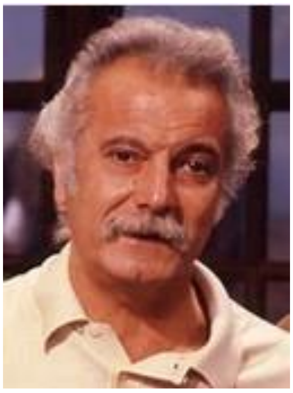
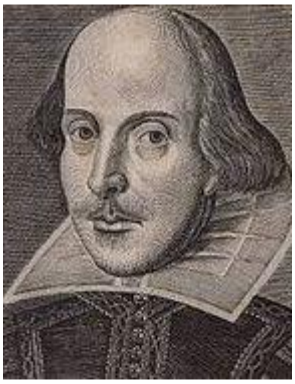
	

Textometry and Poetry

Small essays from Brassens to Shakespeare…

Exploratory analyzes of textual data (which constitute a particular branch of the techniques designated more recently by Textometry or Textometrics) have shown their usefulness in several fields of application.

a) In the case of short, numerous and qualified texts, the typical situation may be the processing of responses to open questions in sample surveys. Several thousand responses on a specific topic on the one hand, hundreds of responses to closed-end questions and respondent characteristics (metadata) on the other, in parallel, allow groupings, contrastive analyzes of texts. Analyzes of tweets, numerous short messages fall into this category.

b) In the case of long but comparable texts from a certain point of view. Typical example: corpuses of novels, political speeches, chronological textual series (evolution, discontinuities, possible attributions of authors).

In these two cases, the lexical frequencies (of both word-forms and word-lemmas) are basic criteria of interest. But sets of poetic texts do not fit into these frameworks.

The two following chapters are very specific contributions to the problem of the confrontation between textometry and poetry. They are obviously far from exhausting the subject.

Chapter 1 is based on a very particular corpus: the collection of 194 songs sung and recorded by the French musician-poet Georges Brassens (1921 – 1981). The simple question asked will be: Can the tools of multidimensional exploratory statistics be applied to a collection of songs (with all the constraints of the genre) and provide new pieces of information?

Chapter 2, on the contrary, deals with a much-studied classical corpus, that of Shakespeare's 154 sonnets. It shows that in the case of collections of formatted poems, the search for themes or topics (*Topic modeling*) can give interesting pieces of information from several statistical tools. We can directly retrieve, through a multidimensional descriptive analysis of the corpus (without *a priori*) the topics identified by the experts during the previous centuries.

To situate the work that will follow, let us specify what it is not... (with respect to a recent series of works relating to French poetry).

It is not a question of studying a specific vocabulary of a work, as was done by M. Bernard (2000), nor of studying or modeling versification techniques (Beaudouin, 2002), nor to study lexical richness and structure (Labbé *et al.*, 1988; Brunet, 1988). Nor is it a scholarly, computer-assisted exploration of an iconic work such as Les fleurs du mal from Baudelaire (Viprey, 2002).

Finally, and more generally, it is not either a question of stylometry following the pioneering work of Yule (1944) or those synthesized by Holmes (1985).

In fact, in the two cases dealt with below, it is most often (and quite simply) a matter of automatic analysis of content, and not pure stylometry (evolution and context of themes for Brassens, identification of already known topics for Shakespeare).

But the link between content and form will sometimes be evoked, since it can arise from the analyzes without having been invited to do so.

Chapter 1.

Poems and Songs. A tentative textometric analysis of the poet-singer Brassens[1]

The corpus that we propose to study here is a complex and elusive material: the collection of 194 songs written and sung by the French musician-poet Georges Brassens (1921-1981).

This author is unique in that he brings together three almost contradictory features:

a) He was a nonconformist and has rubbed shoulders with anarchist movements,

b) In 1967, he received the poetry prize from the very conservative *Académie Française*.

c) He was at the origin of the sale (to date) of 30 million records.

1.1 Problems and challenges of poetic texts

We will begin by showing how the statistical analysis of this type of text constitutes a methodological challenge. Then we will describe the pre-processing of the corpus, the statistical procedures used, and will discuss the first results obtained. We will keep in mind the warning of Brunet (2004) during a remarkable statistical study of the poetic work of Arthur Rimbaud, a warning which applies also to the study of Brassens: *"Besides, the use of tools documentaries and statistics does not go without a certain naivety which gives its faith to the printed words, in their first innocence. But with Rimbaud, the words are often loaded. They are decoys, figureheads, and the reality they designate and hide eludes the most learned exegeses. When esotericism multiplies the traps, how to ensure the semantic constancy of the terms? "*. The studied works of Rimbaud included approximately 40,000 occurrences. The corpus of songs by Brassens used to exemplify the processing here has a comparable size: it contains approximately 52,000 occurrences.

The poetic texts of Brassens are particularly rich in stylistic figures (litotes, metaphors, anaphors, euphemisms, allegories, etc.) which sow doubts about the use of the word (forms or lemmas) as a basic statistical unit. This poet is an expert in the art of diverting locutions (he speaks of the "dark face of the honeymoon", of the "gospel according to Venus") (Lamy, 2004; Poulanges and Tilleu, 2001). It brings up to date popular, outdated or slang expressions (“the poor man's coffee” for: sexual act, etc.). It often uses historical, medieval and even obsolete terms (Rochard, 2009). In the case of songs that may include choruses or partial repetitions, the lexical frequencies no longer have the statistical significance given to them in the usual lexical

[1] An abridged version of this text has been presented at the Jadt2022 conference held in Naples in July 2022.

tables. Versification constraints (e.g.: alexandrines, rhymes) are difficult to integrate into description tools, but influence the choice of words and their frequency.The texts are too short and the corpus too restricted to hope to automatically extract new lexicographic units such as repeated segments (Salem, 1987), patterns, phrases.

The question we are asking is extremely narrow and technical in relation to existing or potential works of literary and musical analyzes of the works of this musician-poet: "Do the typologies and visualizations obtained from the lexical profiles (approximately 1000 word-forms or lemmas) of the 194 songs, confronted with the available metadata, bring new information, notable structural features or new materials likely to interest specialists? ".

1.2 Preprocessing of basic texts and description tools

The 194 units of the corpus will be described by as many groups of words (forms) and groups of lemmas which are complementary statistical units. They are also described by the following metadata: a) belonging to a collection of albums (records) according to 14 categories, variable which has a chronological component of publication (and often of composition/creation); b) the author according to 2 categories (“Brassens” [170 songs] or “Other”, selected and sung by Brassens [24 songs]); c) finally the dominant key (12 categories) of the corresponding musical score. The consideration of tonality is only a sketch here, an attempt. Note that the processing of complete musical scores through correspondence analysis (CA) has been studied in particular in the pioneering works of Morando (1980),and Cocco (2014).

Using both forms and lemmas (for which we used the TreeTagger software [Schmid, 1994]), we will transform each song text into an unweighted vocabulary. In other words, each element will only appear once within a given song (a few lines of Python involving the “dictionary” object of this language easily allow this transformation/reduction of the text). We will therefore have two sets of data corresponding to the raw text file and the lemmatized file. The lemmatized file has the advantage of reducing the diversity of inflections and therefore of allowing lower minimum frequency thresholds. The file of word-forms retains the original diversity of forms, which is fundamental in the case of poetic texts. We will therefore obtain at each stage two different and complementary points of view.

Henceforth, the lexical tables (song x words) will thus be tables of presence – absence. We know that for this type of table, the correlation coefficient r between two songs coincides with the association coefficient Φ of Yule (1912). It is also linked to the χ^2 calculated on the contingency table (2 x 2) crossing the two texts described by *n* words by the formula:

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

The techniques that can be used to describe and detect possible structures will be Principal Component Analysis (PCA, licit on these binary codings), correspondence analysis (CA), classifications by additive tree (Buneman, 1971; Saitou & Nei, 1987).

The metadata defined above will act sometimes as active variables (grouping of songs according to albums, for example), but above all as supplementary variables (*a posteriori* projection, with Bootstrap validation). The software used, free of access, is DtmVic (www.dtmvic.com)[2].

. Let us recall incidently that within the framework of the binary coding used here, the eigenvalues do not have a straitforward interpretation in terms of information. Only statistical

[2] DtmVic makes also use of the software TreeTagger [Schmid, 1994] and SplitsTree [Huson et Briand, 2006].

validations (using *Bootstrap* techniques here) make it possible to judge the statistical validity of both the axes and the locations of points.

1.3 Chronology and collections (discs, albums)

The additive tree in figure 1.1 gives a summary of the links between the records (or albums) identified by their year of publication.

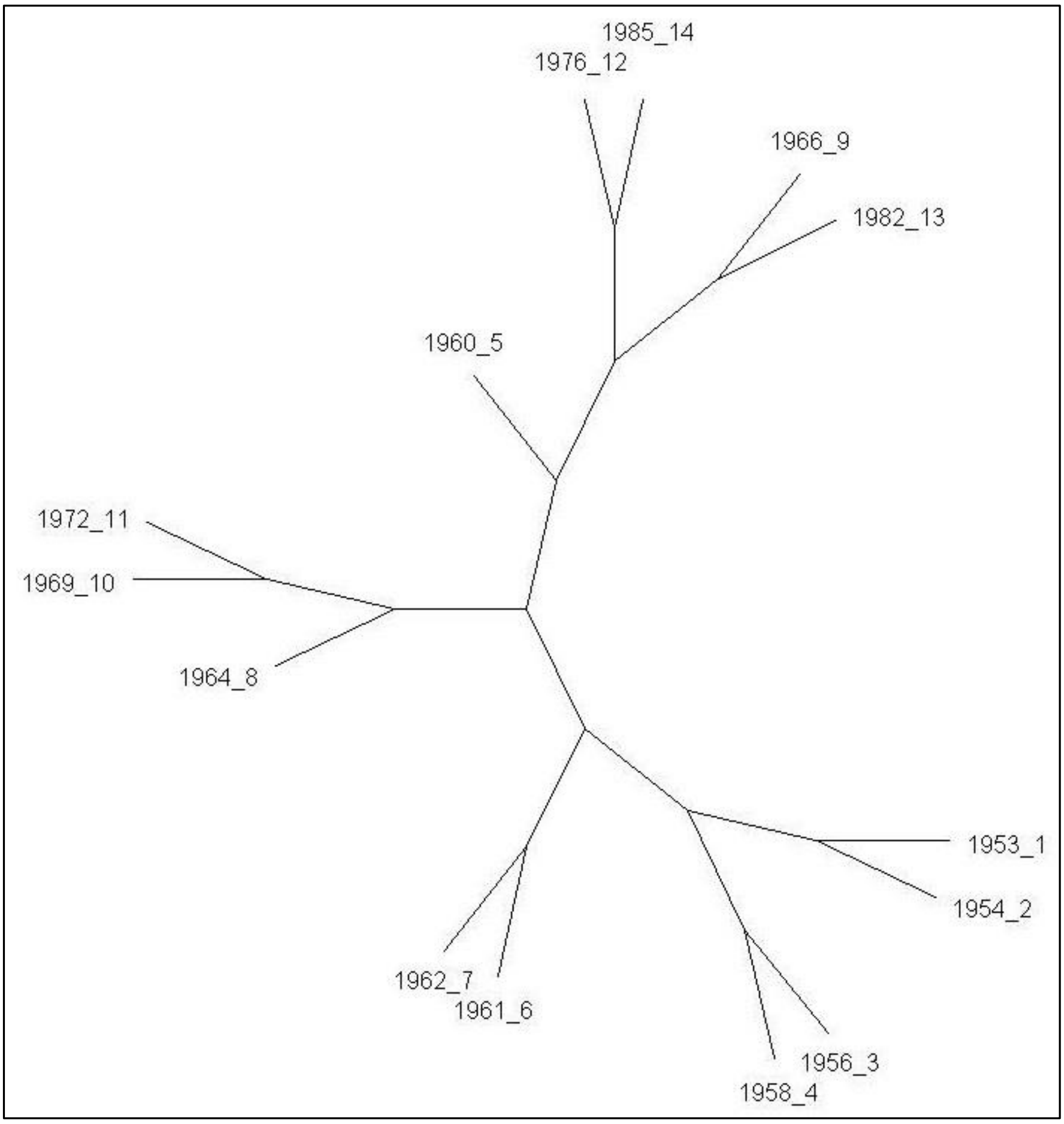


Figure 1.1. Additive tree of the 14 Brassens albums (distances calculated on the presence-absence in each song of 943 forms)..

The last two (years 1982_13 and 1985_14) are posthumous. The first 4 (from 1953_1 to 1958_4) constitute the lower right part of the tree. These albums may include songs composed previously, or assortments suggested by the publisher, or even, like the posthumous albums, early works found and interpreted by others. Despite this, there is still a certain compatibility between the lexical proximities described by this graphic and the chronology. The opposition between the first years and the following ones will be found in all the analyses, whether they involve lemmas or word-forms.

1.4 Brassens and his selected poets

The 24 “external poems” set to music and sung by Brassens constitute a mini-anthology of French poetry very close to his own values and tastes.

Figure 1.2 indeed shows the plane (1, 2) of the CA of the table crossing 194 songs and 901 lemmas (appearing at least four times).

Neither the songs, nor the lemmas could be displayed under the present format. Only a selection of supplementary elements (projected afterward on the plane) is drawn. The two “author” points (**GBrass** and **Autre**) and the two years (**1956** and **1982**)chosen for this display are projected *a posteriori* in the principal plane of the previous analysis, and the Bootstrap replications are obtained by drawings with replacement in the 194 songs (*cf.* Lebart, 2004, 2007). The “external” authors **“Autre”** (Hugo, Lamartine, Paul Fort, Musset, etc.) are closer to the early years in Brassens career.

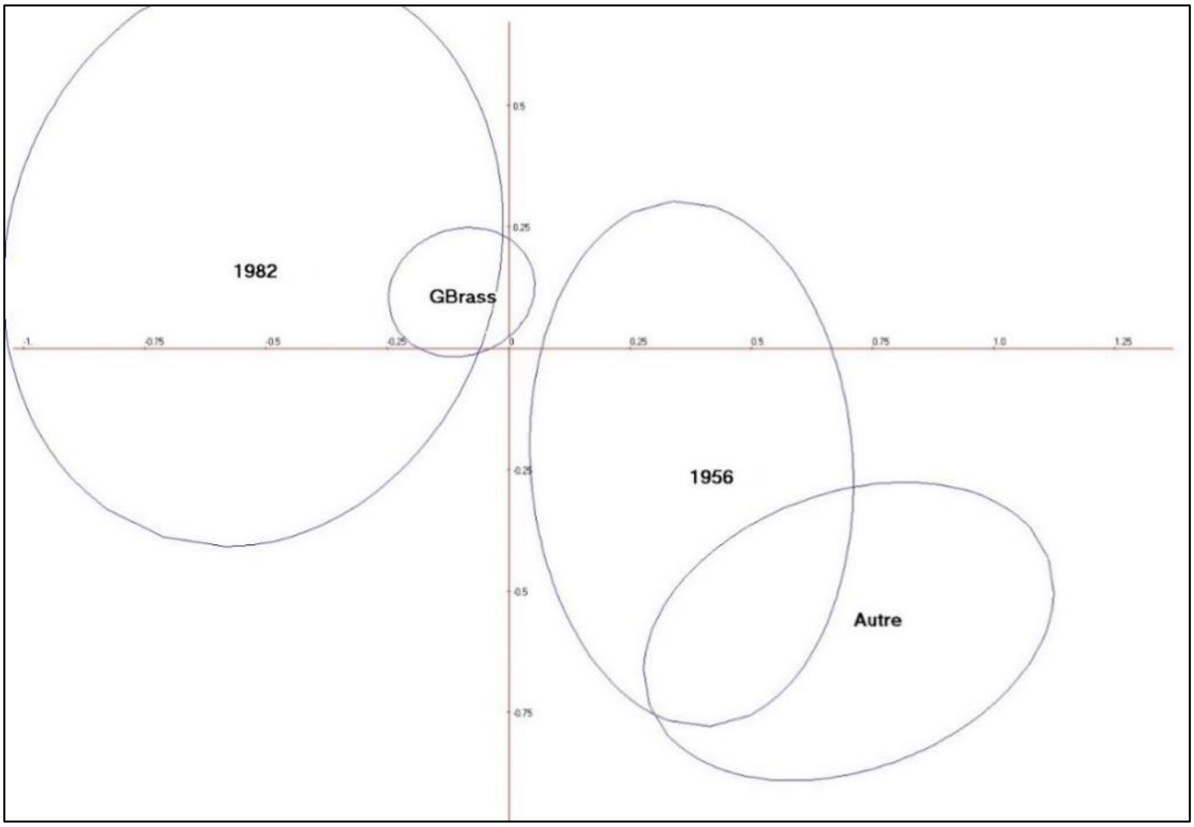


Figure 1.2. Bootstrap confidence areas of four additional categories. GBrass (170 songs written by Brassens) and Autre (24 “external” poems by various authors set to music and sung by Brassens). Bootstrap confidence areas for 2 records (1982 and 1956)..

In fact, one could see on the factorial maps with the position of the 901 lemmas (unpublishable in this publication format, but one can have an idea of this complexity by consulting appendix 2 of this chapter) that the vocabulary has hardened over time, the author calling himself a “pornographer” (from the 5th record).

At the same time, the censorship, active for the first records, was more tolerant during the period.

1.5 Keys (tones)

Brassens was a composer-performer as well as a poet, and one could legitimately wonder if there was a link between the dominant keys of his songs and the characteristics of the songs represented by their lexical profiles..

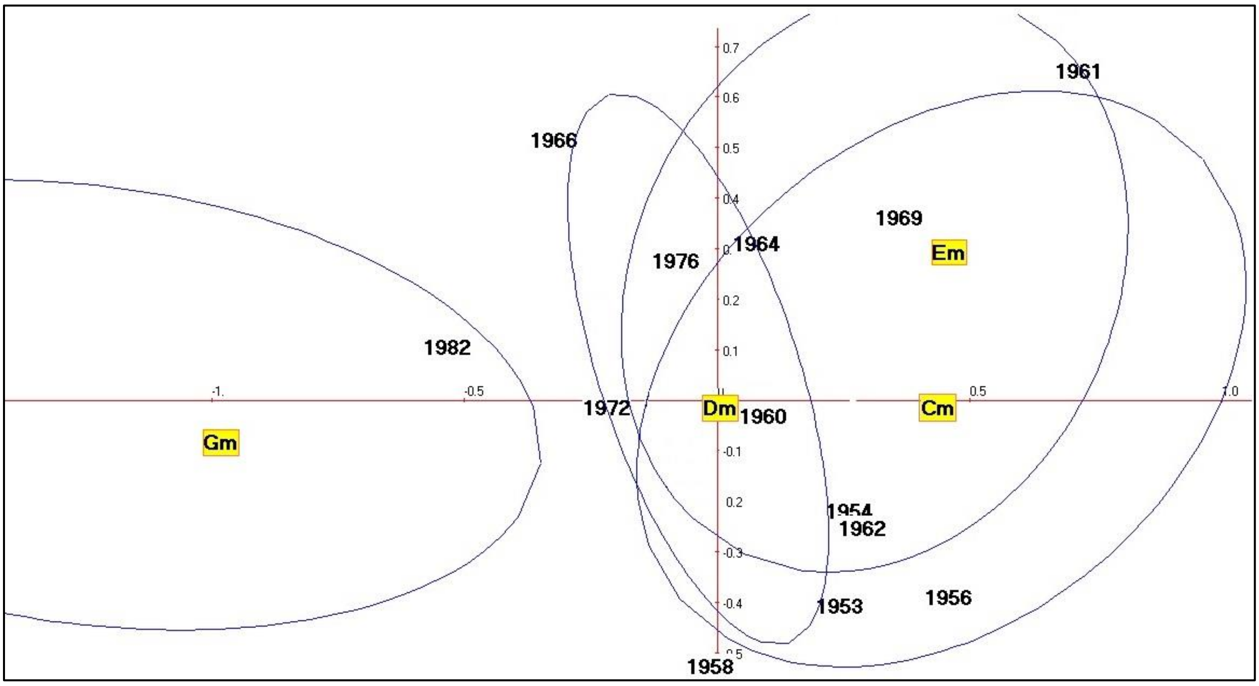


Figure 1.3. Bootstrap confidence zones of 4 keys (Gm, Dm, Em, Cm) among 13 record dates (keys and records are always supplementary elements in the plan (1, 2) of the CA of the binary table (194 songs x 901 lemmas).

In this principal plane from the CA of the table crossing 194 songs and 901 lemmas (figure 1.3), the areas of confidence are indistinguishable except for G minor (Gm) on the left (only 4 areas are published here for readability) . However, this exception only concerns 5 songs, 3 of which were performed by other singers posthumously. We will conclude, in the current state of this research, that there is no clear link between the chosen key and the lexical profiles of the songs.

1.6 Confrontations between lemmas and word forms

This phase is the most essential part when dealing with a poetic text, since it is a question of making content and form compete. We will work on the 170 texts written by the singer himself.

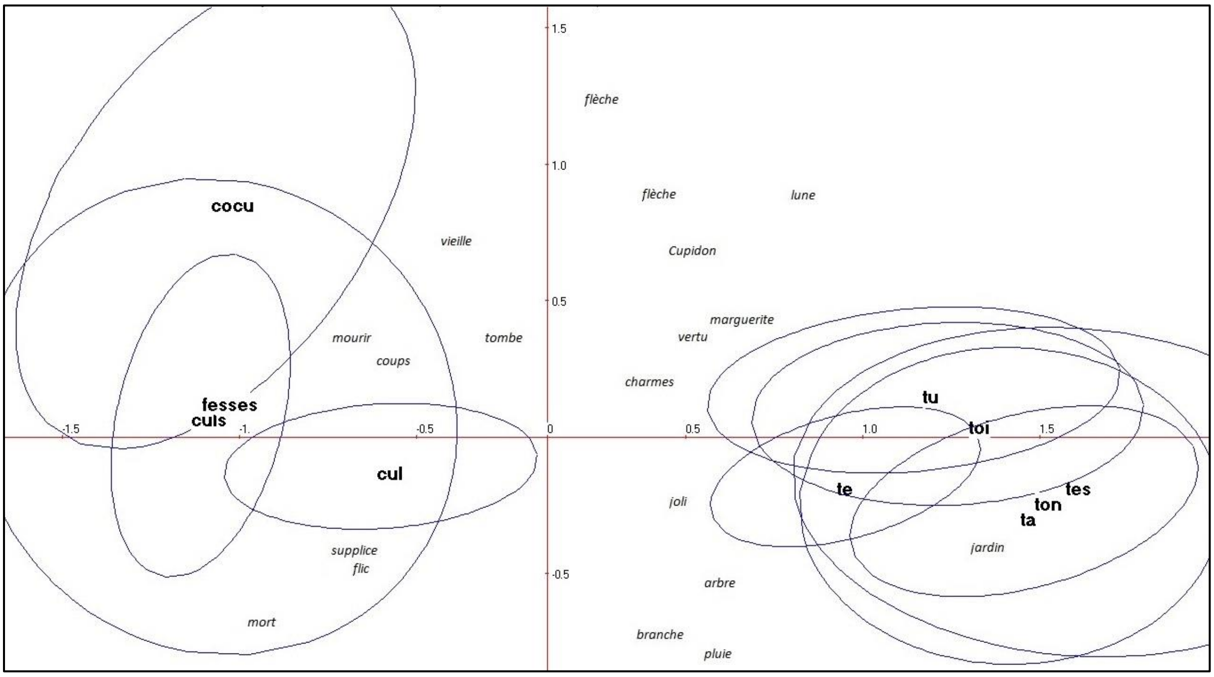


Figure 1.4. Word-forms for Brassens alone (170 songs x 872 forms). Opposition in the principal plane of the CA between the forms of familiarity (te, tu, toi, ton, ta, te, on the right) linked to a refined poetry, and the "more scabrous or coarse vocabulary" (on the left). Some Bootstrap confidence areas are represented. Illustration of the plane with a small subset of word-forms.

For most of the CA-type analyzes bearing on all of the 194 songs, or on the 170 songs of which Brassens is the sole author, a first dimension (horizontal on all the graphs) dominates, whether it concerns lemmas or word-forms: It opposes older texts (four or six first records) to more recent texts. The ancient texts, like the external poems, have a vocabulary that can be described as classic, gentle, even gallant or precious, to force the line. The most recent have a cruder vocabulary, sometimes slang, provocative.

Figure 1.4, which concerns the 170 songs entirely written and composed by Brassens, contains only a small excerpt of the active word-forms.

It shows that the familiarity (six forms concerned, on the right : *te, tu, toi, ta, ton, tes*) means above all intimacy, gentleness for Brassens. Here, the forms reinforce the interpretation of this axis, which is also obtained with the lemmas.

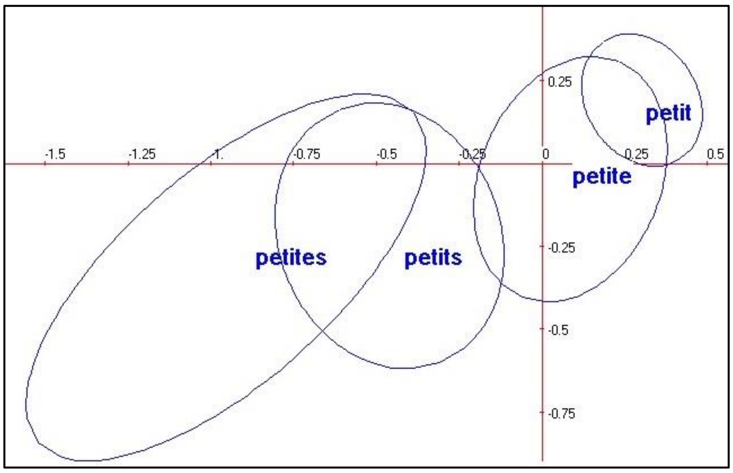


Figure 1.5. Bootstrap confidence zones of four flections of the adjective "petit" ("little") in the CA principal plane (presence absence table crosstabulating 194 songs and 943 word-forms.

The word-forms play a role of magnifying glass for interpretation, but they reveal words that lemmatization makes disappear, as shown in Figure 1.5. The "*petit*" (“small”) lemma, which would replace the four inflections (*petit, petite, petits, petites*) positioned in a first factorial plane similar to the previous ones, would occupy a central position which, in a way, neutralizes its role in the analysis performed on the lemmas. These different inflections are nevertheless linked to this main dimension which opposes, as we have seen, the intimate and the delicate (singular) to the impersonal and the bawdy (plural).

1.7 Conclusion of chapter 1

Despite the limited number of graphical displays presented (necessarily small in size), one can guess that the textometric treatment (multivariate description) of poetic texts brings an original point of view on these texts and also new study materials for specialists. From these first analyses, we were able to detect a general tendency, inextricably linked to age, career, personal development, perhaps to the growing notoriety of the poet and probably to the increasing permissiveness during the considered period. Each time, the use of word-forms amplifies, illustrates and nuances the results obtained from the lemmas.

The unpublished graphs of the large additive trees representing the links between the 194 songs, the displays of the main plans of CA involving approximately 900 word-forms and as many lemmas as well as the songs and the records, also constitute a set of working documents (including figure 1.7 in appendix 2 of this chapter is an example) full of potential, difficult to publish on paper, but fascinating to consult for the specialists or amateurs concerned. Let us then parodying the adage “Garbage in, garbage out” (an adage that is very poorly suited to data analysis techniques which nevertheless have a filtering function) by transforming it into “Poetry in, poetry out”. Indeed, these new documents describing the links and complex patterns between hundreds of words carefully chosen by the great troubadour Brassens are themselves – at least for his faithful admirers – a source of poetic emotions.

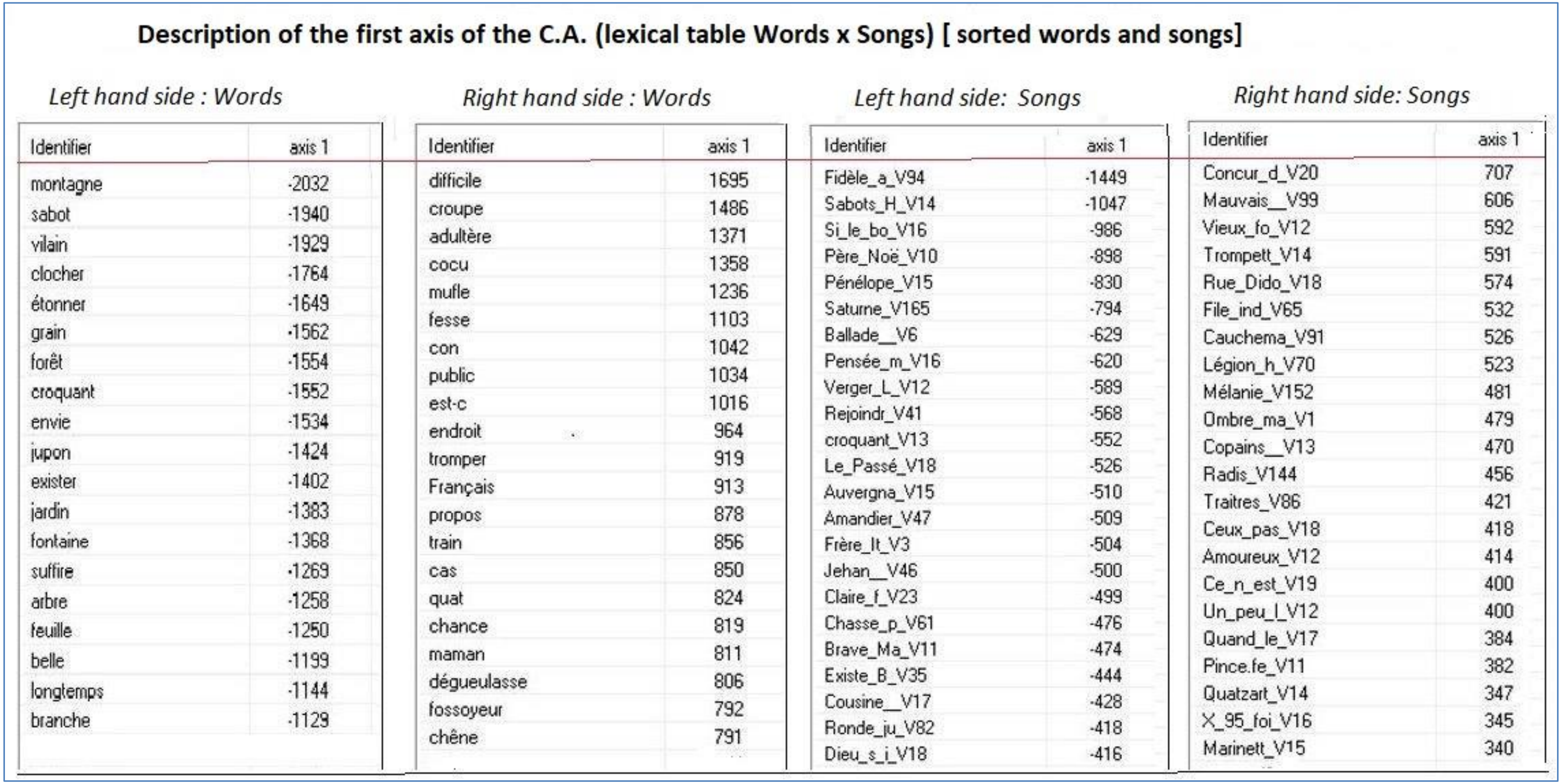
Notes on the appendices:

The first appendix below (Figure 1.6) shows the rankings of the most extreme words on the first axis of a correspondence analysis of the lexical table crossing 703 words (here: lemmas, frequency threshold = 6) and the 194 songs. A very similar first axis (with the same oppositions between words) is observed if we consider only the 170 songs written by Brassens alone (text and music), or if we vary the minimum frequency thresholds from 5 to 30. In a dual way, the technique also allows for classifying the songs (right part of the table). This opposition along the horizontal axis at the level of the lemmas was already visible (and perhaps even more caricatural because of the richness of the word-forms), in Figure 1.4.

The second appendix (Figure 1.7) presents the main plane of the same correspondence analysis (CA) whose table in Figure 1.6 only sketched out the first axis (horizontal axis).
The most extreme points have been brought back to the frame (arrows).

Appendix 1 to Chapter 1:

(Figure 1.6). Table describing the positions of words and songs on the first axis of the correspondence analysis (CA) of the lexical table crossing the 703 most frequent words and the 194 songs.



Description of the first axis of the C.A. (lexical table Words x Songs) [sorted words and songs]

Left hand side : Words

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

Right hand side : Words

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

Left hand side: Songs

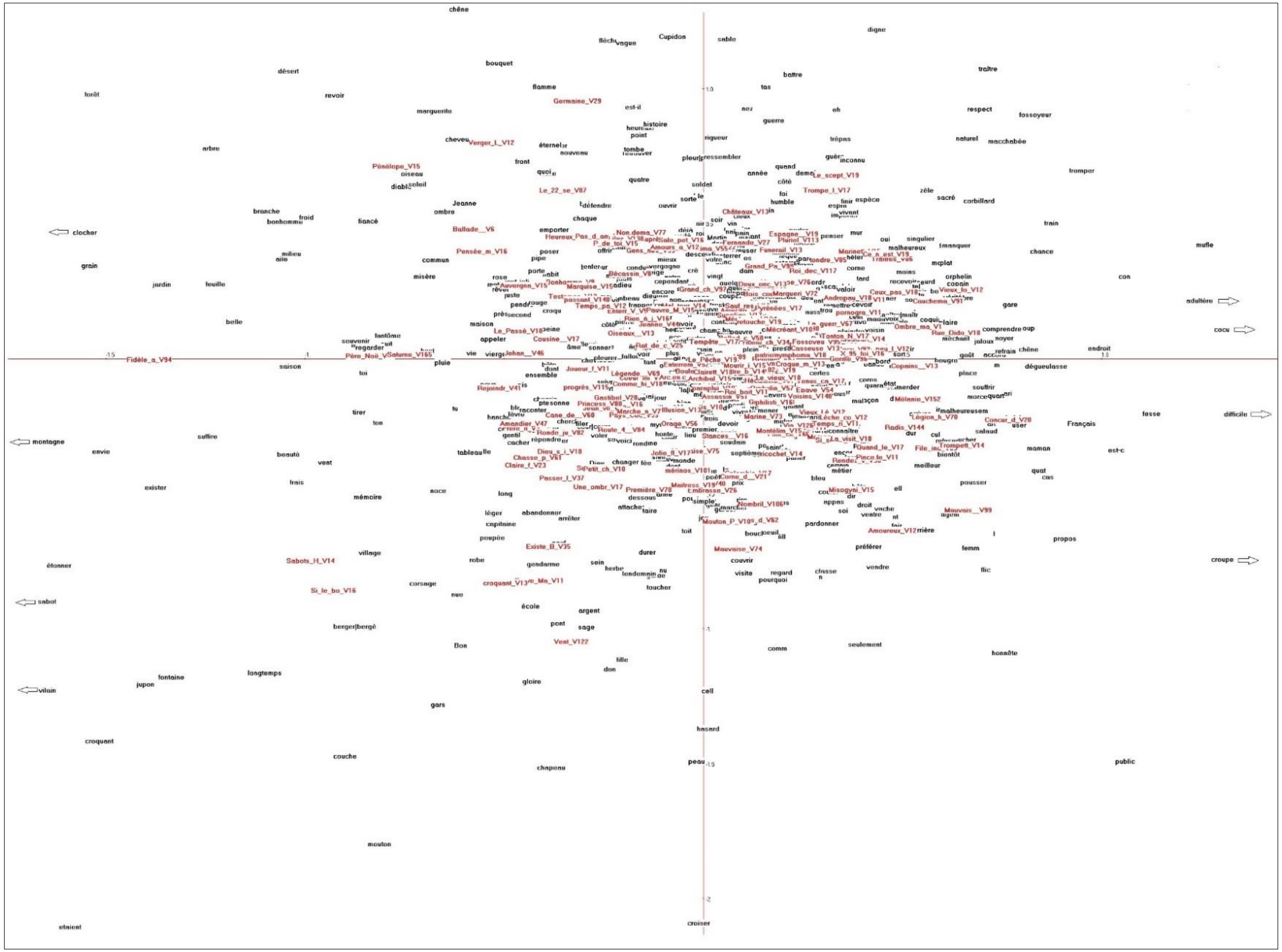
Identifier	axis 1
Fidèle_a_V94	-1449
Sabots_H_V14	-1047
Si_le_bo_V16	-986
Père_Noë_V10	-898
Pénélope_V15	-830
Saturne_V165	-794
Ballade__V6	-629
Pensée_m_V16	-620
Verger_L_V12	-589
Rejoindr_V41	-568
croquant_V13	-552
Le_Passé_V18	-526
Auvergna_V15	-510
Amandier_V47	-509
Frère_It_V3	-504
Jehan__V46	-500
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

Right hand side: Songs

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

(The coding of the titles of the songs, a little complex, allows however an identification. It should be improved in a forthcoming printing)

Appendix 2 to Chapter 1 (Figure 1.7) Main plane (1, 2) of the previous CA. (Words: black, Songs: red)



Chapter 2.

Topic Modeling in Shakespeare's Sonnets.

The field of topic research occupies an intermediate position between exploratory and confirmatory approaches: exploratory procedures are, in a way, in competition with models that must be validated. Some of the results presented below have been the subject of previous publications (see for example: Lebart, 2018; Lebart, Pincemin and Poudat, 2019).

2.1 Introduction

This chapter presents a brief overview of several attempts to identify latent variables (axes or classes) in the case of textual data. These latent variables (classes or axes) are sometimes designated ex ante by the term “topic”. The factor analysis of psychologists at the beginning of the last century (Spearman, 1904) was already an attempt to identify interpretable latent variables. It assumed a model, initially mono-factorial (a single latent variable, the general aptitude factor, or intelligence, to explain a battery of scores), then multifactorial (intelligence, memory, work power, etc.). It is therefore a question of estimating a model at the outset, although this estimation was not related to classical inferential statistics until much later by Lawley and Maxwell (1963). But, like most latent variable models, it is an unsupervised approach: the hidden variables are not directly observable but provided by the model itself.

Expressed more concretely: in the equations that define the model, what is known is only one side of the “=” sign (unlike regression or discriminant analysis, for which one measurement is explained by another measurement, at least for the training samples: there are therefore observations on both sides of the “=” sign). The adventure of the factorial analysis of the psychologists does not stop there, because the model itself became an instrument of observation in the hands of the practitioners, even before it was shown that it was very close to principal component analysis.

The last few years have witnessed a series of algorithmic attempts such as Non-negative Matrix factorization (NMF) or Latent Dirichlet Allocation (LDA). Simultaneously, the topics considered as latent variables could also be identified through several hybridizations and synergies of Principal Axes Methods (such as PCA, CA) and classification techniques.

There is a profusion of new disciplines around industrial applications involving texts, with subsequent proliferations of tools and disparities in terminology. There are also disparities in attitude towards texts, sometimes influenced by the availability and user-friendliness of software. The problems posed by huge collections of newsgroups or tweets are very different from those encountered in the fields of literature, political speeches and psychological investigations.

In this chapter, a single classical corpus of average size will serve as a reference corpus to sketch and compare in a compact way certain characteristics of several methods. Because they are well known, translated into almost every language, deeply studied and commented on, we will use Shakespeare's 154 Sonnets as a reference corpus to briefly compare the ability of several techniques to recognize themes in a corpus.

2.2. An overview of the contents of Shakespeare's sonnets

William Shakespeare's 154 sonnets deal with themes such as love, friendship, the effects of time, beauty, betrayal, lust, death. [1].

2.2.1 Themes, Topic, Subject, Motif

Note that the definition of themes is pragmatic and can also cover the concepts of subject and motif (in the literary sense). Usually, the topic is the subject that is treated, an objective explanation of the content of a text (the main title, for example), while a theme can represent a deeper underlying message. It acts as a foundation of the entire story. A motif, on the other hand, is simply a recurring idea used to reinforce the main theme. Schematically, the topics answer the questions: “What is the story about, who, what, how? and the themes are more like, "Why was history written?" ". The topics of literature are easier to identify than the themes.

Three contiguous series of sonnets are generally recognized as corresponding to three dominant topics:

Sonnets 1 to 17: (*Procreation*). These sonnets celebrate the beauty of a young man who is urged by the poet to marry so as to perpetuate that beauty.

Sonnets 18 to 126: (*Young man*). This longest sequence concerns the same young man (not definitively identified), the destructive effect of time, the force of love, friendship and poetry.

Sonnets 127 to 154: (*Dark Lady*). These sonnets are mostly addressed to a dark haired woman, not without some irony and cynicism (the two last sonnets 153 and 154 are specific epigrams in an ancient style; they should deserve in fact a specific category).

2.2.2 Eight themes derived from expert commentaries

The themes *Young man* and *Dark lady* could contain five sub-themes. While the first theme (*Procreation*) remains untouched, the new *Young man* and *Dark lady* themes will comprise only those sonnets which were not assigned to the five new categories below (*Absence, Storm, Rivalry, Death, Eternal poetry*).

Table 2.1. L List of eight a priori themes/topics with the corresponding sonnets numbers

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

[1] Noe that a French version of Shakespeare's Sonnets by F. Henry (1900) is available on Gallica website (Bnf): http://gallica.bnf.fr/ark:/12148/bpt6k1310005 (with introduction and comments). For the original English version see Shakespeare (1901).

The partition of sonnets given in Table 1 is inspired by the works of Alden (1913) and Paterson (2010) but not explicitly mentioned by these authors. Figure 2.1 shows however that, after a blind correspondence analysis ignoring these themes, most of their locations are statistically significant on the principal plane of visualization.

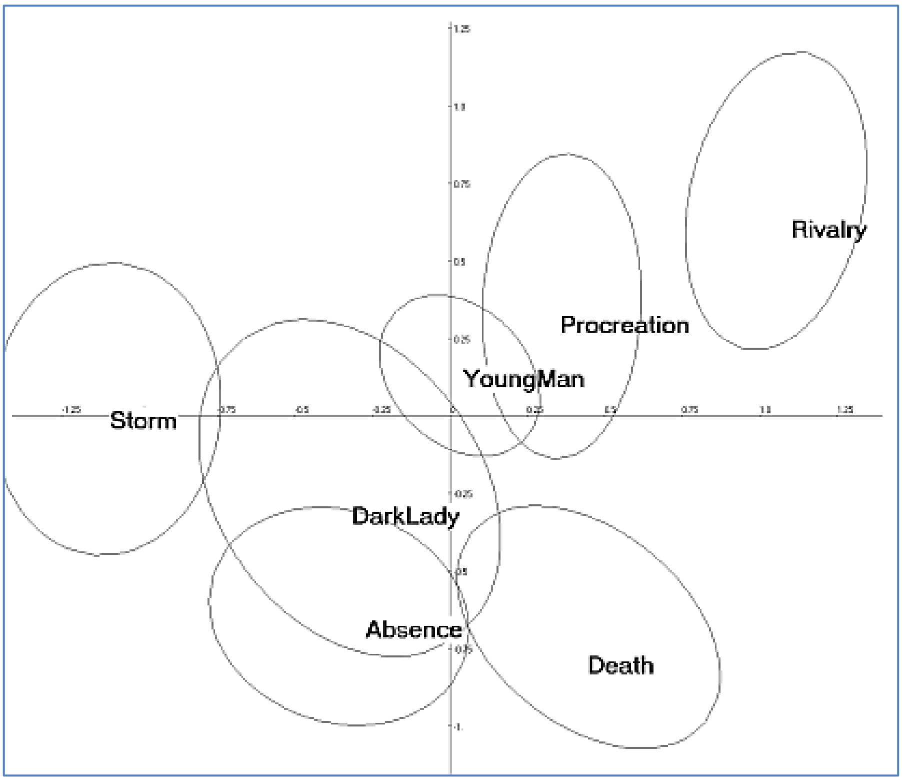


Figure 2.1. Locations of 7 topics/themes in the principal plane of the correspondence analysis of the lexical table (154 sonnets x 173 words, [min. frequency = 10]) as supplementary categorical variables. Conservative bootstrap confidence ellipses [drawing with replacement of sonnets] show the significant distances between several pairs of *a priori* themes. Note that the theme "Eternal Poetry", too much overlapping with others, is missing in this graphical display.

Figure 2.1 shows, however, that after a correspondence analysis ignoring these themes, most of their locations, when projected afterwards with the status of supplementary categories, are statistically significant on the main plane of visualization.

Obviously, the following attempts at automatic highlighting of themes in the corpus of sonnets will ignore this *a priori* partition into themes. Nor do we hope to find these themes automatically. However, knowledge of these themes from literary criticism will provide us with a framework for reading and interpreting the results.

It should be noted that statistical tools, based mainly on frequencies, detect subjects, themes or patterns almost indiscriminately. We will mainly use the term "topic" in the following.

2.3. Six Selected Methods for Finding Topics

Among the six procedures selected in the present application, four (RFA, FCA, ALO, LSA) use the singular value decomposition (SVD). The two remaining methods (NMF, LDA), less geometric, use more complex algorithms (and sometimes much longer in terms of computation time).

RFA (*Rotated Factor Analysis*) is historically the first attempt to identify unobserved “latent factors” (Thurstone, 1947, after the pioneering papers of Spearman, 1904, and Garnet, 1919). RFA involves SVD in most of the recent algorithms used to estimate the model. The topics will be defined by the words characterizing each of the factors retained after a rotation of the axes intended to facilitate their interpretation. Initially designed for numerical values, the method can be adapted to sparse frequency tables (sparse matrices). [R libraries 'psych' and 'GPRotation']

FCA (*Fragmented Correspondence Analysis*), is based on the correspondences analysis of the of fragments of texts [in our case 7 consecutive lines, that is to say a half-sonnet). The principle of this fragmentation into context units was initially proposed by Reinert (1983, 1986a) in his ALCESTE software. (see Ratineau and Déjean, 2009). The main axes of the CA are used to group these fragments, here with a hybrid classification using an Ascending Hierarchical Classification (Ward's criterion), the cut of the tree being optimized by aggregation around moving centers. At the end of the process, the themes are defined by the characteristic words of each class of fragments.

ALO (*Logarithmic Analysis*) (Kazmierczak, 1985) is similar to Spectral Mapping (Lewi, 1976) except for a weighting difference. Both methods, like CA respect the principle of distributional equivalence (stability of results when merging similar columns or rows). Applied to lexical tables, ALO often produces results similar to those of CA, with less sensitivity to outliers, an expected effect of logarithmic transformation. The calculation is operated on the sonnets, and a clustering (similar to that of the FCA) is then carried out. The themes are then the words characterizing each group of sonnets.

LSA (*Latent Semantic Analysis*) is an SVD applied to the table of TF-IDF coefficients (frequency of a term x inverse of the frequency of the documents containing the term). This technique dates back to the work of Furnas *et al.* (1988), Deerwester *et al.* (1990), Bartell *et al.* (1992). Here the documents are the sonnets. A regrouping (similar to those of FCA and ALO) is then carried out. The topics are then the sets of words characterizing each group of sonnets (see the R library: ‘lsa’, by F. Wild).

In the field of text analysis, the following two methods belong more specifically to the field of topic research (Topic Modeling)

NMF (*Non-negative Matrix Factorization*) is initially based on an equation reminiscent of SVD with, however, a constraint of positivity of the coefficients (Lee and Seung, 1999, 2001; Berry *et al.*, 2007, after Paatero and Tapper, 1994; see also Boutsidis and Gallopolous (2008), and, for an R program: Gaujoux, 2010). In the context of topic modeling, the main output of NMF is a set of topics, each of which is characterized by a list of words (“scikit-learn” software [*Python*] by Grisel O., Buitinck L ., Yau CK, In: Pedregosa *et al.* 2011).

LDA (Latent Dirichlet Allocation) (Blei *et al.*, 2003; Griffiths *et al*., 2007) is a generative statistical model (involving latent topics, words and documents) designed to discover the semantic structure of a series of texts or documents (supposed to be a mixture of a small number of topics). The method is based on a hierarchical Bayesian text analysis. (R library: 'topicmodels', and software 'scikit-learn' [*Python*]).

At this stage, we have therefore limited our investigation to six techniques drawn from a large number of approaches likely to identify topics. One could also have used the sequence of direct CA (without fragmentation of the texts) with a subsequent clustering of sonnets, the whole range of clustering techniques applied directly to the sonnets (the obtained clusters being always characterized by their more characteristic words). One could also have used the aforementioned ALCESTE method.

The work presented can obviously be extended at will. Indeed, each method also involves a whole series of parameters (frequency thresholds for words, pre-processing options such as lemmatization / tool words, size of fragments or context units, number of iterations, …). Even limited to the six methods selected, an in-depth application could significantly increase the size of this chapter.

2.4. Excerpts from the list of topics (excerpts limited to two "topics" per method)

Topics are lists of words. Deciding that a list of words deserves the topic name is up to external interpretation. The lists are of variable lengths, and in variable numbers according to the methods. The number of topics detected here by each of the six methods selected [from the implementation of the aforementioned software] varies between six and ten. Just to give an idea of the results provided by these six methods, two themes (i.e. simply two lists of words) are printed below for each method.

The identifiers of topics on future visualizations appear at the beginning of the line: the first three letters indicate the method followed by the number of topics it proposes.

1 Rotated Factor Analysis (Rotation Oblimin): RFA. *(2 topics out of 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 topics out of 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 topics out of 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 topics out of 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 topics out of 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 topics out of 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. A summary of the topics produced

How can we compare complete topic lists, since topic orders are arbitrary, and word orders within a given topic are also arbitrary?

We are here in the presence of a set of "bags of words" illustrated by the lines of section .2.4. We can make a clustering of these lines, considered as answers to an open question posed fictitiously to each topic: “What are the words that characterize you? ". Distances are calculated here as chi-square distances in the lexical contingency table (topics x words).

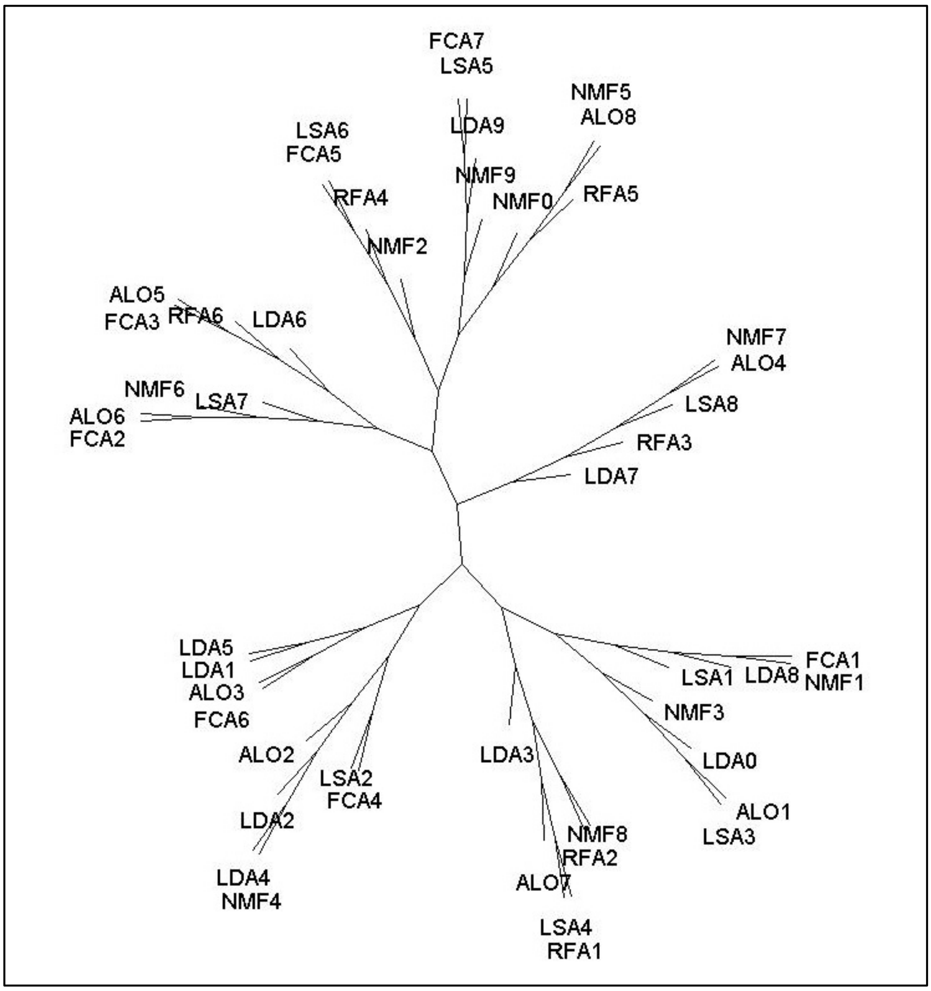


Figure 2.2. Additive tree describing the links between the 49 themes provided by the six selected methods. The identifiers are those of section .4. The distance between two topics is the chi-square distance between the lexical profiles of the topics

The technique of additive trees (Saitou and Nei, 1987) seemed to us to be the most powerful and suggestive tool for synthesizing these 49 themes in a compact form (figure 2.2). Let us recall an important property of additive trees: the real distance between two points (two topics) can be read directly on the tree as the shortest path between the two points. Here, for reasons of readability (figure 2.2) the edges all have the same length, so this property is no longer exactly verified.

We ideally expect to find a tree with as many branches as real topics in the corpus, each branch being characterized by six topics corresponding to the six methods. Such a situation occurs when each method has discovered the same real toics as the others.

The configuration observed is not as good as in this ideal situation, but we can nevertheless distinguish between six and eleven main branches, which gives an idea of the order of magnitude of the number of topics. We also note that several different methods often participate in the same branch, which suggests that this branch corresponds to a real topic discovered simultaneously by several of the methods implemented.

Example of reading figure 2.2: The branch of the tree located in the right part of the tree and at mid-height concerns the points (**NMF7, ALO4, LSA8, RFA3, LDA7**). It probably corresponds to a topic identified by five of the six methods. This branch will be found at the top of the left half of Figure 2.3, which will identify the topic as being that designated by: “Rivalry”.

2.6. Reconciliation with the *a priori* topics

What relationship can there exist between the eight topics resulting from qualitative analyzes of the sonnets by experts in Elizabethan literature and the topics proposed by each of the six methods used?

As shown in Table 2.1 above, these topics are actually groups of sonnets, so they are dominant topics per sonnet, not unconstrained topics. But the sonnets are short enough (14 lines) for this constraint not to be an obstacle to the rapprochement that we are going to attempt.

As these topics correspond to a partition into eight classes of the sonnets, we will start from the aggregate table topics x words (8 x 173) which aggregates the 154 lines of the table sonnets x words into eight classes[2].

Tableau 2.2. List of characteristic words of the eight *a priori* themes
(minimum word frequency threshold: 10, then selection by test values > 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

We will now, from this aggregated table, characterize each theme a priori, therefore each group of sonnets, by its most characteristic words (or “specificities”, see, for example: Lebart and Salem, 1994; Lebart, Salem and Berry, 1998).

The classification by additive tree of figure 2.2 will be redone with these eight a priori additional themes corresponding to a fictitious “seventh method” which we will call “literary analysis”.

An additive tree (like any tree) can give rise to plots of various but equivalent figures.

[2] Remind that the CA of the non-aggregated table (154 x 173) crosstabulating sonnets and words, (with the topics as supplementary elements, i.e.: projected *a posteriori* on to the plane as centroids of hte concerned sonnets) has led to the visualisation of figure 2.1.

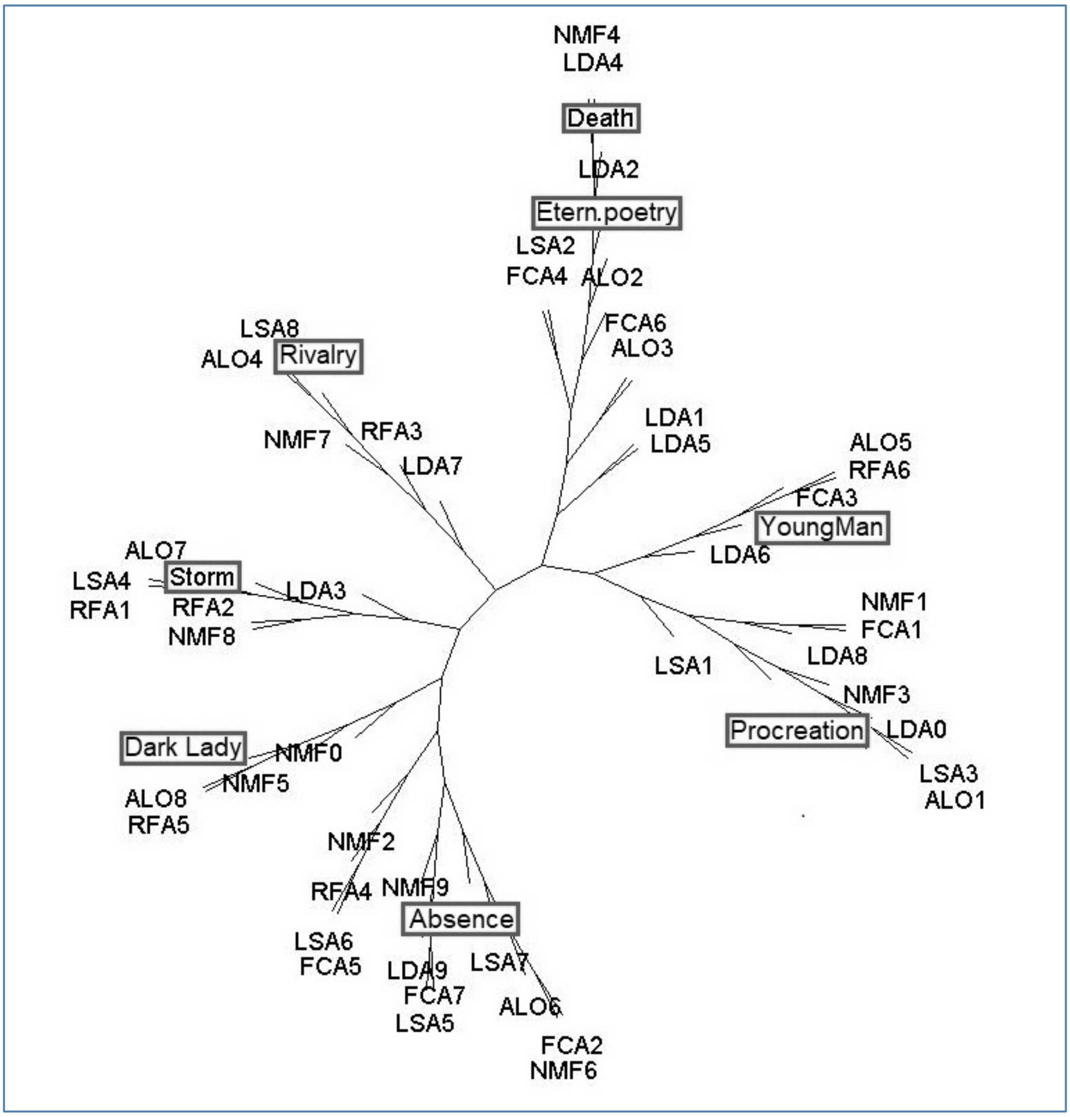


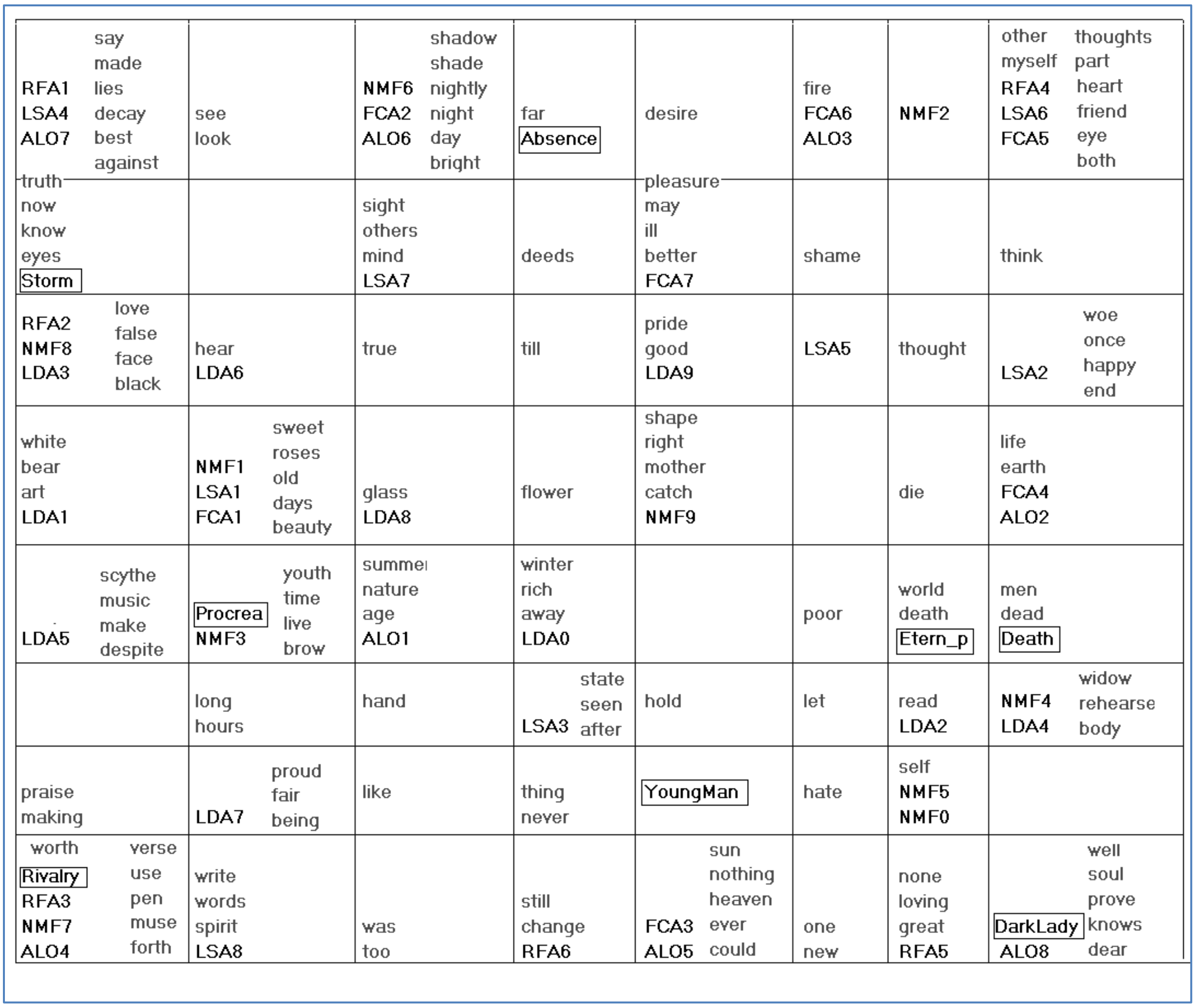
Figure 2.3. Additive tree describing the links between the 57 topics provided by the six technical methods and the “literary method” (49 topics from the previous figure 2.2 + 8 “expert” topics from the table 2.2)

Figure 2.3, which involves the *a priori* topics (or experts) does not have the same general orientations as figure 2.2, but it is satisfactory to rediscover the same main branches.

And it is a pleasant surprise to see that the “expert” topics are distributed in the main great branches of the new tree, without leaving important branches that have escaped the intuitions of the real Shakespeare specialists.

It seemed necessary to publish the two trees of figures 2.2 and 2.3 separately because figure 2.1 corresponds to a normal exploration situation, without the intervention of external information. Since the technique of additive trees does not allow the simple positioning of supplementary variables, the introduction of the "expert / a priori" topics as active elements in the calculation of the tree leading to Figure 2.3 distorted the initial representation.

The publication of figure 2.3 alone would have raised a legitimate question: what is the role of the “expert” topics in the observed structure? With the two representations, we can verify that the groupings, identified (or simply qualified) by the position of the "expert" topics, were already present in the first tree (despite the very different positions of the branches) and therefore that the six research techniques of topics (Topic Modeling) used have made it possible, to varying degrees, to discover these topics.



RFA1 LSA4 ALO7 say made lies decay best against	see look	NMF6 FCA2 ALO6 shadow shade nightly night day bright	far Absence	desire	fire FCA6 ALO3	NMF2	RFA4 LSA6 FCA5 other myself thoughts part heart friend eye both
truth now know eyes Storm		sight others mind LSA7	deeds	pleasure may ill better FCA7	shame		think
RFA2 NMF8 LDA3 love false face black	hear LDA6	true	till	pride good LDA9	LSA5	thought	LSA2 woe once happy end
white bear art LDA1	NMF1 LSA1 FCA1 sweet roses old days beauty	glass LDA8	flower	shape right mother catch NMF9		die	life earth FCA4 ALO2
LDA5 scythe music make despite	Procrea NMF3 youth time live brow	summer nature age ALO1	winter rich away LDA0		poor	world death Etern_p	men dead Death
	long hours	hand	LSA3 state seen after	hold	let	read LDA2	NMF4 LDA4 widow rehearse body
praise making	LDA7 proud fair being	like	thing never	YoungMan	hate	self NMF5 NMF0	
worth Rivalry RFA3 NMF7 ALO4 verse use pen muse forth	write words spirit LSA8	was too	still change RFA6	FCA3 ALO5 sun nothing heaven ever could	one new	none loving great RFA5	DarkLady ALO8 well soul prove knows dear

Figure 2.4 Synthesis by self-organized map of the topics obtained, the expert topics, and the words characterizing these topics from the table (57 x 139) describing the links between the 57 topics (including eight expert topics) and the 139 words describing these topics (words appearing at least twice)

Finally, figure 2.4 summarizes the lexical table crossing all the topics and the words in the form of a self-organized map (Kohonen, 1989). This map summarizes here the simultaneous representation (topics, words) in the space of the first 12 axes of a CA of this table. Simultaneous representation is interpreted as in CA: the proximity between a word and a topic is not read locally by their belonging to the same box, but globally (a topic in relation to the pattern of all the words, a word compared to that of all the topics).

Reading Figure 2.4: Consider the four boxes at the bottom left, which contain the expert topic "Rivalry". They also contain the five topics: RFA3, NMF7, ALO4, LSA8, LDA7. The proximities observed in Figure 2.3 are thus confirmed. Moreover, the words in these boxes are, probably (and probably only) characteristic of these topics (praise, worth, verse). There is therefore confirmation of the relationships between topics, and enrichment by the words that define them**.**

1.7. Conclusion of chapter 2

This empirical and rapid comparison is necessarily partial, as we have said, by the choice of methods and the parameterization of these methods, and also by the choice of the reference corpus.

The choice of methods is however not entirely arbitrary insofar as it is based on experience of exploratory methods (for the first three methods: RFA, FCA, LOA) and on notoriety (measured for example by the number of publications) for the following three methods (LSA, NMF and LDA). The corpus is however from the point of view of the search for topics a difficult corpus, which, in the usual applications on masses of heterogeneous data (in the context of *big data*), would be characterized by a single topic: "Love".

This was not an example of putting the various methods into competition but rather of showing that there can be no unique way to identifying topics. That several fairly classic techniques or arrangements of multidimensional techniques make it possible to honorably achieve this objective, even if one of them, the RFA, as we have pointed out, is markedly over a century old.

But above all we want to show that the exploratory tools (here additive trees and self-organized maps) are essential complements of visualization, validation, support, criticism for the research of topics.

"Exploratory/confirmatory" and "Supervised/Unsupervised".

In the context ofTopic Modeling which occupies a hybrid position between the exploratory and the confirmatory insofar as this research sometimes uses general or probabilistic models, we have been able to see the interest of global visualizations between the words (some of which will constitute the topics) and between the documents or categories of documents. When several methods propose to highlight topics from the same corpus, the visualization of the comparison of results (figures 2.3 and 2.4) seems essential to evaluate and criticize these methods in the context of the current application.

Deep Learning techniques, which correspond to the state of the art of supervised learning methods, require gigantic learning bases and call on several levels of representation, including regularization phases

But the important role of unsupervised approaches is now acknowledged by specialists in these techniques. One can read about the future of deep learning (LeCun *et al*., 2015) that "*Unsupervised learning had a catalytic effect in reviving interest in deep learning, but has since been overshadowed by the successes of purely supervised learning*". These same authors further add: "*…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* ".

It is true that the success of learning (recognition of images, objects, speech, writing) has invaded and facilitated our daily lives as consumers, users and communicators.

But the acquisition of knowledge, the analysis and understanding of natural and social phenomena in all their complexity sometimes require judgment, sensitivity and common sense from users. They must allow the questioning of hypotheses, analogies, intuitions and sometimes feedback on the collection and quality of data, phases for which the exploratory methods presented are irreplaceable.

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.

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.