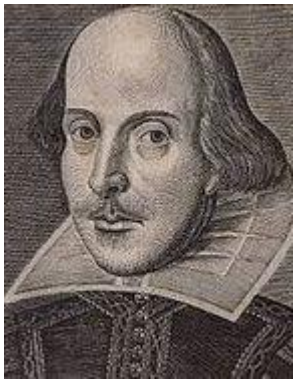
Ludovic Lebart



Textométrie et Poésie

Petits essais de Brassens à Shakespeare…

Les analyses exploratoires de données textuelles (qui constituent une branche particulière des techniques désignées plus récemment par Textométrie) ont montré leur utilité dans plusieurs domaines d'application.

a) Dans le cas de textes courts, nombreux et qualifiés, la situation -type pouvant être les traitements des réponses à des questions ouvertes dans les enquêtes par sondage. Plusieurs milliers de réponses à propos d'un thème précis, avec en parallèle des centaines de réponses à des questions fermées et de caractéristiques des répondants (méta-données) permettent des regroupements, des analyses contrastives de textes. Les analyses de tweets, de messages courts et nombreux entrent dans cette catégorie.
b) Dans le cas de textes longs, mais comparables d'un certain point de vue. Exemple type : corpus de romans, de discours politiques, séries textuelles chronologiques, répondant à une problématique externe (évolution, discontinuités, éventuelles attributions d'auteurs).

Dans ces deux cas, les fréquences lexicales sont des critères d'intérêt et de validation des résultats obtenus. Mais les ensembles de textes poétiques ne rentrent pas dans ces cadres.

Les deux chapitres qui vont suivre sont des contributions très ponctuelles au problème de la confrontation entre textométrie et poésie. Ils sont évidemment loin d'épuiser le sujet.

Le chapitre 1 repose sur un corpus très particulier : le recueil de 194 chansons chantées et enregistrées par le musicien-poète français Georges Brassens (1921 – 1981). La question posée sera : Est-ce que les outils de la statistique exploratoire multidimensionnelle peuvent s'appliquer à un recueil de chansons (avec toutes le contraintes du genre) et donner des éléments d'information nouveaux ?

Le chapitre 2 porte au contraire sur un corpus classique, très étudié, celui des 154 sonnets de Shakespeare. Il montre que dans le cas des recueils de poèmes formattés, la recherche de thèmes

(*Topic modeling*) peut donner des informations intéressantes à partir de plusieurs outils statistiques : il s'agit simplement de retrouver directement par une analyse descriptive (mais multidimensionnelle et sans *a priori*) du corpus les thèmes relevés par les experts au cours des siècles précédents.

Précisons qu'il ne s'agit pas d'étudier un vocabulaire spécifique d'une oeuvre, comme cela a été fait par M. Bernard (2000), ni d'étudier ou de modéliser les techniques de versification (Beaudouin, 2002), ni d'étudier la richesse et la structure lexicale (Labbé *et al.*, 1988 ; Brunet, 1988). Il ne s'agit pas non plus d'une exploration savante et assistée par ordinateur d'une œuvre iconique comme Viprey (2002) l'a fait avec *Les fleurs du mal* de Baudelaire. Enfin il ne s'agit pas non plus de stylométrie suivant les travaux de pionnier de Yule (1944) ou ceux synthétisés par Holmes (1985). Dans les deux cas traités ci-après, il s'agit le plus souvent (et tout simplement) d'analyses automatiques de contenus, et non de stylométrie pure (évolution des thèmes chez Brassens, repérage des thèmes pour Shakespeare). Mais le lien entre contenu et forme sera parfois étudié, parce qu'il peut surgir des analyses sans y avoir été invité.

Chapitre 1.
Poésies et chansons. Le cas de Georges Brassens.
Que peut la textométrie ?

Le corpus que nous nous proposons d'étudier ici est un matériau bien particulier, certes, mais aussi complexe et insaisissable : il s'agit du recueil de 194 chansons chantées et enregistrées par le musicien -poète français Georges Brassens (1921 – 1981).

Cet auteur a ceci de particulier qu'il réunit trois caractéristiques presque antinomiques : a) Il est non-conformiste et a côtoyé les mouvements anarchistes, b) mais il a reçu en 1967 le prix de poésie d'une institution plutôt conservatrice, l'Académie Française. c) et il a été responsable, de son vivant ou après son décès, de la vente de plusieurs dizaines de millions de disques (données de 2021).

1.1 Problèmes et défis des textes poétiques

Nous commencerons par montrer en quoi l'analyse statistique de ce type de textes constitue un défi méthodologique, puis nous décrirons les pré-traitements, les procédures statistiques utilisées, et évoquerons les premiers résultats obtenus.

On gardera en tête l'avertissement de Brunet (2004) lors d'une remarquable étude statistique de l'œuvre poétique d'Arthur Rimbaud, avertissement qui s'applique à l'étude de Brassens : « Au reste, l'emploi des outils documentaires et statistiques ne va pas sans une certaine naïveté qui donne sa foi aux mots imprimés, dans leur innocence première. Or chez Rimbaud, les mots sont souvent pipés. Ce sont des leurres, des prête-noms, et la réalité qu'ils désignent et qu'ils cachent se dérobe aux plus savantes exégèses. Quand l'ésotérisme multiplie les pièges, comment s'assurer de la constance sémantique des termes ? ». Les œuvres étudiées de Rimbaud comportaient environ 40 000 occurrences. Le corpus de chansons de Brassens traité ici, d'un volume assez comparable, contient environ 52 000 occurrences.

On note effectivement que les textes poétiques de Brassens sont particulièrement riches en figures de style (litotes, métaphores, anaphores, euphémismes, allégories, …) qui sèment des doutes sur l'utilisation du mot (graphie ou lemme) comme unité statistique de base. Ce poète est expert dans l'art de détourner des locutions (il parle de la « face cachée de la lune de miel », de l'« évangile selon Vénus », ou déclare : « la loi de la pesanteur est dure, mais c'est la loi »). (Lamy, 2004 ; Poulanges et Tilleu, 2001). Il remet au goût du jour des expressions populaires, désuète ou argotiques (« le café du pauvre » pour : acte sexuel, etc.). Il fait souvent appel à des usages de termes historiques, moyenâgeux, voire obsolètes (Rochard, 2009). Dans le cas des chansons pouvant comporter des refrains ou des répétitions partielles, les fréquences lexicales n'ont plus la signification statistique qu'on leur donne dans les tableaux lexicaux. Les contraintes de versification (alexandrins par exemple, rimes) sont difficiles à intégrer dans les outils de description, mais influencent le choix des mots et leur fréquence.

Les textes sont trop courts et le corpus trop restreint pour espérer en extraire automatiquement de nouvelles unités lexicographiques comme des segments répétés (Salem, 1987), des motifs, des locutions.

La question que nous nous posons est extrêmement étroite et technique par rapport aux travaux existants ou potentiels d'analyses littéraires et musicales des œuvres de ce musicien-poète : « Est-ce que les typologies et les visualisations obtenues à partir des profils lexicaux (environ 1000 graphies ou lemmes) des 194 chansons, confrontées aux métadonnées disponibles, apportent des informations nouvelles, des traits structuraux notables ou des matériaux nouveaux susceptibles d'intéresser les spécialistes ? ».

1.2 Préparation des textes et techniques de base

Les 194 unités du corpus seront décrites par autant de groupes de mots (graphies) et de groupes de lemmes qui sont des unités statistiques complémentaires. Elles sont par ailleurs décrites par les métadonnées suivantes : l'appartenance à un recueil d'albums (disques) selon 14 modalités, variable qui a une composante chronologique de publication (et souvent de composition/création), l'auteur selon 2 modalités (« Brassens » [170 chansons] ou « Autre », choisies et chantées par Brassens [24 chansons]), enfin la tonalité dominante (12 modalités) de la partition musicale correspondante. La prise en compte de la tonalité n'est ici qu'une esquisse, une tentative. Le traitement des partitions musicales complète par analyse des correspondances a été étudié notamment dans les travaux de pionniers de Morando (1980), puis de Cocco (2014).

Qu'il s'agisse des formes graphiques ou des lemmes (pour lesquels on a utilisé le logiciel TreeTagger [Schmid, 1994]), on va transformer chaque texte de chanson en vocabulaire non pondéré, autrement dit, chaque élément n'apparaitra qu'une seule fois à l'intérieur d'une chanson donnée (quelques lignes de Python impliquant l'objet « dictionnaire » de ce langage permettent aisément cette transformation/réduction du texte). On aura donc deux jeux de données provenant d'une part du fichier brut, d'autre part du fichier lemmatisé. Le fichier lemmatisé a l'avantage de réduire la diversité des flexions et donc de permettre des seuils de fréquence minimale plus bas. Le fichier des graphies garde la diversité originale des formes ce qui est fondamental dans le cas de textes poétiques. On obtiendra donc à chaque étape deux points de vue différents et complémentaires.

Dorénavant, les tableaux lexicaux (chanson x mots) seront ainsi des tableaux de présence – absence. On sait que pour ce type de tableau, le coefficient de corrélation r entre deux chansons coïncide avec le coefficient d'association Φ de Yule (1912). Il est par ailleurs lié au χ^2 calculé sur la table de contingence (2 x 2) croisant les deux textes décrits par *n* mots par la formule :

$$r^2 = \phi^2 = \frac{\chi^2}{n}$$

Les techniques utilisables pour décrire et détecter d'éventuelles structures seront l'Analyse en composantes principales (licites sur ces codages binaires), l'analyse des correspondances (AC), les classifications par arbre additifs (Buneman, 1971 ; Saitou & Nei, 1987), les cartes auto-organisées. Les métadonnées définies plus haut interviendront comme variables actives (regroupement de chansons suivant les albums, par exemple), mais surtout comme variables supplémentaires (projection *a posteriori*, avec validation *Bootstrap*). Le logiciel utilisé, libre d'accès, a été DtmVic (www.dtmvic.com)[1]. Rappelons que dans le cadre du codage binaire utilisé ici, les valeurs propres n'ont pas d'interprétation simple en termes d'information. Seules

[1] DtmVic fait aussi appel aux logiciels TreeTagger [Schmid, 1994] et SplitsTree [Huson et Briand, 2006].

les validations statistiques (utilisant ici le *Bootstrap*) permettent de juger la validité statistique des axes.

1.3 Chronologie et recueils (disques, albums)

L'arbre additif de la figure 1.1 donne une synthèse des liens entre les disques (ou albums) repérés par leur année de publication. Les deux derniers (années 1982_13 et 1985_14) sont posthumes. Les 4 premiers (de 1953_1 à 1958_4) constituent la partie basse droite de l'arbre. Ces albums peuvent comporter des chansons composées antérieurement, ou des assortiments suggérés par l'éditeur, ou même, comme les albums posthumes, des œuvres de jeunesse retrouvées et interprétées par d'autres. Malgré cela, il y a quand même une certaine compatibilité entre les proximités lexicographiques décrites par ce graphique et la chronologie. L'opposition entre les premières années et les suivantes se retrouvera dans toutes les analyses, qu'il s'agisse de lemmes ou de graphies.

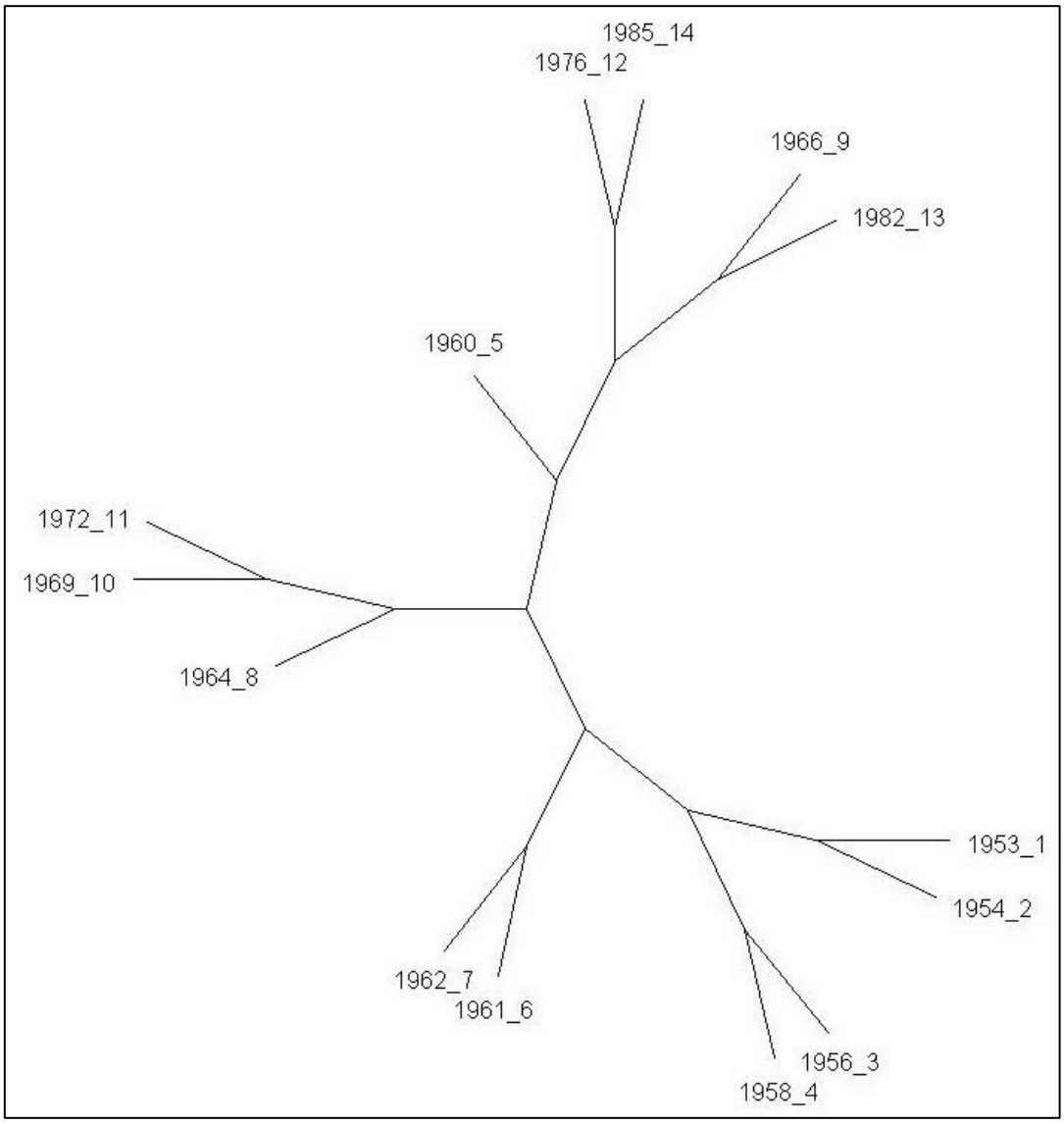


Figure 1.1. Arbre additif des 14 albums Brassens (distances calculées sur la présence-absence dans chaque chanson de 943 formes graphiques).

1.4 Brassens et les poètes qu'il a choisi

Les 24 « poèmes externes » mis en musique et chantés par Brassens sont une mini-anthologie très proche de ses valeurs et de ses goûts.

La figure 1.2 montre en effet le plan (1, 2) de l'AC de la table croisant 194 chansons et 901 lemmes (apparaissant au moins quatre fois).

Les points « auteurs » et les deux années choisies sont projetés *a posteriori* dans le plan de l'analyse précédente, et les réplications Bootstrap s'obtiennent par des tirages avec remise dans les 194 chansons (*cf.* Lebart, 2004, 2007) . Les auteurs « externes » (Hugo, Lamartine, Paul Fort, Musset, etc.) sont plus proches des premières années.

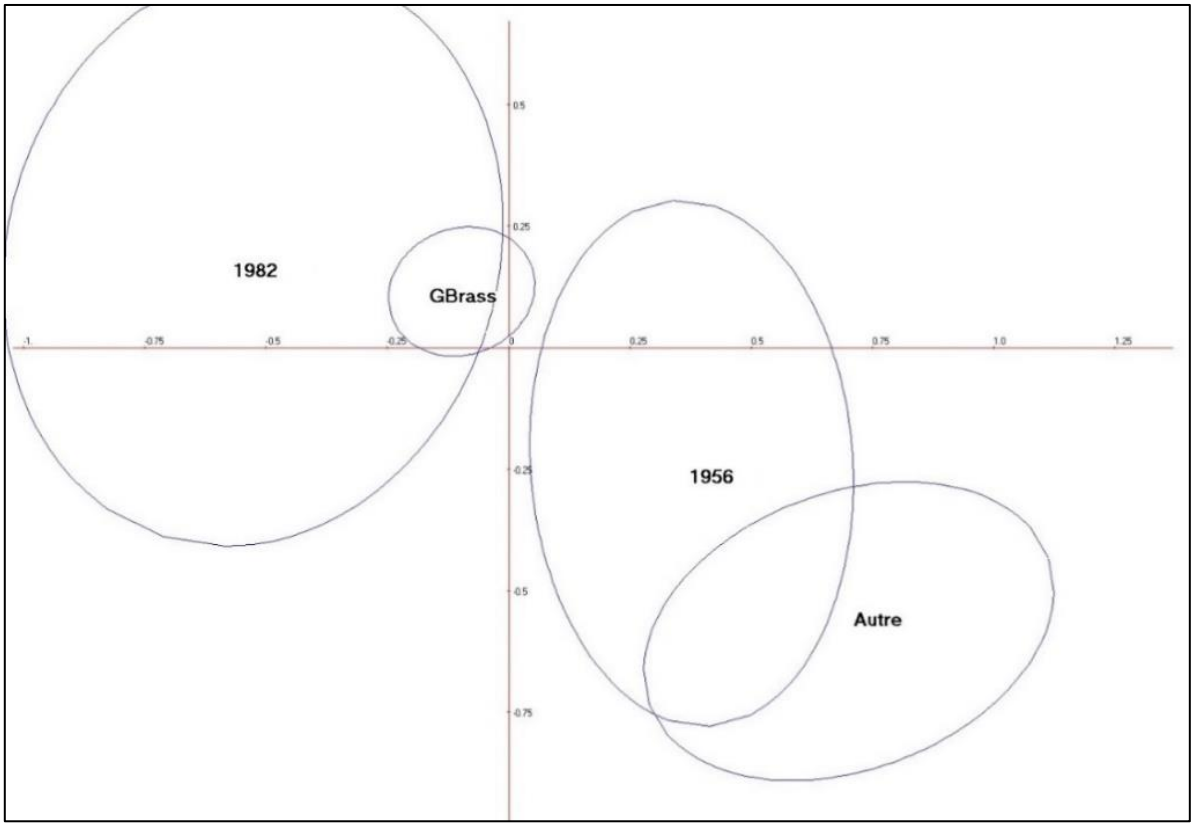


Figure 1.2. Zones de confiance Bootstrap de quatre catégories supplémentaires. GBrass (170 chansons écrites par Brassens) et Autre (24 poèmes « externes » de divers auteurs mis en musique et chantés par Brassens). Zones de confiance bootstrap de 2 albums (1982 et 1956).

En fait, on pourrait voir sur les cartes factorielles avec la position des 901 lemmes (impubliable sous ce format de publication, mais on peut en avoir une idée de cette complexité en consultant l'annexe 2 de ce chapitre) que le vocabulaire s'est durci au cours du temps, l'auteur se qualifiant lui-même de « pornographe » (à partir de l'album 5).

Parallèlement, la censure, active pour les premiers albums, s'est montrée plus tolérante au cours de la période.

1.5 Les tonalités

Brassens était un compositeur-interprète en même temps qu'un poète, et l'on pouvait légitimement se demander s'il y avait un lien entre les tonalités dominantes et les caractéristiques des chansons représentées par leurs profil lexicaux.

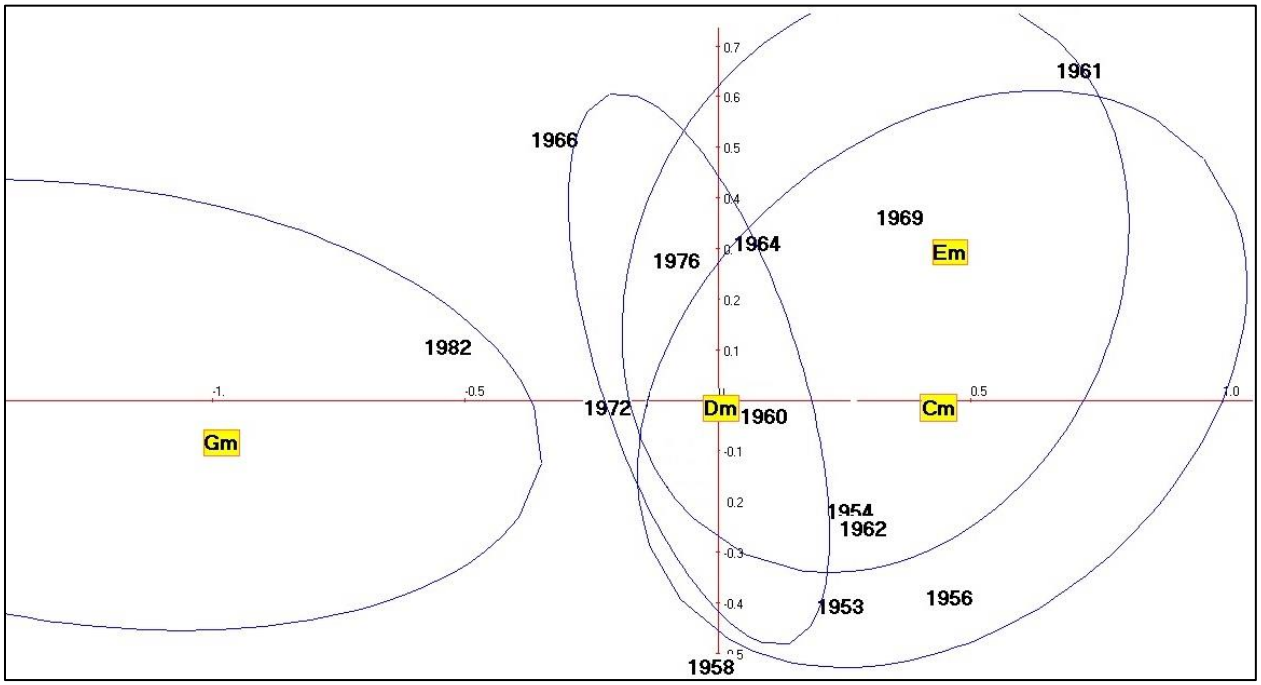


Figure 1.3. Zones de confiance Bootstrap de 4 tonalités (Gm, Dm, Em, Cm) (Sol_min, Ré_min, Mi_min, Do_min) parmi 13 dates d'albums (tonalités et albums sont toujours des éléments supplémentaires dans le plan (1, 2) d'une AC du tableau binaire (194 chansons x 901 lemmes).

Dans ce plan de l'AC du tableau croisant 194 chansons et 901 lemmes (figure 1.3), les zones de confiance sont indiscernables à l'exception de Sol mineur (Gm) sur la gauche (seules 4 zones sont publiées ici pour la lisibilité). Toutefois, cette exception ne concerne que 5 chansons, dont 3 ont été interprétées par d'autres chanteurs de façon posthume. On conclura, en l'état actuel de la recherche, qu'il n'y a pas de lien fragrant entre la tonalité choisie et le profil lexical des chansons.

1.6 Confrontations entre lemmes et formes graphiques

Cette phase est la partie la plus indispensable lorsque l'on a affaire à un texte poétique, puisqu'il s'agit de faire rivaliser le fond et la forme. On travaillera sur les 170 textes écrits par le chanteur.

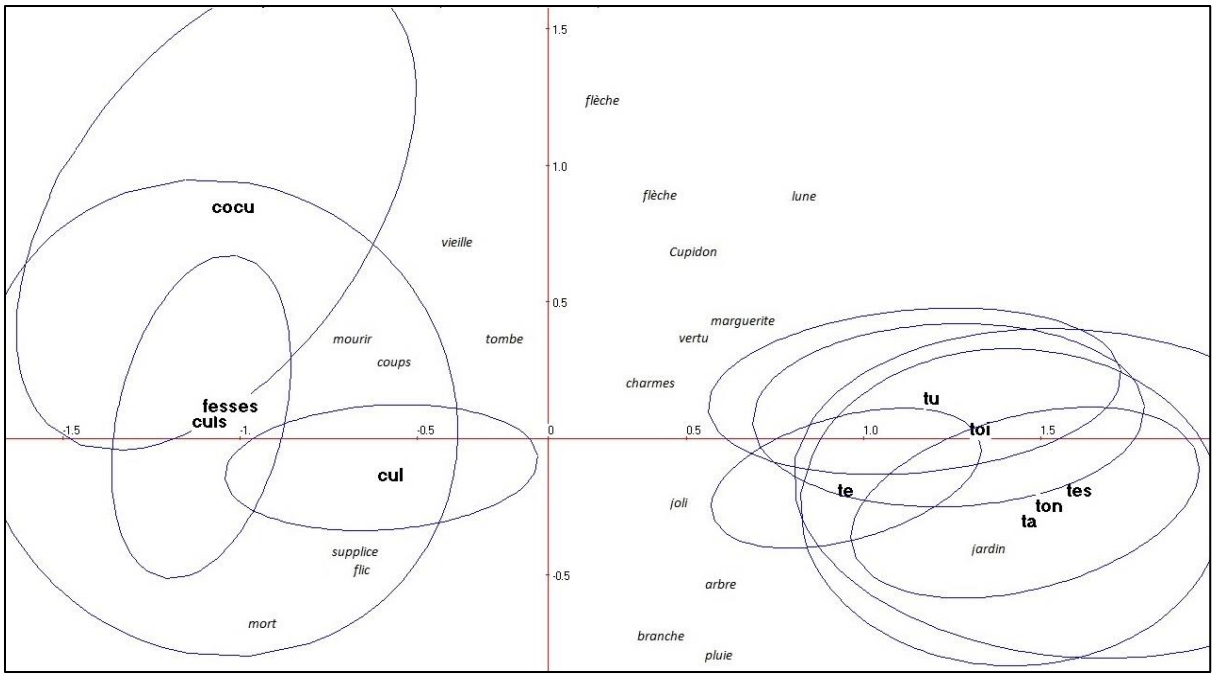


Figure 1.4. Graphies chez Brassens seul (170 chansons x 872 graphies). Opposition dans le premier plan de l'AC entre les formes de tutoiement (te, tu, toi, ton, ta, te, *à droite*) liées à une poésie raffinée, et le « vocabulaire plus scabreux » (*à gauche*). Quelques zones de confiance Bootstrap et illustration du plan par quelques graphies.

Pour la plupart des analyses de type AC portant sur l'ensemble des 194 chansons, ou sur les 170 chansons dont Brassens est le seul auteur, une première dimension (horizontale sur tous les graphiques), domine, qu'il s'agisse de lemmes ou de graphies : Elle oppose les textes anciens (4 ou six premiers albums) aux textes plus récents. Les textes anciens, comme les poèmes externes, ont un vocabulaire que l'on peut qualifier de classique, voire galant ou précieux, pour forcer le trait. Les plus récents ont un vocabulaire plus cru, parfois argotique, provocateur, « salle de garde ». La figure 1.4, qui concerne les 170 chansons entièrement écrites et composées par Brassens, ne comporte qu'un petit extrait des graphies actives.

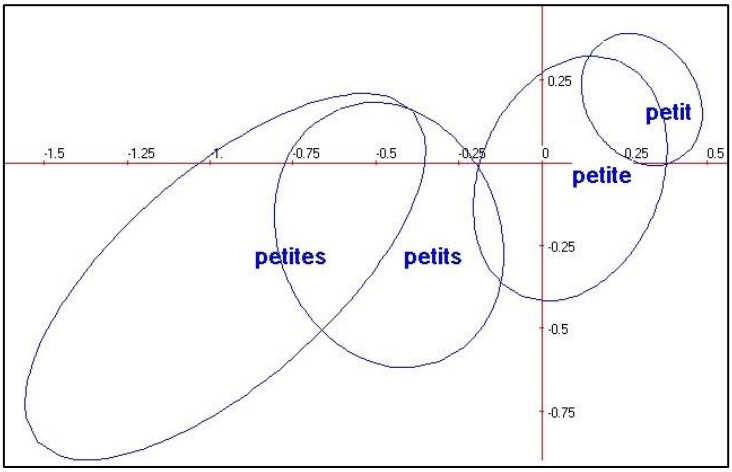


Figure 1.5. Zones de confiance Bootstrap de quatre flexions de l'adjectif « petit » dans le plan principal de l'AC du tableau de présence absence croisant 194 chansons et 943 formes graphiques.

Elle montre que le tutoiement (six formes concernées, à droite) ne signifie pas que familiarité, mais surtout intimité, douceur chez Brassens. Ici, les graphies renforcent l'interprétation de cet axe, que l'on obtient par ailleurs avec les lemmes. Outre les formes (toi, tu, ton, te, ta) on trouve

aussi parmi les 20 points les plus à droite sur cet axe les mots (sabots, fontaine, matin, jupon, jardin, pluie, belle, arbre, feuille, baiser).

Les graphies jouent un rôle de verre grossissant pour l'interprétation, mais elles font apparaître des vocables que la lemmatisation fait disparaitre, comme le montre la figure 1.5. Le lemme « petit », qui remplace les quatre flexions positionnées dans un premier plan factoriel similaire aux précédents, occupe une position centrale qui, en quelque sorte, le neutralise dans l'analyse sur les lemmes. Ces différentes flexions sont pourtant liées à cette dimension principale qui oppose, on l'a vu, l'intime et le délicat (singuliers) à l'impersonnel et au paillard (pluriels).

1.7 Conclusion du chapitre 1

Malgré le nombre limité de graphiques présentés (forcément de petite taille), on peut deviner que le traitement textométrique (descriptif multivarié) des textes poétiques apporte un point de vue original sur ces textes et aussi de nouveaux matériaux d'études pour les spécialistes. Dès ces premières tentatives d'analyse, on a pu déceler une tendance générale, inextricablement liée à l'âge, à la carrière, à l'évolution personnelle, peut-être à la notoriété croissante du poète et probablement à la permissivité croissante au cours de la période considérée. A chaque fois, l'utilisation de formes amplifie, illustre et nuance les résultats obtenus à partir des lemmes. Les graphes inédits des grands arbres additifs représentant les liens entre les 194 chansons, les affichages des plans principaux de CA impliquant environ 900 mots-formes et autant de lemmes ainsi que les chansons et les disques, constituent aussi un ensemble de documents de travail (dont la figure 1.7 de l'annexe 2 de ce chapitre constitue un exemple) plein de potentiel, difficile à publier sur papier, mais fascinant à consulter pour les spécialistes ou les amateurs concernés. On risquera de parodier l'adage « Garbage in, garbage out » (adage d'ailleurs très mal adapté aux techniques d'analyse de données qui ont pourtant une fonction de filtrage) en le transformant en « Poetry in, poetry out ». En effet, ces nouveaux documents décrivant les liens et les schémas complexes entre des centaines de mots soigneusement choisis par le grand troubadour Brassens sont eux-mêmes – au moins pour ses fidèles admirateurs – une source d'émotions poétiques.

Remarques sur les annexes :

La première annexe qui suit (Figure 1.6) montre les classements des mots les plus extrêmes sur le premier axe d'une analyse des correspondances de la table lexicale croisant 703 mots (ici : lemmes, seuil de fréquence = 6) et les 194 chansons analysées. Ce premier axe se maintient avec la même opposition si on ne considère que les 170 chansons écrites par Brassens seul (texte et musique), ou si on fait varier les seuils de fréquence minimales de 5 à 30. De façon duale, la technique permet également de classer les chansons (partie droite du tableau). Cette opposition le long de l'axe horizontal au niveau des lemmes était déjà visible, et peut-être même plus caricaturale en raison de la richesse des graphies, sur la figure 1.4.

La seconde annexe (Figure 1.7) présente le plan principal de la même analyse des correspondances (AC) dont le tableau de la figure 1.6 ne faisait qu'exquisser le premier axe (axe horizontal).
Les points les plus extrêmes on été ramenés vers le cadre (flèches).

Annexe 1 du chapitre 1 : (Figure 1.6). Tableau décrivant les positions des mots et des chansons sur le premier axe de l'analyse des corrrespondances (AC) du tableau lexical croisant les 703 mots les plus fréquents et les 194 chansons.

Description du premier axe de l'A.C. du tableau lexical (mots x chansons) à partir des éléments extrêmes classés

Partie gauche de l'axe 1 *(classement des mots)*

Identifier	axis 1
montagne	-2032
sabot	-1940
vilain	-1929
clocher	-1764
étonner	-1649
grain	-1562
forêt	-1554
croquant	-1552
envie	-1534
jupon	-1424
exister	-1402
jardin	-1383
fontaine	-1368
suffire	-1269
arbre	-1258
feuille	-1250
belle	-1199
longtemps	-1144
branche	-1129

Partie droite de l'axe 1 *(classement des mots)*

Identifier	axis 1
difficile	1695
croupe	1486
adultère	1371
cocu	1358
mufle	1236
fesse	1103
con	1042
public	1034
est-c	1016
endroit	964
tromper	919
Français	913
propos	878
train	856
cas	850
quat	824
chance	819
maman	811
dégueulasse	806
fossoyeur	792
chêne	791

Partie gauche de l'axe 1 *(classement des chansons)*

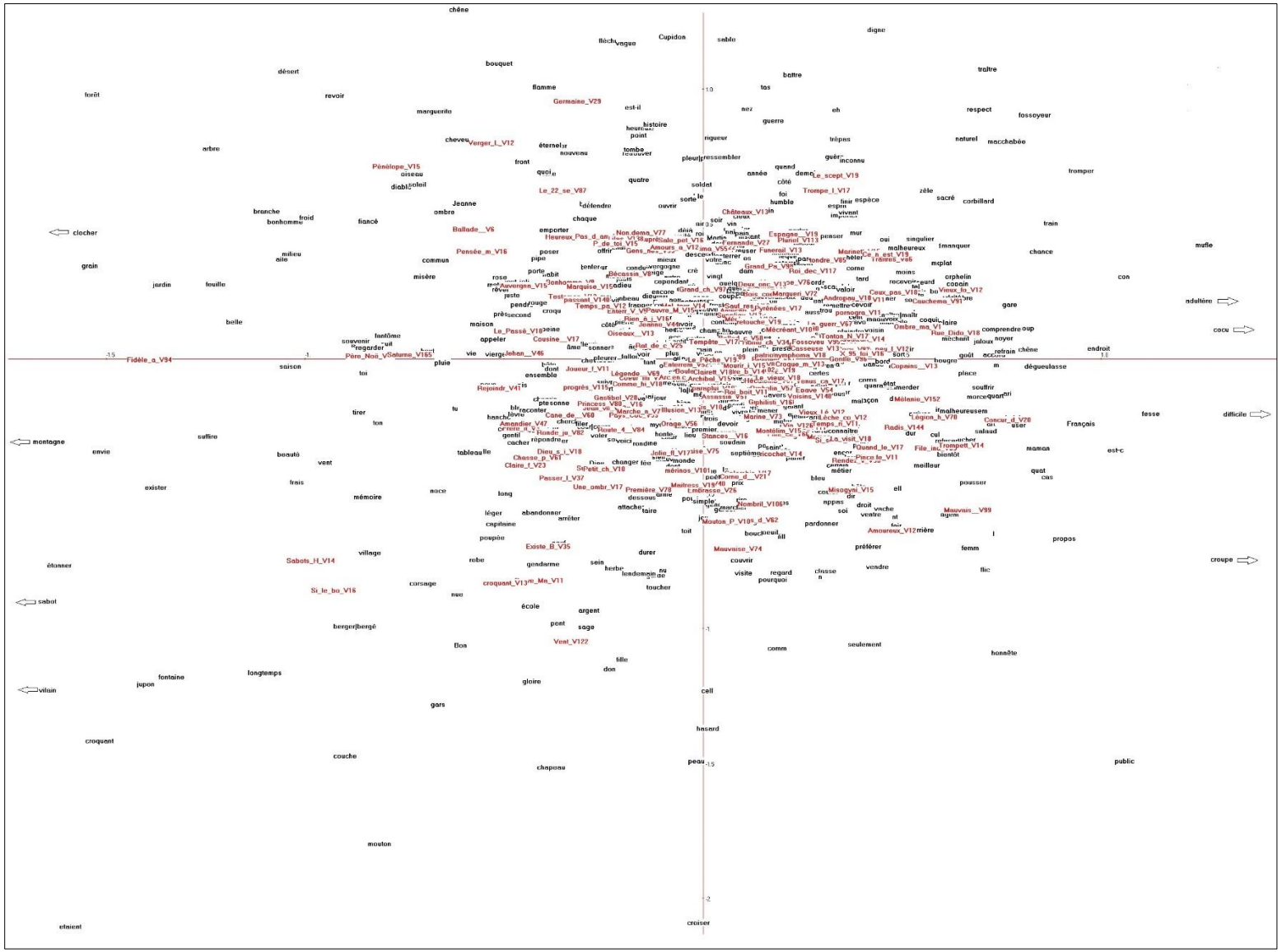
Identifier	axis 1
Fidèle_a_V94	-1449
Sabots_H_V14	-1047
Si_le_bo_V16	-986
Père_Noë_V10	-898
Pénélope_V15	-830
Saturne_V165	-794
Ballade__V6	-629
Pensée_m_V16	-620
Verger_L_V12	-589
Rejoindr_V41	-568
croquant_V13	-552
Le_Passé_V18	-526
Auvergna_V15	-510
Amandier_V47	-509
Frère_It_V3	-504
Jehan__V46	-500
Claire_f_V23	-499
Chasse_p_V61	-476
Brave_Ma_V11	-474
Existe_B_V35	-444
Cousine__V17	-428
Ronde_ju_V82	-418
Dieu_s_i_V18	-416

Partie droite de l'axe 1 *(classement des chansons)*

Identifier	axis 1
Concur_d_V20	707
Mauvais__V99	606
Vieux_fo_V12	592
Trompett_V14	591
Rue_Dido_V18	574
File_ind_V65	532
Cauchema_V91	526
Légion_h_V70	523
Mélanie_V152	481
Ombre_ma_V1	479
Copains__V13	470
Radis_V144	456
Traitres_V86	421
Ceux_pas_V18	418
Amoureux_V12	414
Ce_n_est_V19	400
Un_peu_l_V12	400
Quand_le_V17	384
Pince.fe_V11	382
Quatzart_V14	347
X_95_foi_V16	345
Marinett_V15	340

(Le codage des intitulés des chansons, un peu complexe, permet cependant une identification. Il sera amélioré dans une édition ultérieure)

Annexe 2 du chapitre 1 (Figure 1.7) Plan principal (1, 2) de l'AC précédente.**(Mots :** noir**, Chansons :** rouge**)**



Chapitre 2.

Recherches de thèmes (*Topic Modeling*) dans les Sonnets de Shakespeare.

Le domaine des recherches de thèmes occupe une position intermédiaire entre l'exploratoire et le confirmatoire : les procédures exploratoires sont en quelque sorte en concurrence avec des modèles qui doivent être validés. Certains des résultats présentés ci-dessous ont fait l'objet de publications antérieures (cf. par exemple Lebart, 2018 ; Lebart , Pincemin et Poudat, 2019).

2.1 Introduction

Cette section présente un bref survol de plusieurs tentatives d'identification de variables latentes (axes ou classes) dans le cas des données textuelles. Ces variables latentes (classes ou axes) sont parfois désignées *ex ante* par le terme « *topic* ». L'analyse factorielle des psychologues au début du siècle dernier[1] était déjà une tentative d'identification de variables latentes interprétables. Elle supposait un modèle, au départ mono-factoriel (une seule variable latente, le facteur général d'aptitude, ou intelligence, pour expliquer une batterie de notes), puis multifactoriel (intelligence, mémoire, puissance de travail…). Il s'agit donc bien au départ d'estimer un modèle, bien que cette estimation n'ait été rattachée à la statistique inférentielle classique que beaucoup plus tard par Lawley et Maxwell (1963). Mais, comme la plupart des modèles de variables latentes, il s'agit d'approche non-supervisée : les variables cachées ne sont pas observables directement mais fournies par le modèle lui-même. Exprimé de façon plus concrète : dans les équations qui définissent le modèle, ce qui est connu est d'un seul côté du signe « = » (contrairement à la régression ou à l'analyse discriminante, pour lesquelles on explique une mesure par une autre mesure, au moins pour les échantillons d'apprentissage : il y a donc des observations des deux côtés du signe « = »). L'aventure de l'analyse factorielle des psychologues ne s'arrête pas là, car le modèle est devenu lui-même un instrument d'observation dans les mains des praticiens, avant même que l'on montre qu'il était très proche de l'analyse en composantes principales.

Les dernières années ont été témoin d'une série de tentatives algorithmiques telles que la factorisation matricielle non négative (*Non-negative Matrix factorization* : NMF) ou l'allocation de Dirichlet latente (*Latent Dirichlet Allocation* : LDA). Simultanément, les thèmes considérés comme des variables latentes ont pu également être identifiés à travers plusieurs hybridations et synergies des méthodes en axes principaux et des techniques de classification.

[1] Cette méthode désignée aussi sous le nom de *Analyse en facteurs communs et spécifiques*. Le titre de l'article pionnier de Spearman (1904): *General intelligence, objectively determined and measured* montre l'ambition assez démesurée de la méthode.

Il y a une profusion de nouvelles disciplines autour des applications industrielles impliquant des textes, avec des proliférations subséquentes d'outils et des disparités de terminologies. Il existe également des disparités dans l'attitude envers les textes, parfois influencée par la disponibilité et la convivialité des logiciels. Les problèmes posés par d'énormes recueils de *newsgroups* ou de *tweets* sont très différents de ceux rencontrés dans les domaines de la littérature, des discours politiques et des enquêtes psychologiques.

Dans cette section, un seul corpus classique de taille moyenne servira de corpus référence pour esquisser et comparer de manière compacte certaines caractéristiques de plusieurs méthodes. Parce qu'ils sont bien connus, traduits dans presque toutes les langues, profondément étudiés et commentés, nous utiliserons les 154 Sonnets de Shakespeare comme corpus de référence pour comparer brièvement l'aptitude de plusieurs techniques à reconnaître des thèmes dans un corpus.

2.2. Un aperçu du contenu des sonnets de Shakespeare

Les 154 sonnets de William Shakespeare traitent de thèmes tels que l'amour, l'amitié, les effets du temps, la beauté, la trahison, la luxure, la mort [2].

Thème, Sujet, Motif

Notons que la définition des thèmes est pragmatique et peut également recouvrir les concepts de sujet et de motif (au sens littéraire). Habituellement, le sujet est l'explication objective du contenu d'un texte, alors qu'un thème peut représenter un message sous-jacent plus profond. Un motif, lui, est simplement une idée récurrente utilisée pour renforcer le thème principal. Schématiquement, les *sujets* répondent aux questions : « De quoi parle l'histoire, qui, quoi, comment ? », et les *thèmes* répondent plutôt à : « Pourquoi l'histoire a-t-elle été écrite ? ». Les sujets de la littérature sont plus faciles à identifier que les thèmes.

Trois séries contiguës de sonnets sont généralement reconnues comme correspondant à trois sujets dominants :

Sonnets 1 à 17: (*Procreation*). Ces sonnets célèbrent la beauté d'un jeune homme qui est pressé par le poète de se marier pour perpétuer cette beauté.

Sonnets 18 à 126: (*Young Man*). Cette séquence la plus longue concerne le même jeune homme (non définitivement identifié), l'effet destructeur du temps, la force de l'amour, de l'amitié et de la poésie.

Sonnets 127 à 154: (*Dark Lady*). Ces sonnets sont surtout adressés à une femme aux cheveux sombres. Ils ne sont pas dépourvus d'ironie ni de cynisme (les deux derniers sonnets 153 et 154 sont des épigrammes spécifiques dans un style ancien, et mériteraient en fait une catégorie spéciale).

Huit thèmes inspirés par des commentaires d'experts

Les thèmes *Young Man* et *Dark Lady* pourraient eux-mêmes contenir cinq sous-thèmes. Alors que le premier thème (*Procreation*) reste tel qu'il est, les nouveaux thèmes *Young Man* et *Dark Lady* ne comprennent plus que les sonnets qui ne sont pas assignés aux cinq nouvelles catégories ci-dessous (*Absence, Storm, Rivalry, Death, Eternal poetry*).

---

[2] On peut consulter une version française des Sonnets de Shakespeare par F. Henry (1900) sur le site de Gallica (Bnf) : http://gallica.bnf.fr/ark:/12148/bpt6k1310005 (avec introduction et commentaires). Pour une version anglaise voir Shakespeare (1901).

Tableau 2.1. Liste de huit thèmes / topics a priori avec les numéros de sonnets correspondants [Auteurs]

Procreation	1 - 17
YoungMan	20-25, 33-38, 40-42, 46, 47, 49, 53-55, 59-60,62-70, 75-77, 88-106, 108-112, 115-125,
DarkLady	127-136, 139, 140, 143-146, 153,154
Absence	26-32, 39, 43-45, 48, 50-52, 56-58, 61, 113-114
Storm	141,142,147-152
Rivalry	78-87
Death	71-74
Etern_poetry	18, 19, 81

La partition des sonnets donnée dans le tableau 2.1 est inspirée des travaux d'Alden (1913) et de Paterson (2010) mais pas explicitement mentionnée par ces auteurs.

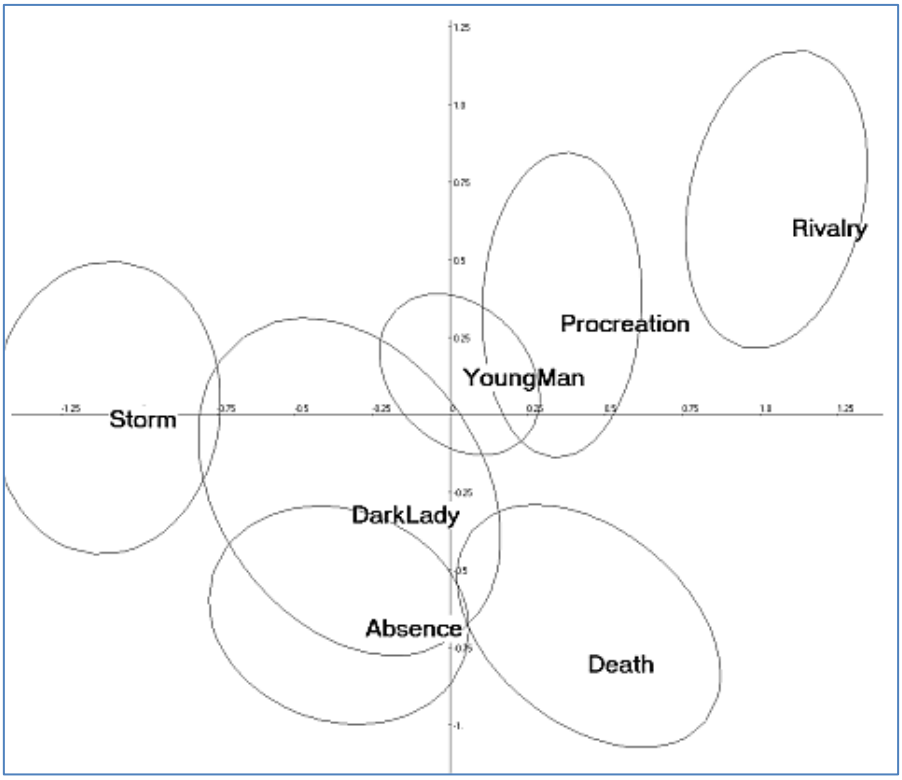


Figure 2.1. Emplacements de 7 thèmes / topics a priori dans le plan principal de l'analyse des correspondances de la table lexicale (154 sonnets x 173 mots), [fréquence minimale = 10]

Les thèmes ont le statut de variables nominales supplémentaires, et sont projetés a posteriori dans ce plan. Les ellipses de confiance « sévères » [validation spécifique : tirage avec remise des sonnets entiers] montrent des distances significatives entre plusieurs paires de thèmes a priori. Le thème Eternal Poetry, qui chevauchait plusieurs thèmes, est absent de cet affichage graphique.

La figure 2.1 montre cependant qu'après une analyse des correspondances ignorant ces thèmes, la plupart de leurs localisations, après projection avec le statut de **catégories supplémentaires**, sont statistiquement significatives sur le plan principal de visualisation.

Evidemment, les tentatives suivantes de mises en évidence automatique des *thèmes* dans le corpus des sonnets vont ignorer cette partition *a priori* en thèmes. Nous n'espérons pas non plus retrouver automatiquement ces thèmes. Cependant, la connaissance de ces thèmes issus de la critique littéraire nous fournira une grille de lecture et d'interprétation des résultats.

Notons que les outils statistiques, fondés principalement sur les fréquences, détectent presque indifféremment des *sujets*, des *thèmes* ou des *motifs*. Nous utiliserons principalement le terme « thème » par la suite.

2.3. Six méthodes sélectionnées pour la recherche de thèmes

Parmi les six procédures sélectionnées dans la présente application, quatre (RFA, FCA, ALO, LSA) utilisent la décomposition en valeurs singulières (SVD). Les deux méthodes restantes (NMF, LDA), moins géométriques, utilisent des algorithmes plus complexes (et parfois beaucoup plus longs en terme de temps de calcul).

La RFA (*Rotated Factor Analysis*) est historiquement la première tentative d'identifier des «facteurs latents» non observés (Thurstone, 1947, après les articles pionniers de Spearman, 1904, et Garnet, 1919). La RFA implique la SVD dans certains des algorithmes utilisés pour estimer le modèle. Les *thèmes* seront définis par les mots caractérisant chacun des facteurs conservés après une rotation des axes destinée à faciliter leur interprétation. Initialement conçu pour des valeurs numériques, la méthode peut être adaptée à des tables de fréquences clairsemées ou « creuses » (*sparse matrices*). [librairies **R** 'psych' et 'GPArotation'].

La FCA (*Fragmented Correspondence Analysis*), est fondée sur l'Analyse des correspondances de fragments de textes [dans notre cas 7 lignes consécutives, soit un demi-sonnet). Le principe de cette fragmentation en unités de contexte a été proposé initialement par Reinert (1983, 1986a) dans son logiciel ALCESTE. (*cf.* Ratineau et Déjean, 2009). Les axes principaux de l'AC servent à regrouper ces fragments, ici avec une classification hybride utilisant une Classification Ascendante Hiérarchique (critère de Ward), la coupure de l'arbre étant optimisée par agrégation autour de centres mobiles. À la fin du processus, les *thèmes* sont définis par les mots caractéristiques de chaque classe de fragments.

La ALO (Analyse LOgarithmique) (Kazmierczak, 1985) est similaire au Spectral Mapping (Lewi, 1976) à une différence de pondération près. Les deux méthodes, comme l'AC, respectent le principe d'équivalence distributionnelle (stabilité des résultats vis-à-vis des fusions de colonnes ou de lignes similaires). Appliquée aux tables lexicales, la ALO produit souvent des résultats similaires à ceux de l'AC, avec moins de sensibilité aux valeurs aberrantes, effet attendu de la transformation logarithmique. Le calcul est opéré sur les sonnets, et un regroupement (similaire à celui de la FCA) est ensuite effectué. Les *thèmes* sont alors les mots caractérisant chaque groupe de sonnets.

La LSA (*Latent Semantic Analysis*) est une SVD appliquée au tableau des coefficients TF-IDF (fréquence d'un terme **x** inverse de la fréquence des documents contenant le terme). Cette technique remonte aux travaux de Furnas et al.(1988), Deerwester et al. (1990), Bartell et al. (1992). Ici les documents sont les sonnets. Un regroupement (similaire à ceux de la FCA et de la ALO) est ensuite effectué. Les *thèmes* sont alors les mots caractérisant chaque groupe de sonnets (voir la librairie **R** : 'lsa', par F. Wild).

Dans le domaine de l'analyse de texte, les deux méthodes suivantes appartiennent plus spécifiquement au domaine de la recherche de thèmes (*Topic Modeling*).

La NMF (factorisation matricielle non négative : *Non-negative Matrix Factorization* :) se fonde au départ sur une équation qui rappelle la SVD avec cependant une contrainte de positivité des coefficients (Lee et Seung, 1999, 2001 ; Berry *et al.*, 2007, d'après Paatero et Tapper, 1994 ; voir aussi Boutsidis et Gallopolous (2008), et, pour un programme **R** : Gaujoux, 2010). Dans le contexte de la recherche de thèmes, le résultat principal de la NMF est un ensemble de *thèmes*, chacun d'entre eux étant caractérisé par une liste de mots (logiciel «scikit-learn» [*Python*] de Grisel O., Buitinck L., Yau CK, In: Pedregosa *et al.* 2011).

La LDA (Allocation latente de Dirichlet : *Latent Dirichlet Allocation*) (Blei *et al.*, 2003; Griffiths *et al.*, 2007) est un modèle statistique génératif (impliquant des *thèmes*, des mots et des documents latents) conçu pour découvrir la structure sémantique d'une série de textes ou

documents (supposés être un mélange d'un petit nombre de *thèmes*). La méthode est fondée sur une analyse bayésienne hiérarchique des textes. (librairie **R**: 'topicmodels', et logiciel 'scikit-learn' [*Python*]).

A ce stade, nous avons donc limité notre investigation à six techniques issues d'un grand nombre d'approches susceptibles d'identifier des thèmes. On aurait pu utiliser également l'enchaînement de l'analyse des correspondances directe (sans fragmentation des textes) avec une classification des sonnets, tout l'éventail des techniques de classification appliquées directement aux sonnets (les classes obtenues étant toujours caractérisées par leurs mots les plus caractéristiques). On aurait pu aussi utiliser la méthode ALCESTE précitée.

Le travail présenté peut évidemment être étendu à l'envi. En effet, chaque méthode implique également toute une série de paramètres (seuil de fréquence pour les mots, options de prétraitement telles que lemmatisation / mots-outils, taille des fragments ou des unités de contexte, nombre d'itérations). Même limitée aux six méthodes retenues, une application approfondie pourrait augmenter notablement la dimension du présent chapitre.

2.4. Extraits de la liste des thèmes (extraits limités à deux « thèmes » par méthode)

Les thèmes (*topics*) sont des listes de mots. Décider qu'une liste de mots mérite le nom de *thème* relève d'une interprétation externe. Les listes sont de longueurs variables, et en nombres variables selon les méthodes. Le nombre de thèmes détectés ici par chacune des six méthodes sélectionnées [à partir de la mise en œuvre des logiciels précités] varie entre six et dix. Simplement pour donner une idée des résultats fournis par ces six méthodes, deux thèmes (c'est-à-dire simplement deux listes de mots) sont imprimés ci-dessous pour chaque méthode.

Les identificateurs des thèmes sur les futures visualisations figurent en début de ligne : les trois premières lettres indiquent la méthode suivie du numéro des thèmes qu'elle propose.

1 Rotated Factor Analysis (Rotation Oblimin): RFA. *(2 thèmes sur 6)*

RFA1 eyes see bright lies best form say days
RFA2 beauty false old face black now truth seem

2 Fragmented Correspondence Analysis : FCA *(2 thèmes sur 7)*

FCA1 beauty truth muse age youth praise old eyes glass long lies false time days
FCA2 night day bright see look sight

3 Analyse logarithmique (Spectral mapping): ALO *(2 thèmes sur 8)*

ALO1 summer away youth sweet state hand age rich beauty time hold nature death
ALO2 pen decay men live earth verse muse once life hours make give gentle death

4 Latent Semantic Analysis : LSA *(2 thèmes sur 8)*

LSA1 beauty live nature art nothing leave could long summer never days false
LSA2 once hand life think time many must dead happy thought lie end woe

5 Non negative Matrix Factorization : NMF thèmes *(2 thèmes sur 10)*

NMF0 love true new hate sweet dear say prove lest things best like ill let know fair soul tongue knows loves
NMF1 beauty fair praise art eyes old days truth sweet false summer nature brow black live dead youth deep born

6 Latent Dirichlet Allocation : LDA *(2 thèmes sur 10)*

LDA0 summer worse praise nature making time like increase flower let copy rich year die away fast winter old writ cold
LDA1 sing sweets summer hear love music eyes bear single confounds prove shade eternal happy art say sweet

2.5. Une synthèse des *thèmes* produits

Comment comparer les listes complètes de *thèmes*, puisque l'ordre des *thèmes*, et l'ordre des mots pour un *thème* donné sont arbitraires?

Nous sommes ici en présence d'un ensemble de « sacs de mots » illustrés par les lignes de la section .2.4. Nous pouvons effectuer une classification de ces lignes, considérées comme des réponses à une question ouverte posée fictivement à chaque thème : « Quels sont les mots qui vous caractérisent ? ». Les distances sont calculées ici en tant que distances du chi-2 dans la table de contingence lexicale (thèmes x mots).

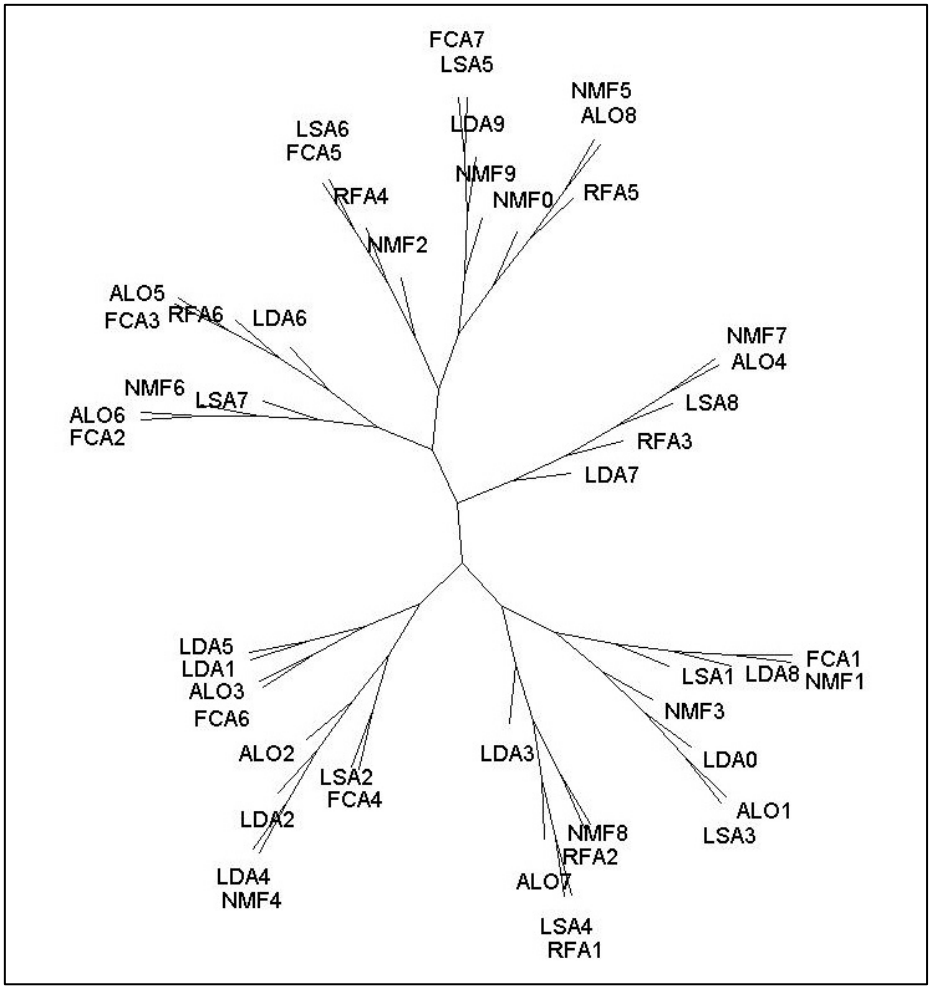


Figure 2.2. Arbre additif décrivant les liens entre les 49 thèmes fournis par les six méthodes sélectionnées. Les identifiants sont ceux de la section .4. La distance entre deux thèmes est la distance du chi-deux entre les profils lexicaux des thèmes

La technique des arbres additifs (voir Saitou et Nei, 1987)) nous a semblé être l'outil le plus puissant et suggestif pour synthétiser sous forme compacte ces 49 thèmes (figure 2.2)[3].

[3] Les figures 2.1 et 2.2 proviennent du logiciel *DtmVic* (régularisation et calculs de distances) qui enchaîne sur le logiciel *SplitsTree* (Huson et Bryant, 2006).

Rappelons une propriété importante des arbres additifs: la distance réelle entre deux points (deux thèmes) peut être lue directement sur l'arbre comme le chemin le plus court entre les deux points. Ici, pour des raisons de lisibilité (figure 2.2) les arêtes ont toutes la même longueur, donc cette propriété n'est plus exactement vérifiée.

Nous nous attendons idéalement à trouver un arbre avec autant de branches que de *thèmes* réels dans le corpus, chaque branche étant caractérisée par six thèmes correspondant aux six méthodes. Une telle situation se produit lorsque chaque méthode a découvert les mêmes thèmes réels que les autres.

La configuration observée n'est pas aussi bonne que dans cette situation idéale, mais on peut cependant distinguer entre six et onze branches principales, ce qui donne une idée de l'ordre de grandeur du nombre de thèmes. On note également que plusieurs méthodes différentes participent souvent à la même branche, ce qui suggère que cette branche correspond à un thème réel découvert simultanément par plusieurs des méthodes mises en œuvre.

Exemple de lecture de la figure 2.2 : La branche de l'arbre située dans la partie droite de l'arbre et à mi-hauteur concerne les points (NMF7, ALO4, LSA8, RFA3, LDA7). Elle correspond vraisemblablement à un thème identifié par cinq des six méthodes. Cette branche se retrouvera en haut de la moitié gauche de la figure 2.3 qui, elle, permettra d'identifier le thème comme étant celui désigné par : « Rivalry ».

2.6. Rapprochement avec les thèmes *a priori*

Quel rapport peut-il exister ente les huit thèmes provenant d'analyses qualitatives des sonnets par des experts de la littérature élisabéthaine et les thèmes proposés par chacune des six méthodes utilisées ?

Comme l'indique le tableau 2.1 précédent, ces thèmes sont en fait des groupes de sonnets, il s'agit donc de thèmes dominants par sonnet, et non de thèmes découverts sans contrainte. Mais les sonnets sont suffisamment courts (14 lignes) pour que cette contrainte ne soit pas un obstacle au rapprochement que nous allons tenter.

Comme ces thèmes correspondent à une partition en huit classes des sonnets, on va partir de la table agrégée thèmes x mots (8 x 173) qui agrège les 154 lignes de la table sonnets x mots en huit classes[4].

Tableau 2.2. Liste des mots caractéristiques des huit thèmes a priori

(seuil de fréquence minimale des mots: 10, puis sélection par valeurs-test > 1.7)

Procreation	beauty self world die age bear another youth live time make
YoungMan	all never heart days time sun ever
DarkLady	black heart soul face one let well friend still
Absence	thought night day mind till being far woe think like
Storm	love eyes hate truth false see know best heart lies
Rivalry	praise worth making verse fair muse therefore was being use others
Death	world death would life
Etern_poetry	men long live world summer death

[4] L'analyse des correspondances de la table non agrégée (154 x 173) croisant sonnets et mots, avec les thèmes en tant que *catégories supplémentaires*, avait donné lieu à la visualisation de la figure 2.1.

Nous allons maintenant, à partir de cette table agrégée, caractériser chaque *thème a priori*, donc chaque groupe de sonnets, par ses mots les plus caractéristiques (ou spécificités, voir, par exemple, Lebart et Salem, 1994 ; Lebart, Salem et Berry, 1998).

La classification par arbre additif de la figure 2.2 va être refaite avec ces huit thèmes *a priori* additionnels correspondant à une « septième méthode » que l'on appelera « analyse littéraire ».

Un arbre additif (comme tout arbre) peut donner lieu à des tracés de figures diverses mais équivalentes.

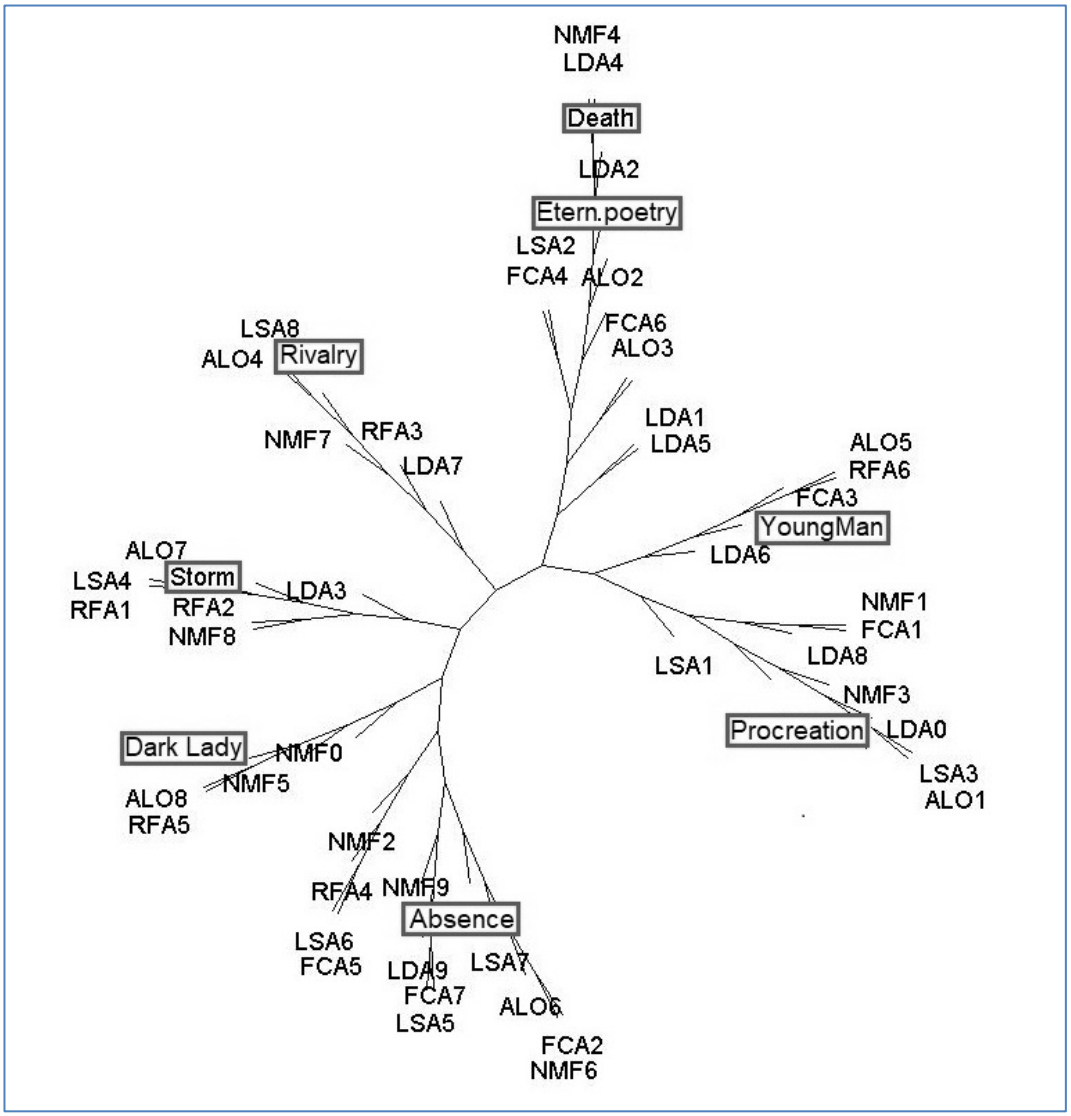


Figure 2.3. Arbre additif décrivant les liens entre les 57 thèmes fournis par les six méthodes techniques et la « méthode littéraire » (49 thèmes de la figure 2.2 précédente + 8 thèmes «experts» du tableau)

La figure 2.3, qui fait intervenir les thèmes *a priori* (ou experts) n'a pas globalement les mêmes orientations que la figure 2.2, mais ce qui est satisfaisant, c'est que l'on retrouve les mêmes grandes branches.

Et c'est une surprise agréable de voir que les thèmes « experts » se distribuent dans les principales grandes branches du nouvel arbre, sans laisser subsister de branches importantes ayant échappé aux intuitions des vrais spécialistes de Shakespeare.

Il a paru nécessaire de publier les deux arbres des figures 2.2 et 2.3 séparément car la figure 2.1 correspond à une situation normale d'exploration, sans intervention d'information externe. La technique des arbres additifs ne permettant pas le positionnement simple de variables supplémentaires, l'introduction des thèmes « experts / a priori » comme éléments actifs dans le calcul de l'arbre conduisant à la figure 2.3 a déformé la représentation initiale. La publication de la seule figure 2.3 aurait suscité une question légitime : quel est le rôle des thèmes « experts » dans la structure observée ? Avec les deux représentations, nous pouvons vérifier que les regroupements, identifiés (ou simplement qualifiés) par la position des thèmes « experts » étaient déjà présents dans le premier arbre (malgré les positions fort différentes des branches) et donc que les six techniques de recherche de thèmes (*Topic Modeling*) utilisées ont bien permis, à des degrés divers, de découvrir ces thèmes.

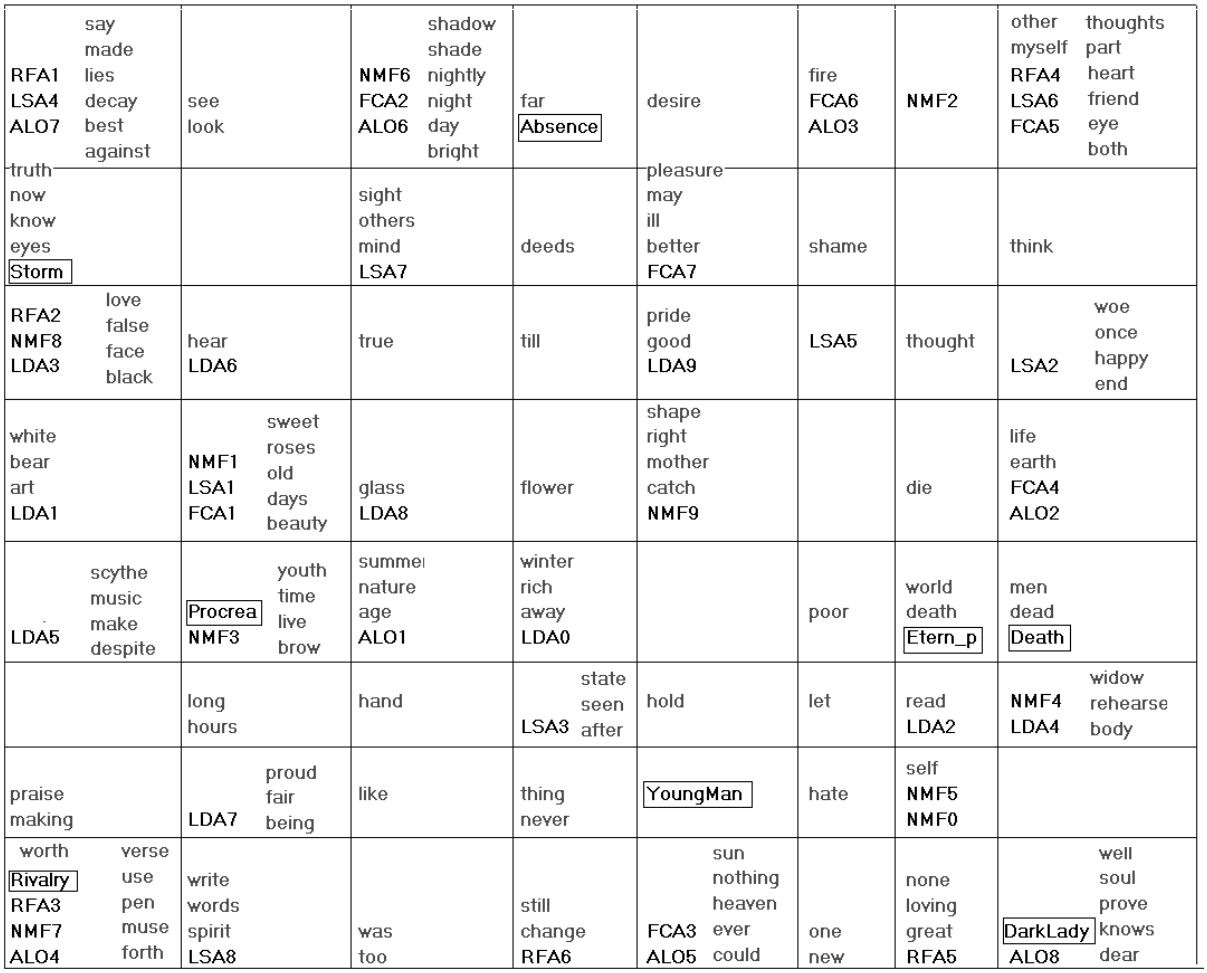


Figure 2.4 Synthèse par carte auto-organisée des thèmes obtenus, des thèmes experts, et des mots caractérisant ces thèmes à partir de la table (57 x 139) décrivant les liens entre les 57 thèmes (dont huit thèmes experts) et les 139 mots décrivant ces thèmes (mots apparaissant au moins deux fois)

Enfin la figure 2.4 synthétise la table lexicale croisant l'ensemble des thèmes et les mots sous forme de carte auto-organisée (Kohonen, 1989). Cette carte résume ici la représentation simultanée (thèmes, mots) dans l'espace des 12 premiers axes d'une AC de cette table. La représentation simultanée s'interprète comme en AC : la proximité entre un mot et un thème ne se lit pas localement par leur appartenance à une même case, mais globalement (un thème par rapport à la disposition de l'ensemble des mots, un mot par rapport à celle de l'ensemble des thèmes).

Lecture de la figure 2.4 : considérons les quatre cases en bas à gauche, qui contiennent le thème expert « Rivalry ». Elles contiennent aussi les cinq thèmes : RFA3, NMF7, ALO4, LSA8, LDA7. Sont ainsi confirmées les proximités observées sur la figure 2.3. De plus, les mots dans ces cases sont, probablement, et probablement seulement, caractéristiques de ces thèmes (praise, worth, verse). Il y a donc confirmation des parentés entre thèmes, et enrichissement par les mots qui les définissent.

1.7. Conclusion du chapitre 2

Cette comparaison empirique et rapide est nécessairement partielle, nous l'avons dit, par le choix des méthodes et du paramétrage de ces méthodes, et aussi par le choix du corpus de référence.

Le choix des méthodes n'est cependant pas tout à fait arbitraire dans la mesure où il repose sur une expérience des méthodes exploratoires (pour les trois premières méthodes : RFA, FCA, LOA) et sur la notoriété (mesurée par exemple par le nombre des publications) pour les trois suivantes (LSA, NMF et LDA). Le corpus est cependant du point de vue de la recherche des *thèmes* un corpus difficile, qui, dans les applications usuelles sur des masses de données hétérogènes du contexte des données massives (*big data*), serait caractérisé par un seul thème : « *Love* »**.**

Il ne s'agissait pas ici sur un exemple de mettre en compétition les diverses méthodes mais plutôt de montrer qu'il ne peut exister de voie royale pour identifier des thèmes. Que plusieurs techniques ou agencements de techniques multidimensionnelles assez classiques permettent d'atteindre honorablement cet objectif, même si l'une d'entre elles, la RFA, comme nous l'avons souligné, a plus d'un siècle.

Mais surtout on souhaite montrer que les outils exploratoires (ici arbres additifs et cartes auto-organisées) sont des compléments indispensables de visualisation, de validation, d'accompagnement, de critique pour les recherches de thèmes.

Remarques sur « Exploratoire / confirmatoire » et « Supervisé / Non-supervisé ».

Dans le cadre des recherches de thèmes (*Topic Modeling*) qui occupent une position hybride entre l'exploratoire et le confirmatoire dans la mesure où ces recherches utilisent parfois des modèles généraux ou probabilistes, on a pu voir l'intérêt des visualisations globales entre les mots (dont certains vont constituer les thèmes) et les documents ou les catégories de documents. Lorsque plusieurs méthodes se proposent de mettre en évidence des thèmes à partir d'un même corpus, la visualisation de la confrontation des résultats (figures 2.3 et 2.4) paraît incontournable pour évaluer et critiquer ces méthodes dans le contexte de l'application en cours.

Les techniques d'apprentissage profond (*Deep Learning*) qui correspondent à l'état de l'art des méthodes supervisées d'apprentissage demandent des bases d'apprentissage gigantesques et font appel à plusieurs niveaux de représentation, incluant des phases de régularisation.

Mais le rôle important des approches non-supervisées n'échappe pas aux spécialistes de ces techniques. Ainsi peut-on lire à propos du futur de l'apprentissage profond (LeCun *et al.*, 2015) que « l'apprentissage non supervisé a joué le rôle de catalyseur pour renouveler l'intérêt envers l'apprentissage profond, mais a été éclipsé depuis par les succès de l'apprentissage purement supervisé »[5]. Ces mêmes auteurs ajoutent plus loin : « On s'attend à ce que l'apprentissage non

[5] *"Unsupervised learning had a catalytic effect in reviving interest in deep learning, but has since been overshadowed by the successes of purely supervised learning".*

supervisé devienne beaucoup plus important à long terme. L'apprentissage humain et animal est largement non supervisé : on découvre la structure du monde en l'observant, et non avec l'indication du nom de chaque objet[6] »

Il est vrai que le succès de l'apprentissage (reconnaissance d'images, d'objets, de la parole, de l'écriture) a envahi et facilité notre vie quotidienne de consommateurs, d'usagers, de communicants.

Mais l'acquisition des connaissances, l'analyse et la compréhension des phénomènes naturels et sociaux dans toute leur complexité demandent parfois de la part des utilisateurs des jugements, de la sensibilité, du bon sens. Elles doivent permettre des remises en question d'hypothèses, des analogies, des intuitions et parfois des retours sur le recueil et la qualité des données, phases pour lesquelles les méthodes exploratoires présentées sont irremplaçables.

Références

Alden, R. M. (1913). *Sonnets and a Lover's Complaint*. New York: Macmillan.

Bartell B.T., Cottrell G.W. et Belew R.K. (1992). Latent semantic indexing is an optimal special case of multidimensional scaling, *Proceedings of the 15th Int. ACM-SIGIR Conf. on Res. and Dev. in Information Retrieval*, Belkin N *et al.* Ed.: 161-167, ACM Press, New York.

Beaudouin V. (2002). *Mètre et rythmes du vers classique. Corneille et Racine.* Champion, Paris.

Bernard M. (2000). Le vocabulaire spécifique d'une œuvre. *Colloque Corpus littéraires - Recueil et numérisation, analyses assistées, didactique*, Université Paris VII. Archives sonores en ligne sur la revue *Texto ! :* http://www.revue-texto.net/Archives/Corpus_litteraires/Corpus_litteraires.html.

Berry M.W., Browne M., Langville Amy N., Pauca V.P., et Plemmons R.J. (2007). "Algorithms and applications for approximate nonnegative matrix factorization". In: *Computational Statistics et Data Analysis* 52.1: 155-173.

Blei, D., Ng, A., et Jordan, M. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022

Boutsidis C., Gallopoulos E. (2008). "SVD based initialization: A head start for nonnegative matrix factorization". In: *Pattern Recognition* 41.4: 1350-1362.

Brunet É. (1988). La structure lexicale dans l'œuvre de Hugo. In : *Etudes sur la richesse et la structure lexicale.* Labbé D., Thoiron P., Serant D. Editeurs, Slatkine-Champion : 23-42.

Brunet É. (2004). Statistiques Rimbaldiennes, SI@T, *Les littératures de l'Europe unie*, Cesenatico, Italie, 88-113, hal-01362731.

Buneman P. (1971). The recovery of trees from measurements of dissimilarity. In: Hodson F. R. D. Kendall G., et Tautu P., (Editors). *Mathematics in the archeological and historical sciences.* Edinburgh University Press, Edinburgh: 387-395.

---

[6] *"…we expect unsupervised learning to become far more important in the longer term. Human and animal learning is largely unsupervised: we discover the structure of the world by observing it, not by being told the name of every object".*

Cocco, C (2014). Typologies textuelles et partitions musicales : dissimilarités, classification et autocorrélation. *Thèse*, Université de Lausanne, Faculté des Lettres, Switzerland.

Deerwester S., Dumais S.T., Furnas G.W., Landauer T.K. et Harshman R. (1990). Indexing by latent semantic analysis, *J. of the Amer. Soc. for Information Science*, 41 (6): 391-407.

Furnas G. W., Deerwester S., Dumais S. T., Landauer T. K., Harshman R. A., Streeter L. A., et Lochbaum K. E. (1988). Information retrieval using a singular value decomposition model of latent semantic structure, *Proceedings of the 14th Int. ACM Conf. on Res. and Dev. In : Information Retrieval*: 465-480.

Garnett J.-C. (1919). General ability, cleverness and purpose. *British J. of Psych.*, 9, 345-366.

Gaujoux R.*et al.* (2010). A flexible R package for nonnegative matrix factorization. In: BMC Bioinformatics 11.1 (2010). 367.

Griffiths T.,L., Steyvers M., and Tenenbaum J.,B. (2007). Topics in Semantic Representation. *Psychological Review*, 114, 2, 211-244.

Henry F. (1900). *Sonnets de Shakespeare (avec Introduction, notes et bibliographie)* . Librairie Paul Ollendorff, Paris.

Holmes D.I. (1985). The analysis of literary style - A Review, *J. R. Statist. Soc.*, 148, Part 4: 328-341.

Huson D.H. and Bryant D. (2006). Application of Phylogenetic Networks in Evolutionary Studies, *Molecular Biology and Evolution*, vol. (23), 2: 254-267.

Kazmierczak J.-B. (1985). Analyse logarithmique : deux exemples d'application. *Revue de Statist. Appl.* , 33, (1): 13-24.

Kohonen T. (1989). *Self-Organization and Associative Memory*, Springer-Verlag, Berlin.

Labbé D., Thoiron P. et Serant D. (Ed.) (1988). *Etudes sur la richesse et la structure lexicales*, Slatkine-Champion, Paris-Genève.

Lamy J.C. (2004). *Brassens, le mécréant de Dieu.* Albin Michel, Paris.

Lawley D. N., Maxwell A. E. (1963). *Factor Analysis as a Statistical Method*, Methuen, London.

Lebart L. (2004). Validation techniques in Text Mining. In: *Text Mining and its Application*, S. Sirmakensis (ed.), Berlin- Heidelberg, Springer Verlag: 169-178.

Lebart L. (2007). Which *bootstrap* for principal axes methods? In: *Selected Contributions in Data Analysis and Classification*, P., Brito *et al.*, editors, Springer: 581 – 588.

Lebart L. (2018). Looking for Topics, a brief review. In: *Text Analytics, Advances and Challenges.* Iezzi D. F., Mayaffre D.& Misuraca M. (Eds), Springer, Cham, Switzerland, 215-224.

Lebart L., Pincemin B., & Poudat C. (2019). *Analyse des Données Textuelles*. P.U.Q. Québec, Canada.

Lebart L., Salem A. (1994). *Statistique textuelle*. Dunod, Paris. *Téléchargement :* http://www.dtmvic.com/ST.html.

Lebart L., Salem A. & Berry E. (1998). *Exploring Textual Data*, Springer, Netherland.

LeCun Y., Bengio Y., & Hinton G. (2015). Deep Learning. *Nature*, 521, 436-444.

Lee D. D. et Seung H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, 401: 788-791.

Lee D.D. et Seung H. S. (2001). Algorithms for nonnegative matrix factorization. In T. G. Dietterich and V. Tresp, editors, *Advances in Neural Information Processing Systems*, volume 13. The MIT Press.

Lewi P.J. (1976). Spectral mapping, a technique for classifying biological activity profiles of chemical compounds. Arzneim. Forsch. in: *Drug Res.* 26, 1295-1300.

Morando B. (1980). L'analyse statistique des partitions de musique, *Les cahiers de l'analyse des données*, tome 5, no 2 (1980), 213-228,

Paatero P., Tapper U. (1994). Positive matrix factorization : A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics,* 5 : 111-126.

Paterson D. (2010). *Reading Shakespeare Sonnets.* Faber and Faber Ltd. London.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot M. et Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python, *Journal of Machine Learning Research* , 12, 2825-2830.

Poulanges A. and Tilleu A. (2001). *Les manuscrits de Brassens*. Textuel, Paris.

Ratinaud P., Déjean S. (2009). IRaMuTeQ : implémentation de la méthode ALCESTE d'analyse de texte dans un logiciel libre. In *Colloque Modélisation Appliquée aux Sciences Humaines et Sociales (MASHS2009),* Toulouse, http://repere.no-ip.org/Members/pratinaud/mes-documents/articles-et-presentations/presentation_mashs2009.pdf/view

Reinert M. (1983). Une méthode de classification descendante hiérarchique : Application à l'analyse lexicale par contexte, *Les Cahiers de l'Analyse des Données*, 3, Dunod: 187-198.

Reinert, M. (1986a). Un logiciel d'analyse lexicale: [ALCESTE]. *Cahiers de l'Analyse des Données*, 4 : 471–484.

Rochard L. (2009). *Les mots de Brassens*, Edition du Cherche Midi, Paris.

Saitou N., Nei M. (1987). The neighbor joining method: a new method for reconstructing phylogenetic trees, *Molecular Biology and Evolution*, vol. (4), 4: 406-425.

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

Schmid H. (1994). Probabilistic part of speech tagging using decision trees. *Proceedings of the International Conference on New Methods in Language Processing*, Manchester, UK.

Shakespeare, W. (1901). *Poems and sonnets: Booklover's Edition*. Ed. The University Society and Israel Gollancz. New York: University Society Press. Shakespeare Online. Dec. 2017.

Spearman C. (1904). General intelligence, objectively determined and measured. *Amer. Journal of Psychology*, 15: 201-293.

Thurstone L. L. (1947). *Multiple Factor Analysis.* The Univ. of Chicago Press, Chicago.

Viprey J.-M. (2002). *Analyses textuelles et hypertextuelles des Fleurs du mal (avec le texte intégral et un moteur de recherche sur CD-Rom).* Champion, Paris.

Yule G.U. (1912). On the methods of measuring the association between two attributes. *J.R. Stat. Soc.*75, 579-642.

Yule G.U. (1944). *The Statistical Study of Literary Vocabulary*, Cambridge University Press, Reprinted in 1968 by Archon Books, Hamden, Connecticut.