*
Possessifs	*21.91*	*26.73*	*+22.0*
Démonstratifs	*10.85*	*10.70*	*-1.5*
Indéfinis	*8.73*	*9.68*	*+10.9*
Adverbes	59.39	59.87	+0.8
Prépositions	133.16	127.67	-4.1
Conjonctions	53.33	56.72	+6.3
Conjonctions de coordination	*32.96*	*34.13*	*+3.5*
Conjonctions de subordination	*20.37*	*22.58*	*+10.9*

Les noms propres (toponymes et patronymes) augmentent de manière importante. Même phénomène chez Flaubert et Maupassant où la fréquence des noms propres croît de 46%. Les noms propres n'appartiennent pas réellement à la langue mais assurent son interface avec la réalité extérieure, en ancrant le texte dans l'espace géographique et social.

De même, l'augmentation des nombres est spectaculaire (+ 270% chez Flaubert et Maupassant). Ils ancrent le texte dans le temps (dates, durées), l'espace (les dimensions), les relations marchandes et financières (prix, sommes d'argent)...

L'augmentation des conjonctions de subordination - spécialement de *que* - est à mettre en relation avec le suremploi de « dire », « vouloir »,…

A l'opposé, on observe un recul important du nom et de ses satellites : adjectifs (-11% chez Hugo, - 23% chez Flaubert et Maupassant), articles (respectivement -24 et -25%) et prépositions.

Pour résumer ces mouvements contradictoires, on peut rassembler les parties du discours en deux grands groupes : nominal et verbal. Le premier comporte les substantifs, les adjectifs, les déterminants et les prépositions. Le second, les verbes, les pronoms, les adverbes et les conjonctions de subordination. Certes, le partage n'est pas mécanique : on trouve des adverbes dans le groupe nominal (notamment devant l'adjectif) ; il y a des prépositions dans le groupe verbal, etc.

Cette réserve admise, le regroupement révèle des tendances intéressantes (tableau 12).

Tableau 12. Poids des groupes nominaux et verbaux dans les lettres d'Hugo, Flaubert et Maupassant comparées au reste de leurs œuvres.

	A-B Autres (‰)	B Correspondance (‰)	(B-A)/A (%)
Hugo			
GV	368,3	408,7	+11,0
GN	625,4	586,6	-6,2
Flaubert-Maupassant			
GV	383,5	429,2	+11,9
GN	613,5	567,8	-7,5

La correspondance se caractérise par un usage plus intensif du groupe verbal.

Dans le corpus « français écrit » (du XVIIe au XXe siècle, comprenant 14 millions de mots), le groupe verbal couvre 369 ‰ de la surface du texte et le groupe nominal 630 ‰. Ces proportions ne diffèrent guère des valeurs enregistrées pour Hugo (deuxième colonne du tableau ci-dessus). Autrement dit, dans son œuvre imprimée, Hugo se situe dans la moyenne des écrivains. En revanche, dans sa correspondance, le verbe progresse et le nom recule. Cela est également vrai pour Flaubert et Maupassant malgré leur préférence commune pour le verbe dans le reste de leurs écrits.

Il existe une vision un peu triviale selon laquelle le verbe est l'action - mais aussi le risque - et le groupe nominal est le domaine des idées mais aussi de la conservation. Dans ce cas, il faudrait admettre une sorte de paradoxe : ce serait lorsque Hugo est exilé, réduit à l'impuissance, qu'il développerait un goût accru pour l'action ou une orientation plus forte vers la réalité ? Cette interprétation n'est pas tenable puisque le même mouvement est enregistré chez Flaubert et Maupassant.

Pour la linguistique, le verbe a une double fonction : la "fonction cohésive" qui organise "en une structure complète les éléments de l'énoncé" et la fonction assertive qui "dote l'énoncé d'un prédicat de réalité" car l'élément verbal implique une référence à un ordre qui n'est plus simplement celui du discours mais aussi celui de la réalité (Benveniste 1981, 1, p. 154).

En fait, là encore, il faut ajouter un élément clef : la comparaison entre l'écrit et l'oral. Suivant l'expérience déjà évoquée, lorsqu'une personne passe de l'écrit à l'oral, la densité du groupe verbal augmente de manière assez nette (Labbé & Monière 2008, p. 31-38). Autrement dit, la correspondance serait un genre proche de l'oral.

Longueur et structure de phrases

Pour les méthodes permettant d'étudier la longueur des phrases, les différents types de phrases et leurs fonctions respectives dans la communication, voir : Monière, Labbé & Labbé 2008. Les auteurs contemporains, déjà observés à propos de la richesse du vocabulaire, sont utilisés comme éléments de comparaison. Les valeurs caractéristiques sont présentées dans le tableau 13 ci-dessous.

La phrase d'Hugo est brève (15 mots en moyenne). C'est la plus basse valeur par rapport à tous ses contemporains et c'est une singularité majeure de son style (du moins si l'on accepte de considérer les textes dépouillés comme représentatifs de l'ensemble de l'œuvre). Seul Rimbaud est proche des valeurs observées chez Hugo, puis viennent Flaubert et Maupassant.

Tableau 13. Comparaison des longueurs de phrase chez Hugo, Flaubert, Maupassant et leurs contemporains.

	Mode	Médiane	Médiale	Moyenne	Variation relative (%)
Hugo					
Contemplations (poésie en vers)	10	15	36	23.0	104.0
Théâtre (vers)	1	5	12	7.9	88.1
Misérables (roman)	6	12	24	16.6	96.8
Notre-Dame de Paris (roman)	6	13	27	18.7	97.0
Correspondance	3	10	18	13.0	86.6
Total Hugo	3	11	22	15.1	101.2
Flaubert et Maupassant					
Lettres de Flaubert à Maupassant	8	10	16	12.4	72.5
Lettres de Maupassant à Flaubert	9	15	24	18.1	76.8
Flaubert (tous les romans)	7	15	26	18.7	79.6
Maupassant (tous les romans)	6	15	27	19.0	79.9
Autres auteurs					
Baudelaire (toutes les poésies)	19	25	38	29,2	76,5
Coppée *Bonne souffrance* (poésie vers)	8	24	41	28,7	83,3
Dumas *Monte Cristo* (roman)	9	18	32	22,6	86,3
Flaubert (tous les romans)	7	15	26	18,7	79,6
Gautier (2 recueils poésies vers)	22	23	30	26,0	77,4
Gautier *Avatar* (roman)	21	30	47	35,6	75,5
Huysmans *A rebours* (roman)	34	45	72	55,4	73,1
Leconte de l'Isle *Poèmes antiques* (vers)	1	17	33	20,7	87,8
Maupassant (tous romans)	6	15	27	19,0	79,9
Rimbaud (toutes les poésies vers)	7	11	25	16,0	95,3
Sand (2 romans)	17	20	33	24,5	70,0
Verlaine (toutes les poésies vers)	5	16	32	21,0	96,8
Zola (4 romans)	8	17	29	21,7	76,3

Le coefficient de variation relative[2] (dernière colonne du tableau) indique que la longueur des phrases est fortement dispersée autour de cette moyenne. En effet, la dernière colonne signifie que, dans les *Contemplations*, la longueur d'un peu plus des deux tiers des phrases est comprise dans un intervalle de plus ou moins 1.04 fois la moyenne (c'est-à-dire de 1 à 46 mots) et qu'il y a donc encore près d'un tiers des phrases dont la longueur est supérieure à deux fois la moyenne (46 mots). La distribution est donc asymétrique et fortement étalée. Ces valeurs signifient que la population étudiée n'est pas homogène et qu'elle résulte du mélange de plusieurs types différents de phrases (Labbé, Labbé et Monière, 2008). Dans un cas de ce genre, trois autres valeurs centrales doivent être considérées :

- La *longueur modale* principale est la longueur que l'on rencontre le plus souvent chez Hugo. Le mode le plus élevé se trouve dans la poésie (10 mots) ; le plus petit dans le théâtre (un mot : ce sont les exclamations et interjections, très nombreuses dans ses pièces). Ce phénomène se retrouve seulement chez Leconte de l'Isle pour les mêmes raisons. Cette valeur est de 3 dans ses lettres qui contiennent donc un grand nombre de phrases brèves.

[2] Ecart type divisé par la moyenne arithmétique. L'écart type est la racine carrée de la variance (somme des carrés des écarts à la moyenne arithmétique, divisée par le nombre de valeurs composant la série). Si la distribution des valeurs autour de cette moyenne n'est due qu'au hasard, environ les deux tiers de valeurs sont comprises dans un intervalle de ±1 écart type.

La *longueur médiane* sépare le nombre de phrase en deux : dans les lettres d'Hugo, la moitié des phrases ont moins de 10 mots. Elles sont donc réduites à un noyau très simple. Là encore le théâtre d'Hugo est singulier : la moitié des phrases ont 5 mots ou moins.

La *longueur médiale* sépare la population étudiée en deux : dans les lettres d'Hugo, la moitié de la surface est occupée par des phrases de 18 mots ou moins. Là encore le théâtre est singulier avec une longueur médiale de 12 mots que l'on ne retrouve nulle part dans la littérature depuis le XVIIe siècle, du moins dans ce que nous en avons dépouillé.

Si on laisse de côté les deux pièces d'Hugo, sa correspondance se caractérise par des phrases plus brèves que le reste des écrits. Pour le reste, la longueur des phrases n'est pas une caractéristique propre à l'auteur. Elle fluctue en fonction du genre. Elle change aussi selon la chronologie : nette réduction entre *Notre Dame de Paris* (1830) et les *Misérables* (1860). Dès lors quel facteur influe sur la longueur de phrase ? A priori, la distinction prose/vers n'est pas pertinente : les phrases les plus longues se trouvent dans les *Contemplations* (vers) et les plus courtes dans le théâtre (également en vers).

Là encore, l'oral et l'écrit apportent un éclairage intéressant : quand une personne passe de l'écrit à l'oral, la longueur moyenne de ses phrases diminue (de même que leur complexité). On peut donc supposer que le texte épistolaire est, avec certaines pièces de théâtre, le genre le plus proche de l'oral.

La ponctuation

On présente ci-dessous les premiers résultats d'un travail en cours sur la structure des phrases, limités pour l'instant à deux aspects : les fins de phrase (tableau 14) et la ponctuation interne à la phrase (tableau 15).

Tableau 14. Comparaison des fins de phrases dans le corpus Hugo (en %)

	point	interrogation	exclamation	suspension
Poésie	56,2	11,7	31,3	0,8
Notre-Dame de Paris	82,7	6,3	10,1	0,9
Misérables	86,4	7,1	6,0	0,5
Théâtre	43,9	11,1	40,5	4,6
Correspondance	90,8	4,7	4,4	0,1

Dans la correspondance, 90,8% des phrases – soit 9 sur 10 - se terminent par le point « classique » qui est la manière normale de terminer la phrase, comme dans les romans. En revanche, la poésie et les deux pièces de théâtre se singularisent. On se rappelle que les pièces se caractérisent par des phrases très courtes et que la longueur modale est d'un mot (le plus souvent une exclamation ou une interjection). Ceci explique qu'il y ait, dans le théâtre d'Hugo, plus de phrases se terminant de manière « non conventionnelle » - surtout par une interrogation ou une exclamation -, mais aussi une proportion non négligeable de phrases inachevées (près de 5% se finissent par des points de suspension). Nous avons déjà formulé l'hypothèse selon laquelle, théâtre et correspondance sont des manières de représenter des échanges verbaux. Il faudrait ajouter que, dans *Hernani* et *Ruy Blas*, ces échanges sont très tendus et emprunts de passions alors que dans la correspondance, ils sont plus paisibles.

L'étude de la ponctuation interne confirme ce dernier point (tableau 15).

En moyenne, dans une phrase des lettres, il y a un peu moins d'une virgule (0.921) alors qu'au théâtre cette proportion tombe à 0.375, en dessous du point d'interrogation. Interrogation et exclamation sont les caractéristiques du drame, la virgule et la parenthèse, celles des lettres.

Tableau 15. Comparaison de la densité des ponctuations internes à la phrase dans les lettres et le théâtre d'Hugo

Ponctuations internes à la phrase	Correspondance	Théâtre
virgules	**0,921**	0.375
points virgules	0.064	0.043
deux points	0.038	0.013
exclamation	0.048	**0.464**
interrogation	0.050	0.125
parenthèses	0.126	-
tirets	0.037	0.069
guillemets	0 .066	0.011
Total ponctuations par phrase	1.291	1.101
Longueur moyenne (mots)	13.04	7.91
ponctuations par mots	0.10	0.14

Conclusions

Quelles leçons tirer de ces expériences ?

Une langue n'est pas seulement la combinaison d'une phonétique, d'une syntaxe et d'un lexique. Elle fournit aussi à ses usagers des « genres », c'est-à-dire des règles pour échanger les signes linguistiques en fonction des situations de communication. Le français est caractérisé par une opposition forte entre ce qui se dit oralement et ce qu'on peut s'écrire. Pour l'écrit, il existe une série de genres particuliers. Nous avons mis en valeur : la fiction romanesque, la poésie versifiée, le théâtre et la correspondance. Naturellement, il existe d'autres « genres » - le texte administratif, juridique, scientifique, etc. - et des sous-genres. Par exemple, le théâtre se divise entre la tragédie et la comédie. Ces codes n'ont pas la rigidité de la syntaxe et encore moins celle du code civil. Ce sont des coutumes assez lâches qui durent tant qu'elles sont adaptées à la société et aux buts poursuivis par les usagers de la langue quand ils parlent ou écrivent. Dans une certaine mesure, la création est aussi une transgression de ces codes comme le prouve la fameuse « querelle d'*Hernani* ».

Les lettres se caractérisent par un emploi élevé de noms propres et de chiffres qui assurent leur ancrage dans le temps, l'espace, la société, l'économie. Elles sont dominées par la tension interpellative (je:vous) et il y est moins question des tiers. On y observe une augmentation significative du groupe verbal par rapport aux autres écrits et une baisse du groupe nominal ; enfin, les phrases « épistolaires » sont plus brèves et plus simplement construites que dans les autres genres écrits.

On conclura que le texte épistolaire appartient bien à un genre singulier dont le ressort essentiel serait de simuler une conversation entre l'auteur et le destinataire, en adoptant certaines caractéristiques de l'oral, tout comme on le fait dans le théâtre.

Ces conclusions peuvent sembler banales a posteriori mais elles n'étaient pas acquises au départ.

Enfin, on ne peut conclure sans évoquer une dernière question : peut-on affirmer que ces caractéristiques se retrouvent peu ou prou dans toute lettre écrite en français quel que soit l'auteur ? Ou encore, peut-on tirer des conclusions définitives alors qu'on a étudié que trois auteurs, même si ces expériences portent sur plus de 1 300 lettres et deux millions de mots ? Cette objection ne pourrait être levée que si l'on disposait de vaste corpus représentatifs des usages du français moderne (sur le modèle du British National Corpus). Un tel outil n'existant pas, les usages du français moderne ne peuvent être qu'imparfaitement approchés par des expériences comparables à celle que nous venons de présenter…

Références

Benveniste E. (1981). *Problèmes de linguistique générale*. Paris : Gallimard.

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

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

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

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.

Hubert P. & Labbé D. (1994). Vocabulary Richness, *Communication au congrès de l'ALLC-ACH*. Paris, La Sorbonne (Reproduit dans *Lexicometrica*, 0, 1997).

Labbé C. & Labbé D. (1994). *Que mesure la spécificité du vocabulaire ?*. Grenoble: CERAT. Décembre 1994 & juin 1997. 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, p. 213-231.

Labbé C. & Labbé D. (2003). "La distance intertextuelle". *Corpus*. 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, p. 311-326.

Labbé C. & Labbé D. (2007a). "Baudelaire, Rimbaud et Verlaine". VIIIe journées de l'ERLA, *Aspects linguistiques du texte poétique*. Brest 16-17 novembre 2007.

Labbé C. et Labbé D. (2007b). « Annexe 7. Flaubert et Maupassant » in *Corneille a écrit 16 pièces représentées sous le nom de Molière. Réponses à : VIPREY Jean-Marie et LEDOUX Claude-Nicolas, 'About Labbé's "Inter-textual Distance"*. Grenoble : PACTE-IEP, p. 144-145.

Labbé C., Labbé D. & Monière D. (2008). "Les styles discursifs des premiers ministres québécois de Jean Lesage à Jean Charest". *Revue canadienne de science politique*. 41-1, p. 43-69.

Labbé D. (1990). *Normes de saisie et de dépouillement des textes politiques*. Grenoble : Cahier du CERAT.

Labbé D. (1998). "La richesse du vocabulaire politique : de Gaulle et Mitterrand". *Mots chiffrés et déchiffrés: mélanges offerts à Étienne Brunet*, Paris, Champion, p. 173-186.

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. & Monière D. (2008). *Les mots qui nous gouvernent. Le discours des premiers ministres québécois (1960-2005)*. Montréal : Monière-Wollank Editeurs.

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.

Lyne A. A. (1985). *The vocabulary of french business correspondance : word frequencies, collocations and problems of lexicometric method*. Genève-Paris : Slatkine-Champion.

Molinié G. (1986). *Eléments de stylistique française*. Paris : PUF.

Molinié G. et Cahné P. eds (1994). *Qu'est-ce que le style ?* Paris : PUF.

Monière D. & Labbé D. (2002). "Essai de stylistique quantitative. Duplessis, Bourassa et Lévesque". In Morin A. & Sébillot P. (eds). *VIe Journées Internationales d'Analyse des Données Textuelles (Saint-Malo 13-15 mars 2002*). Rennes: IRISA-INRIA, 2002, vol. 2, p 561-569.

Richaudeau F. (1988). *Ce que révèlent leurs phrases*. Paris, Retz.

Sauval H. (1724). *Histoire et recherches des antiquités de la ville de Paris*, Paris, Charles Moette et Jacques Chardon. (Reproduit à Paris/Genève : éd. du Palais/Minkoff en 1974).

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

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

Annexe
Les corpus

La correspondance d'Hugo a été déchargée en mai 2009 sur le site WikiSource. Il s'agit du texte de l'édition de référence : A.Michel-Ollendorf-Imprimerie Nationale, tomes II et III (1950 et 1952).

La correspondance entre Flaubet et Maupassant a été déchargée en mai 2009. Les lettres de Flaubert sur le site de l'Université de Rouen (http://flaubert.univ.rouen.fr/). Celles de Maupassant sur le site de T. Selva (http://maupassant.free.fr/).

Les autres œuvres utilisées dans ces expériences ont été déchargées sur les sites de l'Association des Bibliophiles Universels (ABU) et sur Gallica à la fin des années 1990.

Tableau 1. Le corpus des lettres de V. Hugo (1849-1870)

Années	Longueur (mots)	Formes	Vocables	Nombre de lettres	Longueur moyenne (mots)
1849	2 721	922	734	12	227
1850	2 146	768	603	11	195
1851	6 773	1 676	1 265	26	261
1852A	17 645	2 964	2 074	38	464
1852B	18 501	3 045	2 098	40	463
1853	17 337	3 290	2 314	30	578
1854	8 838	2 130	1 583	25	354
1855	19 105	3 481	2 464	56	341
1856	7 704	1 872	1 411	24	321
1857	4 682	1 298	1 007	23	204
1858	3 205	994	773	16	200
1859	7 363	1 889	1 419	26	283
1860	6 398	1 658	1 282	28	229
1861	8 103	1 978	1 498	26	312
1862	23 065	3 869	2 694	72	320
1863	8 468	2 095	1 609	40	212
1864	8 274	2 037	1 537	35	236
1865	10 096	2 240	1 666	45	224
1866	20 664	3 604	2 609	81	255
1867A	15 111	2 969	2 186	60	252
1867B	18 389	3 371	2 458	101	182
1868	17 650	3 180	2 364	100	177
1869A	13 910	2 723	2 054	60	232
1869B	13 673	2 768	2 066	77	178
1870	12 352	2 560	1 882	74	167
Total	292 173	17 085	10 001	1 126	259

Tableau 2 Corpus Victor Hugo

	Nombre de textes	Nombre de mots	Formes graphiques	Vocables
Poésie : *Contemplations*	6	91 890	10 154	5 946
Théâtre : *Hernani*, *Ruy Blas*	2	38 450	5 402	3 490
Roman : *Notre-Dame de Paris*	11	185 483	17 872	10 755
Roman : *Les Misérables* (tome1)	6	120 717	12 965	7 896
Correspondance	25	292 173	17 085	10 001
Corpus Hugo	53	734 046	35 887	19 138

Tableau 3. Corpus de la correspondance entre Flaubert et Maupassant.

	Longueur (mots)	Formes graphiques	Vocables différents
De Flaubert à Maupassant (1875-78)	7 361	1 733	1 341
De Flaubert à Maupassant (1879-80)	7 798	2 092	1 634
Total Lettres de Flaubert à Maupassant	15 159	3 061	2 299
De Maupassant à Flaubert (1875-78)	10 985	2 292	1 685
De Maupassant à Flaubert (1879-80)	10 749	2 172	1 588
Total Lettres de Maupassant à Flaubert	21 734	3 564	2 497
Total correspondance Flaubert – Maupassant	36 893	5 185	3 571

Tableau 4. Corpus des œuvres de Flaubert et Maupassant

		Date de publication	Longueur (mots)	Formes graphiques	Vocables
Gustave Flaubert					
Madame Bovary	Roman	1857	122 660	13 979	8 235
L'Education sentimentale	Roman	1869	152 890	16 196	9 648
Salambô	Roman	1862	109 378	11 883	6 569
Bouvard et Pécuchet	Roman	1881	95 238	14 686	9 210
Un cœur simple	Nouvelle	1877	12 161	3 291	2 531
Hérodias	Nouvelle	1877	10 599	3 177	2 453
Total Flaubert			502 926	31 956	16 939
Guy de Maupassant					
Une vie	Roman	1883	78 665	9 868	5 684
Bel Ami	Roman	1885	112 729	11 609	6 537
Mont Oriol	Roman	1887	85 905	10 166	5 854
Pierre et Jean	Roman	1888	45 560	6 816	4 231
Fort comme la mort	Roman	1889	76 869	9 593	5 484
Notre cœur	Roman	1890	60 589	8 499	5 108
Boule de suif	Nouvelle	1880	10 920	3 113	2 294
106 nouvelles	Nouvelles	1880-90	267 501	20 942	10 712
Total Maupassant			738 738	32 072	15 131