UNIL | Université de Lausanne

Unicentre
CH-1015 Lausanne
http://serval.unil.ch

Year : 2014

Typologies textuelles et partitions musicales : dissimilarités, classification et autocorrélation.

Cocco Christelle

Cocco Christelle, 2014, Typologies textuelles et partitions musicales : dissimilarités, classification et autocorrélation.

Originally published at : Thesis, University of Lausanne

Posted at the University of Lausanne Open Archive http://serval.unil.ch
Document URN : urn:nbn:ch:serval-BIB_C6F578E3BCFB5

Droits d'auteur
L'Université de Lausanne attire expressément l'attention des utilisateurs sur le fait que tous les documents publiés dans l'Archive SERVAL sont protégés par le droit d'auteur, conformément à la loi fédérale sur le droit d'auteur et les droits voisins (LDA). A ce titre, il est indispensable d'obtenir le consentement préalable de l'auteur et/ou de l'éditeur avant toute utilisation d'une oeuvre ou d'une partie d'une oeuvre ne relevant pas d'une utilisation à des fins personnelles au sens de la LDA (art. 19, al. 1 lettre a). A défaut, tout contrevenant s'expose aux sanctions prévues par cette loi. Nous déclinons toute responsabilité en la matière.

Copyright
The University of Lausanne expressly draws the attention of users to the fact that all documents published in the SERVAL Archive are protected by copyright in accordance with federal law on copyright and similar rights (LDA). Accordingly it is indispensable to obtain prior consent from the author and/or publisher before any use of a work or part of a work for purposes other than personal use within the meaning of LDA (art. 19, para. 1 letter a). Failure to do so will expose offenders to the sanctions laid down by this law. We accept no liability in this respect.

UNIL | Université de Lausanne
Faculté des lettres

FACULTÉ DES LETTRES

SECTION DES SCIENCES DU LANGAGE ET DE L'INFORMATION

Typologies textuelles et partitions musicales : dissimilarités, classification et autocorrélation.

THÈSE DE DOCTORAT

présentée à la

Faculté des lettres
de l'Université de Lausanne

pour l'obtention du grade de
Docteur ès lettres

en Informatique et
Méthodes Mathématiques

par

Christelle Cocco

Directeur de thèse

François Bavaud

Jury

Frédéric Kaplan, EPFL
Ludovic Lebart, TELECOM-ParisTech
Aris Xanthos, UNIL

LAUSANNE
2014

UNIL | Université de Lausanne
Faculté des lettres

FACULTÉ DES LETTRES

SECTION DES SCIENCES DU LANGAGE ET DE L'INFORMATION

Typologies textuelles et partitions musicales : dissimilarités, classification et autocorrélation.

THÈSE DE DOCTORAT

présentée à la

Faculté des lettres
de l'Université de Lausanne

pour l'obtention du grade de
Docteur ès lettres

en Informatique et
Méthodes Mathématiques

par

Christelle Cocco

Directeur de thèse

François Bavaud

Jury

Frédéric Kaplan, EPFL
Ludovic Lebart, TELECOM-ParisTech
Aris Xanthos, UNIL

LAUSANNE
2014

UNIL | Université de Lausanne
Faculté des lettres

IMPRIMATUR

Le Décanat de la Faculté des lettres, sur le rapport d'une commission composée de :

Directeur de thèse :	
Monsieur François Bavaud	Professeur, Faculté des lettres, Université de Lausanne
Membres du jury :	
Monsieur Ludovic Lebart	Professeur, Télécom ParisTech, France
Monsieur Frédéric Kaplan	Professeur, EPFL
Monsieur Aris Xanthos	MER, Faculté des lettres, Université de Lausanne

autorise l'impression de la thèse de doctorat de

MADAME CHRISTELLE COCCO

intitulée

Typologies textuelles et partitions musicales :
dissimilarités, classification et autocorrélation.

sans se prononcer sur les opinions du candidat / de la candidate.

La Faculté des lettres, conformément à son règlement, ne décerne aucune mention.

Lausanne, le 2 juillet 2014

François Rosset
Doyen de la Faculté des lettres

Résumé

Axée dans un premier temps sur le formalisme et les méthodes, cette thèse est construite sur trois concepts formalisés : une table de contingence, une matrice de dissimilarités euclidiennes et une matrice d'échange. À partir de ces derniers, plusieurs méthodes d'Analyse des données ou d'apprentissage automatique sont exprimées et développées : l'analyse factorielle des correspondances (AFC), vue comme un cas particulier du *multidimensional scaling* ; la classification supervisée, ou non, combinée aux transformations de Schoenberg ; et les indices d'autocorrélation et d'autocorrélation croisée, adaptés à des analyses multivariées et permettant de considérer diverses familles de voisinages. Ces méthodes débouchent dans un second temps sur une pratique de l'analyse exploratoire de différentes données textuelles et musicales.

Pour les données textuelles, on s'intéresse à la classification automatique en types de discours de propositions énoncées, en se basant sur les catégories morphosyntaxiques (CMS) qu'elles contiennent. Bien que le lien statistique entre les CMS et les types de discours soit confirmé, les résultats de la classification obtenus avec la méthode K-means, combinée à une transformation de Schoenberg, ainsi qu'avec une variante floue de l'algorithme K-means, sont plus difficiles à interpréter. On traite aussi de la classification supervisée multi-étiquette en actes de dialogue de tours de parole, en se basant à nouveau sur les CMS qu'ils contiennent, mais aussi sur les lemmes et le sens des verbes. Les résultats obtenus par l'intermédiaire de l'analyse discriminante combinée à une transformation de Schoenberg sont prometteurs. Finalement, on examine l'autocorrélation textuelle, sous l'angle des similarités entre diverses positions d'un texte, pensé comme une séquence d'unités. En particulier, le phénomène d'alternance de la longueur des mots dans un texte est observé pour des voisinages d'empan variable. On étudie aussi les similarités en fonction de l'apparition, ou non, de certaines parties du discours, ainsi que les similarités sémantiques des diverses positions d'un texte.

Concernant les données musicales, on propose une représentation d'une partition musicale sous forme d'une table de contingence. On commence par utiliser l'AFC et l'indice d'autocorrélation pour découvrir les structures existant dans chaque partition. Ensuite, on opère le même type d'approche sur les différentes voix d'une partition, grâce à l'analyse des correspondances multiples, dans une variante floue, et à l'indice d'autocorrélation croisée. Qu'il s'agisse de la partition complète ou des différentes voix qu'elle contient, des structures répétées sont effectivement détectées, à condition qu'elles ne soient pas transposées. Finalement, on propose de classer automatiquement vingt partitions de quatre compositeurs différents, chacune représentée par une table de contingence, par l'intermédiaire d'un indice mesurant la similarité de deux configurations. Les résultats ainsi obtenus permettent de regrouper avec succès la plupart des œuvres selon leur compositeur.

Abstract

Focused on formalism and methods in its first part, this thesis is constructed from three basic formalised concepts, namely : a contingency table, an Euclidean dissimilarity matrix and an exchange matrix. Those concepts permit the expression and development of several Data Analysis or Machine Learning methods : Correspondence Analysis (CA), interpreted as a particular case of Multidimensional Scaling ; classification and clustering, combined with Schoenberg transformations ; and the autocorrelation and cross-autocorrelation indices, adapted to multivariate analysis and allowing the consideration of various neighbourhood families. In the second part of the thesis, these methods lead to an Exploratory Data Analysis of textual and musical data of various types.

For textual data, we are interested in clustering clauses into discourse types, based upon the distribution of part-of-speech (POS) tags in the clauses. Although the statistical link between POS tags and discourse types is significant, the results obtained with the K-means algorithm or a fuzzy variant of it, possibly combined with a Schoenberg transformation, remain difficult to interpret. We also deal with multi-label classification into dialog acts of turns, again based on the POS tags they contain, but also on lemmas and on the meaning of verbs. Results obtained by means of discriminant analysis combined with a Schoenberg transformation are promising. Finally, we examine the textual autocorrelation, in terms of similarities between various positions in a text, thought as a sequence of localized units. In particular, the phenomenon of word length alternation in a text is studied for a family of neighbourhoods of variable span. We also consider presence-absence similarities, according to the apparition of specific POS, as well as the semantic similarities between textual positions.

Regarding musical data, we propose to represent a musical score as a contingency table. We begin by using CA and the autocorrelation index to discover underlying structures within each score. Then, we apply the same approach on the different voices in a musical score, with a procedure alike to a fuzzy variant of multiple correspondence analysis and making use of the cross-autocorrelation index. Whether in the whole musical scores or in different voices they contain, repeated structures are actually detected, provided they are not transposed. Finally, we propose to cluster twenty musical scores by four different composers, each represented by a contingency table, by introducing a similarity index between the pairs of configurations. A majority of scores turn out to be thus successfully regrouped according to their composer.

Remerciements

J'aimerais remercier tous les gens que j'ai rencontrés durant cette thèse, ceux avec qui j'ai pu échanger, même brièvement, ainsi que ceux qui m'ont donné leur avis ou qui m'ont motivée. Parmi toutes ces personnes, et j'espère que vous serez nombreux à vous reconnaître dans cette description, un grand merci à ma famille et mes amis.

Pour rester succincte, je ne vais pas nommer tout le monde, mais simplement revenir sur les gens sans qui cette thèse n'aurait pas pu exister. Pour commencer, j'aimerais remercier Pathé Barry, un ami de longue date, et Jérémie Mariller, mon compagnon, sans qui je ne me serais jamais lancée dans l'aventure d'une thèse. Ils m'ont tous deux encouragée à postuler pour ce doctorat en informatique et méthodes mathématiques de la faculté des Lettres, domaine relativement éloigné de mes études de master. J'aimerais particulièrement remercier Jérémie pour m'avoir soutenue durant tout mon doctorat.

Merci encore à François Bavaud et Aris Xanthos sans qui rien n'aurait commencé. Ils ont tous deux consacré beaucoup de temps à partager leurs expériences avec moi, ce qui m'a permis de me familiariser avec ce nouveau domaine, la recherche et le monde académique. En particulier, je remercie François pour sa disponibilité, ses conseils et ses nombreuses relectures.

J'aimerais aussi remercier tous les membres de l'ancienne section d'Informatique et Méthodes Mathématiques, ainsi que tous ceux de la nouvelle section des Sciences du Langage et de l'Information. Parmi eux, je voudrais en particulier remercier Jérôme Jacquin avec qui le projet qui m'a permis d'écrire le chapitre 4 a débuté. Concernant ce même chapitre, mes remerciements vont à Gilles Merminod pour ses conseils et nos discussions qui m'ont éclairée sur la dimension linguistique de ce sujet. Merci aussi à Guillaume Guex et Théophile Emmanouilidis avec qui nous avons non seulement partagé un bureau, mais aussi des idées, des avis et des discussions.

Concernant les chapitres 7 et 8, j'aimerais remercier Jamil Alioui, qui m'a aidée à me familiariser avec les fichiers MIDI ; ainsi que le Dr. Daniel Müllensiefen, pour les différentes pistes de départ à propos des recherches actuelles dans le domaine, et surtout des formats symboliques, qu'il a proposées à François et dont j'ai bénéficié.

Merci aussi à l'équipe de la Formation Doctorale Interdisciplinaire de la faculté des Lettres, et à l'équipe des Humanités Digitales de l'UNIL et l'EPFL, pour m'avoir donné l'occasion de présenter mon travail et d'échanger des idées. Merci enfin, particulièrement, aux membres du jury, dont les remarques, toutes pertinentes, m'ont permis de prendre du recul sur ma thèse, d'améliorer différents points et d'aboutir à un ensemble plus clair et cohérent. J'espère que vous aurez du plaisir à lire cette thèse.

Table des matières

Introduction 1

I Méthodes et formalisme 7

1 Table de contingence et analyse factorielle des correspondances 9

1.1 Table de contingence et matrice documents-termes . . . 9

1.2 Lien entre deux variables catégorielles . . . 10

1.2.1 Test d'indépendance du khi2 . . . 10

1.2.2 Cas des variables binaires . . . 10

1.3 Dissimilarité du χ^2 et dissimilarités euclidiennes carrées . . . 12

1.3.1 Dissimilarité du χ^2 et dualité . . . 12

1.3.2 Dissimilarités euclidiennes carrées . . . 12

1.3.3 Principe de Huygens . . . 14

1.3.4 Transformations de Schoenberg . . . 15

1.4 Analyse factorielle des correspondances . . . 15

1.4.1 MDS . . . 16

2 Classification supervisée et non supervisée 17

2.1 Classification non supervisée . . . 17

2.1.1 Classification ascendante hiérarchique, critère de Ward . . . 19

2.1.2 K-means sur les dissimilarités . . . 20

2.1.3 K-means flou sur les dissimilarités . . . 21

2.2 Classification supervisée . . . 22

2.2.1 Analyse discriminante sur les dissimilarités . . . 23

2.3 Évaluation . . . 24

2.3.1 Accord entre partitions . . . 24

2.3.2 Précision, rappel et F-mesure . . . 25

3 Indices d'autocorrélation et d'autocorrélation croisée 29

3.1 Matrice d'échange . . . 29

3.1.1 Exemples . . . 30

3.2 Indice d'autocorrélation . . . 31

3.2.1 Test d'autocorrélation . . . 32

3.3 Indice d'autocorrélation croisée . . . 32

II Applications textuelles 35

4 Classification non supervisée en types de discours 37
4.1 Données . . . 37
4.1.1 Types de discours et annotation . . . 38
4.1.2 Corpus . . . 42
4.1.3 Prétraitement . . . 43
4.1.4 Analyse préliminaire . . . 44
4.2 Visualisation . . . 48
4.2.1 Propositions et CMS . . . 48
4.2.2 Types de discours et CMS avec bootstrap . . . 51
4.3 Classification non supervisée et résultats . . . 55
4.3.1 K-means . . . 55
4.3.2 K-means flou . . . 58
4.4 Discussion . . . 64

5 Classification supervisée multi-étiquette en actes de dialogue 67
5.1 Données . . . 68
5.2 Liens entre étiquettes . . . 69
5.2.1 Traitements . . . 69
5.2.2 Résultats . . . 69
5.3 Classification supervisée . . . 71
5.3.1 Prétraitements et caractéristiques . . . 71
5.3.2 Traitements . . . 72
5.3.3 Résultats . . . 74
5.4 Discussion . . . 78

6 Autocorrélation textuelle 81
6.1 Longueur des mots . . . 81
6.1.1 Principe . . . 81
6.1.2 Traitements et résultats . . . 81
6.2 Parties du discours . . . 84
6.2.1 Dissimilarités binaires relatives à une partie du discours . . . 84
6.2.2 Traitements et résultats . . . 84
6.3 Sens des mots selon WordNet . . . 85
6.3.1 Dissimilarités sémantiques . . . 85
6.3.2 Autocorrélation sémantique . . . 87
6.3.3 MDS et autocorrélation sur les premiers facteurs . . . 89
6.4 Discussion . . . 94

III Applications musicales 97

7 Formats symboliques de données musicales 99
7.1 Partitions . . . 99
7.2 Format MIDI en bref . . . 101
7.3 Formats « textuels » . . . 101
7.3.1 Le format Melisma . . . 101
7.3.2 Le format ABC . . . 104
7.3.3 Le format Humdrum . . . 105
7.3.4 Comparaison de ces trois formats . . . 107

8 Analyse de données musicales 109
8.1 Représentation des données . . . 109
8.1.1 Formalisme . . . 109
8.1.2 Pré-traitement . . . 111
8.2 Analyses d'une partition . . . 112
8.2.1 Traitements . . . 112
8.2.2 Partition monophonique . . . 113
8.2.3 Partitions polyphoniques avec un seul instrument . . . 115
8.2.4 Partition polyphonique avec plusieurs instruments . . . 119
8.3 Analyses inter-voix . . . 121
8.3.1 Traitements . . . 121
8.3.2 Un canon . . . 122
8.3.3 Un quatuor à cordes . . . 124
8.4 Analyses inter-partitions . . . 126
8.4.1 Données . . . 126
8.4.2 Traitement et résultat . . . 127
8.5 Discussion . . . 129

Conclusion et discussion 131

Annexes 139

A Textes de Maupassant annotés 141
A.1 L'Orient . . . 141
A.2 Le Voleur . . . 148
A.3 Un Fou ? . . . 155
A.4 Un Fou . . . 166

B Liens entre types de discours et CMS 179
B.1 Tables des effectifs croisés . . . 180
B.2 Khi2 ponctuel. . . . 182

C Classification non supervisée en types de discours 185
C.1 K-means. . . . 185
C.1.1 Indices d'accord entre partitions . . . 186
C.1.2 V de Cramer . . . 188
C.2 K-means flou . . . 190

Bibliographie 202

Introduction

Cette thèse se propose d'étudier et de révéler certaines structures existant dans des données de type textuel ou musical, par l'intermédiaire de méthodes standard ou novatrices en Analyse des données. En d'autres termes, elle adopte essentiellement l'approche de l'analyse exploratoire des données, par opposition aux approches inférentielles ou basées sur des modèles *a priori*. Alors que ces dernières sont basées sur des hypothèses ou des postulats *a priori* qu'il s'agira de confirmer ou de rejeter, le but est ici de « laisser parler les données » à l'aide d'algorithmes et d'ordinateurs, *i.e.* d'extraire la structure des données qui pourra être *ensuite* interprétée. En d'autres termes :

> [...] La notion de forme ou de modèle devrait émerger d'une mer de données, non par des postulats nominalistes ou des axiomes *a priori*, ni par des mesures trop fragmentaires de faits isolés, en eux-mêmes dénués de sens puisqu'ils dépendent du milieu ambiant et se réorganisent sans cesse, mais par la synthèse simultanée (synthèse pris au sens éthymologique [*sic*] : mettre ensemble) d'un bon nombre de faits élémentaires qui nous aide à gravir les échelons de la hiérarchie des causes. Mais un cerveau humain ne peut accomplir une synthèse multidimensionnelle sans faire de nombreux choix arbitraires qui ôtent souvent toute signification au résultat. Il faut donc l'aide d'une calculatrice pour appliquer aux données préalablement rassemblées un ensemble de calculs ou plutôt de transformations telles qu'on puisse lire avec sûreté à la sortie ce qui, à l'entrée, était indéchiffrable. (Benzécri *et al.*, 1973, pp. 15-16)

Dans le passage ci-dessus, extrait de Benzécri *et al.* (1973), on trouve une expression centrale pour l'analyse des données : « synthèse multidimensionnelle ». Effectivement, le terme d'Analyse des données regroupe plusieurs méthodes, toutes basées sur des statistiques *multidimensionnelles* et descriptives, avec pour objectif de *synthétiser* l'information contenue dans les données en réduisant le nombre de dimensions effectives, grâce à la redondance générée par les relations entre les descripteurs. L'ensemble de ces méthodes peut être divisé en deux grandes familles principales. La première famille de méthodes permet de représenter graphiquement l'information synthétisée, en deux dimensions par exemple, ce qui la rend intelligible pour un être humain. Quant à la seconde famille de méthodes, elle vise à classifier automatiquement les observations en les regroupant de la manière la plus homogène possible, selon leurs profils. Benzécri *et al.* (1973) les commentent ainsi lorsqu'ils abordent la question de la reconnaissance de formes dans l'introduction générale du premier tome sur « L'Analyse des Données » :

> [...] C'est le problème de la reconnaissance des formes : traiter mécaniquement des informations qui ne soient ni réduites à une expression logique séquentielle et définie à l'avance [...], ni représentées analogiquement par des grandeurs physiques [...], mais gardent la multidimensionnalité présente presque partout dans la nature. [...]
>
> La portée de telles recherches dépasse en fait l'objectif initial limité que nous leur avons assigné : réussir dans une ambiance multidimensionnelle et d'abord confus [*sic*], des tâches

> de discrimination accessibles aux moins doués des hommes ou aux animaux. On ne résoudra sur machine de tels problèmes qu'au moyen d'algorithmes de classification et de réduction du nombre de dimensions, i.e. d'algorithmes qui à partir d'un vaste ensemble d'individus (de nature quelconque...), chacun décrit par un grand nombre de mesures numériques ou de relations, reconnaissent les propriétés structurellement importantes et les dimensions selon lesquelles se répartissent continûment les membres de l'ensemble étudié : or, ces propriétés et ces dimensions ne sont généralement aucune de celles que comportait la description initiale, elles en sont des fonctions souvent complexes [...] (Benzécri *et al.*, 1973, pp. 3-4)

Pour pouvoir exprimer les méthodes d'Analyse des données spécifiques qui seront utilisées dans cette thèse, il est nécessaire de définir clairement un formalisme. Le formalisme mathématique adopté ici est relativement succinct et repose sur trois concepts formalisés, à savoir :

— une table de contingence,
— une matrice de dissimilarités euclidiennes carrées et
— une matrice d'échange.

Techniquement, chaque objet est caractérisé par un certain nombre d'attributs (ou caractéristiques). La table de contingence, connue aussi sous le terme de table documents-termes en statistique textuelle, compte le nombre de chacun des attributs contenu dans chaque objet et constitue ainsi le premier concept du formalisme. Le second concept consiste en une matrice de similarités ou de dissimilarités construite entre les objets *en fonction de* leurs attributs. Concernant la matrice d'échange, elle sert à modéliser le voisinage spatial ou temporel qui peut exister entre les différents objets.

Partant de ces trois concepts, trois types de méthodes sont développés, les deux premiers correspondant aux deux grandes familles de méthodes d'Analyse des données décrites ci-dessus. Premièrement, pour visualiser l'information synthétisée que contient une table de contingence, on utilisera l'analyse factorielle des correspondances qui permet de représenter simultanément les objets (ou documents) et les attributs (ou termes) sur un graphique. Ainsi, il est possible de visualiser les correspondances des objets par rapport aux attributs, et inversement. Cette méthode, spécifique aux tables de contingence, est très connue et très populaire, comme en témoignent Benzécri *et al.* (1980) qui y consacrent la totalité de leur deuxième tome sur « L'Analyse des Données » et déclarent :

> [...] l'analyse des correspondances, méthode qui bien mieux que toute autre nous a permis de découvrir les faits de structure que recèle un tableau de données quel qu'il soit. (Benzécri *et al.*, 1980, p. VII)

Deuxièmement, on traitera de la classification en se basant sur la matrice de dissimilarités euclidiennes carrées. Alors que l'Analyse des données vise, comme il a déjà été mentionné, à faire émerger la structure des données sans *a priori*, ce qui correspond à ce que l'on appelle la classification non supervisée, on traitera *en plus* de classification supervisée. Dans ce deuxième cas, l'information contenue dans les données est également synthétisée, mais le but est alors de créer un algorithme capable d'identifier l'appartenance à des groupes définis *a priori*, en se basant sur un échantillon d'apprentissage.

Troisièmement, on s'intéressera à mesurer la proximité entre des objets, en fonction des attributs qui les composent, du point de vue de leur voisinage spatial ou temporel. Pour ce faire, on utilise deux des concepts formalisés présentés plus haut : la matrice de dissimilarités euclidiennes carrées et la matrice d'échange. Leur interaction est à la base de la construction des indices d'autocorrélation et d'autocorrélation croisée.

Si le formalisme et les méthodes associées sont centraux dans ce travail, leurs applications textuelles et musicales le sont également. La suite de l'étude vise ainsi à extraire des structures existant dans les données, structures qui peuvent être attendues, ou au contraire nouvelles, justement découvertes grâce au formalisme. Dans certains cas, l'analyse a été poussée au-delà de la phase exploratoire stricte : il s'agit alors de s'assurer que les structures révélées ne soient pas le fruit du hasard en recourant alors à l'approche inférentielle.

Sans être spécialiste de l'étude des textes ou de la musique, on se positionne ici un peu comme le microscope d'un biologiste qui lui permet d'observer un objet ou une substance de plus près et différemment, donc d'un autre point de vue. Cette thèse n'ambitionne donc pas de développer ou de proposer de nouvelles théories dans des domaines, tels que la littérature, la linguistique, la musicologie ou encore la psychologie, mais plutôt d'offrir un nouveau point de vue à l'une ou l'autre de ces disciplines. En effet, pour reprendre encore les mots de Benzécri et ses collaborateurs dans l'avant-propos de leur premier tome :

> [...] La puissance du calcul électronique permet au statisticien d'aborder d'un point de vue unique les ensembles de faits les plus vastes et les plus divers. Aussi ne s'étonnera-t-on pas qu'il doive être traité ici tant des sciences de la nature [...], que des sciences de l'homme : Psychologie, Linguistique, Economie, Politique [...] ; cependant que la méthode même de la connaissance est l'objet ultime de cette recherche.
>
> Dans chaque volume, on s'efforcera de placer simultanément des exposés théoriques, des programmes de calcul, des exemples d'application. Nous ne croyons pas devoir dissimuler que c'est à ces exemples que va notre prédilection. Nous sommes en effet convaincu que le statisticien a tout à apprendre de la nature et que la statistique, refrénant son vol mathématique, doit s'honorer d'être une science expérimentale. Bien mieux qu'à des modèles conjecturaux, c'est à l'observation qu'on doit demander quel est l'ordre de la réalité : le mérite du calculateur étant de découvrir sans parti pris, sans *a priori*, quels courants de lois traversent l'océan des faits. (Benzécri *et al.*, 1973, p. V)

Ou encore, lorsque Lebart, Morineau et Piron (1995) expliquent la différence qu'il existe entre statistique descriptive et statistique descriptive *multidimensionnelle* dans l'introduction de leur ouvrage :

> Mais le passage au multidimensionnel induit un changement qualitatif important. On ne dit pas en effet que des microscopes ou des appareils radiographiques sont des instruments de description, mais bien des instruments d'observation ou d'exploration, et aussi de recherche. La réalité multidimensionnelle n'est pas seulement simplifiée parce que complexe, mais aussi explorée parce que cachée.
>
> Le travail de préparation et de codage des données, les règles d'interprétation et de validation des représentations fournies par les techniques utilisées dans le cas multidimensionnel n'ont pas la simplicité rencontrée avec la statistique descriptive élémentaire. Il ne s'agit pas seulement de présenter, mais d'analyser, de découvrir, parfois de vérifier et prouver, éventuellement de mettre à l'épreuve certaines hypothèses. (Lebart *et al.*, 1995, p. 1)

En résumé, il s'agira donc de revisiter, à partir de leur définition de base, des méthodes bien connues en Analyse des Données, tout en les combinant avec des éléments théoriques moins balisés ou plus originaux ; et, aussi, de les appliquer sur de nouveaux types de données, c'est-à-dire sur des données sur lesquelles ces analyses n'ont pas (ou peu) encore été pratiquées, à notre connaissance.

En particulier, on s'intéressera, pour les applications textuelles, à la classification automatique (ou non supervisée) de propositions énoncées en types de discours, à la classification supervisée de tours de parole en actes de dialogue, ainsi qu'à la mesure de l'indice d'autocorrélation sur différents textes, en considérant différents attributs et différents voisinages.

Concernant la musique, on se concentrera sur trois niveaux différents. Premièrement, on observera la structure existant dans des partitions considérées séparément. Deuxièmement, on s'intéressera à la structure des différentes voix qui composent une partition, ainsi qu'aux liens qui existent entre elles. Troisièmement, on traitera plusieurs partitions que l'on regroupera à l'aide d'une méthode de classification non supervisée.

Il faut remarquer que le nombre d'applications présenté ici est clairement restreint par rapport aux possibilités du formalisme et des méthodes exposées. C'est pourquoi ces derniers ont volontairement était présentés de manière systématique, avec un effort de clarté et de simplicité, pour pouvoir être envisagés sur d'autres données. Plus précisément, ces méthodes se veulent entièrement transparentes, tout à l'opposé des « boîtes noires ».

Afin de mener à bien ce programme, la thèse est structurée ainsi : la partie I présente les méthodes principales utilisées dans ce travail, toutes basées sur des *dissimilarités euclidiennes carrées*, extraites le plus souvent à partir d'une *table de contingence*. Ensuite, la partie II expose plusieurs applications de ces méthodes sur diverses données textuelles. Finalement, la partie III présente une exploration de ces méthodes sur des données musicales.

En particulier, la **Partie I** expose les définitions et le formalisme utilisés dans le reste de la thèse. Bien que cette première partie expose des concepts généraux, elle n'a pas pour ambition de donner une revue complète des méthodes existantes, mais plutôt de définir les méthodes essentielles qui serviront de base aux applications présentées dans le reste de la thèse. Elle se compose de trois chapitres.

Pour commencer, le **chapitre 1** rappelle les notions relatives à une table de contingence, telles que le quotient d'indépendance ou les dissimilarités du khi2. Plus précisément, les dissimilarités du khi2 constituent des dissimilarités euclidiennes carrées, dont les propriétés fondamentales qui en découlent sont rappelées, parmi lesquelles deux éléments essentiels : d'une part la notion de *transformations de Schoenberg*, transformant des dissimilarités euclidiennes carrées en d'autres dissimilarités euclidiennes carrées dans un espace de plus haute dimensionnalité ; d'autre part l'analyse factorielle des correspondances, *obtenue comme un cas particulier* du *multi-dimensional scaling*. Ces deux derniers points, sans être entièrement originaux, s'écartent toutefois des exposés couramment rencontrés dans la littérature « ordinaire ».

Ensuite, le **chapitre 2** expose différentes techniques de classification, supervisée ou non, toutes basées sur les dissimilarités euclidiennes carrées présentées précédemment, ainsi que les approches permettant d'évaluer ces différentes classifications. En particulier, un autre point original de ce travail est qu'en exposant les différentes techniques de classification à partir du concept formalisé de la matrice des dissimilarités, il est possible de les combiner aux transformations de Schoenberg et donc d'en étendre la portée.

Finalement, le **chapitre 3** introduit le concept de matrice d'échange, issue de la notion de poids spatiaux en statistique spatiale et formalisant la notion de voisinage. Deux mesures peuvent alors être définies, à savoir l'indice d'autocorrélation, basé sur la relation entre une matrice de dissimilarités euclidiennes carrées et une matrice d'échange, ainsi que l'indice d'autocorrélation croisée. Ces deux indices permettent, d'une part, d'étendre les notions d'autocorrélation et de corrélation croisée des séries temporelles ordinaires à des séries multivariées ; et, d'autre part, de généraliser la notion de décalage à une notion de voisinage. Bien que ces deux indices ne soient pas complétement originaux, ils n'ont été que très peu utilisés sous cette forme en analyse textuelle ou musicale *multivariée*.

Trois types d'applications textuelles sont étudiées dans la **Partie II**. Les chapitres 4 et 5 s'intéressent à la classification d'unités linguistiques. Plus exactement, le **chapitre 4** traite de la classification non supervisée de propositions énoncées en types de discours ; et le **chapitre 5**, de la classification supervisée multi-étiquette de tours de parole en actes de dialogue. Dans ces deux chapitres, les données sont représentées sous la forme de tables de contingence inédites. Le **chapitre 6** s'intéresse à mesurer les similarités entre diverses positions d'un texte, compris comme une séquence d'unités, par l'intermédiaire de l'indice d'autocorrélation, ce qui semble constituer un point de vue novateur.

Enfin, la **Partie III** comprend deux chapitres : le **chapitre 7**, qui présente différents formats symboliques de données musicales, et le **chapitre 8**, qui propose une analyse exploratoire de partitions musicales polyphoniques. Dans ce dernier, les partitions polyphoniques sont représentées, une fois de plus, par des tables de contingence (peu ou pas exploitées sous la forme spécifique présentée dans le chapitre pour des données musicales symboliques), ce qui permet d'utiliser le formalisme et les méthodes de la partie I. Spécifiquement, les partitions seront d'abord étudiées dans leur ensemble, grâce à l'analyse factorielle des correspondances et à l'indice d'autocorrélation, pour une unité temporelle donnée. Ensuite, les différentes voix d'une même partition seront analysées, à l'aide de l'indice d'autocorrélation croisée et d'une variante

de l'analyse des correspondances multiples. Finalement, à partir de la représentation en tables de contingence, une approche originale de classification non supervisée de plusieurs partitions est proposée.

Il faut encore préciser que le matériel exposé dans ce travail reprend, en bonne partie, du matériel déjà publié ou en voie de l'être, à savoir :

— Cocco, C., Pittier, R., Bavaud, F. et Xanthos, A. (2011). Segmentation and Clustering of Textual Sequences: a Typological Approach. *In Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*, pp. 427–433. Hissar, Bulgaria: RANLP 2011 Organising Committee.

— Cocco, C. (2012*a*). Catégorisation automatique de propositions textuelles en types de discours. *In Lire demain : des manuscrits antiques à l'ère digitale = Reading tomorrow : from ancient manuscripts to the digital era*, pp. 689–707. Lausanne: PPUR.

— Cocco, C. (2012*b*). Discourse Type Clustering using POS n-gram Profiles and High-Dimensional Embeddings. *In Proceedings of the Student Research Workshop at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 55–63. Avignon, France: Association for Computational Linguistics.

— Bavaud, F., Cocco, C. et Xanthos, A. (2012). Textual autocorrelation: formalism and illustrations. *In JADT 2012: 11èmes Journées internationales d'Analyse statistique des Données Textuelles*, pp. 109–120.

— Cocco, C. (2014). Classification supervisée multi-étiquette en actes de dialogue : analyse discriminante et transformations de schoenberg. *In JADT 2014: 12èmes Journées internationales d'Analyse statistique des Données Textuelles*, pp. 147–160.

— Cocco, C. et Bavaud, F. (accepté pour publication). Correspondence Analysis, Cross-Autocorrelation and Clustering in Polyphonic Music. *In Data Analysis, Learning by Latent Structures, and Knowledge Discovery*, Studies in Classification, Data Analysis, and Knowledge Organization. Berlin; Heidelberg: Springer.

— Bavaud, F., Cocco, C. et Xanthos, A. (accepté pour publication). Textual navigation and autocorrelation. *In* G. Mikros et J. Mačutek (Eds.), *Sequences in Language and Text*, Quantitative Linguistics. Berlin: De Gruyter.

Bien qu'il ne soit que très peu exploité dans cette thèse, du matériel connexe aux méthodes utilisées dans ce travail a également été développé :

— Bavaud, F. et Cocco, C. (accepté pour publication). Factor Analysis of Local Formalism. *In Data Analysis, Learning by Latent Structures, and Knowledge Discovery*, Studies in Classification, Data Analysis, and Knowledge Organization. Berlin; Heidelberg: Springer

Partie I

MÉTHODES ET FORMALISME

CHAPITRE 1

Table de contingence et analyse factorielle des correspondances

Ce premier chapitre, dont le but est de fixer les notations, traite des tables de contingence et de l'analyse factorielle des correspondances. Cette dernière permet d'analyser les dépendances entre deux variables catégorielles tout en les visualisant. Alors que la plupart des ouvrages proposent de pratiquer l'analyse des correspondances directement sur les tables de contingences, la méthode exposée ici (section 1.4) se base sur un MDS (*multi-dimensional scaling*) pondéré des dissimilarités du khi2 (section 1.3.1) obtenues sur la table de contingence (section 1.1). Bien que ces deux méthodes aboutissent au même résultat, la seconde permettra d'introduire plus simplement les concepts des prochains chapitres et d'exploiter les transformations de Schoenberg (section 1.3.4). On reviendra aussi sur les différentes mesures possibles du lien entre deux variables catégorielles (section 1.2) et sur les propriétés des dissimilarités du khi2 (section 1.3) qui sont aussi euclidiennes carrées (section 1.3.2).

1.1 Table de contingence et matrice documents-termes

Soit deux variables catégorielles X et Y avec, respectivement, m_1 et m_2 modalités. La table de contingence $N = (n_{jk})$ compte les effectifs n_{jk} de la modalité $j = 1, \ldots, m_1$ de X et de la modalité $k = 1, \ldots, m_2$ de Y. Le profil marginal de la ligne j est défini comme $n_{j\bullet} = \sum_k n_{jk}$; celui de la colonne k, comme $n_{\bullet k} = \sum_j n_{jk}$; et la taille de l'échantillon, comme $n_{\bullet\bullet} = \sum_{jk} n_{jk}$. La table 1.1 propose un résumé de ces différentes notations.

		Modalités de Y					
		1	...	k	...	m_2	
Modalités de X	1	n_{11}	...	n_{1k}	...	n_{1m_2}	$n_{1\bullet}$
	...	...	...	...	...	...	...
	j	n_{j1}	...	n_{jk}	...	n_{jm_2}	$n_{j\bullet}$
	...	...	...	...	...	...	...
	m_1	$n_{m_1 1}$	...	$n_{m_1 k}$	...	$n_{m_1 m_2}$	$n_{m_1\bullet}$
		$n_{\bullet 1}$	...	$n_{\bullet k}$	...	$n_{\bullet m_2}$	$n_{\bullet\bullet}$

TABLE 1.1 – Vue synthétique des notations d'une table de contingence $N = (n_{jk})$.

La matrice documents-termes, qui est souvent utilisée en analyse textuelle, est un cas particulier de la table de contingence. Dans ce cas, les modalités j de X représentent différents documents ; et les modalités k de Y, différents termes (voir par exemple Lebart et Salem, 1994,

section 2.4.5 sur les tableaux lexicaux et chapitre 3 sur l'analyse des correspondances des tableaux lexicaux). Les n_{jk} représentent généralement les effectifs, soit le nombre d'occurences de chaque terme dans chaque document. Cependant, en statistique textuelle, ils peuvent aussi correspondre à la présence ou l'absence (1/0) de chaque terme dans chaque document ou encore à différents poids de chaque terme dans chaque document, comme, par exemple, la fréquence inverse de document (*idf - inverse document frequency*) (voir par exemple Salton et McGill, 1983, figure 1-12 et chapitre 3).

1.2 Lien entre deux variables catégorielles

A partir d'une table de contingence, il est possible de tester si les deux variables catégorielles sont significativement liées. Le test le plus utilisé est celui du khi2 (section 1.2.1). Cependant, il existe d'autres coefficients et tests, spécifiques à la quantification du lien entre deux variables catégorielles binaires (section 1.2.2).

1.2.1 Test d'indépendance du khi2

Les effectifs de la table de contingence sous indépendance théorique sont définis comme $n_{jk}^{\text{th}} = \frac{n_{j\bullet}n_{\bullet k}}{n_{\bullet\bullet}}$. Ainsi, l'écart des effectifs observés à l'indépendance est mesuré par la variable de décision du khi-carré :

$$\text{khi2} = \sum_{j=1}^{m1}\sum_{k=1}^{m2}\frac{(n_{jk} - n_{jk}^{\text{th}})^2}{n_{jk}^{\text{th}}} \tag{1.1}$$

Pour en tester la significativité (hypothèse H_0 : X et Y sont indépendantes) la variable de décision est comparée à la valeur critique $\chi^2_{1-\alpha}[(m_1-1)(m_2-1)]$, c'est-à-dire au $(1-\alpha)$ème quantile de la loi du χ^2 à $(m_1-1)(m_2-1)$ degrés de liberté.

1.2.1.1 Quotient d'indépendance

Alors que le khi2 mesure le lien entre les variables X et Y, le quotient d'indépendance, aussi connu sous le nom de quotient de localisation (*location quotient*) en géographie et en économie (voir par exemple Hildebrand et Mace, 1950), permet de mesurer le lien entre deux modalités j et k. Il se calcule comme :

$$q_{jk} = \frac{n_{jk}}{n_{jk}^{\text{th}}} = \frac{n_{jk}n_{\bullet\bullet}}{n_{j\bullet}n_{\bullet k}} \tag{1.2}$$

Les deux modalités sont en attraction mutuelle si $q_{jk} > 1$, en répulsion mutuelle si $q_{jk} < 1$ et en neutralité mutuelle si $q_{jk} \cong 1$.

1.2.2 Cas des variables binaires

Les variables binaires (ou bimodales) sont des variables pour lesquelles il n'y a que deux modalités possibles. Ceci engendre un tableau de contingence de taille 2×2. Fréquemment, on utilise une variable binaire pour représenter une modalité et son complémentaire, *i.e.* l'ensemble des *autres* modalités, et ce sera toujours le cas dans ce qui suit (table 1.2). Ainsi, chacune des modalités de la table 1.1 peut être transformée en modalité binaire (cf. table 1.3).

Pour quantifier le lien entre deux variables binaires, il est possible d'utiliser les mêmes coefficients que ceux proposés ci-dessus. Alors que le quotient d'indépendance reste identique (pour une formulation basée sur le principe de la table 1.2, voir Li, Luo et Chung, 2008, équation 6), le khi2 peut être reformulé (section 1.2.2.1). Il existe aussi d'autres indices particulièrement adaptés au calcul de l'accord entre deux partitions binaires (voir par exemple Warrens, 2008), dont deux seront présentés ici : le *coefficient phi* (section 1.2.2.2) et le *Q de Yule* (section 1.2.2.3).

		Y		
		Présence de k	Absence de k	
X	Présence de j	n_{11}	n_{10}	
	Absence de j	n_{01}	n_{00}	
				$n_{\bullet\bullet}$

TABLE 1.2 – Table de contingence pour deux variables binaires, avec $n_{\bullet\bullet} = n_{11}+n_{00}+n_{10}+n_{01}$.

1.2.2.1 Khi2 ponctuel

En appliquant la formule du khi2 (1.1) à une table de contingence 2×2, on obtient, avec les notations de la table 1.2, le khi2 ponctuel entre les paires de modalités j de X et k de Y (voir par exemple Yang et Pedersen, 1997 ; Saporta, 2006, p.152 ; Li *et al.*, 2008) :

$$\chi^2_{jk} = \frac{n_{\bullet\bullet}(n_{11}n_{00} - n_{01}n_{10})^2}{(n_{11}+n_{01})(n_{10}+n_{00})(n_{11}+n_{10})(n_{01}+n_{00})} \tag{1.3}$$

Ce dernier est significatif lorsqu'il est plus grand que $\chi^2_{1-\alpha}[1]$. Par exemple : $\chi^2_{1-0.001}[1] = 10.83$.

Pour pourvoir calculer ce khi2 ponctuel pour toutes les paires de modalités d'une table de contingence et obtenir ainsi une matrice du khi2, les termes de (1.3) sont remplacés par ceux de la table 1.3, ce qui permet finalement de trouver :

$$\chi^2_{jk} = \frac{n_{\bullet\bullet}(n_{jk} - n^{th}_{jk})^2}{n^{th}_{jk}(n_{\bullet\bullet} - n_{j\bullet} - n_{\bullet k} + n^{th}_{jk})} = \frac{n_{\bullet\bullet}(q_{jk}n^{th}_{jk} - n^{th}_{jk})^2}{n^{th}_{jk}(n_{\bullet\bullet} - n_{j\bullet} - n_{\bullet k} + n^{th}_{jk})} = \frac{n_{\bullet\bullet}(q_{jk}-1)^2\, n^{th}_{jk}}{n_{\bullet\bullet} - n_{j\bullet} - n_{\bullet k} + n^{th}_{jk}}$$

		Y		
		Présence de k	Absence de k	
X	Présence de j	$\boldsymbol{n_{jk}}$	$n_{j\bullet} - n_{jk}$	$\boldsymbol{n_{j\bullet}}$
	Absence de j	$n_{\bullet k} - n_{jk}$	$n_{\bullet\bullet} - n_{j\bullet} - n_{\bullet k} + n_{jk}$	$n_{\bullet\bullet} - n_{j\bullet}$
		$\boldsymbol{n_{\bullet k}}$	$n_{\bullet\bullet} - n_{\bullet k}$	$\boldsymbol{n_{\bullet\bullet}}$

TABLE 1.3 – Transformation d'une paire de modalités (j et k) de deux variables multimodales en variables binaires (présence/absence). Les termes écrits en gras sont ceux identiques aux termes de la table de contingence multimodale (table 1.1). Les autres termes se déduisent des termes en gras.

1.2.2.2 Coefficient phi

Le *coefficient phi* équivaut à la corrélation de Pearson appliquée à deux variables binaires (Yule, 1912). Cet indice, en rapport avec le chi carré ($\phi^2_{jk} = \frac{\chi^2_{jk}}{n_{\bullet\bullet}}$, voir (1.3) et (1.17)) se définit comme :

$$\phi_{jk} = \frac{n_{11}n_{00} - n_{10}n_{01}}{\sqrt{(n_{11}+n_{10})(n_{01}+n_{00})(n_{11}+n_{01})(n_{10}+n_{00})}} \tag{1.4}$$

$\phi_{jk} = 1$ si et seulement si chaque élément présent (respectivement absent) dans X est présent (respectivement absent) dans Y ($n_{01} = 0$ et $n_{10} = 1$). Inversement, $\phi_{jk} = -1$, indique que les éléments présents dans X ne le sont pas dans Y, et vice-versa ($n_{11} = 0$ et $n_{00} = 1$). Lorsque $\phi_{jk} = 0$, il n'y a pas de lien entre les deux variables X et Y. La significativité de ce coefficient peut être testée en le comparant à $\sqrt{\chi^2_{1-\frac{\alpha}{2}}[1]}$, qui vaut 0.059 pour $\alpha = 0.05$.

1.2.2.3 Q de Yule

Le *Q de Yule* et est défini comme (Yule, 1900) :

$$Q_{jk} = \frac{n_{11}n_{00} - n_{10}n_{01}}{n_{11}n_{00} + n_{10}n_{01}} \tag{1.5}$$

Si $Q_{jk} = 1$, tous les éléments présents dans X sont présents dans Y ou/et inversement ($n_{01} = 0$ ou/et $n_{10} = 1$). Tandis que si $Q_{jk} = -1$, soit aucun élément n'est simultanément présent dans les deux variables X et Y ($n_{11} = 0$), soit tous les éléments sont présents dans au moins une des deux variables ($n_{00} = 0$), ou les deux. $Q_{jk} = 0$ a la même interprétation que $\phi_{jk} = 0$.

1.3 Dissimilarité du χ^2 et dissimilarités euclidiennes carrées

En se basant sur une table de contingence (section 1.1), il est possible de calculer des dissimilarités entre les modalités (section 1.3.1). Ces dissimilarités ont la propriété d'être des dissimilarités euclidiennes carrées (section 1.3.2). Ce dernier point permet d'utiliser le principe de Huygens (section 1.3.3) et d'appliquer les transformations de Schoenberg à ces dissimilarités (section 1.3.4).

1.3.1 Dissimilarité du χ^2 et dualité

La dissimilarité du χ^2 entre les modalités i et j de X se calcule comme :

$$\hat{D}^{\chi}_{ij} = \sum_{k=1}^{m_2} \rho_k (q_{ik} - q_{jk})^2 \tag{1.6}$$

avec $\rho_k := \frac{n_{\bullet k}}{n_{\bullet\bullet}}$, le poids des colonnes. Par la dualité existant entre les lignes et les colonnes d'une table de contingence, il est possible de calculer la dissimilarité du χ^2 entre les modalités k et l de Y de manière analogue, soit :

$$\check{D}^{\chi}_{kl} = \sum_{j=1}^{m_1} f_j (q_{jk} - q_{jl})^2 \tag{1.7}$$

avec, cette fois, $f_j := \frac{n_{j\bullet}}{n_{\bullet\bullet}}$, le poids des lignes.

Dans la suite de ce chapitre, les équations en prise avec cette dualité seront toujours données par paire, soit celle pour les lignes de la table de contingence et sa duale pour les colonnes.

1.3.2 Dissimilarités euclidiennes carrées

Soit un ensemble d'individus $i = 1, \ldots, n$ possédant des caractéristiques $k = 1, \ldots, p$ et dont les coordonnées sont représentées par $X = (x_{ik})$. Les individus sont munis de poids f_i positifs ($f_i > 0$) et normalisés ($\sum_i f_i = 1$) [1] ; la pondération uniforme s'obtient avec $f_i = 1/n$.

On définit la matrice $D = (D_{ij})$ des *dissimilarités euclidiennes carrées* entre des individus i et j comme [2] :

$$D_{ij} := \sum_{k=1}^{p} (x_{ik} - x_{jk})^2 = \|x_i - x_j\|^2 \tag{1.8}$$

1. Dans le cas particulier de la table de contingence, les poids, f_j pour les lignes et ρ_k pour les colonnes, sont définis selon les équations de la section 1.3.1.

2. Dans cette thèse, D_{ij} désignera toujours une dissimilarité euclidienne carrée entre les objets i et j.

La dissimilarité du khi2 est aussi une distance euclidienne carrée, car les équations (1.6) et (1.7) peuvent être reformulées comme (voir par exemple Bavaud, 2004) :

$$\hat{D}_{ij} = \sum_{k=1}^{m_2} ({}^*x_{ik} - {}^*x_{jk})^2 \qquad \check{D}_{kl} = \sum_{j=1}^{m_1} ({}^*y_{jk} - {}^*y_{jl})^2 \tag{1.9}$$

où

$${}^*x_{ik} = \sqrt{\rho_k}(q_{ik} - 1) \qquad \text{et} \qquad {}^*y_{jk} = \sqrt{f_j}(q_{jk} - 1) \tag{1.10}$$

sont *les coordonnées brutes* ou *de haute dimensionnalité*. Celles-ci, directement calculées à partir de la table de contingence, s'opposent aux coordonnées factorielles (1.25) qui ont la propriété d'exprimer une proportion maximale d'inertie (1.17) dans les basses dimensions.

La matrice $B = (b_{ij})$ des produits scalaires entre i et j, pour les dissimilarités euclidiennes carrées (1.8) et relativement à la pondération f, se définit comme :

$$b_{ij} = \sum_{k=1}^{p} (x_{ik} - \bar{x}_k^f)(x_{jk} - \bar{x}_k^f) \qquad \text{avec} \qquad \bar{x}_k^f = \sum_{i=1}^{n} f_i x_{ik} \tag{1.11}$$

Dans le cas particulier des dissimilarités du khi2, les produits scalaires entre les lignes $\hat{B} = (b_{ij})$ et entre les colonnes $\check{B} = (b_{kl})$ peuvent, par conséquent, s'écrire comme :

$$\hat{b}_{ij} = \sum_{k=1}^{m_2} \rho_k (q_{ik} - 1)(q_{jk} - 1) \qquad \check{b}_{kl} = \sum_{j=1}^{m_1} f_j (q_{jk} - 1)(q_{jl} - 1)$$

Aussi, il est possible de reformuler toute dissimilarité euclidienne carrée (1.8) en se basant sur les produits scalaires correspondant (1.11) [3] :

$$D_{ij} = b_{ii} + b_{jj} - 2b_{ij} \tag{1.12}$$

Cette dernière relation peut aussi s'obtenir à partir du théorème du cosinus. Ce denier peut se reformuler, avec des distances et des produits scalaires, de la manière suivante (Young et Householder, 1938) :

$$D_{ij} = D_{if} + D_{jf} - 2b_{ij} \tag{1.13}$$

où $D_{if} = D_{i\bar{x}_f} = \sum_k (x_{ik} - \bar{x}_k^f)^2$ est la dissimilarité euclidienne carrée entre un point i et la moyenne pondérée des points $\bar{x}_k^f$ (1.11). Comme $D_{if} = \sum_k (x_{ik} - \bar{x}_k^f)^2 = \sum_k (x_{ik} - \bar{x}_k^f)(x_{ik} - \bar{x}_k^f) = b_{ii}$, alors (1.13) est équivalente à (1.12).

L'équation (1.13) permet aussi de déterminer les produits scalaires à partir des distances (Young et Householder, 1938) :

$$b_{ij} = \frac{1}{2}(D_{if} + D_{jf} - D_{ij}) \tag{1.14}$$

Young et Householder (1938) montrent, en partant de l'équation (1.14), que la matrice D représente des dissimilarités euclidiennes carrées si et seulement si la matrice B est semi-définie positive.

3. **Preuve** pour des dissimilarités euclidiennes carrées entre i et j :

$$\begin{aligned} D_{ij} &= \sum_{k=1}^{p} (x_{ik} - x_{jk})^2 = \sum_{k=1}^{p} ((x_{ik} - \bar{x}_k^f) - (x_{jk} - \bar{x}_k^f))^2 \\ &= \sum_{k=1}^{p} (x_{ik} - \bar{x}_k^f)^2 - 2\sum_{k=1}^{p} (x_{ik} - \bar{x}_k^f)(x_{jk} - \bar{x}_k^f) + \sum_{k=1}^{p} (x_{jk} - \bar{x}_k^f)^2 \\ &= b_{ii} - 2b_{ij} + b_{jj} \end{aligned}$$

1.3.3 Principe de Huygens

Étant donné que les distances du khi2 sont euclidiennes carrées (1.9), le principe de Huygens s'applique.

Le principe (fort) de Huygens, également connu sous le nom de « théorème de Steiner » en mécanique des solides, s'écrit, pour toute matrice de dissimilarités euclidiennes carrées (1.8) et un formalisme pondéré, comme :

$$\sum_j^n f_j D_{ij} = \Delta_f + D_{if} \tag{1.15}$$

En d'autres termes, la dispersion du nuage de points *par rapport* à un point i équivaut à la dispersion du nuage de point (par rapport au centre) Δ_f, additionné de la dissimilarité entre le point i et le centre de gravité de l'ensemble des points.

À partir de ce premier principe, découle le principe (faible) de Huygens qui définit l'inertie (ou la variance, ou la dispersion) de l'ensemble des individus comme :

$$\Delta := \Delta_f = \frac{1}{2}\sum_{ij} f_i f_j D_{ij} = \sum_i f_i D_{if} \tag{1.16}$$

Ainsi, la dispersion du nuage de points peut s'exprimer de manière équivalente comme la dissimilarité moyenne entre toutes les paires de points ou comme la dissimilarité moyenne entre chaque point et le centre de gravité de l'ensemble des points.

Dans le cas particulier du khi2, l'inertie est égale au khi2 divisé par l'effectif total de la table de contingence $n_{\bullet\bullet}$, nommé ϕ^2 :

$$\begin{aligned}\Delta = \frac{1}{2}\sum_{ij} f_i f_j \hat{D}_{ij} = \sum_i f_i \hat{D}_{if} = \frac{1}{2}\sum_{kl} \rho_k \rho_l \check{D}_{kl} = \sum_k \rho_k \check{D}_{k\rho} \\ = \frac{1}{n_{\bullet\bullet}} \sum_{jk} \frac{(n_{jk} - n_{jk}^{\text{th}})^2}{n_{jk}^{\text{th}}} = \frac{\text{khi2}}{n_{\bullet\bullet}} = \phi^2\end{aligned} \tag{1.17}$$

où $\hat{D}_{if}$ est la dissimilarité du khi2 entre la modalité i et la moyenne des modalités de X, soit ${}^*\bar{x}_k^f = \sum_i f_i {}^*x_{ik}$; et $\check{D}_{k\rho}$, la dissimilarité du khi2 entre la modalité k et la moyenne des modalités de Y, soit ${}^*\bar{x}_i^\rho = \sum_k \rho_k {}^*x_{ik}$.

Soit un groupe g et une matrice d'appartenance $Z = (z_{ig})$ qui détermine la probabilité [4] que l'individu i appartienne au groupe g, telle que $\sum_g z_{ig} = 1$. Alors, le poids du groupe vaut $\rho_g = \sum_i f_i z_{ig}$, tel que $\sum_g \rho_g = 1$; et la distribution des individus i du groupe, $f_i^g = f_i z_{ig}/\rho_g$, telle que $\sum_i f_i^g = 1$. Avec $\bar{x}_k^g = \sum_i f_i^g x_{ik}$ pour la moyenne du groupe et $D_{ig} = D_{i\bar{x}^g}$, le principe fort de Huygens (1.15) devient :

$$\sum_j^n f_j^g D_{ij} = D_{ig} + \Delta_g \tag{1.18}$$

et le principe faible de Huygens (1.16), pour l'inertie du groupe g :

$$\Delta_g = \frac{1}{2}\sum_{ij} f_i^g f_j^g D_{ij} = \sum_i f_i^g D_{ig} \tag{1.19}$$

Ce qui précède, et en particulier (1.15) et (1.16), permet de trouver que la dissimilarité euclidienne carrée D_{fg} entre les moyennes $\bar{x}_k^f = \sum_i f_i x_{ik}$ et $\bar{x}_k^g = \sum_i g_i x_{ik}$ de deux groupes ou deux

4. Dans le cas particulier d'un partitionnement dur des données, la matrice d'appartenance détermine la présence $z_{ig} = 1$ ou l'absence $z_{ig} = 0$ d'un individu dans un groupe.

distributions f et g peut se calculer uniquement grâce aux distributions et aux dissimilarités entre les individus D_{ij}, soit (Bavaud, 2011) :

$$D_{fg} = D_{\bar{x}^f \bar{x}^g} = -\frac{1}{2}\sum_{ij}(f_i - g_i)(f_j - g_j)D_{ij} \tag{1.20}$$

En remplaçant les termes de (1.14) par ceux des principes de Huygens (1.15) et (1.16), on peut facilement montrer que, avec une matrice de dissimilarités euclidiennes carrées, la matrice des produits scalaires, relativement à f, (1.11) peut aussi s'obtenir matriciellement par [5] :

$$B = -\frac{1}{2}H^f D (H^f)' \qquad \text{avec} \qquad H^f = (h^f_{ij}) = I - \mathbf{1}f' \tag{1.21}$$

1.3.4 Transformations de Schoenberg

Les *transformations de Schoenberg* (Schoenberg, 1938) transforment les dissimilarités euclidiennes carrées originales, D, en d'autres dissimilarités euclidiennes carrées, $\tilde{D} = \varphi(D)$ (Bavaud, 2011, et références y incluses). Tout comme les méthodes à noyaux, les transformations de Schoenberg s'appuient sur un plongement de haute dimensionnalité des objets de départ. Une liste non exhaustive des diverses transformations de Schoenberg possibles se trouve dans l'article de Bavaud (2011). Parmi ces possibilités, une seule est envisagée dans la suite de ce travail, à savoir, la *transformation de puissance* (Schoenberg, 1937), telle que :

$$\varphi(D) = \tilde{D} = D^q \tag{1.22}$$

où $0 < q \leq 1$.

Cette transformation permet de rappeler que toute distance euclidienne est aussi une dissimilarité euclidienne carrée [6], mais que l'inverse n'est pas toujours vrai.

1.4 Analyse factorielle des correspondances

Soit $c_{ik} = \sqrt{f_i}\,{}^*x^c_{ik} = \sqrt{f_i}({}^*x_{ik} - \bar{x}^f_k)$, avec ${}^*x_{ik}$, les coordonnées de haute dimensionnalité, telles que définies dans l'équation (1.10). Il existe alors deux méthodes afin de pratiquer l'analyse factorielle des correspondances (AFC) permettant de visualiser simultanément les modalités de X et de Y.

La première se base sur la décomposition spectrale de la matrice des variances-covariances, soit $\Sigma = C'C$. Cette technique est largement décrite dans la littérature (voir par exemple Greenacre, 1984, en particulier le chapitre 4, pp. 83-125 ; Lebart *et al.*, 1995, section 1.3, pp. 67-107 ; Le Roux et Rouanet, 2004, chapitre 2, pp. 23-74 ; Saporta, 2006, chapitre 9, pp. 201-217). En outre, dans le logiciel R (R Core Team, 2013), il existe plusieurs packages qui produisent des AFC, tels que « ca » (Nenadic et Greenacre, 2007) ou « FactoMineR » (Husson, Josse, Le et Mazet, 2013).

Une seconde méthode consiste à appliquer un MDS et se base alors sur la décomposition spectrale de la matrice des produits scalaires pondérés $K = CC'$. Elle sera exposée dans la section suivante.

Les deux matrices Σ et K étant duales (Bavaud et Cocco, accepté pour publication), ces deux méthodes produisent des résultats complètement équivalents. Cependant, la seconde a l'intérêt d'être plus générale, car applicable à toute dissimilarité euclidienne carrée, et d'introduire des quantités utiles dans la suite de ce travail.

5. Dans le cas des dissimilarités du khi2, la matrice des produits scalaires entre les modalités de X se calcule de manière analogue comme $\hat{B} = -\frac{1}{2}H^f \hat{D}(H^f)'$. Par dualité, la matrice des produits scalaires entre les modalités de Y est définie comme $\check{B} = -\frac{1}{2}H^\rho \check{D}(H^\rho)'$, avec $H^\rho = I - \mathbf{1}\rho'$, la matrice de centration.

6. Soit d_{ij}, la distance euclidienne entre deux individus i et j, alors $d_{ij} = \sqrt{(d_{ij})^2} = \sqrt{D_{ij}} = D^{0,5}_{ij}$. Comme $\varphi(D_{ij}) = \tilde{D}_{ij} = D^{0,5}_{ij}$ est aussi une dissimilarité euclidienne carrée, alors d_{ij} est une dissimilarité euclidienne carrée.

1.4.1 MDS

Le but du MDS est de reconstituer les coordonnées d'un nuage de points dont on connaît les dissimilarités. Le MDS *classique* (ou *métrique*), contrairement au MDS *ordinal* (ou *non-métrique*), s'applique exclusivement à des dissimilarités euclidiennes carrées et va créer des coordonnées qui reproduisent exactement ces dissimilarités. Pour pouvoir appliquer le MDS classique, la matrice des produits scalaires B, calculée par exemple par (1.21), doit donc être semi-définie positive (cf. section 1.3.2), ce qui est bien le cas des dissimilarités du khi2 (cf. 1.11), étant donné qu'elles sont euclidiennes carrées (1.9).

Le MDS, dans sa version ordinaire, (voir par exemple Mardia, Kent et Bibby, 1979) se base sur la décomposition spectrale de $B = U\Lambda U'$ dont découlent les nouvelles coordonnées de l'objet j sur le facteur α, soit $x_{j\alpha} = \sqrt{\lambda_\alpha} u_{j\alpha}$.

Par extension, le MDS pondéré (voir par exemple Cuadras et Fortiana, 1996; Bavaud, 2010) est effectué grâce à la matrice $K = (k_{ij})$ des *produits scalaires pondérés* définis comme :

$$k_{ij} = \sqrt{f_i f_j} b_{ij} \tag{1.23}$$

La décomposition spectrale de $K = U\Lambda U'$, qui est semi-définie positive ssi B l'est aussi, permet alors de calculer les nouvelles coordonnées comme $x_{j\alpha} = \frac{\sqrt{\lambda_\alpha}}{\sqrt{f_j}} u_{j\alpha}$.

Dans le cas particulier des dissimilarités du khi2, les produits scalaires entre les modalités de X, $\hat{K} = (\hat{k}_{ij})$, et ceux entre les modalités de Y, $\check{K} = (\check{k}_{kl})$, sont définis, de manière analogue à (1.23), comme :

$$\hat{k}_{ij} = \sqrt{f_i f_j}\, \hat{b}_{ij} \qquad\qquad \check{k}_{kl} = \sqrt{\rho_k \rho_l}\, \check{b}_{kl} \tag{1.24}$$

La décomposition spectrale de $\hat{K}$ (respectivement de $\check{K}$) engendre les vecteurs propres $u_{j\alpha}$ (respectivement $v_{k\alpha}$) et les valeurs propres λ_α (identiques pour les deux matrices de produits scalaires), où $\alpha = 1, \ldots, r$ et $r \leq \min(m_1, m_2) - 1$. Avec ces derniers, les coordonnées factorielles des modalités de X et de Y sont reliées par des formules de transition et calculées comme :

$$\begin{cases} y_{k\alpha} = \dfrac{\sqrt{\lambda_\alpha}}{\sqrt{\rho_k}} v_{k\alpha} = \dfrac{1}{\sqrt{\lambda_\alpha}} \displaystyle\sum_{j=1}^{m_1} f_j q_{jk} x_{j\alpha} & \text{(1.25a)} \\ x_{j\alpha} = \dfrac{\sqrt{\lambda_\alpha}}{\sqrt{f_j}} u_{j\alpha} = \dfrac{1}{\sqrt{\lambda_\alpha}} \displaystyle\sum_{k=1}^{m_2} \rho_k q_{jk} y_{k\alpha} & \text{(1.25b)} \end{cases}$$

Avec les coordonnées factorielles, il est alors possible de réécrire les dissimilarités du χ^2, (1.6) et (1.7), comme les distances euclidiennes carrées entre ces nouvelles coordonnées :

$$D_{ij} = \sum_{\alpha=1}^{r} (x_{i\alpha} - x_{j\alpha})^2 \qquad D_{kl} = \sum_{\alpha=1}^{r} (y_{k\alpha} - y_{l\alpha})^2$$

CHAPITRE 2

Classification supervisée et non supervisée

Il existe de nombreuses méthodes de classification, supervisée ou non, et de nombreuses distinctions entre ces méthodes. Plutôt que d'en donner une vue exhaustive, ce chapitre vise à expliciter quelques méthodes de classification (section 2.1 et 2.2) ainsi que des méthodes d'évaluation des résultats obtenus (section 2.3), toutes utilisées dans les applications des parties II et III.

Les méthodes de classification peuvent se diviser en deux groupes principaux : les méthodes dites supervisées (section 2.2) et celles dites non supervisées (section 2.1). Ces deux types de méthodes se distinguent par le fait que les groupes (ou classes) sont connus *a priori* dans le premier cas, alors qu'ils ne le sont pas pour le second. Ainsi, l'avantage de la méthode supervisée est que les groupes de départ ont, par construction, un sens clair pour l'utilisateur, ce qui n'est pas garanti avec les méthodes non supervisées. En contre-partie, l'inconvénient principal de la méthode supervisée est la nécessité de disposer de données dont on connaît le groupe. Cela implique, pour le traitement informatique des textes, de créer un corpus annoté conséquent, tâche exigente en ressources. L'avantage de la seconde méthode est donc de pouvoir être appliquée directement aux corpus avec un minimum de traitement.

Généralement, ces méthodes considèrent un jeu de données $X = (x_{ik})$ multivarié, donnant les caractéristiques $k = 1, \ldots, p$ des individus $i = 1, \ldots, n$. La classification supervisée contient une colonne supplémentaire spécifiant le groupe $g = 1, \ldots, m$ auquel appartient l'individu i.

Dans ce chapitre, toutes les méthodes de classification seront présentées en utilisant une matrice de **dissimilarités euclidiennes carrées** $D = (D_{ij})$ entre les individus. Cette dernière peut, typiquement, être calculée par (1.8) si les données de départ sont sous la forme d'un jeu de données numériques X, ou par (1.6) ou (1.7) si elles sont sous la forme d'une table de contingence. Les données sous forme de table de contingence seront les plus courantes dans l'ensemble de ce travail. Ce choix, consistant à travailler sur des dissimilarités euclidiennes carrées, va permettre de combiner les méthodes de classification abordées ici avec les transformations de Schoenberg présentées dans la section 1.3.4.

2.1 Classification non supervisée

Comme déjà mentionné, pour les méthodes de classification **non supervisée** (*clustering*), les groupes ne sont pas connus *a priori*.

Il existe de nombreuses méthodes de classification non supervisée (voir par exemple Jain, Murty et Flynn, 1999). Pour résumer celles qui sont utilisées dans cette thèse, on peut d'abord

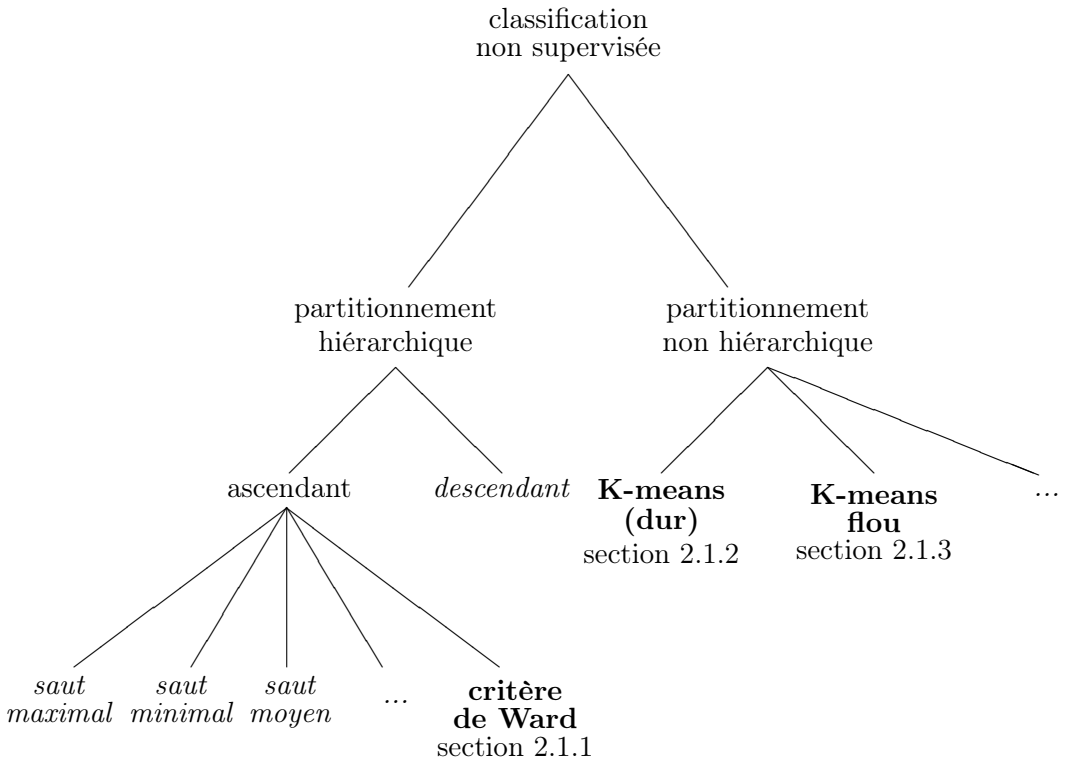


FIGURE 2.1 – Principales méthodes de classification non supervisée, avec, en gras, les méthodes traitées dans ce travail, et, en italique, celles qui ne le sont pas.

opposer les algorithmes de *partitionnement hiérarchique* à ceux de *partitionnement non hiérarchique* (figure 2.1).

Parmi les méthodes de partitionnement hiérarchique, on peut distinguer deux grandes familles : la classification ascendante hiérarchique et la classification descendante hiérarchique. La première est agglomérative, c'est-à-dire que l'on commence avec les n individus qui sont successivement regroupés à chaque étape jusqu'à n'obtenir finalement plus qu'un seul groupe. À l'inverse, la seconde est divisive : l'ensemble des individus est successivement fractionné à chaque étape, pour aboutir finalement à n groupes formés chacun d'un seul individu. Seule la première de ces familles sera traitée ici, et plus particulièrement la classification ascendante hiérarchique avec le critère de Ward (section 2.1.1).

Concernant le partionnement non hiérarchique, seules deux méthodes seront abordées ici : la méthode K-means (section 2.1.2) et la méthode K-means flou (section 2.1.3). La différence principale entre ces deux méthodes est que la première effectue un partitionnement dur des individus i en groupes g, alors que la seconde effectue un partitionnement flou. Pour rappel (cf. section 1.3.3), dans le cas d'une matrice d'appartenance $Z = (z_{ig})$ dure, z_{ig} vaut 1 ou 0 selon que l'individu i appartient ou non au groupe g ; alternativement, pour une matrice d'appartenance floue, z_{ig} est la probabilité que l'individu i appartienne au groupe g.

Il existe deux distinctions importantes entre la classification ascendante hiérarchique et la méthode K-means (flou ou non). Premièrement, l'algorithme K-means implique de choisir un nombre de groupes initial, contrairement à la classification ascendante hiérarchique. Deuxièmement, la classification ascendante hiérarchique s'appuie avant tout sur une matrice de dissimilarités entre paires d'objets, alors que pour la méthode K-means, c'est une matrice de dissimilarités objet-groupe. Dans le premier cas, les données les plus similaires seront regroupées dans les mêmes groupes et, par suite, les plus dissimilaires seront classées dans des groupes différents (section 2.1.1). Pour les dissimilarités objet-groupe, un nombre de centroïdes (ou

centres de gravité), correspondant au nombre de groupes choisis initialement, sera sélectionné. Ensuite, itérativement, les données seront attribuées au groupe le plus proche et les centroïdes re-positionnés (sections 2.1.2 et 2.1.3).

Au final, le point commun de toutes ces méthodes est qu'elles ont pour but de minimiser l'inertie intra-groupe (ou intra-classe), et donc de maximiser l'inertie inter-groupe (ou inter-classe), créant ainsi des groupes homogènes. L'inertie (1.16) s'écrit aussi :

$$\Delta = \Delta_W + \Delta_B \tag{2.1}$$

où Δ_W, pour des groupes $g = 1, \ldots, m$, est l'inertie intra-groupe :

$$\Delta_W = \sum_{g=1}^{m} \rho_g \Delta_g \tag{2.2}$$

avec Δ_g, l'inertie du groupe g, définie en (1.19) ; et Δ_B est l'inertie inter-groupe, soit :

$$\Delta_B = \sum_{g=1}^{m} \rho_g D_{gf} \tag{2.3}$$

Dans cette équation, D_{gf} est la dissimilarité euclidienne carrée entre le centroïde du groupe g, $\bar{x}_k^g = \sum_i f_i^g x_{ik}$, et la moyenne pondérée de l'ensemble des individus, $\bar{x}_k^f = \sum_i f_i x_{ik}$. De plus, pour rappel (cf. section 1.3.3), $\rho_g = \sum_i f_i z_{ig}$ est le poids du groupe g ; et $f_i^g = f_i z_{ig}/\rho_g$, la distribution des individus i dans le groupe g.

2.1.1 Classification ascendante hiérarchique, critère de Ward

Soit une matrice de dissimilarités, euclidiennes ou non, de composantes d_{ij}. La classification ascendante hiérarchique regroupe les individus (ou objets) les plus similaires, qui vont former de nouveaux individus agrégés, dont les plus similaires sont à nouveau regroupés pour créer, au final, un *dendrogramme*. Le point crucial consiste à définir la dissimilarité entre le nouvel individu formé par le regroupement de deux individus a et b, et un autre individu i, noté comme $d((a, b), i)$. Plusieurs critères d'agrégation, bien connus, ont été proposés pour calculer cette nouvelle dissimilarité, tels que *le saut maximal*, *le saut minimal*, *la moyenne des distances*, etc. (voir par exemple Lebart *et al.*, 1995, section 2.2 ; Jain *et al.*, 1999, section 5.1 ; Le Roux et Rouanet, 2004, section 3.6 ; Saporta, 2006, section 11.3). Toutes ces méthodes constituent des cas particuliers de la formule de Lance et Williams généralisée (voir par exemple Saporta, 2006, section 11.3.2.2). Parmi ces dernières, seul le critère de Ward, utilisé dans le chapitre 8, est présenté ici.

Étant donné une matrice de dissimilarités euclidiennes carrées $D = (D_{ij})$, le critère de Ward consiste à minimiser l'inertie intra-groupe et donc à maximiser l'inertie inter-groupe à chaque étape. À la première étape, tous les individus représentent un groupe, et donc l'inertie intra-groupe est nulle ($\Delta_W^0 = 0$) et l'inertie inter-groupe est égale à l'inertie totale ($\Delta_B^0 = \Delta$). Après la première agrégation, l'inertie intra-groupe Δ_W^1 augmente, et l'inter-groupe Δ_B^1 diminue, et ce jusqu'à la dernière étape, r, lorsque tous les individus ne forment plus qu'un groupe. L'inertie intra-groupe est alors maximale ($\Delta_W^r = \Delta$) ; et l'inter-groupe, minimale ($\Delta_B^r = 0$).

Plus précisément, si à la première étape, les individus a et b sont regroupés, alors la différence d'inertie intra-groupe vaudra $\Delta_W^1 - \Delta_W^0$ qui, en vertu de (2.1), sera équivalente à $\Delta_B^0 - \Delta_B^1$. Cette différence s'écrit, avec (2.3), comme :

$$\begin{aligned} \Delta_B^0 - \Delta_B^1 &= \rho_1 D_{1f} + \rho_2 D_{2f} + \cdots + \rho_a D_{af} + \rho_b D_{bf} \\ &\quad - \rho_1 D_{1f} - \rho_2 D_{2f} - \cdots - (\rho_a + \rho_b) D_{(ab)f} \\ &= \rho_a D_{af} + \rho_b D_{bf} - (\rho_a + \rho_b) D_{(ab)f} \end{aligned} \tag{2.4}$$

Par le principe fort de Huygens (1.15), avec $f = (f_1, f_2) = \left(\frac{\rho_a}{\rho_a+\rho_b}, \frac{\rho_b}{\rho_a+\rho_b}\right)$, on obtient le théorème de la médiane :

$$D_{(ab)f} = \frac{1}{\rho_a + \rho_b}(\rho_a D_{af} + \rho_b D_{bf} - \frac{\rho_a \rho_b}{\rho_a + \rho_b} D_{ab})$$

En remplaçant $D_{(ab)f}$ dans (2.4), la perte d'inertie inter-groupe, qui s'exprime finalement comme :

$$\delta(a, b) = \frac{\rho_a \rho_b}{\rho_a + \rho_b} D_{ab} \tag{2.5}$$

constitue le critère d'agrégation de la méthode de Ward.

Pratiquement, à la première étape, la matrice des dissimilarités, D, est transformée en une nouvelle matrice de pertes d'inertie inter-groupe, $\mathcal{D}^0 = (\delta(i, j))$, qui donne, pour chaque paire d'individus (i, j), la valeur du critère d'agrégation (2.5). Comme avec les autres critères, la paire d'individus dont la valeur est la plus petite (a et b par exemple) sont regroupés pour former un nouvel individu. Puis, pour recalculer $\mathcal{D}^1$, on peut soit recalculer $\delta((a, b), i)$ par (2.5) en obtenant $D_{(ab)i}$ par (1.20), soit utiliser la formule de Lance et Williams avec les paramètres adéquats (voir par exemple Le Roux et Rouanet, 2004, équation 3.14 ; Saporta, 2006, p. 259 ; Murtagh et Legendre, 2011).

Il existe de légères variantes de cette méthode (Murtagh et Legendre, 2011). Il faut noter qu'avec la fonction « hclust » du logiciel R et l'option « method = "ward" », qui a été utilisée dans ce travail, les dissimilarités transmises à la fonction *doivent* être euclidiennes carrées (Murtagh et Legendre, 2011).

2.1.2 K-means sur les dissimilarités

La méthode K-means (ou méthode des centres mobiles), déjà brièvement présentée avec les dissimilarités objet-groupe au début de cette section, est relativement intuitive et sa paternité n'est pas clairement établie. Lebart *et al.* (1995) proposent cependant quelques pistes dans l'introduction de leur section 2.1. On peut, entre autres, noter que l'algorithme K-means présenté par MacQueen (1967) diffère de la procédure ci-dessous, car la position des centroïdes (ou centres de gravité) est recalculée après chaque nouvelle attribution d'un individu, et non après l'attribution de tous les individus à tous les centroïdes.

Généralement, l'algorithme K-means est proposé en travaillant directement sur la table des coordonnées $X = (x_{ik})$ et se compose de quatre étapes (voir par exemple Lebart *et al.*, 1995, section 2.1 ; Manning et Schütze, 1999, section 14.2.1 ; Saporta, 2006, section 11.2.1).

La première opération, étape *0)*, consiste à choisir un nombre de groupes m. Ensuite, les m centres provisoires sont positionnés aléatoirement, bien que souvent sélectionnés parmi les individus. Puis, l'algorithme se poursuit itérativement :

1) les distances entre les individus et les centroïdes (ou centres provisoires lors du premier tour), D_{ig}, sont calculées,

2) chaque individu est attribué au centroïde le plus proche,

3) les positions des centroïdes (moyennes pondérées $\bar{x}_k^g = \sum_i f_i^g x_{ik}$ ou non des individus attribués à un groupe) sont recalculées.

L'itération se poursuit jusqu'à convergence de la solution. Pour une justification de l'algorithme montrant que l'inertie intra-groupe diminue à chaque itération, voir, par exemple, la section 2.1.2 de Lebart *et al.* (1995).

Avec le formalisme choisi ici, qui se base sur $D = (D_{ij})$, une matrice de dissimilarités qui *doivent* être euclidiennes carrées , les étapes sont un peu différentes. Lors de l'initialisation, soit lors de l'étape *0)*, on commence par décider d'un nombre de groupes m, comme dans la version « ordinaire » de l'algorithme. Puis, une matrice d'appartenance dure Z de taille $n \times m$ est créée,

où chaque individu est attribué aléatoirement à un des m groupes (d'autres variantes existent). À ce stade, on décide d'opérer deux vérifications supplémentaires pour effectivement avoir m groupes à la fin des itérations. Premièrement, on contrôle qu'aucun groupe ne soit vide et on réinitialise la procédure avec une nouvelle matrice Z le cas échéant. Deuxièmement, on vérifie qu'il n'y ait pas une configuration des positions des centroïdes particulière qui engendrerait la disparition d'un ou plusieurs groupes au premier tour d'itération. Pour ce faire, une première itération est exécutée et si l'un des groupes disparaît, la matrice Z est recréée. Pendant cette étape d'initialisation, on calcule aussi la matrice des dissimilarités euclidiennes carrées $D = (D_{ij})$ entre tous les individus.

Les dissimilarités euclidiennes carrées entre les individus et le centroïde d'un groupe de l'étape *1)* sont déduites indirectement des dissimilarités D_{ij} et de l'inertie d'un groupe (1.19) grâce au principe fort de Huygens (1.18) :

$$D_{ig} = \sum_j f_j^g D_{ij} - \Delta_g \tag{2.6}$$

Ces valeurs sont calculées pour chaque groupe, produisant ainsi une matrice de taille $n \times m$.

Puis, l'étape *2)* consiste à actualiser la matrice d'appartenance comme :

$$z_{ig} = \begin{cases} 1 & \text{si } \; g = \underset{h}{\operatorname{argmin}}\, D_{ih} \\ 0 & \text{sinon} \end{cases} \tag{2.7}$$

Quant à l'étape *3)*, elle n'est plus nécessaire dans ce formalisme, car la position des centroïdes est indirectement déduite de (2.7) dans (2.6).

Pour terminer, on choisit d'arrêter l'algorithme soit quand la matrice Z n'est plus modifiée, soit lorsqu'un certain nombre d'itérations $N_{\max}$ est atteint. Il faut noter que la solution finale dépend de la position initiale des centres à l'étape *0)*.

Finalement, il est possible de combiner simplement la méthode K-means avec les transfomations de Schoenberg (cf. section 1.3.4) en remplaçant, lors de l'initalisation, D par $\tilde{D} = \varphi(D)$. Comme déjà mentionné, la seule transformation utilisée dans ce travail est celle de la puissance (1.22).

2.1.3 K-means flou sur les dissimilarités

Les étapes de l'algorithme K-means flou sont presque identiques à celles de l'algorithme K-means présenté ci-dessus. Une première différence est qu'à l'étape *0)*, au lieu de créer une matrice d'appartenance dure, on décide de créer un matrice d'appartenance Z floue. Pour ce faire, une matrice de taille $n \times m$ est créée avec des valeurs aléatoires extraites d'une loi uniforme et comprises entre 0 et 1. Puis, les lignes sont normalisées pour que $\sum_g z_{ig} = 1$. Pour le reste, cette étape est identique à celle de la méthode K-means, *i.e.* il faut aussi choisir un nombre de groupes m et calculer les dissimilarités euclidiennes carrées D_{ij}.

L'étape *1)* est strictement identique à l'étape *1)* décrite en 2.1.2.

Naturellement, à l'étape *2)*, l'actualisation de la matrice d'appartenance est différente, soit (voir par exemple Rose, Gurewitz et Fox, 1990; Bavaud, 2009) :

$$z_{ig} = \frac{\rho_g \exp(-\beta D_{ig})}{\sum_{h=1}^{m} \rho_h \exp(-\beta D_{ih})} \tag{2.8}$$

où D_{ig} est défini par (2.6), ρ_g est le poids relatif du groupe g (cf. section 1.3.3) et β s'interprète comme une « température inverse » ou l'inverse d'une variance, paramétrée comme $\beta := 1/(t_{\text{rel}} \times \Delta)$ (Bavaud, 2010). Pour cette dernière, Δ représente l'inertie, telle que définie par (1.16) à partir

des dissimilarités D_{ij} ; et t_{rel}, la température relative qui doit être fixée par l'utilisateur en amont, tout comme le nombre de groupes de départ m. Il se trouve que les valeurs « intéressantes » de t_{rel} se situent dans un intervalle compris entre 0.02 et 0.3 environ (cf. section 4.3.2), des valeurs plus basses de t_{rel} générant des instabilités numériques. À l'inverse, des valeurs plus élevées ne produisent qu'un seul groupe final, suite à l'agrégation effectuée lors de l'étape *4)* décrite ci-dessous.

L'équation (2.8) découle de la minimisation de l'inertie intra-groupe Δ_W (2.2), régularisée par un terme d'entropie (Rose *et al.*, 1990) ou d'information mutuelle (Bavaud, 2009). Elle peut aussi être dérivée de l'algorithme d'espérance-maximisation (EM) associé au modèle gaussien multivarié isotrope (Celeux et Govaert, 1992; McLachlan et Krishnan, 1997).

À nouveau, on choisit d'itérer les étapes *1)* et *2)* jusqu'à la convergence de la solution ou jusqu'à ce qu'un nombre maximum d'itérations, $N_{\max}$, soit atteint.

Ensuite, une étape supplémentaire d'agrégation entre les groupes dont les profils sont assez similaires est effectuée, soit l'étape *4)*, réduisant le nombre de groupes de m à M. En effet, la valeur de β contrôle l'étendue moyenne de chaque groupe et donc le nombre de groupes final M. Ainsi, avec $m \leq n$ choisi assez grand, le nombre de groupes est indirectement, mais entièrement, déterminé par le choix de t_{rel}. Plus précisément, plus β sera élevé, plus M le sera aussi.

Concrètement, à l'étape *4)*, l'agrégation entre deux groupes similaires s'effectue en additionnant les appartenances des individus, *i.e.* $z_i^{[g\cup h]} = z_i^g + z_i^h$. Pour déterminer si deux groupes sont assez similaires, on peut utiliser, comme critère de fusion des groupes : $\theta_{gh}/\sqrt{\theta_{gg}\theta_{hh}} \geq 1-10^{-5}$, où $\theta_{gh} = \sum_{i=1}^n f_i z_i^g z_i^h$ mesure le chevauchement entre g et h (Bavaud, 2010). Cette approche produit généralement de bons résultats, sans toutefois empêcher l'apparition d'instabilités numériques pour quelques valeurs de t_{rel} (voir section 4.3.2).

Finalement, une dernière étape *5)* consiste à attribuer chaque individu au groupe le plus probable, soit $\underset{g}{\operatorname{argmin}}\, z_{ig}$.

Cette méthode floue, un peu plus complexe à implémenter que la méthode K-means (dur), est plus robuste par rapport au choix de la partition initiale. Elle a de plus l'avantage de ramener le problème épineux de la détermination du nombre de groupes à celui de la dispersion β de ces mêmes groupes, un paramètre plus facile à interpréter et indépendant de la taille des données.

2.2 Classification supervisée

Pour la classification **supervisée** (*classification* en anglais), on dispose d'un ensemble de données (échantillon d'objets ou d'individus) dont on connaît les profils ou caractéristiques, ainsi que le groupe (ou classe ou étiquette) de chaque individu. Dans un premier temps (phase d'apprentissage), l'algorithme « apprend » des règles sur l'ensemble des données. Ensuite (phase de test), on soumet de nouvelles données à l'algorithme, sans lui spécifier les groupes auxquels ces données appartiennent, et il attribue un groupe à chaque donnée selon les règles élaborées durant la phase d'apprentissage. Puisque l'on connaît les groupes auxquels les nouvelles données appartiennent, la phase de test permet de vérifier si l'algorithme fonctionne correctement ou, en d'autres termes, sa capacité à produire des règles généralisables.

Parmi les nombreuses méthodes de classification supervisée existantes, telles que le « classifieur Bayésien naïf », les « séparateurs à vastes marges » (*Support Vector Machine*), les arbres de décisions, les réseaux de neurones, la méthode des k plus proches voisins (kNN), etc. (voir par exemple Yang, 1999; Sebastiani, 2002, et références y incluses), seule l'analyse discriminante (Fisher, 1936) sera présentée ici.

2.2.1 Analyse discriminante sur les dissimilarités

Soit, comme ensemble d'apprentissage, un jeu de données, $X = (x_{ik})$, donnant les caractéristiques $k = 1, ..., p$ des individus $i = 1, ..., n$. Alors les dissimilarités euclidiennes carrées, D_{ij}, entre deux individus i et j peuvent être caculées par (1.8) .

L'ensemble de test est formé d'individus supplémentaires. Les dissimilarités euclidiennes carrées D_{xj} entre un individu x de l'ensemble de test et un individu j de l'ensemble d'apprentissage sont, à nouveau, calculées selon (1.8).

Dans le cas particulier d'une table de contingence, dont les modalités en lignes sont des individus $i = 1, ..., n$; et celles en colonnes, des caractéristiques $k = 1, ..., p$, les dissimilarités du khi2 entre deux individus de l'ensemble d'apprentissage ou entre un individu de l'ensemble de test et un autre de l'ensemble d'apprentissage sont calculées par (1.6). Il est important de remarquer que, dans ces deux cas, les poids des caractéristiques, ρ_k, sont déterminés à partir de l'ensemble d'apprentissage uniquement. Ainsi, les colonnes qui ne seraient présentes que dans l'ensemble de test devraient être supprimées.

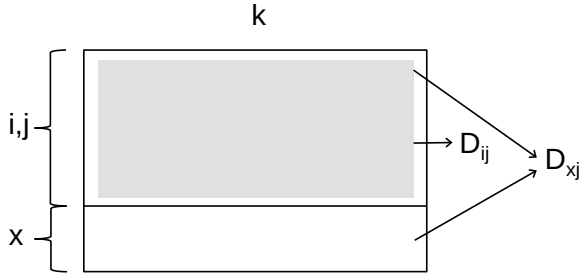


FIGURE 2.2 – Principe du calcul des dissimilarités entre deux individus de l'ensemble d'apprentissage, D_{ij}, et entre un individu de l'ensemble de test et un autre de l'ensemble d'apprentissage, D_{xj}.

Il est possible d'utiliser deux critères d'analyse discriminante. Le premier (*plus proches voisins*) attribue le nouvel individu x de l'ensemble de test au groupe contenant les individus d'apprentissage les plus proches de x en moyenne (Cocco, 2014), *i.e.* :

$$\underset{g}{\operatorname{argmin}} \sum_{j=1}^{n_g} f_j^g D_{xj} \tag{2.9}$$

où $f_j^g = \mathbb{1}(j \in g)/n_g$ est la distribution des individus j dans le groupe g, contenant n_g individus [1].

Le second critère (*plus proche centroïde*) attribue l'individu test x au groupe d'apprentissage dont le centroïde est le plus proche (Bavaud, 2011; Cocco, 2014), soit :

$$\underset{g}{\operatorname{argmin}}\, D_{xg} \tag{2.10}$$

où g est le profil moyen des n_g individus constituant le groupe g. Ces deux critères sont liés par le théorème de Huygens (1.18) qui permet de calculer les dissimilarités D_{xg} à partir des dissimilarités D_{xj} et de l'inertie du groupe Δ_g, calculée sur l'ensemble d'apprentissage (D_{ij}) par (1.19). Ainsi, si les inerties de tous les groupes sont égales, alors les critères (2.9) et (2.10) sont identiques ; sinon, l'attribution d'un nouvel individu au groupe g dépend tant de la position du centroïde que de l'inertie du groupe pour le critère des plus proches voisins, alors qu'il ne dépend que de la position du centroïde pour le critère du plus proche centroïde

Comme pour l'algorithme K-means, les deux critères ci-dessus peuvent être étendus en considérant des transformations de Schoenberg (section 1.3.4), et en particulier la transformation

1. $\mathbb{1}(A)$ représente la *fonction caractéristique* de A qui vaut 1 si A est vrai, et 0 sinon.

de puissance (1.22). Cette transformation est appliquée, pour le premier critère (2.9), sur D_{xj} transformé en $\tilde{D}_{xj} = \varphi(D_{xj})$; et pour le second critère (2.10), sur D_{xj} et D_{ij}, transformés en $\tilde{D}_{xj} = \varphi(D_{xj})$ et $\tilde{D}_{ij} = \varphi(D_{ij})$. Il faut noter que $\tilde{D}_{ij} := \varphi(D_{ij})$, donc $\tilde{\Delta}_g := 1/2 \sum_{ij} f_i^g f_j^g \tilde{D}_{ij}$, mais que $\tilde{D}_{xg} := \sum_j f_j^g \tilde{D}_{xj} - \tilde{\Delta}_g \neq \varphi(D_{xg})$. En d'autres termes, appliquer la transformation de Schoenberg sur D_{ij} et sur D_{xj}, *avant* d'utiliser le principe de Huygens pour obtenir D_{xg}, n'est pas équivalent à utiliser le principe de Huygens pour déterminer D_{xg}, *puis* à y appliquer la transformation de Schoenberg.

2.3 Évaluation

Il est souvent nécessaire de pouvoir évaluer une classification, qu'elle soit supervisée ou non. En général, dans le cas d'une classification non supervisée, les groupes auxquels appartiennent les individus ne sont pas connus a priori et les méthodes d'évaluation, basées sur des critères internes (*internal criteria*), consistent alors principalement à vérifier l'homogénéité des groupes (voir par exemple Estivill-Castro, 2002; Halkidi, Batistakis et Vazirgiannis, 2002). Elles ne seront pas présentées ici. Cependant, pour une des applications présentée dans ce travail (chapitre 4), basée sur un corpus restreint, une classification non supervisée a été effectuée, bien que les groupes soient connus a priori. Dans ce cas, les groupes créés par l'algorithme ne possèdent pas de signification particulière et ne sont pas forcément de même nombre que les groupes « recherchés », rendant toute comparaison directe difficile. On peut alors utiliser une mesure d'accord entre partitions (section 2.3.1) pour comparer les résultats connus a priori avec ceux obtenus par l'algorithme (*external criteria*).

Concernant la classification supervisée, la comparaison entre les groupes connus a priori et les résultats de l'algorithme est plus directe et de nombreux indices d'évaluation ont été proposés dans la littérature (voir par exemple Manning et Schütze, 1999, section 8.1 ; Sokolova et Lapalme, 2009). Seuls trois indices seront présentés ici : la précision, le rappel et la F-mesure (section 2.3.2).

2.3.1 Accord entre partitions

On considère deux partitions, X et Y, obtenues soit avec deux classifications non supervisées différentes (deux méthodes différentes ou en changeant un paramètre par exemple), soit par une classification non supervisée et une autre créée par un expert humain. On peut alors construire une table de contingence (section 1.1), dont les composantes n_{jk} comptent le nombre d'objets ou d'individus attribués simultanément au groupe j de la première partition X et au groupe k de la seconde partition Y. Toutes les mesures d'accord entre partitions se basent sur cette table de contingence.

Il existe de nombreux indices servant à mesurer l'accord entre deux partitions (voir par exemple Pfitzner, Leibbrandt et Powers, 2009), tels que l'indice de Meilă (Meilă, 2003) ou, dans le cas de deux partitions binaires, le coefficient phi ou le Q de Yule (sections 1.2.2.2 et 1.2.2.3). Seul deux de ces indices seront présentés et utilisés dans ce travail, à savoir : *l'indice de Jaccard* et *l'indice de Rand corrigé*.

L'indice de Jaccard (Youness et Saporta, 2004; Denœud et Guénoche, 2006), dont les valeurs varient entre 0 et 1, se définit comme :

$$J = \frac{r}{r + u + v} \tag{2.11}$$

où $r = \frac{1}{2} \sum_{jk} n_{jk}(n_{jk} - 1)$ est le nombre de paires simultanément dans un même groupe dans X et dans Y, $u = \frac{1}{2}(\sum_k n_{\bullet k}^2 - \sum_{jk} n_{jk}^2)$ est le nombre de paires qui sont dans des groupes distincts dans X et dans un même groupe dans Y et $v = \frac{1}{2}(\sum_j n_{j\bullet}^2 - \sum_{jk} n_{jk}^2)$ compte le nombre de paires dans le même groupe de X, mais dans des groupes distincts de Y.

Quant à *l'indice de Rand corrigé* (*Adjusted Rand Index*) (Hubert et Arabie, 1985; Denœud et Guénoche, 2006), il se calcule comme :

$$RC = \frac{r - \mathrm{Exp}(r)}{\mathrm{Max}(r) - \mathrm{Exp}(r)} \tag{2.12}$$

Dans cette équation, $\mathrm{Exp}(r) = \frac{1}{2n(n-1)} \sum_i n_{i\bullet}(n_{i\bullet} - 1) \sum_j n_{\bullet j}(n_{\bullet j} - 1)$ représente le nombre attendu de paires d'individus, sous l'hypothèse du hasard, dans un même groupe de X et dans un même groupe de Y et $\mathrm{Max}(r) = \frac{1}{4} \sum_i n_{i\bullet}(n_{i\bullet}-1) + \sum_j n_{\bullet j}(n_{\bullet j}-1)$ calcule la valeur maximum de l'indice de Rand. Ainsi, l'indice de Rand corrigé possède une valeur maximale de 1. De plus, il vaut 0 lorsque les similarités entre les deux partitions correspondent aux valeurs attendues sous l'hypothèse du hasard. Cependant, cet indice peut aussi prendre des valeurs négatives lorsque $r < \mathrm{Exp}(r)$, *i.e.* que l'accord entre les deux partitions est moins bon qu'un accord obtenu au hasard.

2.3.2 Précision, rappel et F-mesure

Comme déjà expliqué dans la section 2.2 sur la classification supervisée, après la phase d'apprentissage, vient la phase de test où l'algorithme attribue un groupe g à chaque individu i de l'ensemble de test (jeu de données de référence). Pour mesurer la performance de l'algorithme, il faut alors comparer, pour chaque individu, le groupe attribué par l'algorithme (décision) à celui déjà connu (référence).

Il existe trois mesures très généralement utilisées qui permettent d'évaluer les méthodes de classification supervisée : la précision, le rappel et la F-mesure, qui combine les deux premières (voir par exemple Manning et Schütze, 1999, section 8.1 ; Sebastiani, 2002 ; Sokolova et Lapalme, 2009 ; Cocco, 2014, dont cette section reprend une partie de l'exposé).

Avant tout, précisons qu'il existe différents problèmes de classification supervisée, à savoir :

binaire Il existe un seul groupe et chaque individu appartient ou non à ce groupe.

multi-classe Il existe m groupes et chaque individu appartient à un de ces m groupes.

multi-étiquette Il existe m étiquettes et chaque individu peut se voir attribuer une *ou plusieurs* étiquettes. Évidemment, si une seule étiquette est attribuée à chaque individu, alors cette classification est complètement équivalente à la classification multi-classe.

Dans le cas d'une classification **binaire** pour un groupe g, on peut construire une matrice de confusion

Groupe g	Référence	
Décision	OUI	NON
OUI	VP_g	FP_g
NON	FN_g	VN_g

dont les composantes comptent :

- les vrais positifs, VP_g, *i.e.* le nombre d'individus attribués au groupe g par la classification supervisée et classés dans le groupe g dans le jeu de données de référence,
- les faux positifs, FP_g, *i.e.* le nombre d'individus attribués au groupe g par la classification supervisée et non classés dans le groupe g dans le corpus de référence ,
- les faux négatifs, FN_g, *i.e.* le nombre d'individus non attribués au groupe g par la classification supervisée et classés dans le groupe g dans le corpus de référence et
- les vrais négatifs, VN_g, *i.e.* le nombre d'individus non attribués au groupe g par la classification supervisée et non classés dans le groupe g dans le corpus de référence.

Alors, la **précision** détermine le rapport entre le nombre d'individus correctement classés par le système dans le groupe g et le nombre total d'individus classés dans ce même groupe g, correctement ou non, soit :

$$P_g = \frac{VP_g}{VP_g + FP_g} \tag{2.13}$$

Quant au **rappel**, il se définit comme le rapport entre le nombre d'individus correctement classés par l'algorithme dans le groupe g et le nombre d'individus appartenant effectivement à ce groupe dans le jeu de données de référence :

$$R_g = \frac{VP_g}{VP_g + FN_g} \tag{2.14}$$

Si la classification est parfaite, alors la précision et le rappel seront tous deux égaux à 1. Un système performant exige des valeurs élevées *pour ces deux mesures*. En effet, il serait simple, de construire un système qui renvoie tous les documents dans le même groupe. Dans ce cas, et pour ce groupe, le rappel serait égal à 1, mais la précision très faible, d'où la nécessité d'étudier ces deux valeurs simultanément.

Dans son chapitre 7, van Rijsbergen (1979) propose de mesurer la proportion de la différence entre les éléments appartenant réellement au groupe g et ceux attribués à ce même groupe par l'algorithme, comme :

$$E = 1 - \frac{1}{\alpha(\frac{1}{P_g}) + (1-\alpha)\frac{1}{R_g}}$$

où α est un paramètre défini comme $\alpha = \frac{1}{(\beta^2+1)}$, dans lequel le nouveau paramètre β permet de spécifier différentes situations, telles que :

— l'utilisateur attache la même importance à la précision et au rappel ($\beta = 1$ et $\alpha = 1/2$),
— l'utilisateur n'attache aucune importance à la précision ($\beta \to \infty$ et $\alpha \to 0$) et
— l'utilisateur n'attache aucune importance au rappel ($\beta \to 0$ et $\alpha \to 1$).

La fonction F_β, communément utilisée, n'est autre que $1 - E$ (Manning et Schütze, 1999, section 8.1), soit :

$$F_\beta = \frac{(\beta^2+1)P_g R_g}{\beta^2 P_g + R_g}$$

La F-mesure, cas particulier de la fonction F_β pour $\beta = 1$, constitue la moyenne harmonique entre la précision et le rappel :

$$F_g = \frac{2P_g R_g}{P_g + R_g} \tag{2.15}$$

Dans le cas d'une analyse **multi-classe** ou **multi-étiquette**, deux types de moyennes des mesures (2.13), (2.14) et (2.15) peuvent être utilisées pour évaluer la performance de la classification sur l'ensemble des groupes (voir par exemple Sebastiani, 2002, section 7), à savoir, la *macro-moyenne* :

$$P_{macro} = \frac{\sum_{g=1}^{m} P_g}{m} \qquad R_{macro} = \frac{\sum_{g=1}^{m} R_g}{m} \qquad F_{macro} = \frac{2P_{macro}R_{macro}}{P_{macro} + R_{macro}} \tag{2.16}$$

et la *micro-moyenne* :

$$P_{micro} = \frac{\sum_{g=1}^{m} VP_g}{\sum_{g=1}^{m}(VP_g + FP_g)} \qquad R_{micro} = \frac{\sum_{g=1}^{m} VP_g}{\sum_{g=1}^{m}(VP_g + FN_g)}$$
$$F_{micro} = \frac{2P_{micro}R_{micro}}{P_{micro} + R_{micro}} \tag{2.17}$$

Dans la macro-moyenne, tous les groupes ont le même poids, alors que dans la micro-moyenne, tous les individus ont le même poids. Ainsi, dans cette dernière, les groupes les plus fréquents

auront plus d'importance (Yang, 1999). On peut aussi remarquer que dans le cas d'une analyse multi-classe, $\sum_{g=1}^{m} FP_g = \sum_{g=1}^{m} FN_g$, ce qui implique que $P_{micro} = R_{micro} = F_{micro}$ (Van Asch, 2012).

CHAPITRE 3

Indices d'autocorrélation et d'autocorrélation croisée

En analyse des séries temporelles (voir par exemple Box et Jenkins, 1976), la corrélation croisée mesure la corrélation entre deux signaux numériques univariés, dont un est décalé d'un certain temps (*lag*) par rapport à l'autre. Quant à l'autocorrélation, elle mesure la corrélation croisée entre un signal et lui-même.

Les indices d'autocorrélation et d'autocorrélation croisée présentés dans ce chapitre ont une double visée : d'une part, étendre l'analyse des séries temporelles à des problèmes numériques multivariés, ainsi qu'à des variables catégorielles multimodales (via la dissimilarité du khi2) ; et d'autre part, généraliser la notion de décalage à une notion de voisinage.

Soit $i, j = 1, \ldots, n$, des *positions* ordonnées, et $D = (D_{ij})$, la matrice des dissimilarités euclidiennes carrées entre ces positions. Plus précisément, ces dissimilarités sont calculées par rapport aux caractéristiques k des unités localisées sur ces positions. En définissant un voisinage par l'intermédiaire d'une *matrice d'échange* $E = (e_{ij})$ (section 3.1), *l'indice d'autocorrélation* (section 3.2) va mesurer la différence entre la variabilité des dissimilarités sur l'ensemble des positions et la variabilité locale dans un voisinage, tel que défini par E. *L'indice d'autocorrélation croisée* (section 3.3) généralise celui d'autocorrélation en considérant deux jeux de données et mesure la similarité entre les positions de ces deux jeux, par rapport aux caractéristiques k de chacun de ces jeux, selon le voisinage défini par E.

3.1 Matrice d'échange

Les voisins j de la position i sont déterminés par une matrice d'échange $E = (e_{ij})$, de taille $n \times n$, qui a pour propriétés d'être :

— non négative,
— symétrique,
— compatible avec le poids des individus $e_{i\bullet} = e_{\bullet i} = f_i$,
— et normalisée $e_{\bullet\bullet} = 1$.

Ainsi, e_{ij} peut s'interpréter comme la probabilité jointe de sélectionner les positions i et j, sans considération de l'ordre de ces positions ; et $e_{i\bullet} = f_i$, comme la probabilité de sélectionner la position i. On peut aussi remarquer que $w_{ij} = \frac{e_{ij}}{f_i}$ correspond aux composantes de la matrice $W = (w_{ij})$ de transition d'une chaîne de Markov de distribution stationnaire f.

3.1.1 Exemples

En toute généralité, les « positions » i, j réfèrent à des objets (localisés dans l'espace, dans le temps, ou plus généralement simplement identifiés par leurs indices i, j) exempts de relations mutuelles particulières *a priori*, ces dernières étant précisément définies par la matrice d'échange E.

Dans cette thèse, le cas particulier des *séries temporelles* est abordé, ce qui signifie que les indices i et j peuvent être mis en correspondance au moyen de relations de la forme $j = i + r$, où r est un entier relatif. Parmi les nombreuses matrices d'échange potentiellement pertinentes dans ce contexte particulier, trois familles seront présentées ici et utilisées par la suite.

La première matrice d'échange $\mathring{E}$, qu'on appellera matrice d'échange *itérée*, considère des voisinages à r itérations avec corrections dans les bords (Bavaud, Cocco et Xanthos, 2012). Pour $r = 1$, la matrice d'échange vaut [1] :

$$\mathring{e}_{ij}^{(1)} := \frac{1}{2n}[\mathbb{1}(j = i \pm 1) + \mathbb{1}(i = j = 1) + \mathbb{1}(i = j = n)] \tag{3.1}$$

Puis, pour $r > 1$, on définit $\mathring{E}^{(r)} = \Pi W^r$, avec $\Pi = \mathrm{diag}(f)$. Étant donné que cette matrice produit des poids uniformes, tels que $f_i = 1/n$, alors $w_{ij} = n\,\mathring{e}_{ij}$, avec $\mathring{e}_{ij} = \mathring{e}_{ij}^{(1)}$, et donc $\mathring{E}^{(r)} = \frac{1}{n}W^r = n^{(r-1)}\mathring{E}^r$.

La deuxième est une matrice d'échange *périodique*, $\breve{E}$, qui considère les voisins j à une distance (*lag*) r (à gauche et à droite) de la position i (Cocco et Bavaud, accepté pour publication) :

$$\breve{e}_{ij}^{(r)} = \frac{1}{2n}[\mathbb{1}(j = (i \pm r) \bmod n) + \mathbb{1}((i \pm r) \bmod n = 0) \cdot \mathbb{1}(j = n)] \tag{3.2}$$

Comme la matrice d'échange itérée, cette matrice d'échange produit des poids uniformes. De plus, comme le voisinage est périodique, alors $\breve{E}^{(r)} = \breve{E}^{(n-r)}$.

Finalement, la matrice d'échange à *fenêtres mobiles*, $\dot{E}$, considère toutes les positions dans deux fenêtres de largeur r, l'une à gauche et l'autre à droite (Bavaud *et al.*, 2012) :

$$\dot{e}_{ij}^{[r]} = \frac{c_{ij}^{[r]}}{c_{\bullet\bullet}^{[r]}} \qquad\qquad c_{ij}^{[r]} := \mathbb{1}(|j - i| \leq r) \cdot \mathbb{1}(i \neq j) \tag{3.3}$$

Contrairement aux deux autres matrices, les poids résultants ne sont pas uniformes, mais plus petits pour les positions de bord que pour les autres.

Toutes ces matrices d'échange dépendent principalement de la *différence* $|j - i|$ des positions i et j (à des effets de bord près), et l'on s'attend à ce que leur utilisation permette de révéler d'autant mieux un phénomène que la loi le gouvernant soit *stationnaire*, *i.e.* invariante par translation $|j - i|$. Ce qui, on peut le préciser, n'affecte en rien la question de la légitimité de leur utilisation dans le cadre d'une *analyse exploratoire de données*, telle qu'effectuée aux chapitres 6 et 8.

Deux exemples ($r = 1$ et $r = 2$) de chacun de ces trois types de matrices d'échange sont présentés dans la table 3.1 pour 5 positions ordonnées. Le réseau non pondéré et non orienté correspondant à chacun de ces six exemples est exposé dans la table 3.2. On remarque que les matrices d'échange périodique et à fenêtres mobiles sont assez similaires, cependant elles présentent deux différences essentielles :

— premièrement, comme son nom l'indique, la matrice d'échange périodique considère que les positions sont périodiques et donc que la position 1 se trouve après la position n, contrairement à la matrice d'échange à fenêtres mobiles ;

1. Comme déjà mentionné (cf. chapitre 2, note 1), $\mathbb{1}(A)$ représente la fonction caractéristique associée à l'événement A.

— deuxièmement, avec la matrice d'échange à fenêtres mobiles toutes les positions des deux fenêtres de largeur r sont considérées, à l'inverse de la matrice d'échange périodique pour laquelle on ne considère que deux positions à une distance r d'une position donnée, sans prendre en compte les positions intercalaires qui les séparent.

r	$\mathring{E}$ (itérée)	$\breve{E}$ (périodique)	$\dot{E}$ (fenêtres mobiles)
$r=1$	$\frac{1}{10}\begin{pmatrix} \mathbf{1} & \mathbf{1} & 0 & 0 & 0 \\ \mathbf{1} & 0 & \mathbf{1} & 0 & 0 \\ 0 & \mathbf{1} & 0 & \mathbf{1} & 0 \\ 0 & 0 & \mathbf{1} & 0 & \mathbf{1} \\ 0 & 0 & 0 & \mathbf{1} & \mathbf{1} \end{pmatrix}$	$\frac{1}{10}\begin{pmatrix} 0 & \mathbf{1} & 0 & 0 & \mathbf{1} \\ \mathbf{1} & 0 & \mathbf{1} & 0 & 0 \\ 0 & \mathbf{1} & 0 & \mathbf{1} & 0 \\ 0 & 0 & \mathbf{1} & 0 & \mathbf{1} \\ \mathbf{1} & 0 & 0 & \mathbf{1} & 0 \end{pmatrix}$	$\frac{1}{8}\begin{pmatrix} 0 & \mathbf{1} & 0 & 0 & 0 \\ \mathbf{1} & 0 & \mathbf{1} & 0 & 0 \\ 0 & \mathbf{1} & 0 & \mathbf{1} & 0 \\ 0 & 0 & \mathbf{1} & 0 & \mathbf{1} \\ 0 & 0 & 0 & \mathbf{1} & 0 \end{pmatrix}$
$r=2$	$\frac{1}{20}\begin{pmatrix} \mathbf{2} & \mathbf{1} & \mathbf{1} & 0 & 0 \\ \mathbf{1} & \mathbf{2} & 0 & \mathbf{1} & 0 \\ \mathbf{1} & 0 & \mathbf{2} & 0 & \mathbf{1} \\ 0 & \mathbf{1} & 0 & \mathbf{2} & \mathbf{1} \\ 0 & 0 & \mathbf{1} & \mathbf{1} & \mathbf{2} \end{pmatrix}$	$\frac{1}{10}\begin{pmatrix} 0 & 0 & \mathbf{1} & \mathbf{1} & 0 \\ 0 & 0 & 0 & \mathbf{1} & \mathbf{1} \\ \mathbf{1} & 0 & 0 & 0 & \mathbf{1} \\ \mathbf{1} & \mathbf{1} & 0 & 0 & 0 \\ 0 & \mathbf{1} & \mathbf{1} & 0 & 0 \end{pmatrix}$	$\frac{1}{14}\begin{pmatrix} 0 & \mathbf{1} & \mathbf{1} & 0 & 0 \\ \mathbf{1} & 0 & \mathbf{1} & \mathbf{1} & 0 \\ \mathbf{1} & \mathbf{1} & 0 & \mathbf{1} & \mathbf{1} \\ 0 & \mathbf{1} & \mathbf{1} & 0 & \mathbf{1} \\ 0 & 0 & \mathbf{1} & \mathbf{1} & 0 \end{pmatrix}$

TABLE 3.1 – Exemples pour les trois matrices d'échanges étudiées, avec $n = 5$.

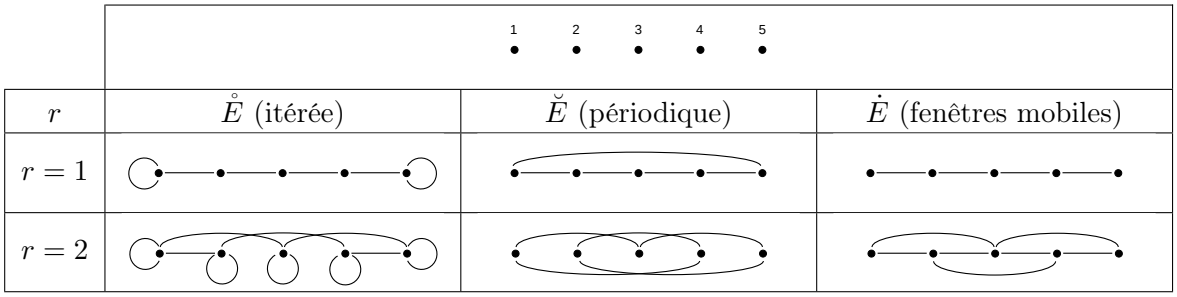


TABLE 3.2 – Réseau non pondéré et non orienté des trois matrices d'échange étudiées, représentant les liens non nuls ($e_{ij} > 0$) entre les $n = 5$ positions.

3.2 Indice d'autocorrélation

L'indice d'autocorrélation se définit comme (Bavaud *et al.*, 2012) :

$$\delta := \frac{\Delta - \Delta_{\text{loc}}}{\Delta} \in [-1, 1] \tag{3.4}$$

où Δ est l'inertie (globale) qui se calcule, à partir de la matrice des dissimilarités euclidiennes carrées entre les positions D_{ij}, par (1.16) et Δ_{loc} est l'inertie locale, telle que :

$$\Delta_{\text{loc}} := \frac{1}{2}\sum_{ij} e_{ij} D_{ij}$$

Cet indice d'autocorrélation généralise le I de Moran (Moran, 1950), la mesure standard de l'autocorrélation spatiale d'une variable numérique, ou son complémentaire, le c de Geary (Geary, 1954; Lebart, 1969), au cas multivarié. Concernant l'interprétation, comme le montre l'équation (3.4), δ sera élevé si les individus sont plus similaires dans le voisinage défini par E qu'en prenant des positions choisies aléatoirement, et inversement.

Rappelons que la dissimilarité euclidienne carrée D_{ij} est basée sur les caractéristiques des unités apparaissant en i et j. Dans le cas de caractéristiques catégorielles, D_{ij} ne sera autre

que la dissimilarité du khi2 entre les lignes (ou les colonnes) i et j de la table de contingence associée, calculée par (1.6) (ou (1.7)), ou encore par (1.9) : voir le chapitre 8. Le chapitre 6 décrit d'autres applications impliquant des dissimilarités euclidiennes carrées distinctes de celles du khi2.

3.2.1 Test d'autocorrélation

L'espérance de l'indice d'autocorrélation sous l'hypothèse H_0 d'absence d'autocorrélation vaut (voir par exemple Bavaud, 2013) :

$$E_0(\delta) = \frac{\mathrm{trace}(W) - 1}{n - 1} \tag{3.5}$$

avec $W = (w_{ij})$, la matrice de transition de Markov, telle que définie dans la section 3.1. Concernant les exemples de la section 3.1.1, l'espérance sous indépendance de la matrice d'échange itérée est variable selon r et vaut $E_0^{(r)} = (\mathrm{trace}(W^r) - 1)/(n-1)$, alors qu'elle a une valeur fixe pour les deux autres matrices d'échange, soit $E_0^{(r)} = -1/(n-1)$.

La variance correspondante s'écrit (voir par exemple Cliff et Ord, 1981) :

$$\mathrm{Var}_0(\delta) = \tfrac{2}{n^2-1}\left[\mathrm{trace}(W^2) - 1 - \tfrac{(\mathrm{trace}(W)-1)^2}{n-1}\right]$$

Sous approximation normale, on peut ainsi évaluer la significativité statistique de l'indice d'autocorrélation au niveau α en effectuant le test suivant :

$$\left|\frac{\delta - E_0(\delta)}{\sqrt{\mathrm{Var}_0(\delta)}}\right| \geq u_{1-\alpha/2} \tag{3.6}$$

où $u_{1-\alpha/2}$ est le α-ème quantile de la loi normale standardisée.

3.3 Indice d'autocorrélation croisée

Soit deux jeux de coordonnées $X = (x_{ik})$ et $Y = (y_{ik})$ munis des mêmes positions $i = 1, \ldots, n$ et des mêmes caractéristiques $k = 1, \ldots, p$, mais dont les valeurs diffèrent [2]. Alors, on définit l'indice d'autocorrélation croisée comme (Cocco et Bavaud, accepté pour publication) :

$$\delta(X, Y) := \frac{\Delta(X, Y) - \Delta_{\mathrm{loc}}(X, Y)}{\sqrt{\Delta(X)\Delta(Y)}} \in [-1, 1] \tag{3.7}$$

Dans cette équation, $\Delta(X)$ représente l'inertie globale de X (1.16), identique à celle utilisée dans (3.4). Puis, en définissant la *dissimilarité croisée* entre deux positions i et j des deux jeux de coordonnées X et Y comme :

$$D_{ij}^{xy} = \sum_k (x_{ik} - x_{jk})(y_{ik} - y_{jk})$$

on peut définir l'inertie croisée entre X et Y comme :

$$\Delta(X, Y) = \frac{1}{2}\sum_{ij} f_i f_j D_{ij}^{xy} = \sum_i f_i \sum_k x_{ik} y_{ik} - \sum_k \bar{x}_k \bar{y}_k$$

et l'inertie croisée locale comme :

$$\Delta_{\mathrm{loc}}(X, Y) = \frac{1}{2}\sum_{ij} e_{ij} D_{ij}^{xy} = \sum_i f_i \sum_k x_{ik} y_{ik} - \sum_{ij} e_{ij} \sum_k x_{ik} y_{jk}$$

2. Il pourrait s'agir, par exemple, de différents indices k concernant la population, tels que le taux de naissance ou d'immigration, pour des régions i, à deux dates différentes, soit X et Y.

Étant donné que $\Delta(X, X) = \Delta(X)$ et que $\Delta_{\text{loc}}(X, X) = \Delta_{\text{loc}}(X)$, il apparaît que l'indice d'autocorrélation croisée est une généralisation de l'indice d'autocorrélation, car $\delta(X, X) = \delta(X) = \delta$, tel que défini dans l'équation (3.4).

L'indice $\delta(X, Y)$ (3.7) est applicable à deux jeux de coordonnées, X et Y, ssi, comme déjà mentionné, les deux jeux de coordonnées sont munis des mêmes positions i et des mêmes caractéristiques k, mais aussi ssi les poids des positions de X, f_i^x, sont identiques à ceux de Y, f_i^y, soit $f_i^x = f_i^y = f_i$. L'autocorrélation croisée $\delta(X, Y)$ peut aussi se concevoir comme une version pondérée du coefficient de codispersion (voir par exemple Matheron, 1965; Rukhin et Vallejos, 2008) utilisé en Géostatistique.

Si les données de départ sont catégorielles, alors l'indice d'autocorrélation croisée entre deux tables de contingence N^α et N^β est $\delta({}^*X^\alpha, {}^*X^\beta)$ (respectivement $\delta({}^*Y^\alpha, {}^*Y^\beta)$), où ${}^*x_{ik}^\alpha$ et ${}^*x_{ik}^\beta$ (respectivement ${}^*y_{ik}^\alpha$ et ${}^*y_{ik}^\beta$) sont les coordonnées de haute dimensionnalité (1.10) des lignes (respectivement des colonnes). Dans ce cas, l'indice d'autocorrélation croisée $\delta({}^*X^\alpha, {}^*X^\beta)$ mesure la similarité entre la distribution des caractéristiques catégorielles k de la table de contingence α et la distribution des caractéristiques de la table β dans un voisinage déterminé par E. Il est ainsi utilisé dans le chapitre 8, section 8.3.

Partie II

APPLICATIONS TEXTUELLES

CHAPITRE 4

Classification non supervisée en types de discours

Le travail présenté dans ce chapitre est à la fois un résumé et une extension de trois articles (Cocco, Pittier, Bavaud et Xanthos, 2011; Cocco, 2012*a*,*b*) et en reprend de larges extraits. Le but de ce chapitre est de catégoriser automatiquement des propositions énoncées par rapport à des séquences textuelles, comprises ici comme des *types de discours*, tels que le narratif, l'argumentatif, l'explicatif, le descriptif, le dialogal et l'injonctif (section 4.1.1).

Pour ce faire, quatre contes de Maupassant ont d'abord été segmentés en propositions et annotés par un expert humain (section 4.1). Ensuite, les propositions ont été représentées à l'aide d'une AFC (section 4.2.1). Puis elles ont été classées automatiquement (classification non supervisée) en se basant sur les catégories morphosyntaxiques (CMS) qu'elles contiennent, et plus précisément sur les n-grammes de CMS et les résultats sont évalués par le biais d'indices d'accords entre partitions (section 4.3).

Les CMS ont été choisies comme caractéristiques de cette classification non supervisée, car elles ont déjà montré leur utilité dans des travaux connexes. En effet, les CMS ont été de plus en plus exploitées, parmi d'autres caractéristiques, pour la catégorisation automatique de textes depuis les travaux de Biber (1988), qui s'intéresse à la détection de types de textes. Par exemple, Malrieu et Rastier (2001) travaillent sur la distinction, d'une part, et la classification automatique, d'autre part, de textes selon les genres (comédie, tragédie, drame, etc.) et selon les discours (littéraire, juridique, politique, etc.) en utilisant des variables majoritairement morphosyntaxiques. Karlgren et Cutting (1994) s'intéressent à la classification supervisée en genres de textes avec des CMS. On peut encore citer Palmer, Ponvert, Baldridge et Smith (2007) qui travaillent, en utilisant des CMS parmi d'autres caractéristiques, sur la classification supervisée de *situation entities*, un élément essentiel des modes de discours (*modes of discourse*) en linguistique anglaise (Smith, 2003), concepts relativement similaires aux types de discours en linguistique française. Pour déterminer si les CMS sont également utiles dans la détection des types de discours traités ici, une analyse préliminaire visant à mesurer le lien entre les CMS et les types de discours est effectuée dans la section 4.1.4. Finalement, la méthode et les résultats obtenus sont discutés dans la section 4.4.

4.1 Données

Les données se composent de quatre contes de Maupassant, du 19$^{\text{ème}}$ siècle, annotées en types de discours par un expert humain. Ce dernier a proposé de travailler sur des contes de Maupassant pour trois raisons : les textes n'étaient pas trop longs et pouvaient être annotés en un temps raisonnable, ils étaient susceptibles de contenir tous les types de discours et ils

étaient disponibles sur Internet. Aussi, un seul auteur et un seul genre sont considérés, car comme déjà expliqué dans l'introduction, les CMS varient en fonction des genres, mais aussi en fonction de l'auteur (voir par exemple Koppel et Schler, 2003). L'expert humain a utilisé des balises XML pour annoter les textes, une pratique standard dans ce domaine (voir par exemple Daoust, Marcoux et Viprey, 2010). Avant de pouvoir annoter les textes en types de discours, il a commencé par segmenter le texte en propositions énoncées, car le niveau des phrases, composées d'une ou plusieurs propositions énoncées, était trop grossier. C'est cette segmentation manuelle qui va servir de base à la classification non supervisée.

Après avoir présenté les critères utilisés par l'expert humain pour l'annotation en types de discours (section 4.1.1), le corpus, ainsi que quelques statistiques descriptives le caractérisant, sont exposés dans la section 4.1.2. Ensuite, le prétraitement pour la création des tables de contingence croisant les propositions et les CMS est expliqué (section 4.1.3). De plus, comme déjà mentionné dans l'introduction de ce chapitre, une analyse préliminaire a été effectuée afin de s'assurer que les CMS sont des caractéristiques utiles à la distinction des types de discours et les résultats sont présentés dans la section 4.1.4.

4.1.1 Types de discours et annotation

Les types de discours retenus pour ce projet sont adaptés des travaux de Jean-Michel Adam, spécialiste en linguistique textuelle et de Jean-Paul Bronckart, spécialiste en psycholinguistique et didactique des langues.

En premier lieu, il faut noter que l'appellation « types de discours » est abusive, mais sera généralement utilisée dans ce qui suit. En effet, même si elle est courante en Français (Filliettaz, 2001), le terme « types de séquences » est plus précis, car il fait référence à des passages de textes et non à des textes entiers, et c'est celui utilisé par Adam (2008*a*,*b*) en général et par Bronckart (1996) lorsqu'il aborde les types traités ici. De plus, lorsque Bronckart (1996, section 5.2) parle de types de discours, il distingue quatre architypes psychologiques : le discours interactif, le discours théorique, le récit interactif et la narration, qu'il différencie des séquences décrites par Adam (2008*a*,*b*). Partant de cela, il définit ensuite des types linguistiques (Bronckart, 1996, section 5.3). Au chapitre suivant, il passe en revue les « Séquences et autres formes de planification » qui sont les éléments traités dans ce projet, (Bronckart, 1996, p. 219, chapitre 6) :

> Dans notre approche, les types de discours constituent les ingrédients fondamentaux de l'*infrastructure générale des textes*, [...] L'infrastructure textuelle se caractérise cependant aussi par une autre dimension, qui est celle de **l'organisation séquentielle** ou **linéaire** de son contenu thématique.

De là, il reprend les séquences décrites par J.-M. Adam auxquelles il ajoute la séquence injonctive.

Les types de discours (ou séquences) considérés par Adam (2008*a*,*b*) sont le narratif, l'argumentatif, l'explicatif, le dialogal et le descriptif. En plus de ces cinq types, on considérera ici le type de discours (ou séquence) injonctif, suggéré par Bronckart (1996), qui, dans les textes traités dans ce projet, est toujours un « sous-type » du type dialogal [1]. Il a été demandé à l'expert humain, Raphaël Pittier, alors étudiant de master en sciences du langage et de la communication, ainsi qu'en français moderne (orientation linguistique française), d'annoter des textes selon ces six types de discours en se basant sur le travail de Adam (2008*a*,*b*) et Bronckart (1996). Dans ce qui suit, les types sont définis selon ces théories, ainsi que selon les critères retenus par l'expert humain, spécialiste dans ce domaine. De plus, il est fait mention des marques linguistiques que ce dernier a trouvé pertinentes.

Il faut aussi noter que Adam (2008*a*,*b*) différencie les périodes et les séquences de chaque type ; les séquences étant plus complexes et étendues que les périodes. Dans le cadre de ce

1. Pour l'anglais, l'appellation courante semble être *Modes of discourse* et selon Smith (2003), il y en a cinq : *narrative, description, report, information* et *argument*.

travail, cette distinction n'a pas été retenue. C'est pourquoi, les parties de textes, annotées comme étant d'un certain type, peuvent être des séquences ou des périodes ; voire même des parties plus courtes que la période comme dans le cas du discours direct pour le type dialogal (voir section 4.1.1.5). Néanmoins, il est important d'envisager les différences entre séquences et périodes dans l'esposé théorique des types de discours.

4.1.1.1 Narratif

Le type de discours narratif correspond au récit raconté. Trois sortes de parties de textes ont été annotées comme étant narratives :

1. **la séquence narrative** qui est composée d'étapes précises, dont certaines sont facultatives (cf. Adam, 2008*a*, schéma 20, p. 147) :
   — Pn0 : entrée-préface ou résumé : *facultative*,
   — Pn1 : situation initiale (orientation),
   — Pn2 : noeud (déclencheur),
   — Pn3 : (ré-)action ou évaluation,
   — Pn4 : dénouement (résolution),
   — Pn5 : situation finale,
   — PnΩ : chute ou évaluation finale (morale) : *facultative*.
   Lorsque les étapes facultatives sont présentes, on ne parle plus de séquence narrative, mais d'intrigue narrative.
2. **la période narrative ou l'épisode narratif** où un état de départ est suivi d'un événement qui transforme cet état initial afin de parvenir à un autre état.
3. **le narratif itératif** qui correspond à une description d'actions répétées ou simplement à des actions répétées, comme par exemple : « Tous les matins, il buvait du café... ». En raison de la répétition, cette catégorie de texte, annotée comme narrative, tend vers le type de discours descriptif.

Marques linguistiques : Pour la séquence narrative (point **1**), tout comme pour la période narrative (point **2**), on note souvent la présence de passé simple, mais ce n'est pas un critère absolu. En plus du passé simple, il peut exister des déclencheurs tels que la conjonction *or* ou la locution adverbiale *tout à coup*. Une autre tendance est la juxtaposition d'actions, soit des groupes qui se suivent dans l'ordre chronologique, comme par exemple : « Il alla à la bibliothèque, prit un livre, lut trois pages... ». Pour le narratif itératif (point **3**), l'imparfait est généralement utilisé. Mais à nouveau, il s'agit plus d'une tendance que d'un critère absolu. Bronckart (1996, pp. 179–181) propose une liste de marques linguistiques pour la narration, dont certaines, listées ci-avant, correspondent à celles utilisées par l'expert humain.

4.1.1.2 Argumentatif

Le type de discours argumentatif correspond à des textes, ou parties de textes, ayant pour but de convaincre l'autre de son argument, c'est-à-dire de démontrer, justifier ou réfuter une thèse.

En résumé, **la séquence argumentative** se compose (cf. Adam, 2008*a*, schéma 21, p. 150) :
— de données (prémisses) ou fait(s), suivies
— d'un étayage qui mène à
— une assertion conclusive.
Une présentation plus complète de cette séquence est exposée dans Adam (2008*a*, schéma 22, p. 151).

Concernant **la période argumentative**, il s'agit d'une « suite de propositions liées par des connecteurs argumentatifs » (Adam, 2008*a*, p. 150). Pour ce projet, nous avons considéré que

lorsque les prémisses sont implicites ou déjà mentionnées en amont, ou que l'étayage est implicite ou douteux, il s'agissait d'une période argumentative.

Marques linguistiques : Présence de connecteurs argumentatifs qui peuvent être (Adam, 2008*a*, p. 120) :
- argumentatifs et concessifs : *mais*, *pourtant*, *cependant*, *certes*, *toutefois*, *quand même*, ... ;
- explicatifs et justificatifs : *car*, *parce que*, *puisque*, *si - c'est que*, ... ;
- de simples marqueurs d'un argument : *même*, *d'ailleurs*, *de plus*, *non seulement*, ... ; et
- le *si* et le *quand* des phrases hypothétiques.

4.1.1.3 Explicatif

Le type explicatif se différencie du type argumentatif par sa fonction, qui n'est pas de convaincre, mais d'expliquer quelque chose de non su. Il s'agit plutôt de délivrer un type de savoir encyclopédique. L'explication répond à la question « Pourquoi ? » (Adam, 2008*b*, pp. 127–138).

La séquence explicative (cf. Adam, 2008*a*, schéma 26, p. 157) :
- commence par une schématisation initiale qui présente un objet complexe ;
- ensuite, par un premier opérateur *pourquoi*, passe à une schématisation qui construit l'objet comme problématique ;
- enfin, par un second opérateur *parce que*, passe à une schématisation explicative.

Quant aux **périodes explicatives**, elles sont souvent composées d'une proposition qui pose un problème et qui est introduite par *si* et d'une explication introduite par *c'est que* ou *c'est parce que* (Adam, 2008*a*, p. 153).

Marques linguistiques : Présence de locutions phraséologiques telles que (Adam, 2008*a*, section 4.5) : *(Si)... c'est parce que/c'est pour (que)/c'est pourquoi/c'est que/c'est en raison de/cela tient à...*, *voilà pourquoi...*, etc.

4.1.1.4 Descriptif

Le type descriptif consiste en un arrêt sur image où le temps de l'histoire s'arrête. Ce type de discours correspond donc à l'attribution des propriétés propres à un sujet, qu'il soit animé ou non. Il peut s'agir, par exemple, d'un personnage, d'un objet, d'un lieu ou d'une action (pour cette dernière, il s'agira plutôt, en général, de narratif itératif). Au plan de l'équilibre textuel, on n'observe pas une forme de séquence, mais plutôt différentes **opérations**, à savoir (Adam, 2008*a*, section 4.2) les opérations :
- de thématisation,
- d'aspectualisation,
- de mise en relation et
- d'expansion par sous-thématisations.

Par exemple, dans les opérations d'aspectualisation, le sujet à décrire peut être *fragmenté* en parties. Puis, ces parties peuvent être *qualifiées* par des adjectifs (Adam, 2008*a*, p. 142). En d'autres termes, des propriétés sont attribuées (essentiellement des adjectifs) au substantif de la description par l'intermédiaire, en général, d'un verbe d'état. Un substantif peut aussi remplacer l'adjectif, comme dans la phrase : « Cette table est un chef-d'œuvre. ».

Il faut encore noter que la description n'est pas, en général, dominante, mais plutôt au service d'un autre type (Bronckart, 1996, p. 238), notamment de la narration (Adam, 2008*b*, p. 100).

Marques linguistiques : Plusieurs marques linguistiques se retrouvent pour ce type :
- utilisation, en général, de verbes au passé et souvent à l'imparfait (cependant, lorsque la narration ou le discours est au présent, la description sera aussi au présent) ;

— forte proportion d'adjectifs, en raison de l'attribution de propriétés par des groupes nominaux de la forme *nom + adjectif* (Adam, 2008*a*, p. 142) ;
— présence d'organisateurs spatio-temporels : *à gauche*, *à droite*, *hier*, *demain*, *en haut*, *en bas*, *au premier plan*, *au second plan*,... ;
— présence de verbes d'état : *être*, *paraître*, *sembler*,... ; et
— présence, parfois, de constructions analogiques par l'intermédiaire de mots, tels que *comme*, *tel*, etc.

4.1.1.5 Dialogal

Le type dialogal se comprend comme la représentation d'un échange verbal se situant à un niveau différent du reste du récit ; il peut aussi se trouver dans un système verbo-temporel différent. Par exemple, un dialogue au présent peut être inclus dans une narration au passé.

Théoriquement, **la séquence dialogale** implique un échange. Typiquement, un texte conversationnel se compose (cf. Adam, 2008*a*, schéma 29, p. 161) :
— d'un échange d'ouverture (*séquence phatique*) ;
— d'une *séquence transactionnelle* comprenant
  — une question,
  — une réponse et
  — une évaluation ; et
— d'un échange de clôture (*séquence phatique*).

Notons que dans l'annotation utilisée pour ce travail, le discours direct a été considéré comme étant de type dialogal.

Marques linguistiques : Présence de guillemets, changement de tiroir verbo-temporel et, souvent, ponctuation forte, telle que le point d'interrogation ou d'exclamation. Parfois, on trouve aussi les points de suspension qui indiquent un discours non terminé ou interrompu. De plus, on note la présence de verbes introducteurs de discours direct tels que *il dit*, *elle demanda*, etc. Bien que ces verbes n'appartiennent pas directement au discours direct, ils permettent de faire la transition entre le récit principal et le discours direct.

4.1.1.6 Injonctif

Le type injonctif représente le fait d'ordonner quelque chose à quelqu'un. C'est une incitation à l'action, dont les formes de textualisation varient selon le genre de cette incitation (Adam, 2008*a*, p. 133). En résumé, le but est de « **faire agir** le destinataire d'une certaine manière ou dans une direction donnée » (Bronckart, 1996, p. 240). Ce type est considéré par Bronckart (1996), mais rejeté par Adam qui reconnaît les propriétés d'incitation à l'action du discours injonctif, mais qui se demande s'il ne s'agit pas d'« actualisations singulières d'un simple genre de description » (Adam, 2008*b*, p. 95).

Il se trouve que, dans le corpus traité ici, le type de discours injonctif est toujours placé dans une séquence dialogale (ou dans du discours direct).

Marques linguistiques : Verbes à l'impératif, points d'exclamation et verbes introducteurs du dialogue tel que *il lui ordonna*.

Remarque : Ce type de discours étant constamment inclus dans le type de discours dialogal dans nos textes, il serait possible de ne pas le considérer et d'attribuer tout ce qui le concerne au type dialogal, réduisant ainsi le nombre de types de discours à cinq. S'il s'agissait d'un texte correspondant à une recette de cuisine et annoté comme injonctif, il faudrait alors lui attribuer le type descriptif selon les séquences décrites par (Adam, 2008*b*, p. 95), mais cette situation ne se produit jamais dans les textes utilisés dans ce travail.

4.1.1.7 Structure hiérarchique et récursive

Il est clair que ces types de discours ne sont pas univoques et que leur interprétation pourrait différer pour un autre expert. Il faut encore ajouter que ces périodes ou séquences sont généralement imbriquées les unes dans les autres. Par exemple, dans un conte, on ne sera pas surpris de trouver une longue séquence narrative, parfois le conte entier, qui contiendra d'autres séquences, explicatives ou descriptives par exemple. Ces dernières pourront à leur tour contenir d'autres séquences du même type ou d'un autre type. Comme déjà expliqué, l'annotateur a utilisé des balises XML pour annoter le texte, ce qui a permis de prendre en compte cette structure hiérarchique (cf. figure 4.1). Cependant, dans la suite de ce chapitre, la structure du texte est considérée comme linéaire et seules les feuilles de l'arbre sont traitées.

4.1.2 Corpus

Comme déjà mentionné, l'expert humain a segmenté et annoté quatre textes de Maupassant qu'il a obtenu sur internet :

— « L'Orient » (de Maupassant, 1883),
— « Le Voleur » (de Maupassant, 1882),
— « Un Fou ? » (de Maupassant, 1884) et
— « Un Fou » (de Maupassant, 1885).

Il a choisi de traiter des contes de Maupassant, car il estimait que ces textes étaient susceptibles de contenir les six types de discours. Il faut aussi préciser que puisque l'annotation a été une tâche difficile qui a nécessité beaucoup de temps, il n'a pu annoter que quatre textes.

Pour annoter ces quatre textes, l'expert a utilisé les balises XML suivantes :

<e>...</e>	Balises ouvrantes et fermantes qui délimitent les propositions.
<cr/>	Balises vides qui marquent la fin des paragraphes (ou les retours chariot).
** ... **	Balises ouvrantes et fermantes qui délimitent les différents types de discours et contiennent un attribut, nommé *type*, indiquant le type de discours. Une valeur supplémentaire, nommée *date*, a été ajoutée à cet attribut pour le texte « Un Fou » ; ceci afin de délimiter les dates, ce texte étant écrit sous la forme d'un journal intime.

Un exemple est présenté dans la figure 4.1 pour le texte « L'Orient » et l'ensemble des quatre textes annotés se trouve dans l'annexe A.

```
<div type="narratif">
  <e>Je le trouvai tantôt couché sur un divan,
   en plein rêve d'opium.</e>
  <e>Il me tendit la main sans remuer le corps,</e>
  <e>et me dit :</e><cr/>
  <div type="dialogal">
    <div type="injonctif">
      <e>Reste là, parle,</e>
    </div>
    <div type="argumentatif">
      <e>je te répondrai de temps en temps,</e>
      <div type="explicatif">
        <e>mais je ne bougerai point,</e>
        <e>car tu sais qu'une fois la drogue avalée</e>
        <e>il faut demeurer sur le dos.</e><cr/>
      </div>
    </div>
  </div>
</div>
```

FIGURE 4.1 – Extrait annoté de « L'Orient » correspondant aux lignes 14 à 29 de l'annexe A.1.

Les statistiques descriptives concernant les quatre textes annotés par l'expert humain sont données dans la table 4.1. Ces valeurs sont basées sur l'utilisation d'unigrammes. Pour les bi- et les trigrammes, on a supprimé les propositions composées respectivement de moins de deux ou trois occurrences selon TreeTagger (Schmid, 1994), l'outil utilisé pour annoter les textes en CMS. Ainsi, pour « L'Orient », trois propositions ont été retirées pour l'analyse basée sur des trigrammes. Concernant « Le Voleur », une proposition a été supprimée pour l'utilisation de trigrammes. Pour le texte « Un Fou ? », treize propositions ont été soustraites, à nouveau pour l'analyse avec des trigrammes. Pour le texte « Un Fou », une étape additionnelle a été effectuée. Comme déjà mentionné, des balises supplémentaires entourant les dates ont été ajoutées, car ce texte est écrit, majoritairement, sous la forme d'un journal intime et il est difficile d'attribuer les dates à l'un des six types de discours proposés. Ces dates ont donc été retirées, réduisant le nombre de proposition de 401 à 376. Finalement, deux propositions ont été retirées pour les bigrammes et dix de plus pour les trigrammes.

Textes	♯ phrases	♯ prop.	♯ occurrences		♯ formes		% de types de discours selon l'expert humain					
			ponct.	s/ ponct.	mot	CMS	nar	arg	expl	descr	dial	inj
L'Orient	88	189	1'749	1'488	654	27	28.04	4.23	19.05	20.11	25.93	2.65
Le Voleur	102	208	1'918	1'582	667	29	61.54	4.81	4.81	12.02	13.94	2.88
Un Fou ?	150	314	2'625	2'185	764	28	33.76	18.15	14.65	10.51	14.65	8.28
Un Fou	242	376	3'065	2'548	828	29	42.55	17.82	11.70	13.83	1.86	12.23

TABLE 4.1 – Statistiques descriptives pour les quatre textes annotés de Maupassant. Pour le texte « Un Fou », les dates ont préalablement été retirées du texte. Nombre de phrases telles que considérées par TreeTagger (Schmid, 1994). Nombre de propositions telles que segmentées par l'expert humain. Nombre d'occurrences (*tokens*) incluant les ponctuations et les mots composés comme TreeTagger les a étiquetés. Nombre d'occurrences sans ponctuations, ni chiffres, et dont les mots composés sont considérés comme des occurrences séparées. Nombre de formes (*types*) de mots. Nombre de formes de CMS. Les dernières colonnes donnent le pourcentage de propositions pour chaque type de discours (nar = narratif, arg = argumentatif, expl = explicatif, descr = descriptif, dial = dialogal et inj = injonctif).

4.1.3 Prétraitement

Par l'intermédiaire d'un programme écrit en Perl, chacun des quatre textes est transformé en trois tables de contingence, $N = (n_{ik})$, comptant, pour chaque proposition i délimitée par l'annotateur, le nombre n_{ik} de chaque uni-, bi- ou tri-gramme de CMS de type k. De surcroît, le type de discours de chaque proposition est extrait des textes annotés et ajouté comme une colonne supplémentaire.

Dans le détail, le texte est d'abord étiqueté par TreeTagger (Schmid, 1994) à l'aide du module Perl `Lingua::TreeTagger` [2]. Ce dernier permet d'obtenir, pour chaque mot ou balise XML rencontrée dans le texte, respectivement la CMS du mot ou la balise XML originale, toutes regroupées sous le terme d'étiquette dans la suite de ce paragraphe. Ensuite, pour chaque étiquette, on vérifie si elle correspond, ou non, à une balise XML. S'il ne s'agit pas d'une balise XML, l'étiquette correspond à une CMS et elle est stockée dans un tableau temporaire. Sinon, l'étiquette correspond à une balise délimitant un type de discours ou une proposition. Étant donné que seules les « feuilles » de la structure hiérarchique de l'annotation sont considérées (cf. section 4.1.1.7), il n'est pas nécessaire de conserver l'entièreté de la structure de l'annotation en types de discours. Ainsi, les types de discours peuvent être sauvegardés sous la forme d'une pile (*stack*) [3]. Dès lors,

2. `http://search.cpan.org/dist/Lingua-TreeTagger`
3. Pour rappel, en informatique, une pile est une structure de données basée sur le principe de « dernier arrivé, premier sorti » (*LIFO : « Last-In-First-Out »*).

— s'il s'agit d'une balise ouvrante délimitant un type de discours ( ), alors le type de discours est conservé dans la pile,
— s'il s'agit d'une balise ouvrante délimitant une proposition (<e>), alors le type de discours conservé dans la pile est attribué à cette proposition,
— s'il s'agit d'une balise fermante délimitant un type de discours ( ), alors de dernier type de discours entré dans la pile est retiré, et
— s'il s'agit d'une balise fermante délimitant une proposition (</e>), alors les n-grammes de CMS contenus dans le tableau temporaire sont comptés et attribués à cette proposition.

Cette procédure est exécutée trois fois pour chaque texte, soit une fois pour chaque longueur de n-gramme de CMS.

4.1.4 Analyse préliminaire

Avant de passer à la classification non supervisée, il convient de s'assurer que la représentation des données choisie est pertinente. Pour ce faire, il va être déterminé, d'une part, s'il existe un lien général entre les types de discours et les CMS, et d'autre part, si certaines CMS sont spécifiquement présentes dans chacun des types de discours. Ceci sera fait pour les quatre textes regroupés, puis pour chaque texte pris séparément.

En premier lieu, des tables de contingence spécifiant le nombre de fois que chaque CMS apparaît dans un des six types de discours sont construites pour chaque texte, puis pour les quatre textes réunis. Ceci est fait en agrégeant les propositions appartenant à un même type de discours dans les tables de contingence propositions - unigrammes de CMS préalablement construites (cf. section 4.1.3). Les cinq tables ainsi créées sont exposées dans l'annexe B, section B.1.

Pour vérifier s'il existe un lien général entre les CMS et les types de discours, un test du khi2 (1.1) est effectué sur chacune de ces cinq tables, conduisant aux résultats suivants :

Texte	ddl	khi2	valeur p
4 textes réunis	150	1100.18	$< 2.2 \times 10^{-16}$
L'Orient	130	304.15	5.46×10^{-16}
Le Voleur	140	587.22	$< 2.2 \times 10^{-16}$
Un Fou ?	135	671.01	$< 2.2 \times 10^{-16}$
Un Fou	140	586.63	$< 2.2 \times 10^{-16}$

Les CMS et les types de discours sont donc significativement dépendants, que ce soit pour les quatre textes réunis ou pour chaque texte étudié séparément. Ainsi, le choix d'utiliser les CMS comme caractéristiques semble pertinent.

Ensuite, pour savoir s'il existe une attraction mutuelle entre certaines CMS et certains types de discours, on calcule le quotient d'indépendance (1.2) et le khi2 ponctuel (1.3) sur les cinq tables pour chaque paire de CMS - type de discours. Les résultats pour le khi2 ponctuel sont présentés dans l'annexe B, section B.2. Quant aux quotients d'indépendance, les résultats pour les quatre textes réunis sont exposés dans la table 4.2 ; et ceux pour chacun des quatre textes, dans la table 4.3. De plus, dans ces tables, les valeurs significatives selon le khi2 ponctuel pour $\alpha = 0.1\%$ sont marquées par une étoile [4]. Une définition de toutes les abréviations de CMS utilisées dans ces tables, ainsi que sur les figures de la section 4.2, se trouve dans l'annexe B.

En considérant les quatre textes réunis (table 4.2), on observe qu'il existe une attraction mutuelle entre des CMS et des types de discours correspondant aux marques linguistiques décrites dans la section 4.1.1. Par exemple, la ponctuation de citations (PUN:cit) est, comme on pouvait s'y attendre, la CMS en attraction la plus forte avec le type dialogal. De plus, ces deux modalités sont significativement dépendantes selon le khi2 ponctuel. On remarque

4. Naturellement, un traitement inférentiel rigoureux devrait tenir compte du problème des *comparaisons multiples* non poursuivi ici.

	nar	arg	expl	descr	dial	inj
ABR	***2.56***	0.00	0.00	0.00	0.00	0.00
ADJ	0.82	0.87	1.06	**1.57***	1.02	0.86
ADV	0.93	1.12	1.13	0.80	1.04	**1.37**
DET:ART	0.99	**1.18**	0.77	1.16	0.90	0.88
DET:POS	**1.21**	0.95	0.95	0.84	0.79	0.70
INT	1.12	1.07	1.16	0.00	**1.56**	1.04
KON	0.94	1.25	**1.30**	0.80	0.93	0.78
NAM	1.21	0.34	0.74	**1.52**	0.70	1.11
NOM	0.95	1.15	0.87	**1.15**	0.90	1.04
NUM	1.05	1.06	0.94	**1.50**	0.71	0.00
PRO	***2.56***	0.00	0.00	0.00	0.00	0.00
PRO:DEM	0.64*	1.49	**1.52**	1.03	1.18	0.58
PRO:IND	0.73	1.53	**1.53**	1.03	1.11	0.00
PRO:PER	**1.24***	0.86	1.06	0.60*	1.02	0.64
PRO:REL	0.78	1.12	**1.28**	1.18	1.03	1.04
PRP	1.01	1.01	1.00	**1.15**	0.87	0.79
PRP:det	0.63*	1.18	0.68	1.28	1.51	**1.83**
PUN	1.03	0.96	0.82	1.12	0.86	**1.27**
PUN:cit	0.00*	0.20*	0.34	0.47	***4.77****	4.24*
SENT	1.02	0.87	1.09	0.83	1.14	**1.16**
VER:cond	1.10	***1.91***	0.88	0.24	1.11	0.00
VER:futu	0.43	0.37	0.46	0.37	**4.55***	1.44
VER:impe	0.00	0.00	0.00	0.00	0.00	***17.33****
VER:impf	1.22	0.50	0.43	***2.27****	0.36	0.10
VER:infi	0.98	0.94	**1.52**	0.93	0.96	0.50
VER:pper	**1.21**	0.75	0.97	1.07	0.82	0.51
VER:ppre	**1.61***	0.25	0.61	0.85	0.72	0.64
VER:pres	0.68*	1.20	1.46*	0.71	1.18	**2.05***
VER:simp	**2.29***	0.25*	0.17*	0.28*	0.04*	0.00*
VER:subi	0.64	0.00	***5.50****	0.55	0.00	0.00
VER:subp	0.26	1.34	1.65	0.00	**3.12**	1.73

TABLE 4.2 – Quotients d'indépendance entre les CMS et les types de discours pour les quatre textes réunis. Les valeurs en gras désignent le quotient d'indépendance maximum pour chaque CMS ; celles en italique, le quotient d'indépendance maximum pour chaque type de discours ; et celles suivies d'une étoile, les valeurs significatives à $\alpha = 0.1\%$ selon le khi2 ponctuel.

aussi qu'il existe une attraction mutuelle entre les interjections (INT) et le type dialogal, ce qui semble cohérent, bien que ces CMS n'aient pas été considérées comme des marques linguistiques. Cependant, la dépendance n'est ici pas significative.

Concernant le narratif, il existe une attraction mutuelle entre ce dernier et le passé simple (VER:simp) et cette dépendance est significative, ce qui correspond aux marques linguistiques retenues par l'expert humain. Il existe aussi une répulsion mutuelle entre le type narratif et la ponctuation de citation, et il s'agit d'une dépendance significative. En effet, en observant les effectifs des CMS dans les types de discours (table B.1), on remarque que c'est le seul type de discours pour lequel la ponctuation de citation n'apparaît jamais. Plus surprenant, pour ce type de discours, l'attraction mutuelle la plus importante a lieu avec les abréviations (ABR), d'une part, et les pronoms (PRO), d'autre part. Néanmoins, aucune de ces deux CMS n'est significativement dépendante de ce type de discours. En se référant à nouveau à la table B.1, on remarque que bien que ces deux CMS n'apparaissent que dans le type narratif, elles sont rares, soit une apparition pour les abréviations et deux, pour les pronoms. Ces deux pronoms apparaissent dans le texte « Un Fou » (cf. table B.2) et correspondent à des pronoms interrogatifs.

<e>		
Qui	PRO	qui
le	PRO:PER	le
croirait	VER:cond	croire
?	SENT	?
</e>		

FIGURE 4.2 – Extrait étiqueté par TreeTagger d'« Un Fou », correspondant à ligne 463 de l'annexe A.4.

Un exemple est présenté dans la figure 4.2.

Pour l'injonctif, l'attraction mutuelle se produit, comme attendu d'après les marques linguistiques, avec l'impératif (VER:impe) et cette dépendance est significative, bien que cette CMS n'apparaisse que dans ce type de discours (table B.1). Concernant le descriptif, il est, sans surprise, en attraction mutuelle avec les adjectifs (ADJ) et les verbes à l'imparfait (VER:impf).

Finalement, concernant l'argumentatif et l'explicatif, les résultats sont moins évidents à interpréter. On peut simplement constater que l'argumentatif possède l'attraction mutuelle la plus importante avec les verbes au conditionnel (VER:cond), sans que cette dépendance ne soit significative. Pour l'explicatif, l'attraction mutuelle la plus élevée est avec les verbes à l'imparfait du subjonctif (VER:subi) et il s'agit d'une dépendance significative. Néanmoins, uniquement huit occurrences de cette CMS apparaissent dans l'ensemble des quatre textes (table B.1).

Au vu de ces premiers résultats, il est clair que le khi2 ponctuel et le quotient d'indépendance donnent des informations différentes, mais complémentaires, qu'il pourrait être avantageux de combiner, en particulier si l'on voulait faire une sélection de caractéristiques (*feature selection*). À titre d'exemple, Li *et al.* (2008) proposent une telle combinaison qui ne sera pas utilisée ici.

Concernant les quatre textes étudiés séparément (table 4.3), une première constatation est que malgré la présence de certains points communs, il existe des différences entre ces quatre textes. En effet, on retrouve que le narratif est en attraction mutuelle avec le passé simple pour les quatre textes, et le descriptif avec l'imparfait, même s'il ne s'agit pas systématiquement de dépendances significatives. Cependant, les autres observations faites sur les quatre textes réunis sont moins évidentes ici.

Par exemple, les adjectifs sont clairement en attraction mutuelle avec le descriptif pour les textes « Un Fou ? » et « Un Fou », mais cette attraction est moins évidente pour le texte « Le Voleur » et, inversement, pour le texte « L'Orient », il y a répulsion mutuelle entre les adjectifs et le descriptif. Néanmoins, dans ces deux derniers textes, on remarque une attraction mutuelle importante entre les adjectifs et l'injonctif.

Aussi, on constate que les interjections sont en attraction mutuelle avec le dialogal pour les textes « L'Orient », « Le Voleur » et « Un Fou », mais avec l'injonctif pour le texte « Un Fou ? ». Quant à la ponctuation de citation, elle est en attraction mutuelle avec le dialogal et l'injonctif pour « L'Orient », mais cette attraction est plus élevée pour l'injonctif, alors qu'il existe une dépendance significative selon le khi2 avec le dialogal. On retrouve une situation analogue pour « Le Voleur », si ce n'est que dans ce texte, la dépendance entre la ponctuation de citation est significative pour les deux types de discours. Cette même CMS est clairement en attraction mutuelle avec l'injonctif pour « Un Fou ? » et avec le dialogal pour « Un Fou ». Il semble donc que ces deux types de discours se confondent, ce qui peut certainement s'expliquer par le fait que dans notre corpus, l'injonctif est, comme déjà mentionné, systématiquement inclus dans le type dialogal.

Finalement, on peut remarquer que les conjonctions (KON) sont en attraction mutuelle avec l'argumentatif pour « L'Orient » et « Un Fou », alors qu'elles sont en attraction mutuelle avec l'explicatif pour « Le Voleur » et « Un Fou ? ».

Ces différences entre les quatre textes peuvent probablement s'expliquer par le fait que, bien que les quatre textes soient des contes du même auteur, leur forme varie. Par exemple, comme déjà mentionné plus haut, « Un Fou » est écrit sous la forme d'un journal intime et comporte

	« L'Orient »						« Le Voleur »					
	nar	arg	expl	descr	dial	inj	nar	arg	expl	descr	dial	inj
ABR												
ADJ	0.93	0.34	1.00	0.93	1.15	**1.67**	1.00	1.21	0.27	1.32	0.73	**2.08**
ADV	**1.56**	1.02	1.09	0.64	0.79	1.01	0.86	**2.11**	0.56	1.20	1.17	0.97
DET:ART	0.84	1.18	0.95	**1.19**	1.05	0.00	1.09	1.01	0.34	**1.23**	0.70	0.58
DET:POS	0.55	2.31	0.86	0.40	1.24	***5.67****	1.03	1.35	***2.40***	0.45	0.93	0.00
INT	0.00	0.00	0.00	0.00	***3.09***	0.00	0.00	0.00	0.00	0.00	***7.21****	0.00
KON	1.11	**1.37**	0.92	0.50	1.19	1.35	0.96	1.30	**2.33**	1.05	0.63	0.00
NAM	0.77	0.00	**1.79**	1.25	0.77	0.00	1.00	0.99	0.89	**1.83**	0.34	0.00
NOM	0.79	1.24	0.89	**1.24**	1.03	0.80	**1.06**	0.91	0.81	0.94	0.91	0.87
NUM	1.43	***2.40***	0.34	1.25	0.77	0.00	1.20	0.00	0.00	**2.00**	0.00	0.00
PRO												
PRO:DEM	0.63	0.00	1.67	***1.73***	0.64	0.00	0.58	1.81	0.00	1.52	**2.51**	0.00
PRO:IND	0.26	0.00	1.49	0.84	**1.55**	0.00	1.01	**4.17**	1.86	0.70	0.00	0.00
PRO:PER	**1.75***	0.94	0.89	0.61	0.83	0.58	1.06	1.08	**1.23**	0.66	1.05	0.45
PRO:REL	0.51	0.00	**1.49**	1.39	1.03	0.00	1.10	**2.41**	0.72	0.81	0.55	0.00
PRP	0.90	**1.17**	1.15	0.86	1.06	0.86	1.09	0.72	**1.28**	1.05	0.54	1.10
PRP:det	0.43	1.09	0.61	**1.51**	1.34	0.00	0.84	0.00	0.72	1.08	**1.94**	1.84
PUN	0.96	0.98	0.79	1.24	0.96	**1.51**	**1.16**	0.53	1.05	1.01	0.48	0.74
PUN:cit	0.00	1.83	0.00	0.48	2.21*	**4.50**	0.00*	0.43	1.52	0.14	5.15*	**7.83***
SENT	1.15	0.87	1.16	1.03	0.77	**1.61**	0.84	0.82	0.73	0.96	1.77	**1.88**
VER:cond	1.53	0.00	***3.59***	0.00	0.00	0.00	0.56	0.00	0.00	2.33	**2.40**	0.00
VER:futu	0.00	0.92	0.26	0.24	**2.65***	0.00	0.00	**6.95**	0.00	0.00	4.81	0.00
VER:impe							0.00	0.00	0.00	0.00	0.00	***47.95****
VER:impf	1.53	0.00	0.90	**1.67**	0.52	0.00	0.73	0.87	1.16	***2.92****	0.30	0.00
VER:infi	0.39	1.10	**2.61***	0.72	0.71	0.00	0.77	1.25	1.49	1.12	**1.73**	0.00
VER:pper	1.00	1.20	0.84	0.78	**1.26**	0.00	1.01	0.39	**1.76**	1.45	0.54	0.00
VER:ppre	0.00	0.00	1.54	0.72	**1.77**	0.00	**1.58**	0.00	0.00	0.41	0.00	0.00
VER:pres	1.05	0.95	1.19	1.31	0.58*	**2.32**	0.26*	1.81	1.62	0.61	3.45*	**5.21***
VER:simp	***4.59****	0.00	0.00	0.00	0.00	0.00	**1.51***	0.58	0.70	0.26	0.00*	0.00
VER:subi							***1.68***	0.00	0.00	0.00	0.00	0.00
VER:subp	0.00	0.00	0.00	0.00	***3.09***	0.00	0.00	***20.85****	0.00	0.00	0.00	0.00

	« Un Fou ? »						« Un Fou »					
	nar	arg	expl	descr	dial	inj	nar	arg	expl	descr	dial	inj
ABR							***2.59***	0.00	0.00	0.00	0.00	0.00
ADJ	0.93	1.11	1.08	**1.90***	0.39	0.18	0.68	0.77	1.06	**2.03***	1.37	0.93
ADV	0.77	1.13	**1.43**	0.60	1.31	0.95	0.94	0.92	0.86	0.86	**1.74**	1.56
DET:ART	1.02	**1.36**	0.45	1.35	0.77	0.78	0.90	**1.17**	1.11	0.96	0.50	1.04
DET:POS	**1.70***	0.93	0.69	0.38	0.00	0.95	1.09	0.70	0.90	**1.75**	1.36	0.20
INT	0.23	1.22	1.09	0.00	2.06	**5.56***	1.65	0.68	1.53	0.00	**3.72**	0.00
KON	0.92	1.21	**1.51**	0.69	0.83	0.17	0.89	**1.22**	0.91	0.99	0.50	1.09
NAM	1.55	0.61	0.00	0.56	1.03	**2.78**	0.32	0.00	1.26	2.27	0.00	***3.33***
NOM	1.06	**1.23**	0.76	1.18	0.56	1.04	0.86	**1.16**	0.98	1.12	0.64	1.06
NUM	1.38	1.46	0.00	**1.79**	0.00	0.00	0.71	0.82	**3.68**	1.10	0.00	0.00
PRO							***2.59***	0.00	0.00	0.00	0.00	0.00
PRO:DEM	0.20*	**1.76**	1.51	1.04	1.56	0.64	1.25	1.01	**1.40**	0.52	0.00	0.61
PRO:IND	1.01	1.03	**1.71**	0.94	0.43	0.00	0.68	***1.96***	0.88	1.32	0.00	0.00
PRO:PER	1.12	0.75	1.19	0.75	**1.21**	0.60	**1.36***	0.94	0.89	0.46*	1.12	0.75
PRO:REL	0.74	1.31	1.27	0.44	**1.61**	0.54	0.77	0.99	1.11	**1.42**	0.00	1.25
PRP	**1.22**	0.93	0.94	0.93	0.78	0.50	0.81	1.15	0.79	**1.54***	0.24	0.87
PRP:det	0.92	**1.44**	0.72	1.33	0.92	0.00	0.61	1.34	0.63	0.85	0.00	**2.36***
PUN	1.01	1.11	0.81	0.96	0.88	**1.50**	0.90	0.87	0.79	1.25	**1.96**	1.26
PUN:cit	0.00*	0.00	0.00	1.49	2.06	***12.98****	0.00	0.00	0.00	0.00	***61.30****	0.00
SENT	0.72	0.88	1.00	1.13	1.60	**1.93**	1.29*	0.76	1.30	0.55	**2.28**	0.70
VER:cond	0.00	***2.78***	0.93	0.00	2.36	0.00	1.73	1.20	0.00	0.00	**4.09**	0.00
VER:futu	0.00	0.00	3.26	0.00	**4.13**	0.00	1.56	0.00	0.00	0.61	0.00	**2.67**
VER:impe												
VER:impf	1.38	0.30	0.39	***2.85****	0.50	0.00	1.30	0.59	0.19	**1.91**	1.14	0.16
VER:infi	1.34	0.30	1.22	0.78	**1.44**	0.00	0.95	1.17	**1.35**	1.03	0.00	0.63
VER:pper	0.90	0.11	1.51	**2.08**	1.54	0.00	**1.40***	0.94	0.73	0.71	0.00	0.56
VER:ppre	**2.10***	0.23	0.62	0.43	0.39	0.00	0.58	0.50	0.00	***2.69****	0.00	1.97
VER:pres	0.40*	1.23	1.42	0.30	1.97*	**2.50***	1.02	1.00	**1.46**	0.46*	0.28	1.44
VER:simp	***2.33****	0.33	0.00*	0.61	0.14	0.00	***2.59***	0.00	0.00	0.00	0.00	0.00
VER:subi	0.50	0.00	***4.74****	0.81	0.00	0.00						
VER:subp	0.00	0.00	0.00	0.00	**8.25***	0.00	0.52	0.90	***4.05***	0.00	0.00	1.78

TABLE 4.3 – Quotients d'indépendance entre les CMS et les types de discours pour chaque texte considéré séparément. Les valeurs en gras désignent le quotient d'indépendance maximum pour chaque CMS ; celles en italique, le quotient d'indépendance maximum pour chaque type de discours ; et celles suivies d'une étoile, les valeurs significatives à $\alpha = 0.1\%$ selon le khi2 ponctuel.

donc un grand nombre de verbes au présent, aussi dans les types narratifs et descriptifs. Le texte « L'Orient » contient aussi plusieurs longs monologues écrits au présent. Au vu de ces

différences, les quatre textes seront systématiquement étudiés séparément et non plus réunis dans la suite de ce chapitre.

4.2 Visualisation

4.2.1 Propositions et CMS

Les données étant représentées sous la forme de tables de contingence croisant les propositions et les n-grammes de CMS (ici seuls les unigrammes de CMS sont traités), il est possible d'y appliquer l'analyse factorielle des correspondances (AFC) (cf. section 1.4). Pour ce faire, il est possible de calculer les dissimilarités du khi2, puis d'effectuer un MDS sur ces dernières. Il est aussi possible d'utiliser directement un logiciel dédié qui se base sur la décomposition spectrale de la matrice des variances-covariances. La seconde solution a été adoptée ici, avec le package « ca » de R (Nenadic et Greenacre, 2007). Plus précisément, afin d'obtenir des résultats identiques à ceux qui seraient produits par le MDS, on utilise, pour créer les biplots, les coordonnées dites *principales*, extraites grâce à la fonction « summary », au lieu des coordonnées dites *standardisées* qui sont produites par défaut (Nenadic et Greenacre, 2007). Les résultats ainsi obtenus sont présentés dans les figures 4.3 à 4.6.

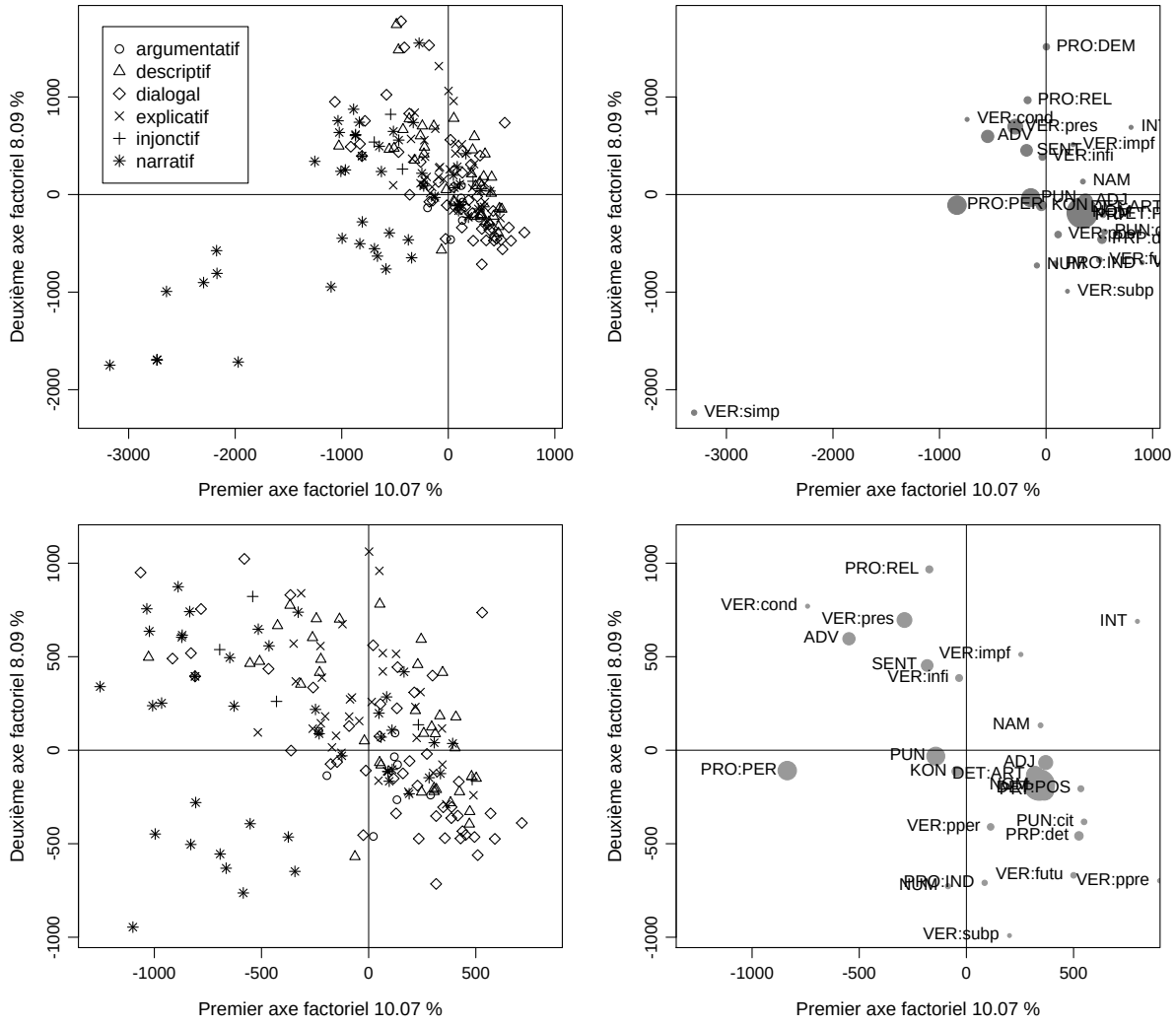


FIGURE 4.3 – AFC sur « L'Orient ». Coordonnées factorielles des propositions (en haut à gauche) et des unigrammes de CMS (en haut à droite). En bas : zoom sur le centre des figures du haut.

Un premier constat est que l'inertie expliquée par les deux premiers facteurs est assez faible pour les quatre textes, systématiquement inférieure à 20 % et il n'est donc pas évident d'interpréter ces biplots. Il est tout de moins possible de remarquer quelques tendances.

La figure 4.3 montre le résultat de l'AFC sur le texte « L'Orient ». Bien qu'il soit difficile de distinguer clairement des groupes, la vue d'ensemble (figures en haut) montre que les deux axes différencient principalement le passé simple des autres CMS (figure droite). Le passé simple, qui marque le narratif (cf. sections 4.1.1.1 et 4.1.4), est en attraction mutuelle avec les propositions narratives (figure gauche, quadrant sud-ouest). Concernant, les AFC agrandies (figures en bas), il est difficile de distinguer des groupes. Néanmoins, on remarque une concentration plus élevée de propositions de type dialogal dans le quadrant sud-est (figure gauche) qui sont certainement en attraction mutuelle avec la ponctuation de citation, les verbes au futur (VER:futu), les verbes au participe présent (VER:ppre) et les verbes au subjonctif présent (VER:subp) (cf. table 4.3).

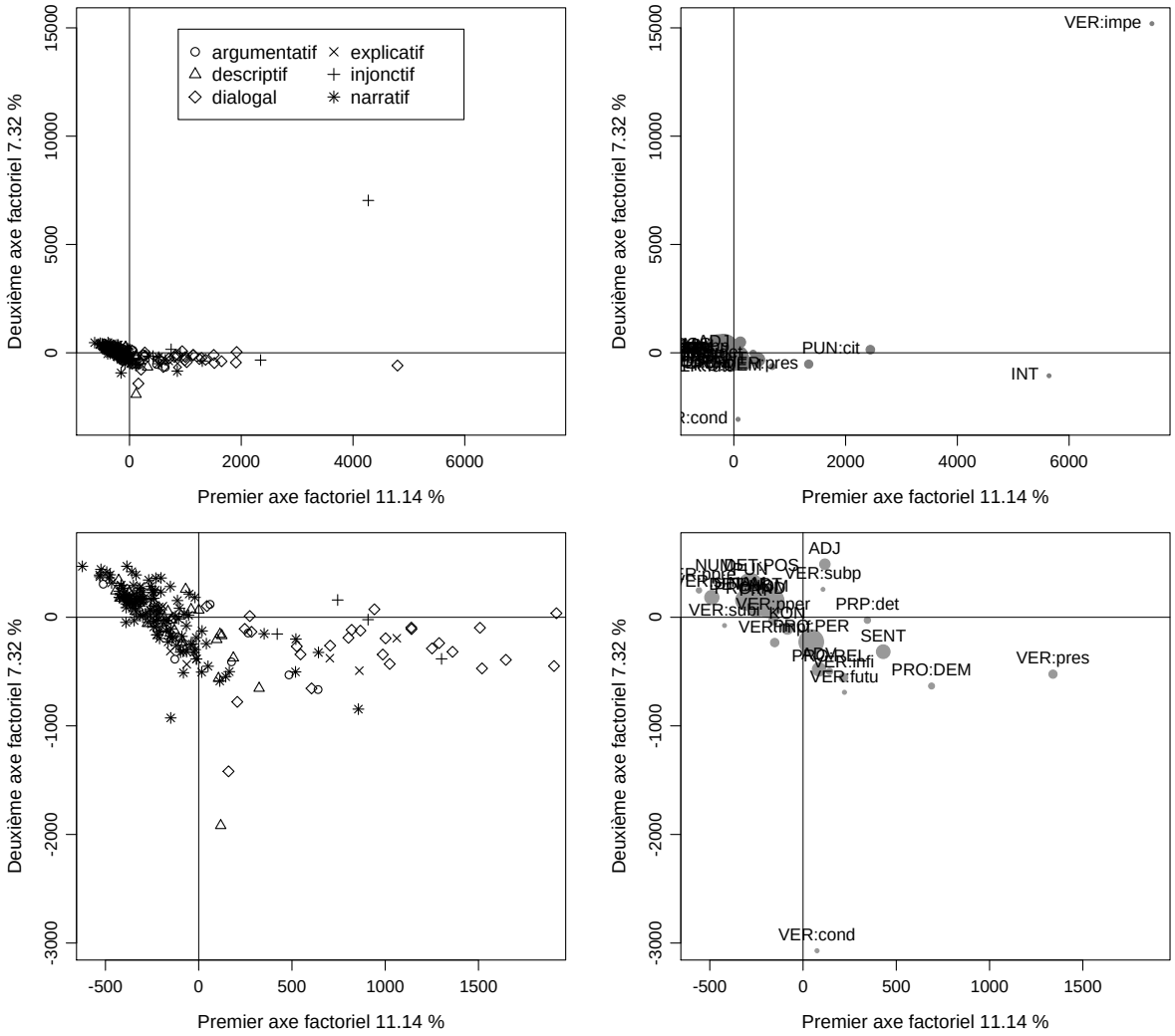


FIGURE 4.4 – AFC sur « Le Voleur ». Coordonnées factorielles des propositions (en haut à gauche) et des unigrammes de CMS (en haut à droite). En bas : zoom sur le centre des figures du haut.

Concernant le texte « Le Voleur » (figure 4.4), on observe, pour les CMS (figure en haut, à droite), que le premier axe différencie les verbes à l'impératif, marque linguistique de l'injonctif (cf. section 4.1.1.6) et les interjections, souvent présentes dans le type dialogal (cf. section 4.1.4), des autres CMS. Quant au second axe, il différencie à nouveau les verbes à l'impératif des autres CMS. Cependant, il est difficile de distinguer des groupes de types de discours (figure en haut, à

gauche). En observant le figure agrandie sur les propositions (en bas, à gauche), le premier axe factoriel différencie les propositions dialogales et injonctives (à l'est) des propositions narratives (à l'ouest). Ce contraste est certainement en relation avec la présence des interjections et de la ponctuation de citation dans la zone est (figure en haut à droite). À nouveau, on constate que le type injonctif et le type dialogal sont difficiles à distinguer dans notre corpus.

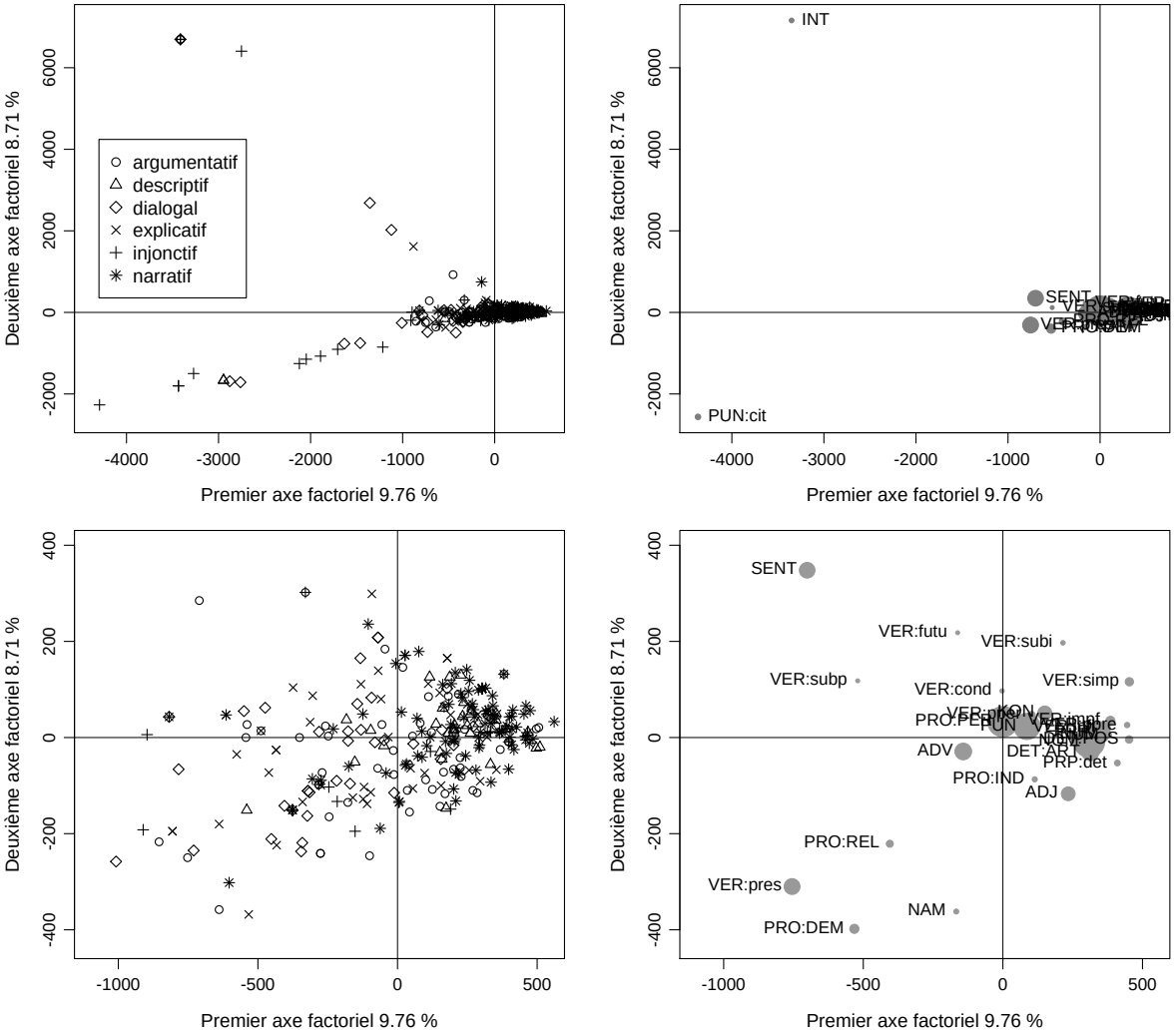


FIGURE 4.5 – AFC sur « Un Fou ? ». Coordonnées factorielles des propositions (en haut à gauche) et des unigrammes de CMS (en haut à droite). En bas : zoom sur le centre des figures du haut.

Sur la figure 4.5, pour le texte « Un Fou ? », on observe sur la vue d'ensemble pour les propositions (figure en haut à gauche) un détachement de propositions injonctives dans le quadrant sud-ouest. On constate aussi la présence de ponctuation de citation dans ce même quadrant (figure en haut à droite), qui est en attraction mutuelle avec ce type dans ce texte (cf. table 4.3). Sur la figure agrandie pour les propositions (figure en bas à gauche), les propositions narratives se concentrent dans la zone est du graphique et sont certainement en attraction mutuelle avec les verbes au passé simple dans le quadrant nord-est (figure en bas à droite).

Concernant le texte « Un Fou » (figure 4.6), il est nettement plus difficile de distinguer les six types de discours que pour les autres textes. On peut tout de même remarquer (figure en haut à droite) que le premier axe différencie les interjections des autres CMS ; et le second axe, les pronoms (PRO), des autres CMS. Aussi, quelques CMS, soit les chiffres (NUM), la ponctuation marquant la fin d'une phrase (SENT), la ponctuation de citation, les abréviations et les verbes

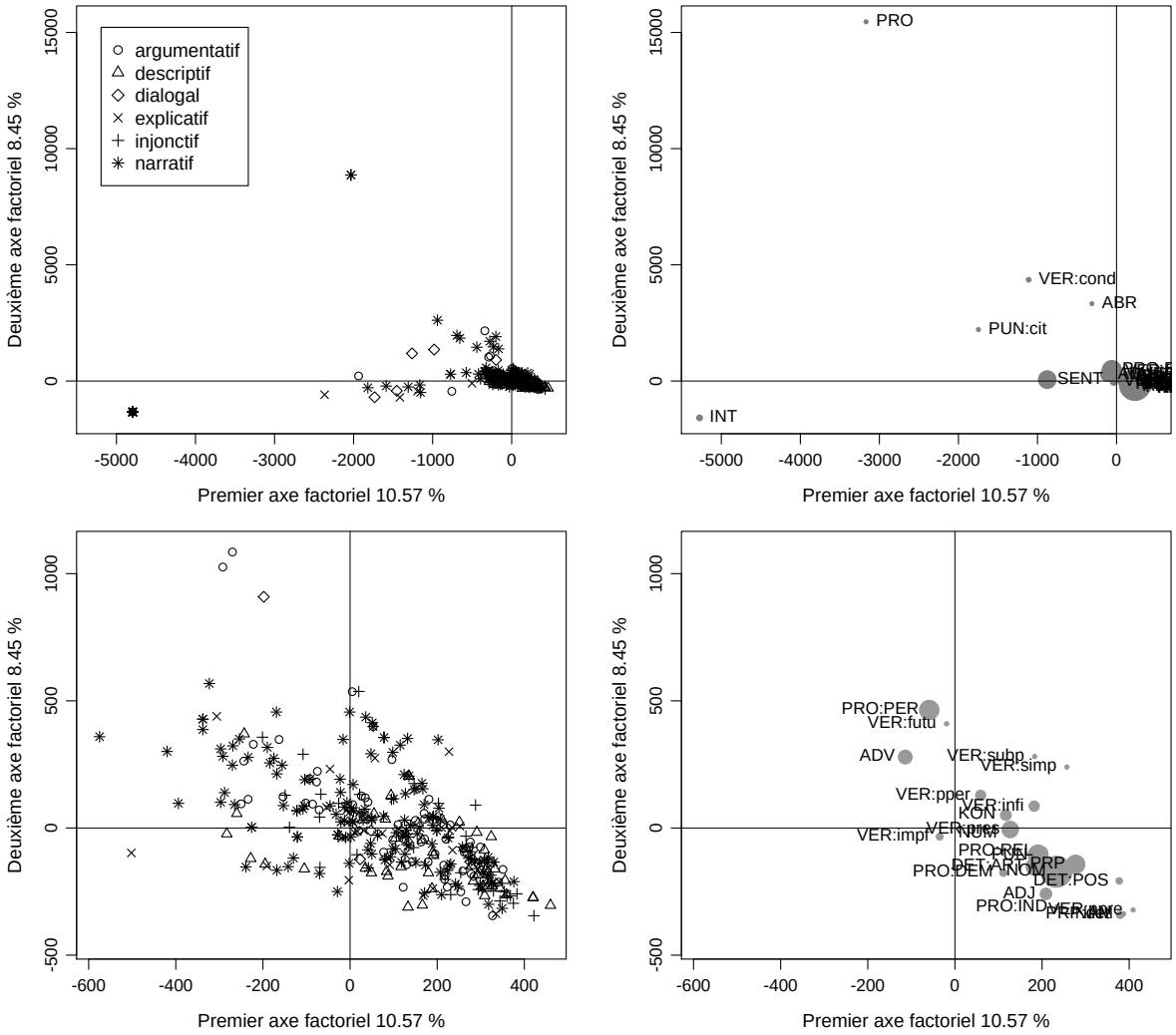


FIGURE 4.6 – AFC sur « Un Fou ». Coordonnées factorielles des propositions (en haut à gauche) et des unigrammes de CMS (en haut à droite). En bas : zoom sur le centre des figures du haut.

au conditionnel, se distinguent du rassemblement compact des autres CMS au centre (figure en bas à droite). Par conséquent (figure en haut à gauche), quelques propositions se détachent du noyau central, mais n'étant pas toutes du même type, il est difficile d'en proposer une interprétation.

En conclusion, cette section a permis de visualiser les observations déjà décrites numériquement dans la section 4.1.4. Ainsi, on constate à nouveau des différences pour ces quatre textes et on remarque, en utilisant uniquement les deux premières dimensions, qu'il n'est pas simple de distinguer les six types de discours et que cette difficulté varie selon les textes, mais aussi selon les types de discours.

Naturellement, il serait aussi possible, sur ces figures, de visualiser les résultats obtenus avec les classifications automatiques présentées dans la section 4.3. Trois exemples de classification non supervisée pour le texte « Un Fou ? » sont présentés dans les articles suivant : Cocco *et al.* (2011) avec l'algorithme K-means flou et 8 groupes après agrégation, et Cocco (2012*a*) avec l'algorithme K-means, dur et flou, et 6 groupes.

4.2.2 Types de discours et CMS avec *bootstrap*

Comme il a été possible de représenter les tables de contingence croisant les CMS et les propositions grâce à l'AFC, il est aussi possible de le faire avec les tables de contingence croisant

les CMS et les types de discours, constitués de groupes de propositions, présentées dans l'annexe B.1 et analysées dans la section 4.1.4. Les graphiques présentés dans cette section ont été créés avec le logiciel Dtm-Vic [5] pour pouvoir valider les résultats par la technique du *bootstrap* qui y est intégrée. Le *bootstrap* est une méthode empirique de validation d'un paramètre (ou estimateur) basée sur le rééchantillonnage (voir par exemple Efron et Tibshirani, 1993). Le principe consiste à créer plusieurs nouveaux échantillons de même taille que l'échantillon de départ par un tirage avec remise dans cet échantillon de départ, puis de calculer le paramètre sur ces nouveaux échantillons afin de simuler sa distribution. Il est alors possible de déterminer l'intervalle de confiance dudit paramètre (voir par exemple Saporta, 2006, section 15.3.1 ; Lebart *et al.*, 1995, section 4.2.2). En particulier, pour l'AFC, plusieurs tables de contingence sont créées en tirant $n_{\bullet\bullet}$ observations de la table de contingence initiale avec remise. Ceci est équivalent à faire un tirage selon une loi multinomiale de probabilité $p_{ij} = n_{ij}/n_{\bullet\bullet}$ (Lebart *et al.*, 1995, section 4.2.3.a). Ensuite, pour construire des intervalles de confiance, qui seront ici des ellipses de confiance, il existe deux possibilités : projeter les modalités des nouvelles tables en tant que variables supplémentaires sur l'AFC produite avec la table initiale (*bootstrap* partiel) ; ou refaire une AFC pour chaque nouvelle table (*bootstrap* total) (voir par exemple Lebart, 2007; Dupuis et Lebart, 2009). La première solution a été adoptée ici en utilisant 30 réplications de la table d'origine. Précisons encore que l'on n'a pas pratiqué le *bootstrap* sur les tables de contingence propositions - CMS utilisées dans la section précédente, à cause du caractère creux de ces dernières, susceptible de ne pas permettre le rééchantillonnage de certaines modalités qui seraient alors supprimées avant l'application du *bootstrap*.

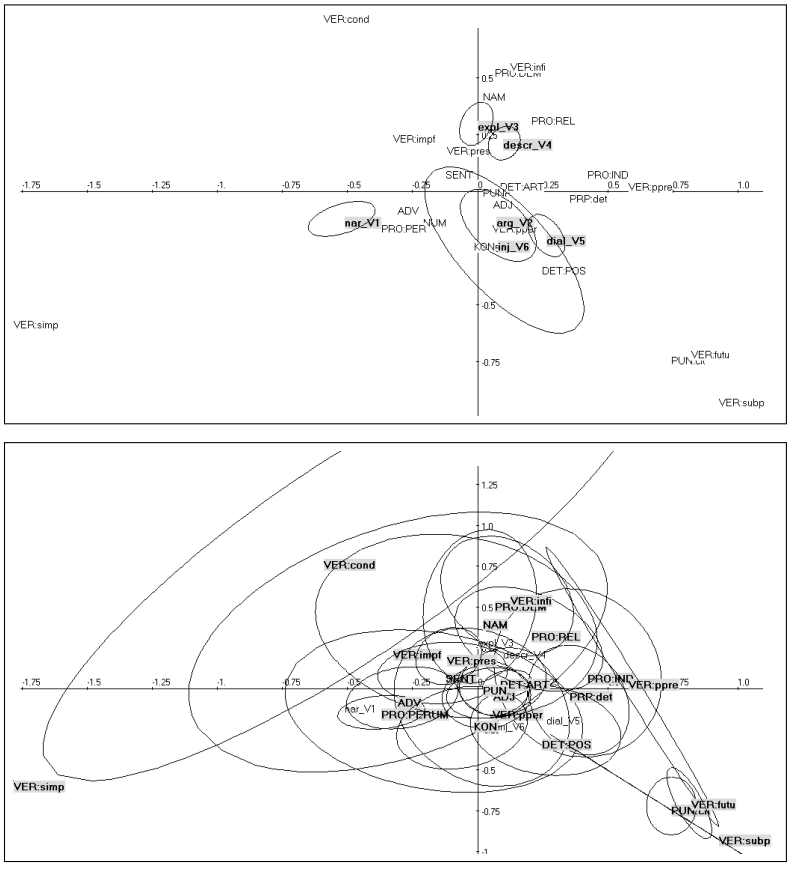


FIGURE 4.7 – AFC sur « L'Orient » entre les CMS et les types de discours avec validation par *bootstrap*. Inertie expliquée par le premier axe factoriel : 46.73% ; et par le deuxième : 24.99%.

Les figures 4.7 à 4.10 présentent les résultats pour les quatre contes étudiés. À nouveau, les

5. Ce logiciel peut être librement téléchargé sur le site de Ludovic Lebart : `http://www.dtmvic.com/`.

visualisations obtenues sont différentes pour chacun des textes. Aussi, étant donné que ces tables de contingence comportent moins de modalités que celles utilisées dans la section 4.2.1, l'inertie expliquée par les deux premiers facteurs est plus élevée, soit systématiquement supérieure à 70%.

Le résultat obtenu pour le texte « L'Orient » est présenté dans la figure 4.7. Pour commencer, on constate que les CMS ont des positions relativement similaires à celles de la figure 4.3. Rappelons que les tables utilisées dans cette section sont des agrégations des tables utilisées dans la section 4.2.1 par rapport aux types de discours. En d'autres termes, les types de discours représentés sur les figures de cette section sont les moyennes des propositions appartenant à ces types de discours.

Concernant les types de discours de ce texte (figure 4.7, haut), la validation nous donne des informations supplémentaires à celles que l'on aurait obtenues par une AFC simple d'une part ; et à celles obtenues dans la section 4.1.4 d'autre part. Par exemple, on constate que les types de discours argumentatif et injonctif ne sont pas significativement différents de l'origine, associée au profil du « type de discours moyen ». Cela signifie que ces types de discours et les CMS ne sont pas significativement dépendants. À l'inverse, les types narratif, explicatif, dialogal et descriptif sont significativement différents de l'origine, et donc significativement dépendants des CMS. De plus, on observe l'absence d'intersection entre les ellipses de confiance de ces types qui sont donc bien différenciés selon les CMS qu'ils contiennent.

Les ellipses de confiance obtenues pour les CMS sont plus difficiles à distinguer, car elles sont nombreuses (figure 4.7, bas). On peut néanmoins remarquer qu'elles ne sont jamais isolées et qu'il existe donc une continuité entre elles. On constate aussi, par exemple, que la ponctuation de citation et les verbes au futur sont significativement différents de l'origine.

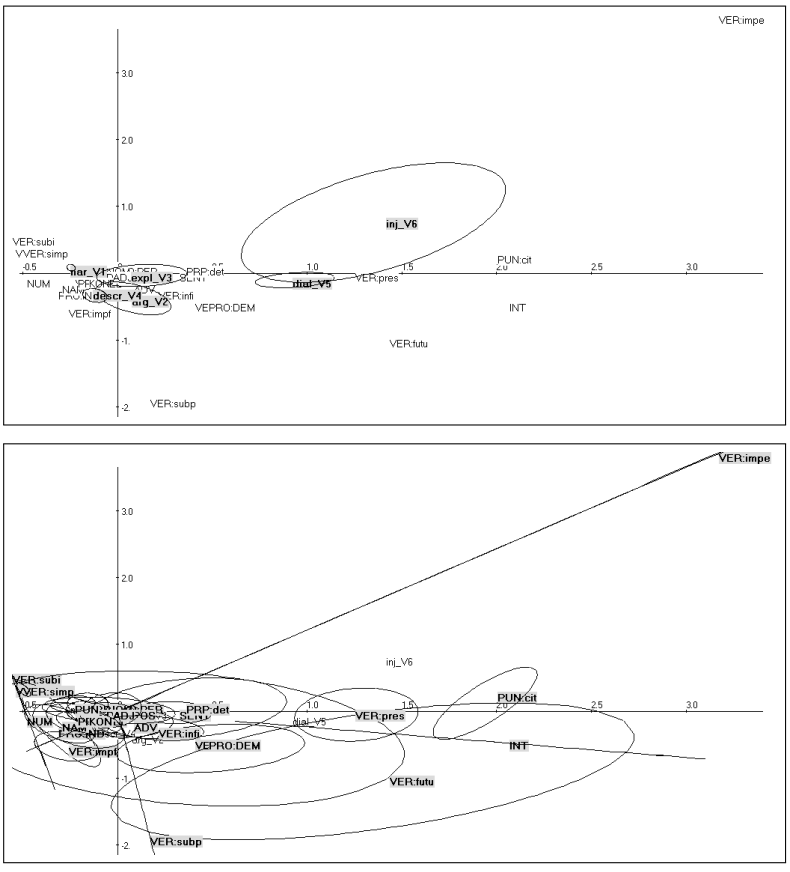


FIGURE 4.8 – AFC sur « Le Voleur » entre les CMS et les types de discours avec validation par *bootstrap*. Inertie expliquée par le premier axe factoriel : 65.29% ; et par le deuxième : 12.53%.

Les positions des CMS pour le texte « Le Voleur » de la figure 4.8 sont à nouveau assez similaires à celles de la figure 4.4, quoique ce soit moins évident que pour le texte « L'Orient ». Au sujet des types de discours (figure 4.8, haut), on observe que les types de discours injonctif, dialogal, argumentatif, descriptif et narratif sont significativement différents de l'origine. Seul le type explicatif ne l'est pas. Aussi, le type narratif est isolé des autres et il est stable, au sens de faiblement variable. L'ellipse de confiance du type dialogal est quasiment incluse dans l'ellipse de confiance du type injonctif, ce qui confirme à nouveau que ces deux types sont relativement similaires. Les ellipses de confiance des types descriptif, argumentatif et explicatif se chevauchent aussi et ne sont donc pas clairement distincts par rapport aux CMS. Concernant les CMS (figure 4.8, bas), il est à nouveau difficile de les distinguer. Cependant, on remarque que les verbes au présent (VER:pres) et la ponctuation de citation sont significativement différents de l'origine et sont en attraction mutuelle avec les types injonctif et dialogal, ce que l'on pouvait déjà observer dans la table 4.3.

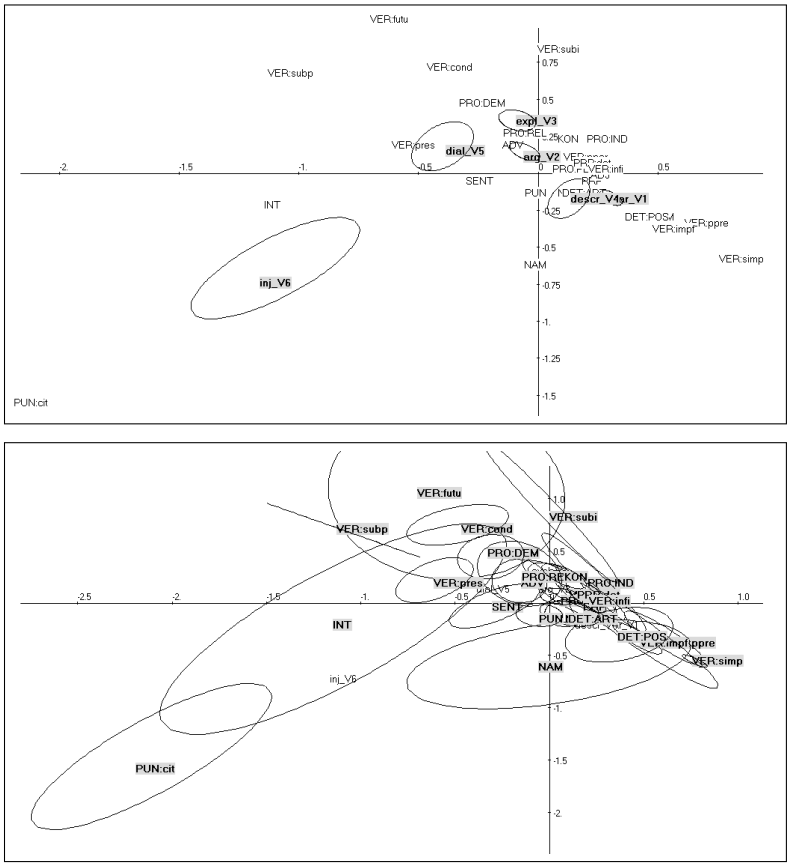


FIGURE 4.9 – AFC sur « Un Fou ? » entre les CMS et les types de discours avec validation par *bootstrap*. Inertie expliquée par le premier axe factoriel : 46.20% ; et par le deuxième : 25.00%.

À nouveau, on remarque que la configuration des CMS de la figure 4.9 partage des similitudes avec celle de la figure 4.5, avec tout de même quelques différences importantes. Concernant les types de discours (figure 4.9, haut), ils sont tous significativement différents de l'origine et donc dépendants des CMS. De plus, il n'y a aucune intersection entre toutes les ellipses de confiance de ces types de discours, ils sont donc clairement distincts. Concernant les CMS, on constate, par exemple, que les interjections et la ponctuation de citation sont significativement différentes de l'origine. Cependant, les ellipses de confiance sont étendues et donc ces CMS ne sont pas stables. Quant aux verbes au présent et au conditionnel, ils sont aussi significativement différents de l'origine, mais ils sont plus stables que les deux autres CMS.

Finalement, la figure 4.10 présente les résultats obtenus pour le texte « Un Fou ». Pour ce

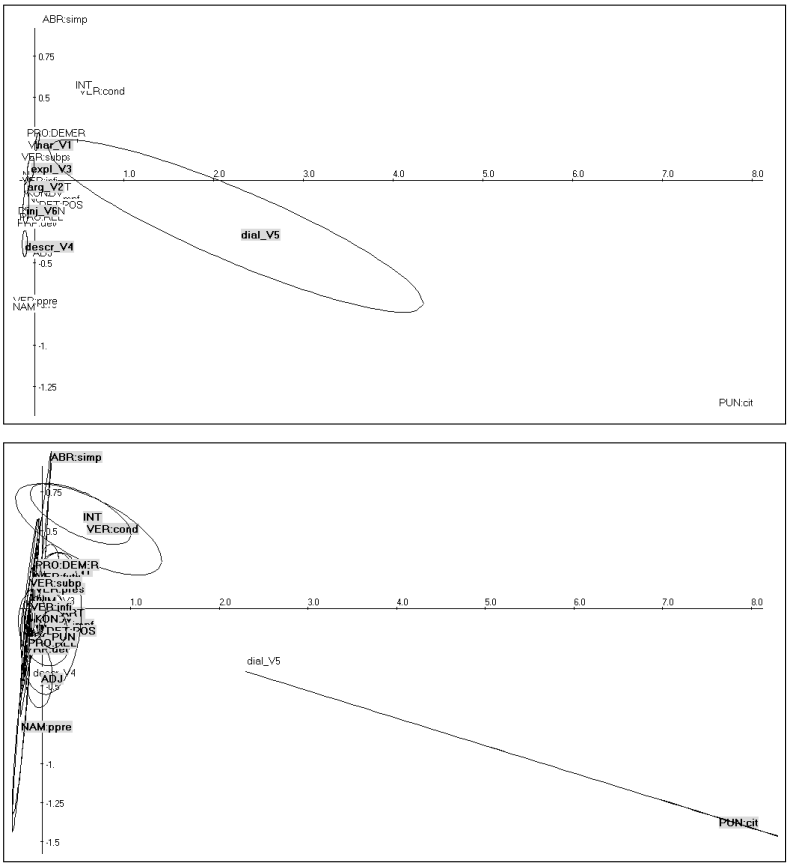


FIGURE 4.10 – AFC sur « Un Fou » entre les CMS et les types de discours avec validation par *bootstrap*. Inertie expliquée par le premier axe factoriel : 47.79% ; et par le deuxième : 27.06%.

dernier texte, le représentation des CMS est très différente à celle de la figure 4.6. Tous les types de discours sont significativement différents de l'origine. La grande différence entre le type dialogal et les autres types est que pour le premier, l'ellipse de confiance est très étendue et donc que ce type n'est pas très stable. Concernant les CMS, on peut distinguer que les interjections, les verbes au conditionnel et les les adjectifs sont significativement différents de l'origine. De plus, ce dernier est en attraction mutuelle avec le type descriptif, comme il avait déjà été observé dans la section 4.1.4.

4.3 Classification non supervisée et résultats

En premier lieu, les tables de contingence, croisant propositions et n-grammes de CMS (cf. section 4.1.3), sont transformées par (1.6) en matrices de dissimilarités du khi2 entre les propositions $D = (D_{ij})$. Cette étape est effectuée pour chacun des quatre textes et pour chaque longueur de n-gramme de CMS, soit les uni-, bi- et trigrammes. Ensuite, deux méthodes de classification non supervisée (présentées dans la section 2.1) sont utilisées : l'algorithme K-means (section 4.3.1) et l'algorithme K-means flou (section 4.3.2). Leurs résultats sont évalués au moyen d'indices d'accord entre partitions.

4.3.1 K-means

4.3.1.1 Choix des paramètres

Pour effectuer l'algorithme K-means, la matrice de dissimilarités du khi2 est utilisée avec l'algorithme tel qu'il est décrit dans la section 2.1.2, en y incluant les transformations de puissance

de Schoenberg.

Plus particulièrement, l'algorithme K-means a été appliqué aux quatre textes, considérés séparément, pour les uni-, bi- et trigrammes de CMS. La principale visée de cette classification non supervisée étant de retrouver les 6 types de discours, on choisit un nombre de groupes $m = 6$. Aussi, le nombre d'itérations maximal est fixé à $N_{\max} = 400$. [6] Concernant la transformation de puissance (1.22), la puissance q varie de 0.1 à 1, avec des incréments de 0.05. Ainsi, la méthode K-means est effectuée pour les 4 textes avec les 3 longueurs de n-gramme différentes, pour 19 valeurs de q, ce qui conduit à 228 cas différents.

Il faut encore noter que, puisque la solution de l'algorithme K-means dépend de la position initiale des centres, déterminée ici par la matrice Z, générée aléatoirement, chaque cas est calculé 300 fois et l'on prend ensuite la moyenne des résultats obtenus pour chacun des cas.

Plus précisément, pour chacun des résultats, l'indice de Jaccard (2.11), J, et l'indice de Rand corrigé (2.12), RC, sont calculés sur la table de contingence croisant les effectifs des propositions catégorisées en 6 groupes par l'annotateur et classifiées en 6 groupes selon l'algorithme. Puis, la moyenne de ces résultats, pour chaque cas, est calculée.

Une version de la méthode K-means, non pondérée, a aussi été testée en posant $f_i = 1/n$ pour le calcul de f_j^g dans (2.6). Les résultats ainsi obtenus sont exposés dans l'article de Cocco *et al.* (2011).

4.3.1.2 Résultats

Les moyennes des résultats, obtenus pour J et RC en fonction de q avec l'algorithme K-means, sont présentées dans les figures 4.11 à 4.14, sans les écarts-types. Les figures « complètes », avec les écarts-types des deux indices d'accord entre partitions, se trouvent dans l'annexe C, section C.1.1.

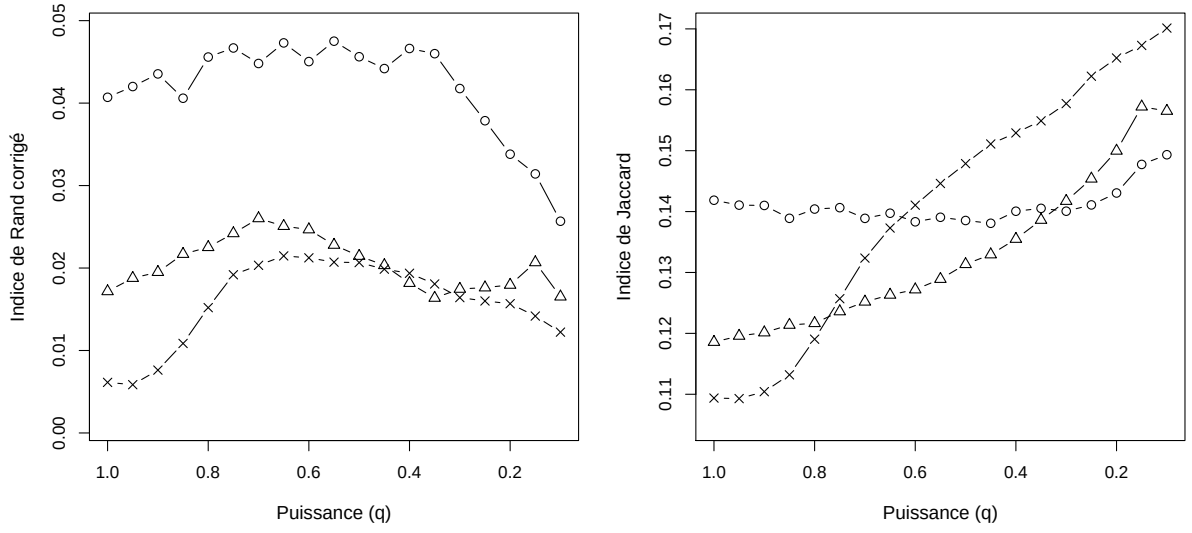


FIGURE 4.11 – « L'Orient » avec l'algorithme K-means. Indice de Rand corrigé (gauche) et de Jaccard (droite) en fonction de la puissance q. ($\circ$ = unigrammes, $\triangle$ = bigrammes et $\times$ = trigrammes). Pour rappel, $q = 1$ est équivalent à ne pas effectuer de transformation.

Deux premières constatations sont évidentes. Premièrement, les résultats obtenus pour les quatre textes sont différents, comme le laisser supposer les liens entre les CMS et les types de discours (cf. sections 4.1.4 et 4.2). Deuxièmement, les deux indices d'accord entre partitions choisis produisent des résultats très différents. Cette différence entre les deux indices découle

6. Cette valeur n'est jamais atteinte, car la solution se stabilise rapidement (le nombre d'itérations maximum observé sur l'ensemble des résultats jusqu'à stabilisation de la solution est de 46).

certainement du fait que l'indice de Jaccard ne considère pas le nombre de paires simultanément séparées dans les deux partitions (Milligan et Cooper, 1986).

Cependant, on remarque aussi des régularités. Par exemple, concernant les textes « L'Orient », « Le Voleur » et « Un Fou ? », avec l'indice de Rand corrigé (graphiques de gauche des figures 4.11, 4.12 et 4.13), les unigrammes produisent les meilleurs résultats ; et les trigrammes, les moins bons résultats en général. Pour le texte « Le Voleur » (figure 4.12, droite), avec l'indice de Jaccard, les unigrammes révèlent aussi de meilleurs résultats.

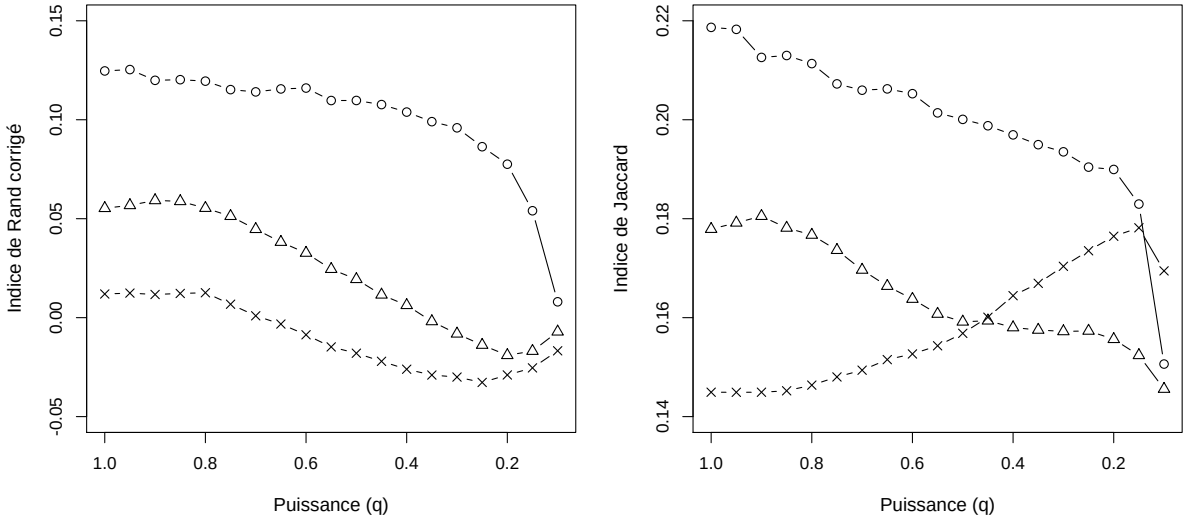


FIGURE 4.12 – « Le Voleur » avec l'algorithme K-means. Indice de Rand corrigé (gauche) et de Jaccard (droite) en fonction de la puissance q. ($\circ$ = unigrammes, $\triangle$ = bigrammes et $\times$ = trigrammes).

Concernant l'indice de Jaccard avec les trois autres textes, soit « L'Orient », « Un Fou ? » et « Un Fou » (graphiques de droite des figures 4.11, 4.13 et 4.14), on constate que les trigrammes engendrent de meilleurs résultats pour des valeurs faibles de q et que la tendance s'inverse, avec les meilleurs résultats pour les unigrammes, lorsque $q > 0.65$, respectivement 0.45 et 0.7.

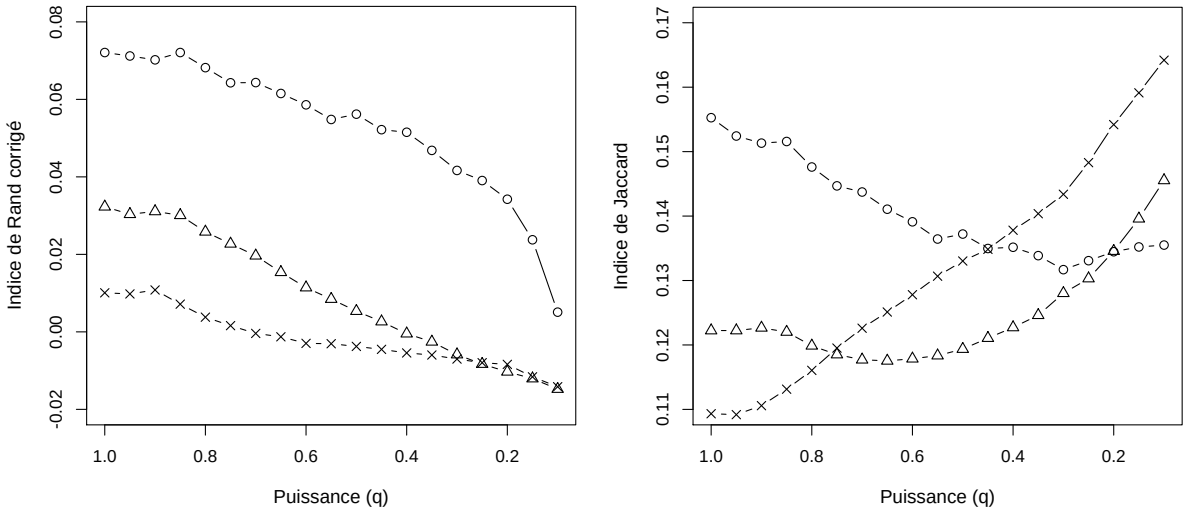


FIGURE 4.13 – « Un Fou ? » avec l'algorithme K-means. Indice de Rand corrigé (gauche) et de Jaccard (droite) en fonction de la puissance q. ($\circ$ = unigrammes, $\triangle$ = bigrammes et $\times$ = trigrammes).

Avec l'indice de Rand corrigé, on observe aussi que les transformations de puissance semblent

améliorer les résultats. En effet, le meilleur résultat obtenu pour « L'Orient » (figure 4.11, gauche), avec les unigrammes, est $RC = 0.048$ pour $q = 0.55$; pour « Le Voleur » (figure 4.12, gauche), $RC = 0.125$ pour $q = 0.95$; pour « Un Fou ? » (figure 4.13, gauche), $RC = 0.072$ pour $q = 0.85$; et pour « Un Fou » (figure 4.14, gauche), $RC = 0.046$ pour $q = 0.25$, mais cette fois pour les trigrammes. De plus, que ce soit pour les uni-, les bi- ou les trigrammes de CMS dans ce dernier texte, l'introduction de la transformation de puissance améliore systématiquement les résultats au regard de l'indice de Rand corrigé.

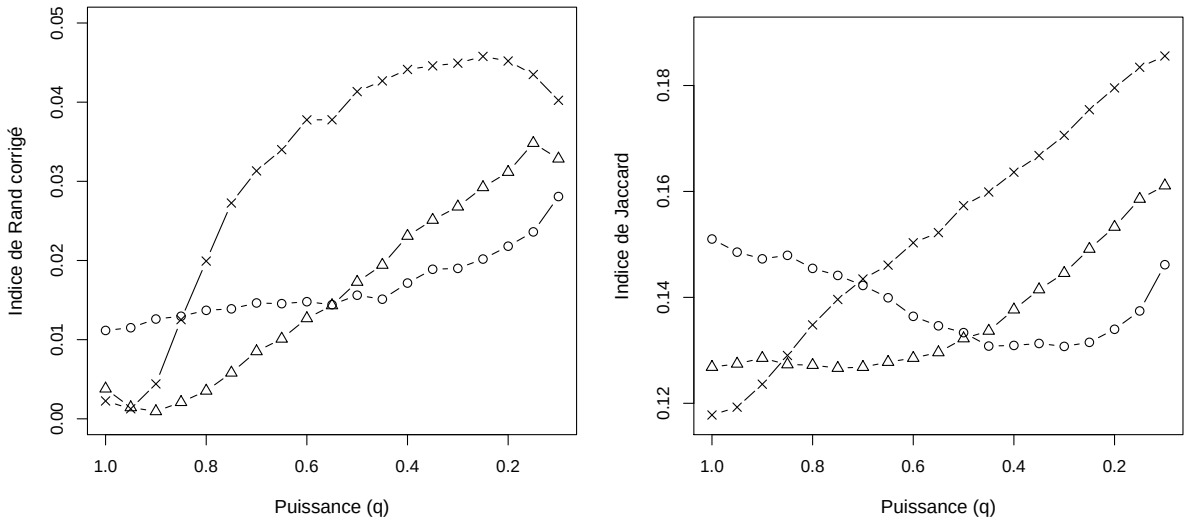


FIGURE 4.14 – « Un Fou » avec l'algorithme K-means. Indice de Rand corrigé (gauche) et de Jaccard (droite) en fonction de la puissance q. ($\circ$ = unigrammes, $\triangle$ = bigrammes et $\times$ = trigrammes).

Finalement, il est difficile de comparer les résultats et beaucoup de différences subsistent entre les indices d'accord entre partitions. Cependant, malgré ces différences, les résultats sont toujours meilleurs pour le texte « Le Voleur », et ce avec les deux indices utilisés. Comme déjà mentionné (cf. section 2.3.1), il existe d'autres indices. À titre d'exemple, les mêmes essais ont été faits en comparant les partitions par le biais du V de Cramer et les résultats sont présentés dans la section C.1.2 de l'annexe. Derechef, les résultats les plus élevés sont obtenus, lorsque l'on considère des unigrammes de CMS, pour le texte « Le Voleur ».

4.3.2 K-means flou

4.3.2.1 Choix de paramètres

Pour appliquer l'algorithme K-means flou, on utilise, à nouveau, pour chaque texte et pour chaque longueur de n-gramme de CMS différente, la matrice des dissimilarités du khi2, D (cf. introduction de cette section 4.3). L'algorithme K-means flou, tel qu'il est présenté dans la section 2.1.3, est appliqué sur chacune de ces matrices D.

En particulier, pour chaque texte, le nombre de groupes de départ, m, est choisi égal au nombre n de propositions présentes dans chacun des textes. Ainsi, le nombre de groupes final après agrégation, M, est déterminé uniquement par la température relative, t_{rel}. Après plusieurs essais, on choisit de faire varier cette dernière entre 0.022 et 0.3, avec des incréments de 0.001. Concernant le texte « Un Fou », qui contient plus de propositions (table 4.1), on choisit de faire varier t_{rel} entre 0.02 et 0.3, avec des incréments de 0.01, pour maintenir un temps de calcul raisonnable.

À nouveau, le nombre d'itérations maximum a été fixé à $N_{\max} = 400$. Contrairement à la méthode K-means (dur), cette valeur est parfois atteinte, car la solution semble se stabiliser

plus lentement, en particulier lorsque les valeurs de t_{rel} sont basses, mais pas forcément pour les valeurs minimales choisies. Finalement, pour chaque t_{rel}, l'algorithme a été exécuté 20 fois, puis les moyennes des indices d'accord entre partitions, J et RC, ont été calculées.

Il faut préciser que les 20 exécutions n'ont pas systématiquement abouti à un résultat, car deux problèmes d'instabilités numériques différents ont été détectés. Le premier se produit lors de la seconde itération si les valeurs de t_{rel} sont trop petites ; et le second, lors de l'agrégation des m groupes en M groupes avec le critère de fusion des groupes (cf. section 2.1.3). Ces instabilités numériques étant rares, les résultats ont simplement été supprimés, sans être recalculés.

4.3.2.2 Résultats

Les figures 4.15 à 4.22 présentent un résumé des résultats obtenus en appliquant l'algorithme K-means flou sur les quatre textes. Sur toutes ces figures, les graphiques de droite présentent un indice d'accord entre partitions en fonction du nombre de groupes final M. En réalité, il s'agit d'une représentation paramétrique de la moyenne de l'indice d'accord entre partitions et de la moyenne de M, sur les 20 exécutions, en fonction de la température relative t_{rel}. Aussi, les résultats pour les moyennes de M et pour les moyennes des indices d'accord entre partitions en fonction de t_{rel} sont présentés dans l'annexe C.2.

En premier lieu, on observe, sur les graphiques de gauche des figures 4.15 à 4.22, ainsi que sur les graphiques du haut des figures de la section C.2, que, comme déjà annoncé dans la section 2.1.3 présentant l'algorithme, le nombre de groupes final M, pour les trois longueurs de n-gramme de CMS, diminue lorsque la température relative augmente. Aussi, comme pour l'algorithme K-means (dur), on remarque que les résultats diffèrent fortement selon l'indice d'accord entre partitions utilisé et selon les textes.

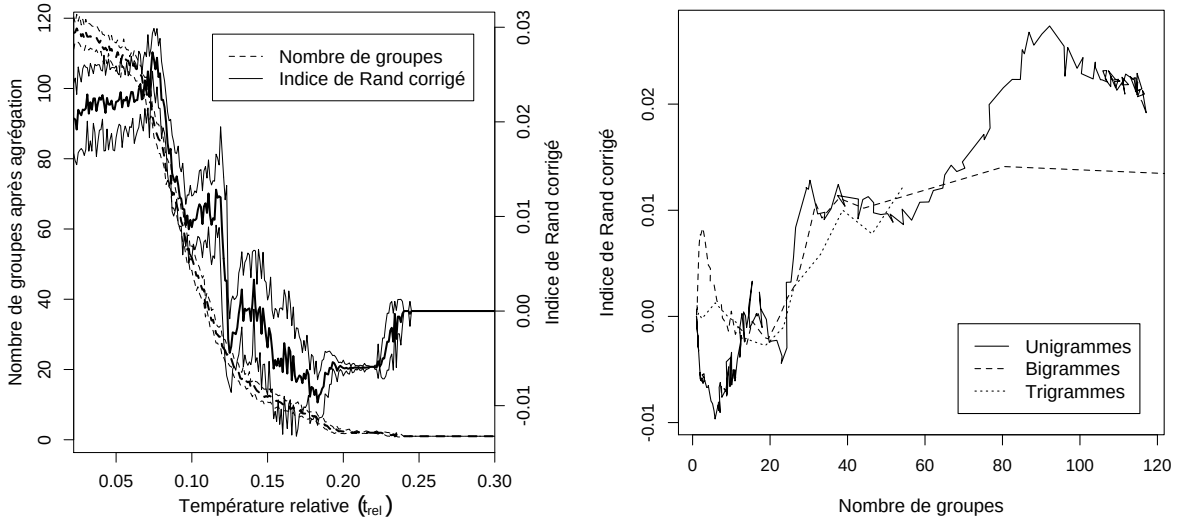


FIGURE 4.15 – « L'Orient » avec l'algorithme K-means flou. Moyenne (ligne épaisse) et écarts-types (ligne fine) de l'indice de Rand corrigé, RC, et du nombre de groupes après agrégation, M, en fonction de la température relative, t_{rel}, pour les unigrammes de CMS (gauche) et moyenne de RC en fonction de la moyenne de M pour les uni-, bi- et trigrammes de CMS (droite).

La figure 4.15 montre que, pour « L'Orient » avec l'indice de Rand corrigé, les meilleurs résultats sont obtenus avec des unigrammes pour un nombre de groupe élevé, alors que pour un nombre de groupes plus petit, en particulier pour $M < 8$ environ, les bigrammes engendrent de meilleurs résultats. La valeur la plus élevée de l'indice, $RC = 0.027$, est obtenue avec les unigrammes, lorsque $M = 92.2$, ce qui correspond à $t_{\text{rel}} = 0.074$. Ainsi, le meilleur résultat s'obtient lorsqu'il y a environ 92 groupes pour 189 propositions (table 4.1), donc les groupes contiennent 2 propositions en moyenne. Aussi, toujours pour les unigrammes, les résultats sont

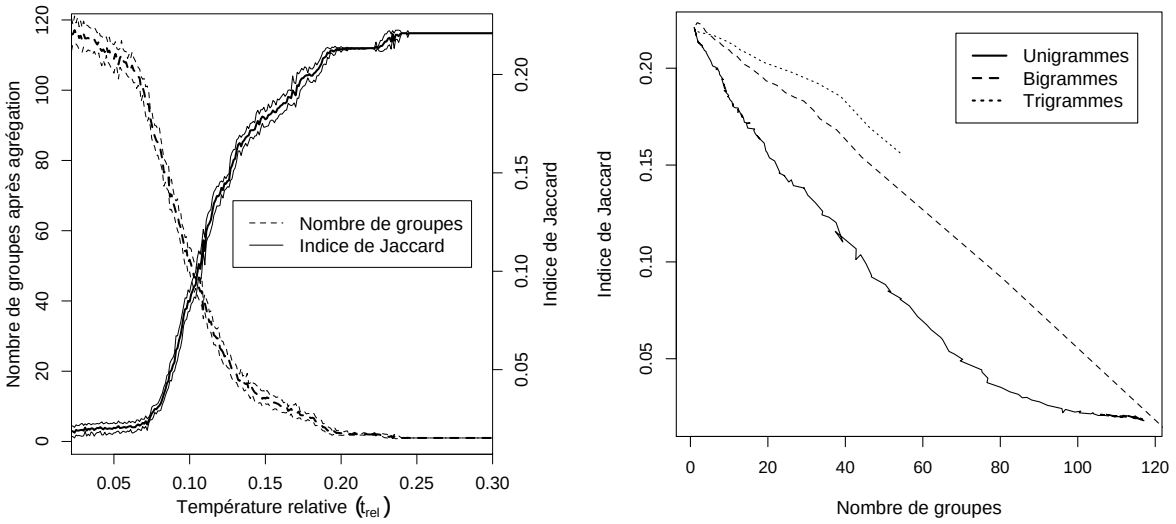


FIGURE 4.16 – « L'Orient » avec l'algorithme K-means flou. Moyenne (ligne épaisse) et écarts-types (ligne fine) de l'indice de Jaccard, J, et du nombre de groupes après agrégation, M, en fonction de la température relative, t_{rel}, pour les unigrammes de CMS (gauche) et moyenne de J en fonction de la moyenne de M pour les uni-, bi- et trigrammes de CMS (droite).

parfois négatifs, ce qui signifie que l'accord entre la partition obtenue par l'algorithme et celle créée par l'expert humain est moins bon qu'un accord qui serait obtenu au hasard (cf. section 2.3.1). Concernant les résultat obtenus avec l'indice de Jaccard (figure 4.16), on constate un petit pic pour les bigrammes, $J = 0.224$, lorsque $M = 1.9$ ($t_{\text{rel}} = 0.066$), qui est le meilleur résultat obtenu pour ce texte. À l'exception de ce pic, les meilleurs résultats sont obtenus avec les trigrammes lorsque le nombre de groupes est petit ($M < 54$ environ). Pour un nombre de groupes plus élevé, l'indice de Jaccard n'a pas pu être calculé en raison d'instabilités numériques dues à des valeurs de t_{rel} (0.022 et 0.023) trop petites.

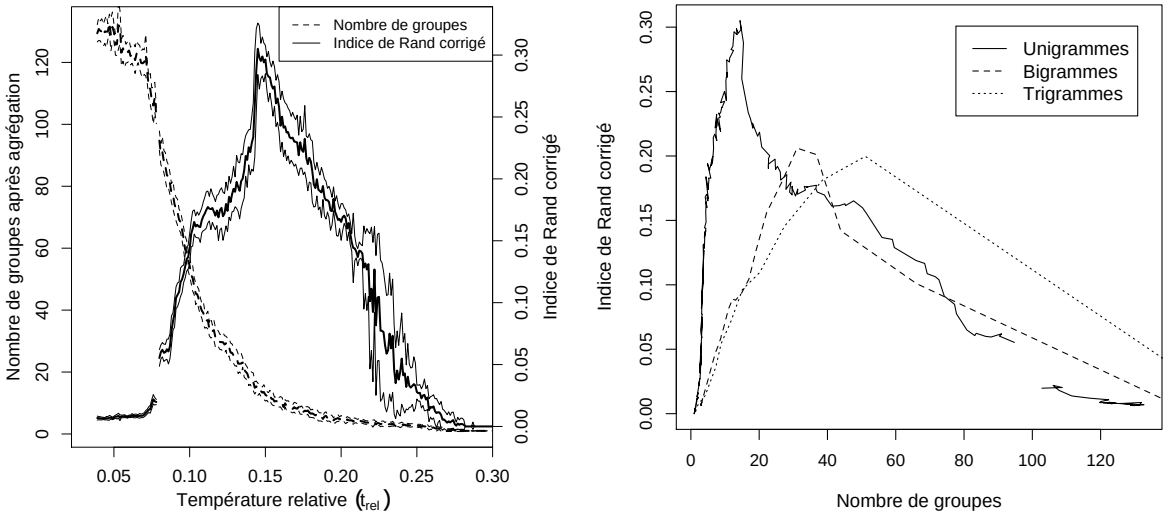


FIGURE 4.17 – « Le Voleur » avec l'algorithme K-means flou. Moyenne (ligne épaisse) et écarts-types (ligne fine) de RC et de M en fonction de t_{rel}, pour les unigrammes de CMS (gauche) et moyenne de RC en fonction de la moyenne de M pour les uni-, bi- et trigrammes de CMS (droite).

Concernant le texte « Le Voleur » (figures 4.17 et 4.18), les meilleurs résultats sont obtenus avec les unigrammes pour les deux indices d'accord entre partitions. Aussi, il existe un pic

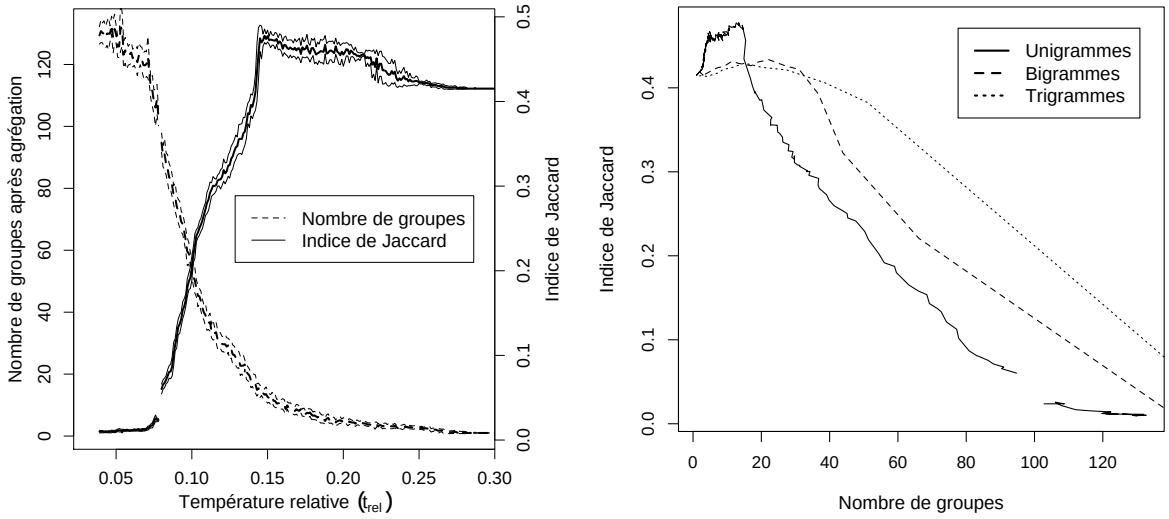


FIGURE 4.18 – « Le Voleur » avec l'algorithme K-means flou. Moyenne (ligne épaisse) et écarts-types (ligne fine) de J et de M en fonction de t_{rel}, pour les unigrammes de CMS (gauche) et moyenne de J en fonction de la moyenne de M pour les uni-, bi- et trigrammes de CMS (droite).

remarquable sur chacune de ces deux figures, bien que plus important avec l'indice de Rand corrigé. Pour l'indice de Rand corrigé (figure 4.17), il atteint une moyenne de $RC = 0.305$, lorsque $t_{\text{rel}} = 0.145$, ce qui correspond à une moyenne de groupes $M = 14.4$. Quant à l'indice de Jaccard (figure 4.18), sa valeur maximale est de $J = 0.478$, pour $M = 13.4$ ($t_{\text{rel}} = 0.148$). Il semble donc que le nombre de groupes optimal soit plutôt de 14 que de 6. À titre d'exemple, la table 4.4 présente une exécution typique de l'algorithme aboutissant à la génération de 14 groupes. En particulier, on constate que le groupe le groupe 1 est clairement dominant et qu'il est associé au type de discours narratif, attribué à plus de 60% des propositions (cf. table 4.1). Aussi, la majorité des propositions classées dans le groupe 11 par l'algorithme correspondent à celles annotées comme descriptives par l'expert humain. Les propositions correspondant aux différents groupes définis par l'algorithme sont fournies dans la table 4.5.

Effectifs

	Algorithme K-means flou													
Expert	1	2	3	4	5	6	7	8	9	10	11	12	13	14
argumentatif	7	0	0	0	1	1	0	0	1	0	0	0	0	0
descriptif	19	2	0	0	1	0	1	0	0	1	0	0	1	0
dialogal	7	0	1	0	0	0	0	2	2	1	15	0	0	1
explicatif	7	0	0	0	0	0	0	0	0	0	3	0	0	0
injonctif	2	0	0	1	0	0	0	0	0	0	2	0	0	1
narratif	116	1	0	0	2	0	3	0	0	1	0	4	1	0

Quotients d'indépendance

	Algorithme K-means flou													
Expert	1	2	3	4	5	6	7	8	9	10	11	12	13	14
argumentatif	0.92	0.00	0.00	0.00	5.20	20.80	0.00	0.00	6.93	0.00	0.00	0.00	0.00	0.00
descriptif	1.00	5.55	0.00	0.00	2.08	0.00	2.08	0.00	0.00	2.77	0.00	0.00	4.16	0.00
dialogal	0.32	0.00	7.17	0.00	0.00	0.00	0.00	7.17	4.78	2.39	5.38	0.00	0.00	3.59
explicatif	0.92	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	3.12	0.00	0.00	0.00
injonctif	0.44	0.00	0.00	34.67	0.00	0.00	0.00	0.00	0.00	0.00	3.47	0.00	0.00	17.33
narratif	1.19	0.54	0.00	0.00	0.81	0.00	1.22	0.00	0.00	0.54	0.00	1.62	0.81	0.00

TABLE 4.4 – Exemple d'un résultat obtenu avec l'algorithme K-mean flou sur le texte « Le Voleur » avec $t_{\text{rel}} = 0.146$, aboutissant à la création de 14 groupes. Pour cet exemple : $RC = 0.322$ et $J = 0.483$.

Groupe	Exemple		Autres membres du groupe
1	Et le vieil artiste se mit à cheval sur une chaise.	21	8, 11, 12, 13, 17, 23, 25, 28, 29, 32, 33, 35, 36, 37, 38, 39, 42, 43, 44, 46, 47, 49, 50, 51, 58, 59, 60, 61, 62, 63, 64, 65, 69, 70, 74, 76, 78,79, 80, 81, 82, 83, 85, 86, 87, 91, 92, 95, 96, 98, 99, 100, 101, 102, 103, 109, 111, 112, 114, 116, 117, 120, 121, 123, 124, 125, 126, 128, 129, 130, 131, 132, 133, 137, 139, 143, 144, 145, 147, 148, 149, 150, 151, 152, 153, 154, 159, 160, 161, 162, 167, 169, 170, 172, 173, 175, 176, 178, 179, 180, 184, 185, 186, 194, 198, 199, 200, 206, 207, 208, 209, 211, 215, 222, 224, 229, 230, 232, 233, 234, 235, 237, 238, 240, 241, 242, 244, 245, 246, 247, 248, 252, 253, 254, 255, 256, 264, 265, 266, 267, 268, 269, 270, 271, 273, 274, 275, 276, 283, 284, 285, 286, 287, 290, 295, 296, 297
2	Il était sombre et profond.	113	34, 158
3	celui-ci doit être livré au bourreau.	225	-
4	"Soyons prudents",	106	-
5	où l'esprit farceur sévissait si bien	16	97, 115, 146
6	qui ont connu cette époque	15	-
7	où il fut englouti.	48	84, 118, 210
8	"Eh bien, mon pauv'vieux, comment ça va-t-il ?"	258	141
9	Les peintres seuls ne s'étonneront point, surtout les vieux	14	6, 93
10	mais je n'oserais affirmer	30	136, 273
11	"Vous voulez rire, sans doute."	182	5, 53, 67, 72, 90, 135, 156, 164, 190, 192, 203, 213, 218, 226, 250
12	Puis il dit :	187	88, 166, 260
13	que c'était lui.	31	45
14	"Au secours !"	196	55

TABLE 4.5 – Propositions énoncées correspondant au résultat présenté dans la table 4.4 et obtenu avec l'algorithme K-means flou. Les nombres dans cette table font référence aux lignes de l'annexe A.2.

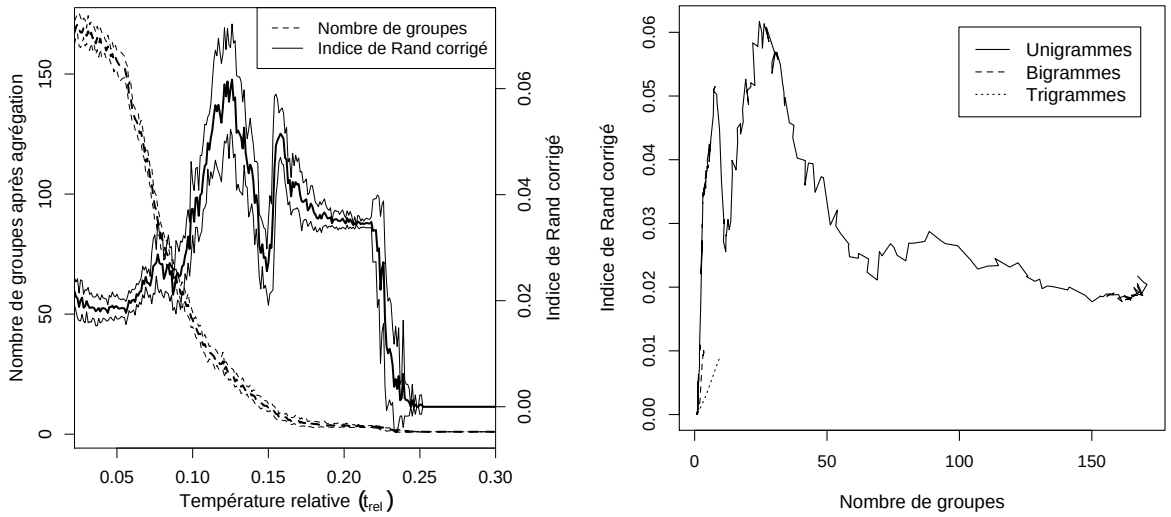


FIGURE 4.19 – « Un Fou ? » avec l'algorithme K-means flou. Moyenne (ligne épaisse) et écarts-types (ligne fine) de RC et de M en fonction de t_{rel}, pour les unigrammes de CMS (gauche) et moyenne de RC en fonction de la moyenne de M pour les uni-, bi- et trigrammes de CMS (droite).

Concernant le texte « Un Fou ? » (figures 4.19 et 4.20), à l'instar du texte « Le Voleur », les unigrammes produisent systématiquement les meilleurs résultats. En fait, pour les bi- et les trigrammes, le nombre de groupes après agrégation M chute rapidement à 1, plus précisément lorsque $t_{\text{rel}} > 0.079$ pour les bigrammes, et lorsque $t_{\text{rel}} > 0.028$ pour les trigrammes. De plus, pour des valeurs basses de t_{rel}, des instabilités numériques se produisent. Ainsi, peu de résultas sont exploitables. Avec l'indice de Rand corrigé (figure 4.19), deux pics apparaissent pour les unigrammes. Le premier vaut $RC = 0.051$, lorsque $M = 7.8$ ($t_{\text{rel}} = 0.158$) ; et le second, plus élevé, $RC = 0.062$, lorsque $M = 24.5$ ($t_{\text{rel}} = 0.126$). Avec l'indice de Jaccard (figure 4.20),

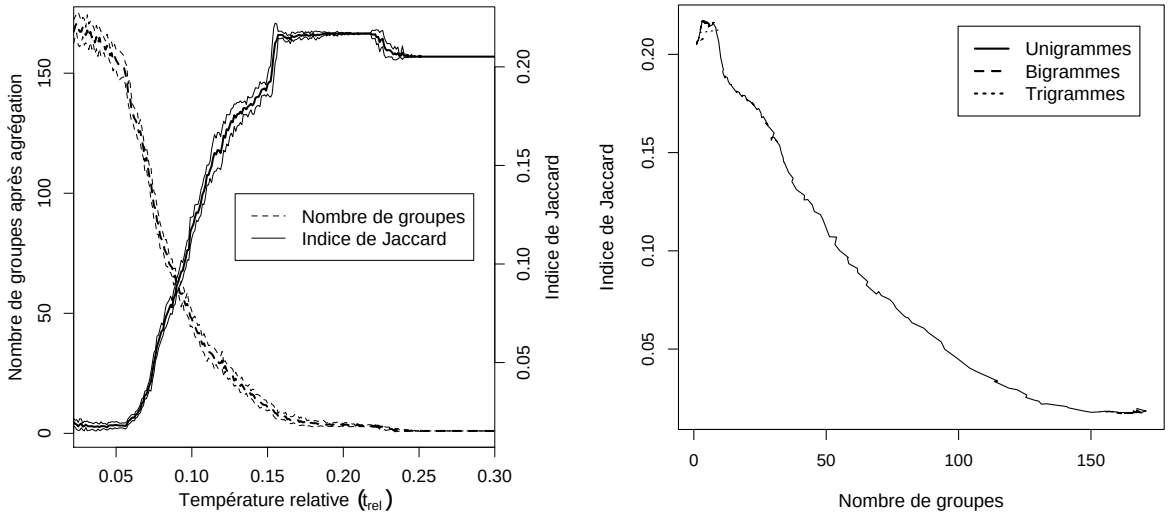


FIGURE 4.20 – « Un Fou ? » avec l'algorithme K-means flou. Moyenne (ligne épaisse) et écarts-types (ligne fine) de J et de M en fonction de t_{rel}, pour les unigrammes de CMS (gauche) et moyenne de J en fonction de la moyenne de M pour les uni-, bi- et trigrammes de CMS (droite).

on observe un petit pic pour les unigrammes, $J = 0.216$, lorsque $M = 7.1$ ($t_{\text{rel}} = 0.157$), donc pour un nombre de groupes proche de celui du premier pic observé avec l'indice de Rand corrigé. Cependant, il ne correspond pas à la valeur maximale obtenue pour ce texte, qui est de $J = 0.217$ pour $M = 3.7$ ($t_{\text{rel}} = 0.202$).

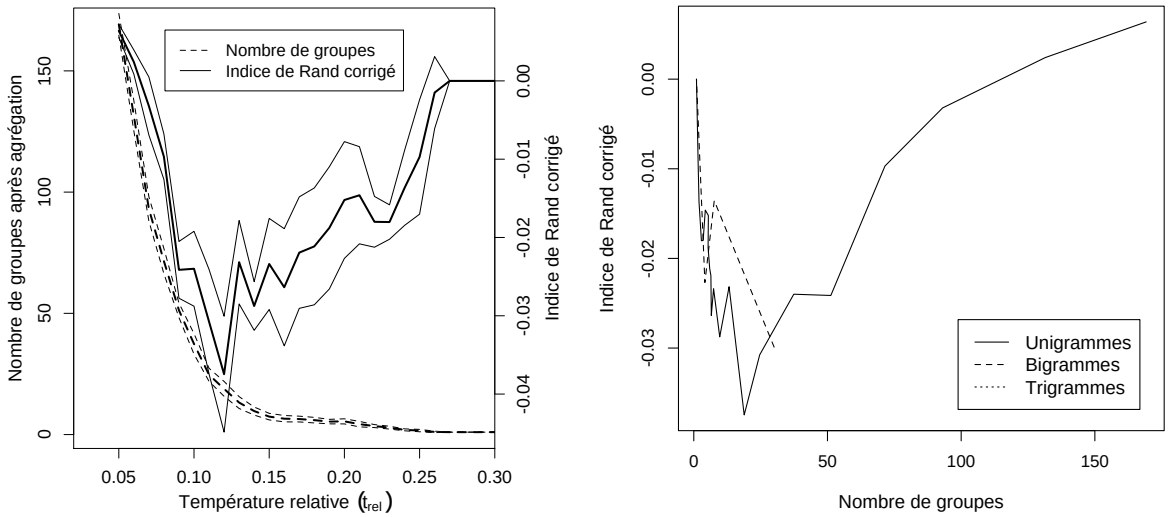


FIGURE 4.21 – « Un Fou » avec l'algorithme K-means flou. Moyenne (ligne épaisse) et écarts-types (ligne fine) de RC et de M en fonction de t_{rel}, pour les unigrammes de CMS (gauche) et moyenne de RC en fonction de la moyenne de M pour les uni-, bi- et trigrammes de CMS (droite).

Comme pour « L'Orient », les résultats pour « Un Fou » prennent des valeurs négatives avec l'indice de Rand corrigé (figure 4.21), mais sur une plus grande étendue pour ce texte, en particulier avec les unigrammes. Il faut noter que peu de résultats obtenus avec les bigrammes sont exploitables, et encore moins avec les trigrammes, car pour ces derniers tous les résultats, indépendamment de la valeur de t_{rel} correspondent à $M = 1$ (cf. graphique du haut de la figure C.12). Concernant les résultats obtenus avec l'indice de Jaccard (figure 4.22), ils sont aussi très similaires à ceux obtenus pour le texte de « L'Orient », sans pics cependant.

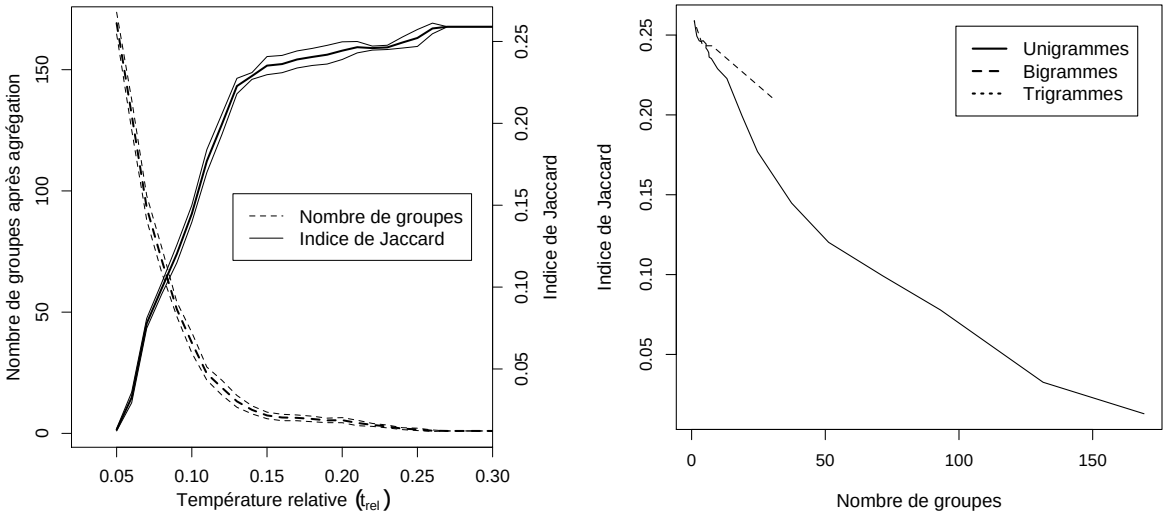


FIGURE 4.22 – « Un Fou » avec l'algorithme K-means flou. Moyenne (ligne épaisse) et écarts-types (ligne fine) de J et de M en fonction de t_{rel}, pour les unigrammes de CMS (gauche) et moyenne de J en fonction de la moyenne de M pour les uni-, bi- et trigrammes de CMS (droite).

Malgré l'hétérogénéité des résultats, on constate, comme avec l'algorithme K-means, plusieurs régularités. En particulier, les unigrammes produisent généralement de meilleurs résultats avec l'indice de Rand corrigé. On remarque aussi que le nombre de groupes semble évoluer différemment en fonction de t_{rel} selon que les uni-, bi- ou trigrammes de CMS sont utilisés. Il pourrait donc être intéressant de faire varier différemment t_{rel} pour les bi- et trigrammes, malgré les résultats souvent moins bons pour ces derniers.

Aussi, les résultats obtenus pour le texte « Le Voleur », quel que soit l'indice d'accord entre partitions utilisé, sont nettement meilleurs, tout comme avec l'algorithme K-means (dur). Les résultats sont plus décevants pour les textes « L'Orient » et « Un Fou » qui, pour rappel (cf. section 4.1.4), ont certainement une structure un peu différente des deux autres contes. On note aussi que, par construction, l'indice de Jaccard (graphiques de droite des figures de la section C.2) prend une valeur constante et positive lorsqu'il ne reste qu'un groupe (pour un exemple de calcul avec « Le Voleur » voir Cocco, 2012*b*).

4.4 Discussion

Concernant la classification non supervisée, il est clair que les résultats sont difficiles à interpréter, que les deux indices choisis ne fournissent pas la même information et que les différents textes engendrent des résultats différents, et ce quelque soit l'algorithme choisi. Cependant, plusieurs régularités ont été observées et permettent quelques conclusions. Premièrement, les unigrammes de CMS donnent généralement de meilleurs résultats que les bi- et les trigrammes de CMS. Cela s'explique peut-être par le fait que pour les bi- et les trigrammes, la table de contingence est creuse, conduisant au problème du « fléau de la dimension » ou de « la malédiction de la dimension » (*curse of dimensionality*), en particulier dans le cadre la classification (voir par exemple Houle, Kriegel, Kröger, Schubert et Zimek, 2010). Deuxièmement, l'introduction de la transformation de puissance peut améliorer les résultats. Aussi, les résultats sont systématiquement meilleurs pour le texte « Le Voleur », peu importe l'algorithme de classification ou la méthode d'évaluation. Il faut noter que ces résultats sont à considérer avec précaution en raison de deux limitations : la taille relativement courte des textes et le fait qu'il n'y ait qu'un seul annotateur.

Au-delà de ces résultats intéressants, il reste de nombreuses pistes à explorer. Au niveau de la

méthode de classification, il est clair qu'il serait intéressant de combiner les transformations de Schoenberg avec l'algorithme K-means flou. Pratiquement, pour ce faire, il suffirait, comme pour l'algorithme K-means, de transformer D en $\tilde{D} = \varphi(D)$ lors de l'initialisation (cf. section 2.1.3). Une toute autre voie serait d'utiliser des algorithmes de classification supervisée (cf. section 2.2) qu'il serait avantageux d'associer avec des méthodes de sélection de caractéristiques (*feature selection*) (voir par exemple Yang et Pedersen, 1997). Il s'agirait alors de les appliquer à des textes plus longs pour se prémunir des risques de sur-paramétrisation.

Concernant le choix des caractéristiques, une première piste de recherche serait d'utiliser spécifiquement les marques linguistiques de chaque type de discours décrites dans la section 4.1.1. Aussi, on pourrait s'intéresser au fonctionnement du logiciel Tropes [7] qui permet, entre autres choses, de classer l'ensemble d'un texte, à condition qu'il soit suffisamment long, dans un des quatre « styles » suivants : argumentatif, narratif, énonciatif et descriptif (basés sur les *modes d'organisation du discours* de Charaudeau, 1992, troisième partie, pp. 631-835). Bien que ces modes de discours ne soient pas strictement identiques aux types de discours étudiés dans ce travail, les caractéristiques retenues par le logiciel sont de même type que celles proposées dans les marques linguistiques (cf. section 4.1.1), mais plus fines que les CMS obtenues avec TreeTagger. Il pourrait ainsi être intéressant de se baser sur ces caractéristiques. Dans cette même perspective, il serait possible d'utiliser un autre étiqueteur morpho-syntaxique, tel que *Cordial Analyseur* [8]. Il serait aussi possible d'utiliser les lemmes, à la place de ou en combinaison avec les CMS. La difficulté principale de toutes ces approches serait le risque d'obtenir des matrices creuses et donc de rencontrer, à nouveau, le problème du fléau de la dimension. Il faudrait aussi, dans la perspective d'obtenir un système totalement indépendant d'un annotateur, définir une méthode de segmentation automatique du texte en propositions.

Une autre étape supplémentaire qu'il faudrait envisager est la prise en compte de la structure hiérarchique des types de discours (cf. section 4.1.1.7), car seules les feuilles de la structure ont été utilisées ici. Par exemple, il serait intéressant de déterminer le type de discours dominant pour chaque proposition, ce qui devrait d'abord être défini par un expert humain. Ainsi, il serait possible de travailler sur des unités plus longues que les propositions. Aussi, le type de discours injonctif étant systématiquement inclus dans le type dialogal à l'intérieur de notre corpus, il pourrait être supprimé pour obtenir un groupe dialogal plus important.

Finalement, il serait idéal d'obtenir plus de textes annotés, ce qui permettrait d'améliorer les résultats et d'utiliser les méthodes proposées ci-dessus. Il faudrait aussi un second annotateur, au minimum, pour pouvoir mesurer la difficulté de la tâche d'annotation pour un expert humain.

7. `http://www.tropes.fr/`
8. `http://www.synapse-fr.com/Cordial_Analyseur/Presentation_Cordial_Analyseur.htm`

CHAPITRE 5

Classification supervisée multi-étiquette en actes de dialogue

Ce chapitre reprend, presque intégralement, l'article Cocco (2014), en présentant quelques résultats supplémentaires. La visée de ce chapitre est la classification supervisée multi-étiquette en actes de dialogue des tours de parole des contributeurs aux pages de discussion de *Simple English Wikipedia* (Wikipédia en anglais simple).

Les articles de Wikipédia sont créés par ses contributeurs, qui partagent leurs informations et leurs critiques sur des pages de discussion, chaque article étant lié à une page de discussion. Ces discussions fournissent une base de données que Ferschke, Gurevych et Chebotar (2012) ont segmentée, pour *Simple English Wikipedia*, en *tours de parole*, définis comme les interventions successives des intervenants. Ils ont ensuite annoté ces tours de parole avec des actes de dialogue (section 5.1).

De nombreux travaux (voir par exemple Stolcke *et al.*, 2000) se sont intéressés à la classification de dialogues écrits ou oraux en actes de dialogue (*dialogue acts*) ou en actes de langage ou de discours (*speech acts*), servant à caractériser la fonction d'un énoncé dans un dialogue (Austin, 1962; Searle, 1969). Les actes de dialogue peuvent être différents selon le but de la classification (pour une comparaison des principaux actes de dialogue et de langage utilisés, voir par exemple Goldstein et Sabin, 2006). Ferschke *et al.* (2012) utilisent leur propre jeu d'étiquettes d'actes de dialogue avec pour but de comprendre les « efforts de coordination pour l'amélioration d'un article ». Dans un second temps, ils ont procédé à une classification multi-étiquette. En général, un acte de dialogue est attribué à chaque énoncé, ce qui conduit à une classification ordinaire mono-étiquette. Dans ce jeu de données, les tours de parole, pouvant être composés de plusieurs énoncés, sont étudiés et peuvent donc se voir attribuer un ou plusieurs actes de dialogue, ce qui conduit à une classification multi-étiquette (cf. section 2.3.2 et 5.3.2.1) des tours de paroles en actes de dialogue. Pour examiner la cohérence de ces annotations et pour déterminer une méthode de classification, on commence ici par analyser les relations entre les étiquettes (section 5.2).

Concernant les actes de dialogue, Colineau et Caelen (1995) distinguent quatre types de marqueurs :

— linguistiques (morphologiques, syntaxiques et lexicaux),
— prosodiques,
— situationnels (phases du dialogue et règles d'enchaînement préférentiel) et
— du geste.

Ici, le jeu de données contient exclusivement des textes écrits, sans annotation des actions qui découlent du dialogue ; ainsi seuls les marqueurs linguistiques et situationnels peuvent être employés. Ferschke *et al.* (2012) utilisent les deux types de marqueurs, *i.e.* des uni-, des bi- et des

trigrammes (linguistiques), d'une part, et le temps entre les tours de parole, leur indentation, etc. (situationnels), d'autre part, puis les combinent. Ce travail propose d'utiliser trois autres caractéristiques (*features*), toutes de nature linguistique, et de les étudier séparément pour mieux comprendre l'impact de chacune d'entre elles, sans visée de performance globale. Les trois types de caractéristiques employées sont (section 5.3.1) :

- les **lemmes** (unigrammes), donnant des résultats légèrement meilleurs que les mots-formes dans la classification en actes de dialogues de messages dans des *chats* (Kim, Cavedon et Baldwin, 2010) ;
- les **catégories morphosyntaxiques** (CMS), dont l'intérêt pour la classification en actes de dialogue est démontré dans plusieurs travaux (voir par exemple Cohen, Carvalho et Mitchell, 2004; Boyer, Ha, Phillips, Wallis, Vouk et Lester, 2010) ; et
- le **sens des verbes selon WordNet** (Fellbaum, 1998). Deux articles, l'un étudiant la classification de messages sur des forums (Qadir et Riloff, 2011), l'autre la classification d'e-mails (Goldstein et Sabin, 2006), concluent que des classes de verbes (selon des listes prédéfinies) aident à la reconnaissance de certains actes de langage. L'idée, un peu différente ici, est de voir si les classes recréées à l'aide de WordNet permettent une telle reconnaissance dans le jeu de données étudié.

Finalement, concernant la méthode de classification, alors que les auteurs du jeu de données ont employé des approches classiques, telles que le classifieur Bayésien naïf ou les Séparateurs à Vastes Marges (SVM), ce travail utilise l'analyse discriminante linéaire, étendue aux transformations de Schoenberg. Les résultats ainsi obtenus sont exposés dans la section 5.3.3, puis les extensions possibles de la méthode sont discutées dans la section 5.4.

5.1 Données

Les données utilisées dans ce projet sont celles de Ferschke *et al.* (2012) et mises librement à disposition sur Internet (`http://www.ukp.tu-darmstadt.de/data/wikidiscourse`). Comme déjà expliqué ci-dessus, elles concernent les pages de discussion de Wikipédia en anglais simple. Une partie de ces pages de discussion ont été extraites, segmentées automatiquement en tours de parole (1450 au total), puis classifiées en actes de dialogue. Pour cette dernière étape, deux annotateurs ont classifié l'ensemble du corpus. Ensuite, dans les cas où les deux annotateurs n'étaient pas d'accord, un troisième annotateur expert a pris la décision finale, ce qui a permis constituer un corpus de référence (pour la structure des données et le détail, voir Ferschke *et al.*, 2012).

Les étiquettes qu'ils ont utilisées se divisent en quatre groupes principaux, lesquelles se subdivisent en un jeu de 17 étiquettes, soit [1] :

- Les étiquettes **interpersonnelles** (*Interpersonal*) « décrivent l'attitude qui est exprimée envers les autres participants dans la discussion et/ou les commentaires ». Ces étiquettes se divisent en trois sous-étiquettes :
  - « une approbation ou un rejet partiel » (ATTP),
  - « une attitude négative envers un autre participant ou un rejet » (ATT-) et
  - « une attitude positive envers un autre participant ou une approbation » (ATT+).
- Les étiquettes de **critique d'articles** (*Article Criticism*) « dénotent les commentaires qui identifient des insuffisances dans l'article. La critique peut porter sur l'article entier ou sur une partie de l'article ». Cet ensemble se subdivise en sept parties :
  - « les insuffisances de langage ou de style » (CL),
  - « un contenu incomplet ou un manque de détail » (CM),
  - « d'autres sortes de critiques » (CO),

1. Les définitions de ce paragraphe sont une traduction personnelle des définitions proposées dans Ferschke *et al.* (2012). Des exemples de tours de parole appartenant à chacune de ces 17 étiquettes et extraites du jeu de données se trouvent dans leur article.

  - « des problèmes objectifs » (COBJ),
  - « des problèmes structurels » (CS),
  - « un contenu inapproprié ou inutile » (CU) et
  - « le manque de précision ou d'exactitude » (CW).
- Les étiquettes sur le **contenu informationnel** (*Information Content*) « décrivent la direction de la communication ». Elles se divisent en trois catégories :
  - « une correction de l'information » (IC),
  - « un apport d'information » (IP) et
  - « une demande d'information » (IS).
- Les étiquettes de **performativité explicite** (*Explicit Performative*) concernent « l'annonce, le rapport ou la suggestion d'activités d'édition ». Elles se divisent en quatre sous-catégories :
  - « un engagement à une action dans le futur » (PFC),
  - « le rapport d'une action accomplie » (PPC),
  - « une référence explicite ou un indicateur » (PREF) et
  - « une suggestion, une recommandation ou une demande explicite » (PSR).

5.2 Liens entre étiquettes

Chaque tour de discussion pouvant avoir plusieurs étiquettes ou appartenir à plusieurs groupes $g = 1, \ldots, m$, il semblait pertinent de commencer par déterminer s'il existe des liens entre ces étiquettes. En plus de permettre une meilleure compréhension de l'annotation et de sa cohérence, cette première étude permet de choisir une méthode de classification multi-étiquette appropriée, *i.e.* prenant en compte ou non le lien entre les étiquettes (cf. section 5.3.2.1).

5.2.1 Traitements

Pour mesurer le lien qui existe entre deux étiquettes (ou classes ou groupes) g et g', on utilise les indices présentés dans la section 1.2.2, et en particulier, le *coefficient phi* (cf. section 1.2.2.2) et le *Q de Yule* (cf. section 1.2.2.3).

Pour ce faire, une table de contingence 2×2 a été créée pour chaque paire d'étiquettes, représentant le nombre d'absences et de présences (codées 0 et 1) simultanées de chaque classe pour chaque étiquette $i = 1, \ldots, n$, comme présenté dans la table 1.2. Dans cette table, la variable catégorielle X possède deux modalités, soit la présence et l'absence de g, et la variable Y, la présence et l'absence de g'. Ceci nous permet de calculer $\phi_{gg'}$ (1.4) et $Q_{gg'}$ (1.5).

Dans un second temps, à partir de la matrice des corrélations entre toutes les classes $\Phi = (\phi_{gg'})$, une analyse en composantes principales (ACP) (voir par exemple Lebart *et al.*, 1995, section 1.2) a été effectuée afin de visualiser les relations entre les différentes étiquettes et étudier la diversité de ces dernières. Pour pratiquer l'ACP, on utilise la fonction « PCA » du package « FactoMineR » (Lê, Josse et Husson, 2008; Husson *et al.*, 2013) de R.

5.2.2 Résultats

Les résultats pour le coefficient phi et le Q de Yule sont présentés dans la table 5.1. Pour les coefficients phi, la valeur maximale de 0.358 est obtenue pour la paire d'étiquettes CS et PSR, ce qui signifie que, souvent, les tours de parole classés comme parlant de problèmes structurels sont aussi classés comme constituant une suggestion, une recommandation ou une demande explicite, et inversement, ce qui semble cohérent. Quant à la valeur minimale de -0.306, elle se produit entre les classes IP et PFC. Cela suggère qu'en général, si un tour de parole apporte de l'information, il ne propose pas en même temps un engagement à une action dans le futur.

En ce qui concerne le Q de Yule, la valeur maximale de 0.925 est atteinte pour les classes IP et IC, ce qui signifie qu'une des classes est presque incluse dans l'autre ; en fait, IC est presque

	ATTP	ATT-	ATT+	CL	CM	CO	COBJ	CS	CU
ATTP		-0.039	-0.051	-0.051	-0.028	-0.028	0.047	-0.049	-0.026
ATT-	**-1**		-0.055	-0.107*	-0.053	-0.047	0.008	-0.071*	-0.026
ATT+	**-1**	-0.527		-0.089*	-0.013	-0.010	0.022	-0.051	-0.030
CL	-0.707	**-1**	-0.532		0.018	-0.046	0.056	0.043	-0.004
CM	-0.477	-0.590	-0.084	0.086		0.031	-0.003	0.123*	0.010
CO	**-1**	**-1**	-0.099	-0.464	0.253		0.003	-0.020	-0.032
COBJ	0.564	0.115	0.229	0.415	-0.042	0.059		-0.009	0.067*
CS	**-1**	-0.809	-0.364	0.183	0.503	-0.222	-0.130		0.001
CU	**-1**	-0.455	-0.383	-0.034	0.098	**-1**	0.632	0.009	
CW	**-1**	-0.381	-0.301	-0.034	-0.064	-0.417	0.229	-0.271	0.473
IC	0.008	0.204	-0.670	0.817	-0.152	0.279	-0.105	-0.118	-0.333
IP	0.842	0.723	0.232	0.722	0.605	0.287	0.638	0.663	0.760
IS	-0.288	-0.358	-0.534	0.132	0.284	0.410	-0.387	0.042	0.281
PFC	0.435	-0.424	0.584	-0.370	0.074	**-1**	0.180	-0.320	-0.059
PPC	-0.196	-0.597	-0.144	-0.742	-0.736	-0.576	-0.311	-0.776	-0.523
PREF	0.347	0.058	-0.415	-0.594	-0.594	-0.207	**-1**	-0.648	-0.139
PSR	-0.722	-0.562	-0.168	0.683	0.810	0.583	0.418	0.845	0.528

	CW	IC	IP	IS	PFC	PPC	PREF	PSR
ATTP	-0.034	0.001	0.080*	-0.026	0.046	-0.023	0.026	-0.075*
ATT-	-0.030	0.033	0.118*	-0.050	-0.034	-0.098*	0.005	-0.099*
ATT+	-0.032	-0.080*	0.056	-0.089*	0.137*	-0.035	-0.033	-0.043
CL	-0.005	0.353*	0.190*	0.036	-0.048	-0.188*	-0.053	0.303*
CM	-0.007	-0.021	0.118*	0.067*	0.010	-0.133*	-0.038	0.309*
CO	-0.024	0.036	0.040	0.072*	-0.044	-0.070*	-0.011	0.125*
COBJ	0.017	-0.008	0.059*	-0.030	0.013	-0.032	-0.025	0.062*
CS	-0.028	-0.018	0.138*	0.009	-0.034	-0.151*	-0.044	**0.358***
CU	0.057	-0.025	0.084*	0.042	-0.004	-0.061*	-0.007	0.103*
CW		0.222*	0.120*	0.021	0.033	-0.084*	0.034	0.060*
IC	0.758		0.176*	-0.072*	-0.053	-0.124*	0.013	0.159*
IP	0.855	**0.925**		-0.115*	0.099*	**-0.306***	0.089*	0.295*
IS	0.128	-0.438	-0.322		-0.023	-0.149*	-0.031	0.002
PFC	0.267	-0.588	0.632	-0.157		-0.064*	-0.007	-0.024
PPC	-0.570	-0.622	-0.627	-0.563	-0.389		-0.066*	-0.293*
PREF	0.329	0.123	0.776	-0.301	-0.096	-0.551		-0.050
PSR	0.277	0.497	0.802	0.005	-0.127	-0.825	-0.366	

TABLE 5.1 – Pour toutes les paires d'étiquettes, g et g', coefficients $\phi_{gg'}$, suivis d'une étoile pour les valeurs significatives au niveau $\alpha = 5\%$ (matrice triangulaire supérieure) et $Q_{gg'}$ (matrice triangulaire inférieure). Les valeurs maximales et minimales de chaque coefficient sont notées en gras.

incluse dans IP, car cette dernière a été assignée à la grande majorité des tours de parole, soit 78,3 % (Ferschke *et al.*, 2012). Ainsi, la plupart des tours de parole proposant une correction de l'information, amènent aussi de l'information. Aussi, la majorité de cette classe IP devrait impliquer que la plupart des autres classes soient, en parties, incluses dans celle-ci. En effet, on observe que le Q de Yule est positif entre la classe IP et chaque autre classe, à l'exception des classes IS et PPC.

Quant à la valeur minimale de -1, elle est obtenue pour plusieurs paires de classes. Cela signifie, pour rappel (cf. section 1.2.2.3), que soit aucun tour de parole n'appartient simultanément aux deux classes, soit tous les tours de parole appartiennent à au moins une des deux classes. En fait, il s'agit du premier cas pour toutes les paires de classes. En particulier, on remarque qu'une approbation ou un rejet partiel (ATTP) exclut une attitude négative (respectivement positive) envers un autre participant ou un rejet (resp. une approbation) (ATT- resp. ATT+), diverses critiques (CO), des problèmes structurels (CS), un contenu inapproprié (CU) ou la manque d'exactitude (CW). Cependant, cette exclusion, qui pourrait sembler utile à la classification, est certainement due au fait que l'étiquette ATTP est peu présente dans le corpus (elle est attribuée à seulement 2.4 % des tours de parole selon Ferschke *et al.*, 2012).

Finalement, comme il a été exposé dans la section précédente, une ACP a été effectuée

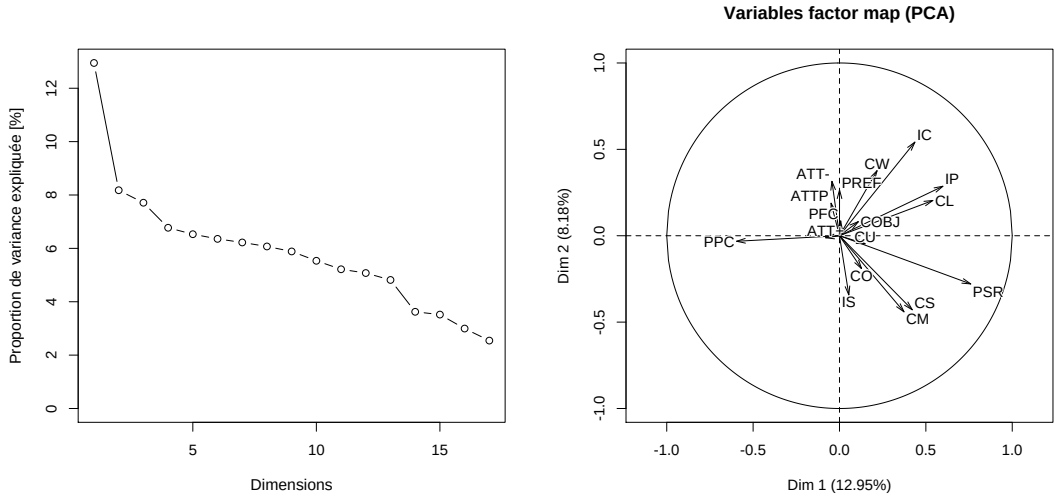


FIGURE 5.1 – ACP sur la matrice des corrélations Φ : proportion de variance expliquée par chaque dimension (gauche) et cercle des corrélations (droite).

sur la matrice des corrélations Φ. Le diagramme des valeurs propres de la figure 5.1 (gauche) montre qu'un faible pourcentage de la variance totale est expliqué par les deux premiers facteurs (moins de 22%), ce qui signifie que les étiquettes sont diversifiées et que l'information qu'elles contiennent peut difficilement être compressée. Le cercle des corrélations (figure 5.1 droite) est difficilement interprétable, un phénomène attendu au vu de la non significativité d'un grand nombre de coefficients phi.

5.3 Classification supervisée

5.3.1 Prétraitements et caractéristiques

Comme déjà mentionné dans l'introduction de ce chapitre, les caractéristiques utilisées dans ce travail sont uniquement linguistiques. La première étape du prétraitement, exécuté à l'aide d'un programme écrit en Perl, a consisté à « nettoyer » les données de Ferschke *et al.* (2012), pour en enlever les balises HTML (concernant principalement la mise en forme) [2], les ponctuations découlant de la mise en forme du texte, les informations concernant les utilisateurs, l'heure à laquelle le tour de parole a été posté, les symboles indiquant l'indentation du tour de parole par rapport au premier tour de parole de la discussion, les éléments concernant les liens et les tags dans Wikipédia et les divers symboles, tels que des coeurs ou des lettres grecques, car ces derniers n'étaient pas compatibles avec la suite du prétraitement. Aussi, les caractères spéciaux en code HTML ont été remplacés par les caractères correspondants en latin1.

Ensuite, trois types de caractéristiques ont été extraites pour chaque tour de parole : les lemmes, les CMS, et le sens des verbes (selon WordNet). Les lemmes et les CMS ont été extraits à l'aide de TreeTagger (Schmid, 1994) [3], à l'aide du même module Perl que celui exposé dans la section 4.1.3.

L'extraction du sens des verbes a été effectuée à l'aide de WordNet et TreeTagger. Dans WordNet, les verbes sont organisés sous forme d'ontologie par des relations sémantiques, dont

2. Par conséquent, les quelques balises faisant partie d'une question ou d'une réponse sur leurs utilisations, par exemple, ont aussi été supprimées.

3. À ce stade, quelques petites modifications ont été apportées au texte pour qu'il soit plus correctement étiqueté par TreeTagger. En particulier, des espaces ont été ajoutés entre certains symboles et les mots qu'ils entouraient ; les guillemets simples ont été remplacés par des guillemets doubles ; et les symboles, tels que « = », répétés deux ou trois fois, ont été remplacés par une seule occurrence de ce même symbole.

l'hyperonymie (pour plus d'information, voir section 6.3.1). Aussi, l'ensemble des concepts de verbes n'ont pas une seule racine commune, consistant en un seul plus petit hyperonyme commun. Ainsi, pour chaque tour de parole, les lemmes des mots considérés comme des verbes par TreeTagger ont été soumis à WordNet, par l'intermédiaire du module Perl `WordNet::QueryData` (Rennie, 2000). En particulier, le premier sens du verbe proposé, pour des raisons d'automatisation, a été retenu, puis l'hyperonyme le plus général a été conservé et ce dernier hyperonyme est retenu comme caractéristique de ce tour de parole. Les verbes modaux ne sont pas traités par WordNet. Cependant, au vu de leur importance supposée pour la classification en actes de dialogue, il semblait intéressant de les ajouter explicitement au même titre que les hyperonymes traités par WordNet.

À ce stade, trois tables de contingence sont créées : tours de parole - CMS, tours de parole - lemmes et tours de paroles - verbes (hyperonyme le plus général ou un des verbe modaux), comptant le nombre d'occurrences de chaque caractéristique par tour de parole. Les tours de parole qui n'étaient pas étiquetés ont été supprimés ; il s'agissait généralement de tours de parole soit trop longs et contenant toutes sortes d'informations, soit écrits en français ou encore mal segmentés. Les tours de parole ne contenant aucune des caractéristiques décrites plus haut ont également été supprimés. Au final, la base de données a été réduite de 1'450 à 1'324 tours de parole, contenant 5'198 lemmes distincts, 57 CMS distinctes et 155 sens de verbes distincts.

5.3.2 Traitements

5.3.2.1 Classification multi-étiquette

Deux types d'approche sont couramment pratiqués pour la classification multi-étiquette (Tsoumakas, Katakis et Vlahavas, 2010) : le premier (*problem transformation*) consiste à recoder le jeu de données pour le transformer en problème de classification ordinaire, sans modification des algorithmes de classification ; le second (*algorithm adaptation*) adapte les algorithmes pour qu'ils puissent directement traiter des données multi-étiquette.

Pour ce travail, il a été choisi d'utiliser le premier traitement, *i.e.* le recodage des données. Parmi les nombreux recodages possibles, celui du recodage binaire (*Binary Relevance* (BR)) a été choisi. Cela signifie que chaque tour de parole sera classé de façon binaire, *i.e.* comme faisant partie ou non d'une classe donnée (avec un classifieur pour chaque étiquette). Bien que ce recodage soit parfois critiqué, car il ne prend pas en compte les dépendances entre les étiquettes, il a ici plusieurs avantages :

— il permet de rendre les résultats comparables à ceux de Ferschke *et al.* (2012) qui utilisent le même principe ;
— il a le mérite, en plus d'avoir une complexité computationnelle faible, d'être simple, intuitif, résistant au surapprentissage des combinaisons d'étiquettes et de pouvoir traiter les étiquetages irréguliers (Read, Pfahringer, Holmes et Frank, 2011) ; et
— il est particulièrement adapté aux situations où il n'y a pas de dépendance entre les étiquettes, ce qui semble être le cas ici (cf. section 5.2.2).

Par ailleurs, Luaces, Dìez, Barranquero, del Coz et Bahamonde (2012) proposent un indice qui mesure la dépendance entre toutes les étiquettes comme la moyenne des corrélations $\phi_{gg'}$ pour chaque paire d'étiquettes g et g', pondérée par le nombre d'individus (ici les tours de parole) communs $|g \cap g'|$:

$$\text{dépendance} = \frac{\sum_{g<g'} \phi_{gg'} |g \cap g'|}{\sum_{g<g'} |g \cap g'|}$$

Pour le jeu d'étiquettes de ce travail, cette dépendance vaut 0.10, ce qui correspond à la valeur la plus faible trouvée par Luaces *et al.* (2012) dans la vingtaine de jeux de données qu'ils examinent, et qui conforte ainsi le choix de la méthode BR.

Étiquette	Taille
ATTP	64
ATT-	174
ATT+	276

Étiquette	Taille
CL	422
CM	228
CO	96
COBJ	54
CS	258
CU	82
CW	140

Étiquette	Taille
IC	260
IP	554
IS	424

Étiquette	Taille
PFC	154
PPC	694
PREF	88
PSR	798

TABLE 5.2 – Taille, en nombre de tours de paroles, de l'échantillon équilibré (exemples positifs et négatifs) pour chaque étiquette.

Finalement, étant donné que le nombre de tours de parole appartenant ou non à une étiquette est très variable, à l'instar de Ferschke *et al.* (2012), des échantillons équilibrés ont été constitués, contenant, pour une étiquette donnée, un nombre identique de tours de parole lui appartenant ou non. Cet échantillon a été constitué une fois pour toutes pour chacune des classes et le nombre de tours de parole retenu pour chaque étiquette se trouve dans la table 5.2. Ce choix a été effectué pour éviter que les étiquettes les plus fréquentes soient plus facilement attribuées lors de la classification et inversement. Cependant, alors que le principe de la méthode BR est de sélectionner un individu, de le faire passer dans les différents classifieurs et d'obtenir toutes les étiquettes de cet individu, le choix des échantillons équilibrés conduit à une série de classifications séparées, ce qui influencera le choix de la méthode d'évaluation des résultats.

5.3.2.2 Algorithme de classification supervisée et évaluation

Pour **la classification**, l'analyse discriminante linéaire, telle qu'elle est présentée dans la section 2.2.1, a été appliquée, combinée avec une validation croisée sur 5 sous-échantillons (contre 10 utilisés par les auteurs de la base de données). Plus précisément, pour chacune des trois tables de contingence, tours de parole ($i = 1, \ldots, n$) - caractéristiques ($k = 1, \ldots, p$), une classification discriminante binaire, avec comme groupes g, le fait d'appartenir ou non à une classe, a été effectuée cinq fois pour chaque étiquette. Pour rappel, lors de l'utilisation de l'analyse discriminante sur des dissimilarités calculées à partir d'une table de contingence, les colonnes non présentes dans l'ensemble d'apprentissage doivent être supprimées. Pour le sens des verbes, cela a engendré la suppression de certains tours de paroles ne contenant plus aucun verbe, réduisant les tailles d'échantillon présentées dans la table 5.2 d'une unité pour les étiquettes suivantes : ATT-, CL, CS, IP, PFC et PPC.

En particulier, les deux critères d'analyse discriminante ont été utilisés, soit celui des *plus proches voisins* (2.9) et celui du *plus proche centroïde* (2.10). De plus, les deux critères ont été combinés aux transformations de puissance de Schoenberg (1.22), selon la procédure décrite à la fin de la section 2.2.1, avec q allant de 0.5 à 1, avec des incréments de 0.1.

Pour **l'évaluation** des classifications multi-étiquette, on distingue deux familles de méthodes (Tsoumakas *et al.*, 2010) : celles basées sur les individus (*example-based*) et celles basées sur les étiquettes (*label-based*). Pour pouvoir utiliser la première famille de méthodes, la classification doit déterminer, pour chaque individu, ici les tours de parole, l'ensemble des étiquettes qu'il possède. En raison du choix consistant à sélectionner des échantillons équilibrés, on obtient, pour un tour de parole, l'appartenance ou non à une classe, et il ne s'agit pas toujours des mêmes tours de parole. Ainsi, la seconde famille de méthodes doit être employée. Parmi les différentes possibilités, les méthodes standards ont été appliquées : précision, rappel et F-mesure (cf. section 2.3.2).

Plus précisément, pour chaque étiquette g, les résultats des 5 validations croisées ont été agrégés pour former la matrice de confusion. Ensuite, la précision (2.13), le rappel (2.14) et la

F-mesure (2.15) ont été calculés pour chaque étiquette. Puis, pour évaluer la performance de la classification sur l'ensemble des classes, la macro-moyenne (2.16) et la micro-moyenne (2.17) des ces trois mesures ont été calculées.

5.3.3 Résultats

Les figures 5.2 à 5.7 présentent les résultats, pour les deux critères d'analyse discriminante, de la F-mesure pour chaque étiquette (gauche) et de la micro- et la macro-moyenne de la précision, du rappel et de la F-mesure pour l'ensemble des étiquettes (droite).

Globalement, dans les 6 figures, il est remarquable que la classification de l'étiquette IP donne la meilleure F-mesure[4]. Ce résultat semble cohérent avec ceux de Ferschke *et al.* (2012) qui obtiennent toujours la meilleure F-mesure pour cette étiquette, quelle que soit la méthode utilisée. Un autre point récurrent pour les trois cas et les deux critères, lors de l'évaluation pour l'ensemble des étiquettes (graphiques de droite), est que les micro- et macro- moyennes donnent des résultats très similaires pour la précision, le rappel et la F-mesure. Ceci est dû au fait que les résultats pour chacune des étiquettes sont très proches, indépendamment de la taille de l'échantillon. Aussi, le rappel est toujours plus élevé que la précision. Cela signifie que le nombre de faux positifs (tours de parole n'appartenant pas à une classe mais étiquetés comme y appartenant) est plus élevé que le nombre de faux négatifs (tours de parole appartenant à une classe mais étiquetés comme n'y appartenant pas).

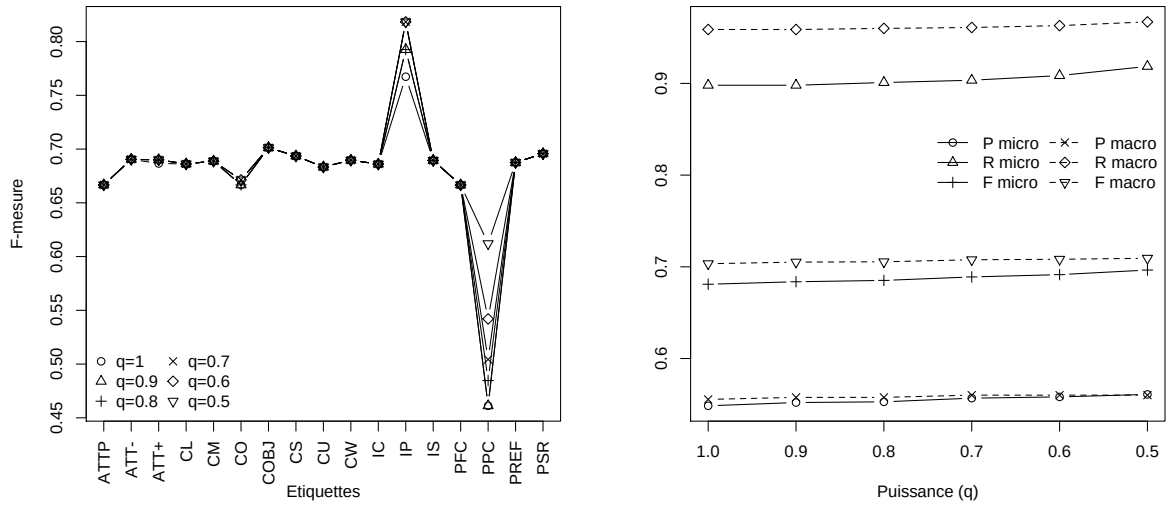


FIGURE 5.2 – Classification avec les lemmes et le critère des plus proches voisins. F-mesure pour chaque étiquette selon la puissance q (gauche). Macro- et micro-moyennes pour la précision, le rappel et la F-mesure (droite).

Avec les lemmes comme caractéristiques (figures 5.2 et 5.3), l'étiquette PPC révèle clairement un résultat moins bon que ceux obtenus avec les autres étiquettes (graphiques de gauche). Aussi, on remarque, pour l'ensemble des étiquettes (graphiques de droite), qu'à l'exception du rappel, le critère du plus proche centroïde produit des résultats systématiquement plus élevés que le critère des plus proches voisins.

Plus spécifiquement, avec le critère des plus proches voisins (figure 5.2), on observe que l'utilisation de la puissance q influence peu les résultats pour la plupart des étiquettes, à l'exception

4. Précisons que ce résultat n'est pas dû à la fréquence élevée de cette étiquette dans le jeu de données, car des échantillons équilibrés ont été considérés (cf. table 5.2).

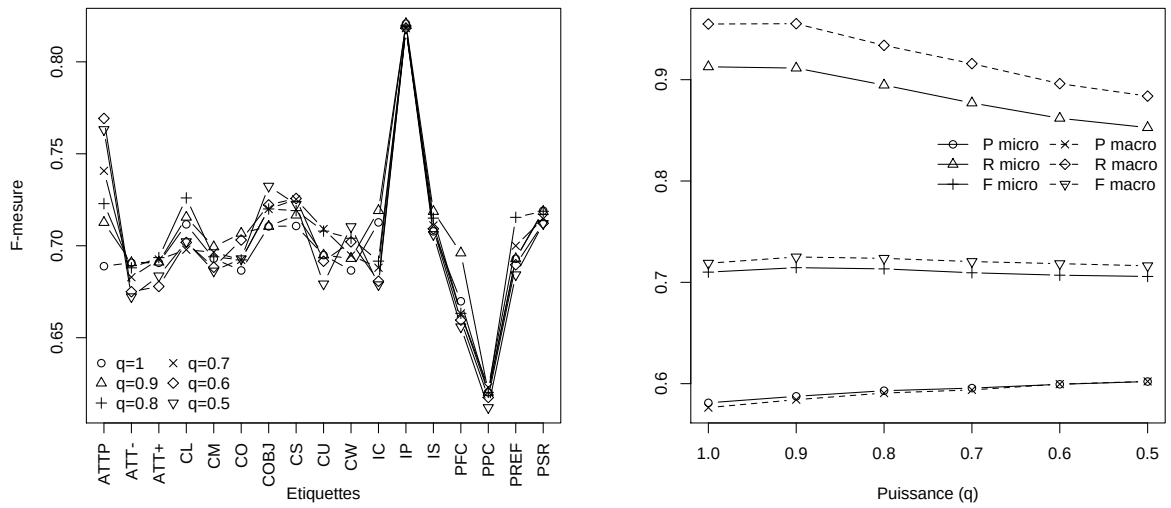


FIGURE 5.3 – Classification avec les lemmes et le critère du plus proche centroïde. F-mesure pour chaque étiquette selon la puissance q (gauche). Macro- et micro-moyennes pour la précision, le rappel et la F-mesure (droite).

des étiquettes IP et PPC, correspondant toutes deux aux résultats les plus extrêmes (graphique de gauche). Pour ces deux étiquettes, l'influence de q est positive, car la valeur maximale de la F-mesure, $F = 0.82$ (respectivement $F = 0.61$), est obtenue lorsque $q \leq 0.7$ (resp. $q = 0.5$) pour l'étiquette IP (resp. PPC). Au niveau de l'ensemble de la classification (graphique de droite), la puissance q améliore tous les résultats, bien que de façon peu prononcée. Au final, le meilleur résultat, obtenu avec la macro-moyenne de la F-mesure, vaut 0.71 lorsque $q = 0.5$.

À l'inverse, avec le critère du plus proche centroïde (figure 5.3), l'utilisation de la transformation de puissance q influence les résultats de façon marquée en général. Pour l'ensemble des étiquettes (graphique de droite), elle améliore la précision de la classification ($P_{macro} = 0.58$ pour $q = 1$ et $P_{macro} = 0.60$ pour $q = 0.5$), mais diminue le rappel. Au final, la meilleure $F_{macro} = 0.73$ est obtenue pour $q = 0.9$. En regardant les résultats obtenus pour chaque étiquette (graphique de gauche), l'intérêt de la puissance q est plus marqué : $q = 0.5$ donne les meilleures F-mesures pour les étiquettes COBJ et CW ; $q = 0.6$, pour ATTP et CS ; $q = 0.7$, pour CU et PPC ; $q = 0.8$, pour ATT+, CL et PREF ; $q = 0.9$, pour CM, CO, IC, IS et PFC. En particulier, l'amélioration de la F-mesure obtenue pour l'étiquette ATTP est importante passant de 0.69 pour $q = 1$ à 0.77 pour $q = 0.6$.

Avec, comme caractéristiques, les CMS (figures 5.4 et 5.5), les moins bons résultats sont obtenus pour l'étiquette PPC (graphiques de gauche), comme pour les lemmes. Et à nouveau, pour l'ensemble de la classification (graphiques de droite), le critère du plus proche centroïde engendre de meilleurs résultats, excepté pour le rappel.

Concernant le critère des plus proches voisins (figure 5.4), la puissance q n'influence absolument pas les résultats. Cela signifie que malgré la transformation, les tours de parole des ensembles de tests sont toujours proches des mêmes tours de parole des ensembles d'apprentissage lorsque q varie. Ainsi, F_{macro} vaut constamment 0.72.

Avec le critère du plus proche centroïde (figure 5.5), comme pour les lemmes, l'amélioration apportée par la transformation de puissance q est difficilement visible sur le résultat global (graphique de droite), même si l'on remarque que le meilleur résultat, $F_{macro} = 0.74$, est obtenu pour $q = 0.6$. Ce dernier est aussi le meilleur sur l'ensemble des résultats. Au niveau des étiquettes (graphique de gauche), la transformation q améliore les résultats pour la plupart des

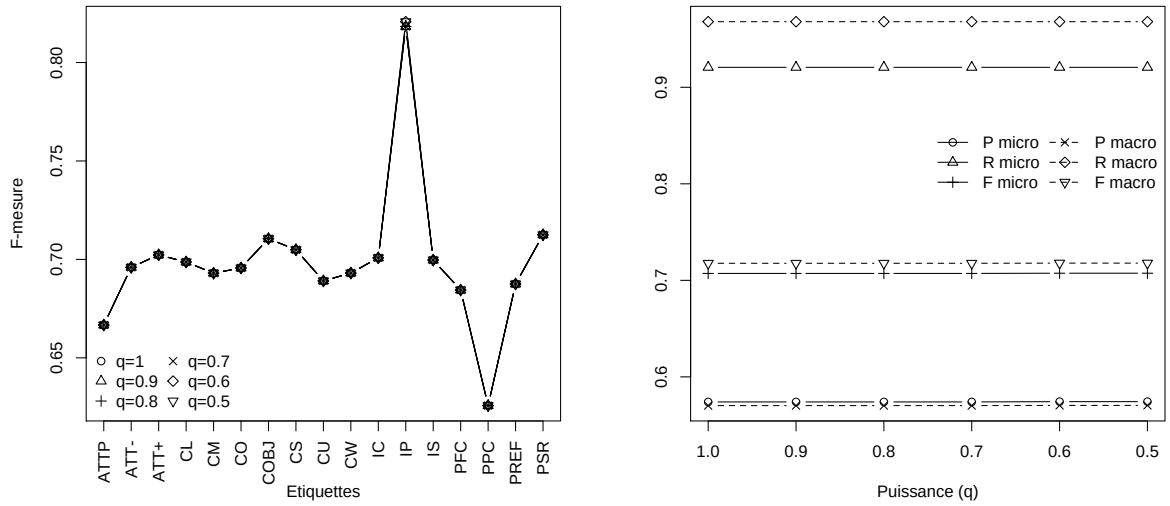


FIGURE 5.4 – Classification avec les CMS et le critère des plus proches voisins. F-mesure pour chaque étiquette selon la puissance q (gauche). Macro- et micro-moyennes pour la précision, le rappel et la F-mesure (droite).

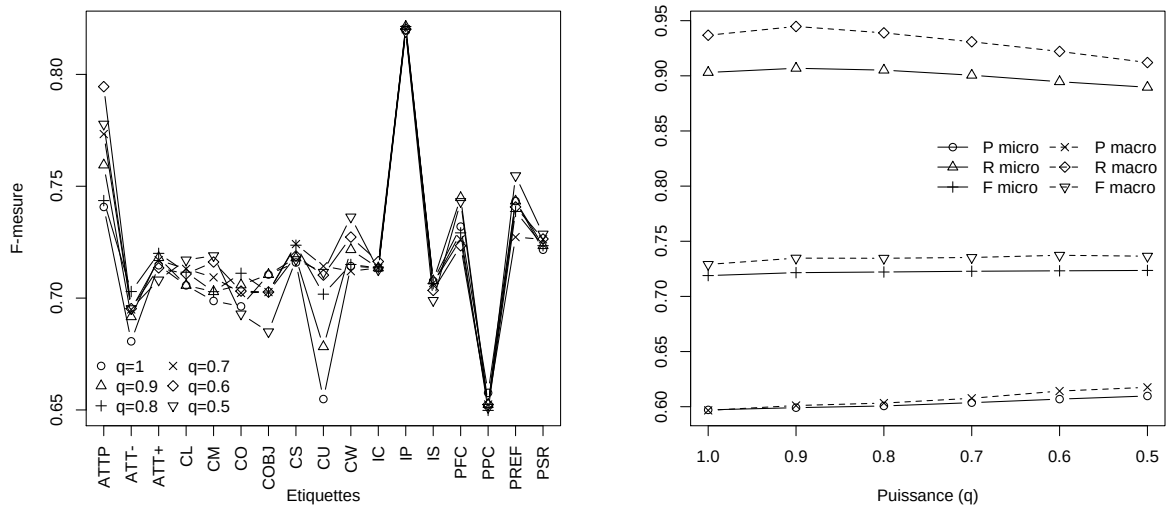


FIGURE 5.5 – Classification avec les CMS et le critère du plus proche centroïde. F-mesure pour chaque étiquette selon la puissance q (gauche). Macro- et micro-moyennes pour la précision, le rappel et la F-mesure (droite).

étiquettes (toutes sauf les étiquettes COBJ, IS et PPC), et particulièrement pour les étiquettes ATTP ($F = 0.74$ pour $q = 1$ et $F = 0.79$ pour $q = 0.6$) et CU ($F = 0.65$ pour $q = 1$ et $F = 0.71$ pour $q = 0.7$).

Au niveau de la classification pour l'ensemble des étiquettes avec le sens des verbes selon WordNet (figures 5.6 et 5.7, à droite), la première différence avec les autres caractéristiques est qu'ici les micro-moyennes donnent de meilleurs résultats que les macro-moyennes en général. Pour rappel, les micro-moyennes donnent plus d'importance aux classes comptant le plus d'individus et l'échantillon équilibré de l'étiquette PPC compte 694 tours de parole, ce qui en fait le

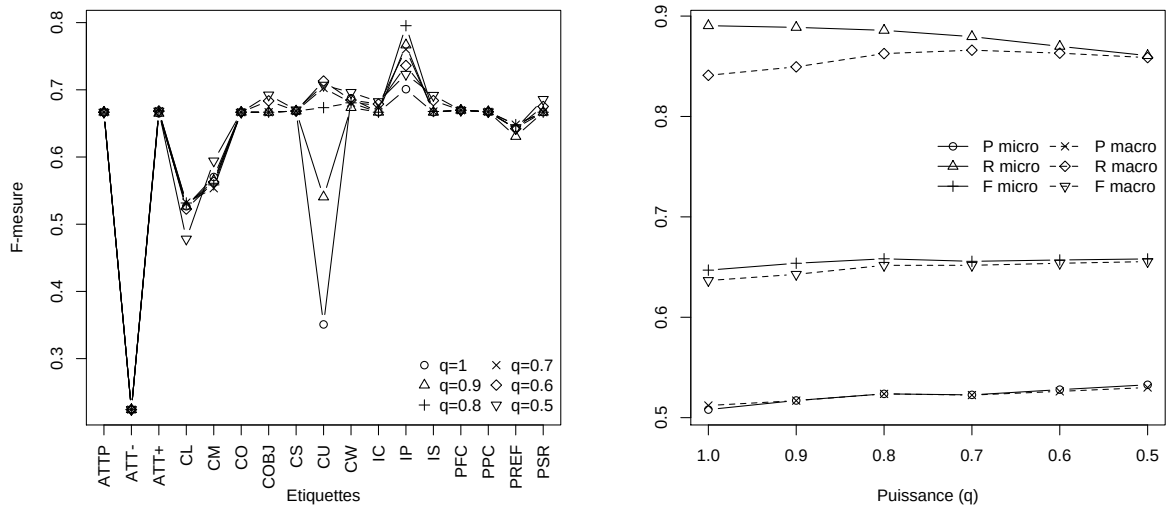


FIGURE 5.6 – Classification avec le sens des verbes et le critère des plus proches voisins. F-mesure pour chaque étiquette selon la puissance q (gauche). Macro- et micro-moyennes pour la précision, le rappel et la F-mesure (droite).

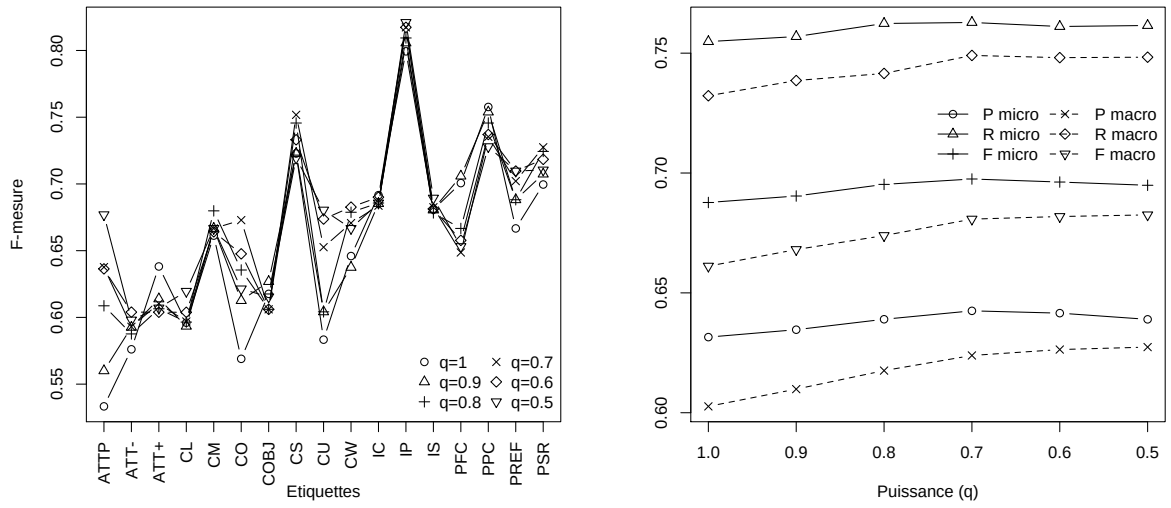


FIGURE 5.7 – Classification avec le sens des verbes et le critère du plus proche centroïde. F-mesure pour chaque étiquette selon la puissance q (gauche). Macro- et micro-moyennes pour la précision, le rappel et la F-mesure (droite).

deuxième plus grand. Contrairement aux résultats obtenus avec les deux autres caractéristiques, cette étiquette n'est pas moins bien classée (graphiques de gauche) que les autres avec le sens des verbes, ce qui explique certainement cette différence. De plus, comme pour les autres caractéristiques, pour l'ensemble des étiquettes, les résultats obtenus avec le critère de plus proche centroïde sont meilleurs que ceux obtenus avec celui des plus proches voisins, hormis pour le rappel. Cependant, à l'inverse des autres caractéristiques, les résultats de la classification par étiquette sont très différents selon le critère d'analyse discriminante utilisé.

À propos des résultats avec le critère des plus proches voisins (figure 5.6), on remarque,

pour la classification sur l'ensemble des étiquettes (graphique de droite), que la micro-moyenne du rappel diminue lorsque q diminue, alors que sa macro-moyenne augmente lorsque $q \leq 0.7$ ($R_{macro} = 0.87$), puis diminue faiblement. Pour la précision, le meilleur résultat ($P_{micro} = 0.53$) est obtenu lorsque $q = 0.5$, et pour la F-mesure, le meilleur résultat ($F_{micro} = 0.66$) apparaît pour $q = 0.8$. Concernant la classification par étiquette (graphique de gauche), la puissance q influence peu certaines étiquettes, telles que ATT- ou PPC, mais influence clairement les étiquettes CU et IP. Pour la première (respectivement la seconde), $F = 0.35$ (resp. $F = 0.70$) lorsque $q = 1$, alors que $F = 0.71$ (resp. $F = 0.80$) lorsque $q = 0.6$ (resp. $q = 0.8$).

Avec le critère du plus proche centroïde (figure 5.7, à droite), la meilleure classification est obtenue pour $q = 0.7$ ($F_{micro} = 0.70$). Concernant la classification par étiquette, comme pour les autres caractéristiques avec ce critère d'analyse discriminante, la figure 5.7 de gauche montre que la puissance q améliore les résultats pour plusieurs étiquettes (toutes sauf ATT+, IC, et PPC). Un autre point plus important, comme déjà mentionné, est que l'étiquette PPC, qui obtient des résultats assez faibles avec les lemmes ($F = 0.62$ pour $q = 0.7$) et les CMS ($F = 0.66$ pour $q = 1$), est bien mieux classifiée ici ($F = 0.76$ pour $q = 1$). Il en est de même pour l'étiquette CS ($F = 0.75$ pour $q = 0.7$, contre $F = 0.73$ pour $q = 0.6$ avec les lemmes et $F = 0.72$ pour $q = 0.7$ avec les CMS). Ainsi, même si le sens des verbes donne de moins bons résultats sur l'ensemble de la classification, cette caractéristique est plus discriminante pour ces deux étiquettes.

En résumé, il semble que le critère du plus proche centroïde produise, en général, de meilleurs résultats. Selon ce critère, la valeur maximale de la F-mesure vaut 0.74 (macro avec les CMS). Comme attendu, cette dernière est plus faible que celle obtenue par Ferschke *et al.* (2012), qui trouvent une F-mesure maximale de 0.82 par micro-moyenne (et de 0.73 par macro-moyenne). Cependant, elle reste tout à fait comparable et élevée, étant donné qu'ici les résultats se calculent avec une caractéristique à la fois, et sans combiner les meilleurs résultats obtenus pour chaque caractéristique [5].

5.4 Discussion

La première partie de ce chapitre concernant le lien entre les étiquettes (section 5.2) a montré que l'annotation semblait cohérente et que les liens, mêmes s'ils existent, ne sont en majorité pas significatifs. Cette dernière constatation a permis de choisir la méthode de classification multi-étiquette, à savoir BR. En associant ce choix à l'analyse discriminante et aux transformations de puissance de Schoenberg, la classification, avec le critère du plus proche centroïde, a donné de bons résultats pour les trois caractéristiques linguistiques choisies, mais plus particulièrement avec les CMS. Cependant, au vu des meilleurs résultats du sens des verbes selon WordNet avec certaines étiquettes, et des lemmes sur d'autres étiquettes, il serait intéressant de combiner ces caractéristiques, par exemple en mélangeant les distances avec différents poids β :

$$D_{ij}^{\text{tot}} = \beta_{\text{lemme}} D_{ij}^{\text{lemme}} + \beta_{\text{CMS}} D_{ij}^{\text{CMS}} + (1 - \beta_{\text{lemme}} - \beta_{\text{CMS}}) D_{ij}^{\text{verbe}} \tag{5.1}$$

De la même manière, il serait possible d'y ajouter des caractéristiques situationnelles, telles que celles utilisées par Ferschke *et al.* (2012), soit le temps entre les tours de parole, l'indentation entre les tours de parole, etc. De plus, la transformation de puissance montre une amélioration pour toutes les caractéristiques qui pourrait être aussi utilisée avant de mélanger les différentes distances. Il serait aussi intéressant d'explorer d'autres transformations de Schoenberg, susceptibles de donner de meilleurs résultats.

Une toute autre approche consisterait à explorer les liens entre étiquettes en utilisant une méthode d'adaptation de l'algorithme (au sens de la section 5.3.2.1) d'analyse discriminante (Park et Lee, 2008) afin qu'il puisse traiter globalement l'entièreté des étiquettes de chaque

5. Par contraste, Ferschke *et al.* (2012) assemblent toutes les caractéristiques, en font une sélection (*feature selection*) et combinent les meilleurs résultats obtenus avec différentes méthodes.

tour de parole. Ce choix pourrait être intéressant, car bien que ces liens entre étiquettes ne soient pas très importants selon le coefficient phi, le Q de Yule a permis de déterminer des exclusions entre certaines étiquettes qu'il faudrait exploiter.

Finalement, pour apprécier l'impact des caractéristiques proposées dans ce travail sur la performance, il faudrait les ajouter à celles utilisées par Ferschke *et al.* (2012) en utilisant les algorithmes employés par ces auteurs et disponibles dans WEKA (Hall, Frank, Holmes, Pfahringer, Reutemann et Witten, 2009).

CHAPITRE 6

Autocorrélation textuelle

Dans ce chapitre, l'indice d'autocorrélation (3.4) exposé dans la section 3.2 est appliqué à différentes caractéristiques mesurées sur des unités textuelles. Les résultats exposés sont adaptés de ceux présentés dans deux articles : Bavaud *et al.* (2012) et Bavaud, Cocco et Xanthos (accepté pour publication), dans lesquels se trouvent de plus amples détails concernant le formalisme, ainsi que d'autres résultats.

Comme expliqué dans la partie théorique, les deux éléments nécessaires pour calculer l'indice d'autocorrélation δ sont une matrice de dissimilarités euclidiennes carrées D et une matrice d'échange E (cf. section 3.1), qui définit le voisinage.

Le premier traitement proposé se base sur la différence de longueurs des mots (section 6.1). Ensuite, un deuxième exemple considère des dissimilarités basées sur la présence et l'absence de certaines parties du discours dans un voisinage donné (section 6.2). Finalement, un dernier exemple, plus sophistiqué, utilise les dissimilarités entre les sens des mots (section 6.3). Différents textes on été utilisés dans ce chapitre, en raison de leur disponibilité, l'objectif étant d'observer les propriétés génériques, plutôt que spécifiques, des textes.

6.1 Longueur des mots

6.1.1 Principe

Soit un texte composé de n mots, avec l_i, la longueur du mot à la position $i = 1, \ldots, n$. La dissimilarité euclidienne carrée entre les longueurs de deux mots aux positions i et j vaut $D_{ij}^{\text{long}} = (l_i - l_j)^2$. En considérant un texte comme linéaire, c'est-à-dire lu en continu de gauche à droite, les positions i représentent les positions successives des mots dans le texte, avec n, le dernier mot.

Par suite, avec la matrice des dissimilarités entre les longueurs des mots D^{long}, il est possible de calculer l'indice d'autocorrélation pour différents voisinages, tels que définis par une matrice d'échange (section 3.1).

6.1.2 Traitements et résultats

Pour illustrer l'alternance de la longueur des mots dans un texte, les $n = 2'000$ premiers mots (sur $180'610$ au total) de *Notre-Dame de Paris* de Victor Hugo, paru en 1831, ont été considérés. Concernant le voisinage, les trois matrices d'échange proposées dans la section 3.1.1 seront utilisées. Ainsi, ce premier cas permettra, d'une part, d'analyser l'autocorrélation qui

existe entre les longueurs des mots d'un texte ; et d'autre part, de comparer les trois différentes matrices d'échange. Les résultats sont exposés dans les figures 6.1, 6.2 et 6.3.

La première constatation est que les résultats semblent, de prime abord, très différents selon le voisinage choisi. En effet, les trois matrices d'échange considèrent différents types de voisinage et, par conséquent, révèlent des informations différentes.

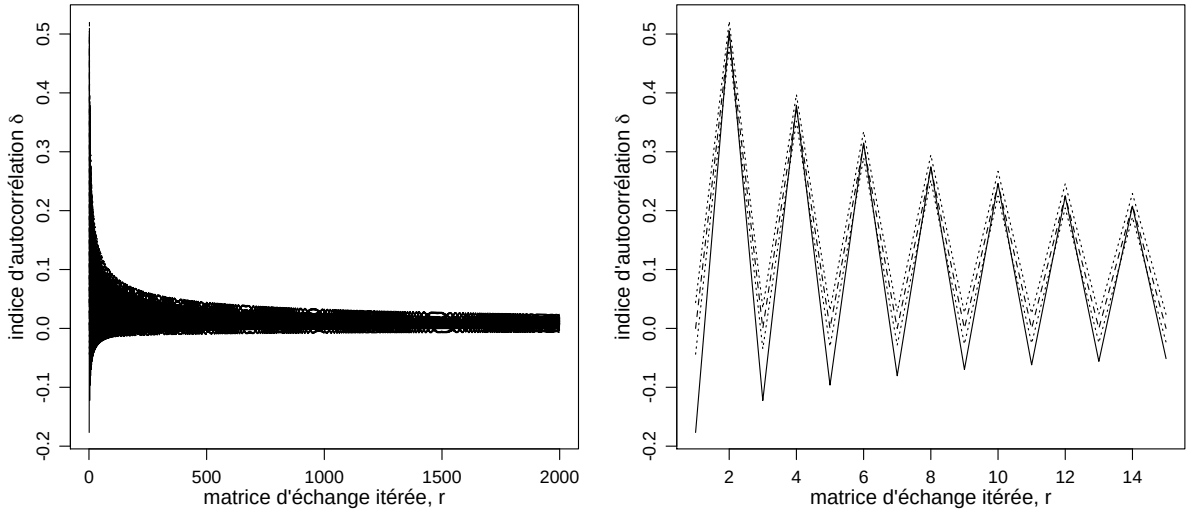


FIGURE 6.1 – Indice d'autocorrélation, δ, en fonction du décalage, r, pour l'alternance de la longueur des mots avec une matrice d'échange itérée (3.1). Gauche : r variant de 1 à $n = 2000$; droite : zoom avec r variant de 1 à 15. Dans cette figure, ainsi que dans les suivantes de ce chapitre, la ligne continue représente l'indice d'autocorrélation ; la ligne traitillée, la valeur attendue $E_0(\delta)$ (3.5) ; et les lignes pointillées, l'intervalle de confiance à 95% (3.6).

La figure 6.1 expose le résultat obtenu avec la matrice d'échange itérée (3.1). Le graphique de gauche, qui présente l'ensemble des valeurs obtenues pour r compris entre 1 et n, montre que les valeurs maximales obtenues pour δ diminuent lorsque r augmente. Ce phénomène provient du fait que lorsque r augmente, de plus en plus de voisins sont considérés. En effet (cf. table 3.1), lorsque $r = 1$, alors on regarde l'autocorrélation entre la position i et deux voisins, soit un à la position $j = i-1$ et un autre à $j = i+1$; lorsque $r = 2$, on considère deux fois la position i avec elle-même et deux voisins, soit un à $j = i-2$ et un autre à $j = i+2$; lorsque $r = 3$, on considère trois fois les positions $j = i-1$ et $j = i+1$, et une fois les positions $j = i-3$ et $j = i+3$; etc. Cette alternance entre le fait de considérer la position i avec elle-même pour chaque valeur de r paire et de ne pas le faire pour chaque valeur de r impaire conduit à une courbe en dents de scie visible sur le graphique de droite, car pour chaque r pair, il y a un élément exactement identique à lui-même, conduisant à une autocorrélation plus élevée. Ce phénomène, systématique pour toutes les applications, rend le graphique difficile à interpréter, c'est pourquoi la matrice itérée ne sera plus utilisée dans les exemples suivants.

Concernant l'interprétation spécifique à la longueur des mots, l'autocorrélation est négative, inférieure à la valeur attendue sous l'hypothèse d'absence d'autocorrélation, $E_0(\delta)$, et significative pour les r impairs, et n'est pas significative et environ égale à $E_0(\delta)$ pour les r pairs. On retrouve donc, comme attendu, le fait que deux mots qui se suivent auront des longueurs contrastées. En particulier, l'autocorrélation la plus négative, $\delta = -0.1767$, est obtenue pour $r = 1$, représentant probablement l'alternance entre les mots pleins (longs) et les mots outils (courts).

La figure 6.2 présente les résultats obtenus avec une matrice d'échange périodique (3.2). Contrairement aux deux autres matrices d'échange, cette dernière permet aussi de ne pas considérer de décalage ($r = 0$), ce qui implique, par définition, que $e_{ii} = 1/n$ et $\delta = 1$. Comme déjà mentionné dans la section 3.1.1 : $\breve{E}^{(r)} = \breve{E}^{(n-r)}$, ce qui conduit à la symétrie que l'on observe

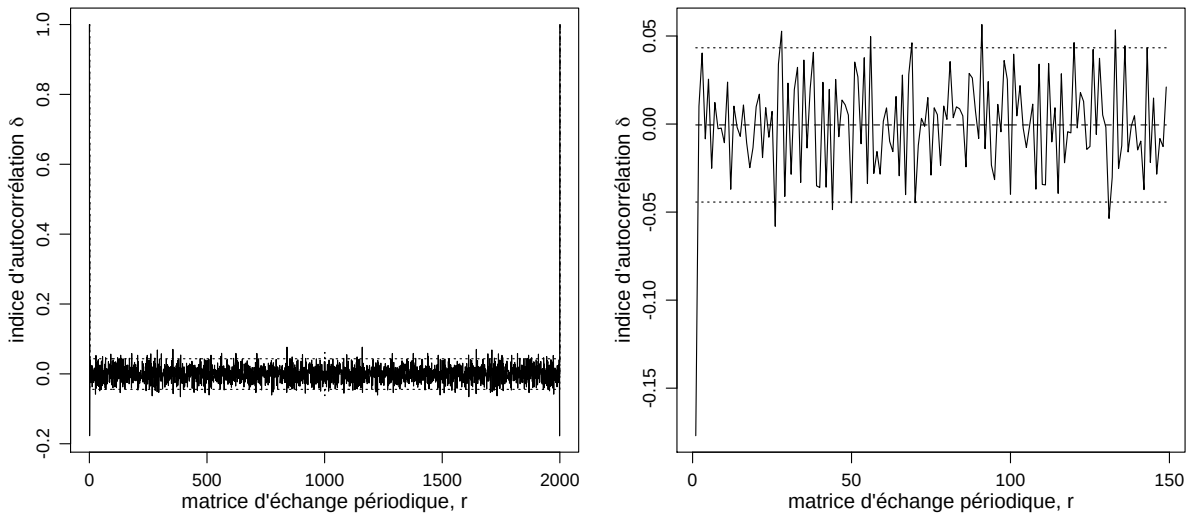


FIGURE 6.2 – Indice d'autocorrélation, δ, en fonction du décalage, r, pour l'alternance de la longueur des mots avec une matrice d'échange périodique (3.2). Gauche : r variant de 0 à $n = 2000$; droite : zoom avec r variant de 1 à 150.

sur le graphique de gauche. Pour cette application particulière, on découvre aussi que pour $r = 1$, δ est significatif et vaut -0.1770, conduisant à la même interprétation que celle obtenue dans l'exemple précédent d'alternance entre mots outils et mots pleins. De plus, l'alternance irrégulière de δ entre des valeurs positives et négatives, parfois significatives, bien que difficile à interpréter, semble cohérente avec l'hypothèse d'alternance entre mots longs et courts. La matrice d'échange périodique, particulièrement utile pour l'analyse de partitions musicales (cf. section 8.2), paraît moins pertinente pour le texte et ne sera donc plus utilisée dans la suite de ce chapitre.

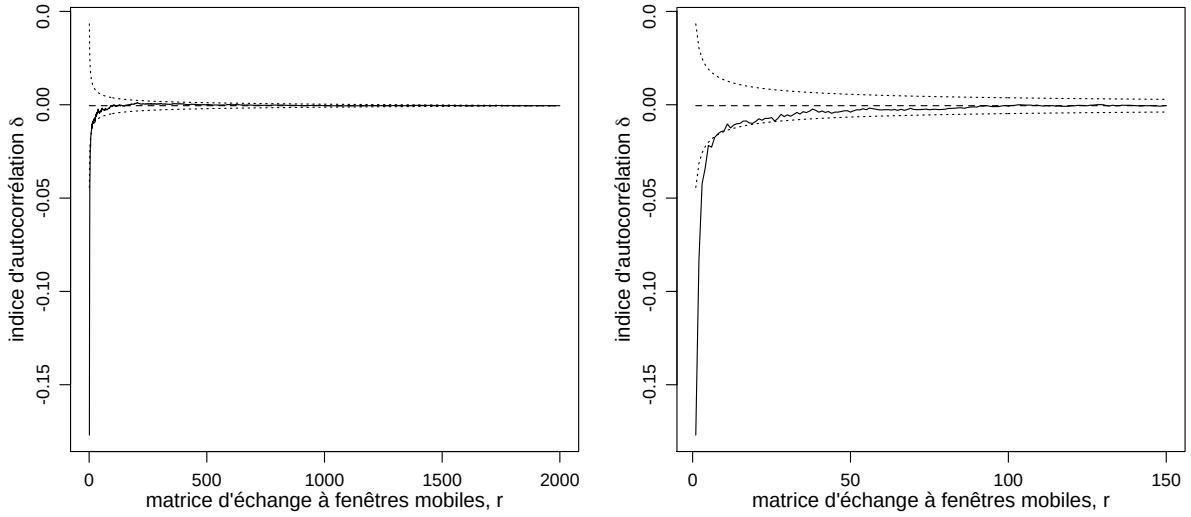


FIGURE 6.3 – Indice d'autocorrélation, δ, en fonction du décalage, r, pour l'alternance de la longueur des mots avec une matrice d'échange à fenêtres mobiles (3.3). Gauche : r variant de 1 à $n = 2000$; droite : zoom avec r variant de 1 à 150.

Finalement, l'indice d'autocorrélation a été calculé avec la matrice d'échange à fenêtres mobiles (3.3) qui considère un voisinage, à droite et à gauche de la position i, croissant avec r (figure 6.3). Ainsi, lorsque $r \to n$, alors $\delta \to E_0(\delta)$. À nouveau, lorsque $r = 1$, δ est significatif et vaut -0.1770. De plus, $\delta < E_0(\delta)$ pour $n \leq 92$, et est significatif lorsque $n \leq 7$, ce qui signifie

que, comme attendu, les mots ont des longueurs plus contrastées dans un voisinage proche qu'à des positions sélectionnées aléatoirement.

Comme il a déjà été suggéré dans les trois applications, lorsque $r = 1$, les résultats obtenus avec les trois matrices d'échange sont quasi identiques, soit $\delta \cong -0.177$, ce qui paraît évident au vu de la ressemblance des trois matrices pour ce décalage (cf. table 3.1).

6.2 Parties du discours

6.2.1 Dissimilarités binaires relatives à une partie du discours

Soit un terme k et une matrice indicatrice de termes $X = (x_{ik})$, telle que :

$$x_{ik} = \begin{cases} 1 & \text{si le terme } k \text{ apparaît à la position } i \\ 0 & \text{sinon} \end{cases}$$

Par construction, $x_{i\bullet} = 1$. La dissimilarité de termes $D_{ij}^{\text{terme}} = \frac{1}{2}\sum_k (x_{ik} - x_{jk})^2$ vaut 0 si les termes aux positions i et j sont identiques, et 1 s'ils sont différents. Un exemple avec des termes est exposé dans Bavaud *et al.* (2012).

Une variante de cette dissimilarité, qu'on nommera dissimilarité binaire d'une *partie du discours* (PDD), consiste à ne considérer que deux « termes » : présence et absence d'une partie du discours donnée. Dans cet exemple, on se limite à considérer quatre parties du discours : les noms, les verbes, les adjectifs et les adverbes. Ainsi, X est de taille $n \times 2$ et la dissimilarité D_{ij}^{PDD} d'une partie de discours est égale à 1 si la partie du discours est présente en i et absente en j, et 0 sinon (*i.e* si il y a co-absence ou co-présence de cette partie du discours aux positions i et j). Au final, quatre matrices de dissimilarité de termes binaires sont ainsi construites : D^{nom}, D^{verbe}, D^{adj} et D^{adv}.

6.2.2 Traitements et résultats

Le texte utilisé dans cet exemple est la *Déclaration des droits de l'homme et du citoyen de 1789*. Les parties du discours sont extraites du texte avec TreeTagger (Schmid, 1994), puis celles correspondant à la ponctuation ou aux balises de phrase sont supprimées. Au final, le total des occurrences s'élève à $n = 668$. Ensuite, quatre matrices de dissimilarités sont créées, soit une pour chaque type de discours. Quant à la matrice d'échange, on utilise celle à fenêtres mobiles (3.3).

Les résultats sont présentés dans la figure 6.4. Pour le cas des noms, l'autocorrélation est négative et significative pour $r = 1, 2$. Ceci semble cohérent, car un nom est rarement suivi ou précédé par un autre nom ($r = 1$), et parfois suivi ou précédé, avec un décalage de deux, par un nom ($r = 2$), comme par exemple : « conservation des droits » ou « toutes dignités, places et emplois ». A contrario, pour les verbes, δ est significativement positif pour $r = 1, 2$. En effet, les verbes peuvent fréquemment se suivre, lors de l'emploi de temps composés (« a prescrites », « ait été déclaré ») ou lorsqu'un infinitif suit un auxiliaire modal (« doit être », « peuvent être fondées »). Quant aux adjectifs et aux adverbes, l'autocorrélation n'est presque jamais significative. Cependant, dans les deux applications, elle est négative pour $r = 1$ et positive pour $r = 2$. En effet, des adjectifs ou des adverbes se suivent rarement directement, mais parfois dans un voisinage de largeur deux lorsqu'ils sont liés par une conjonction, comme par exemple : « établie et promulguée » ou « juste et préalable » pour les adjectifs et « strictement et évidemment » pour les adverbes.

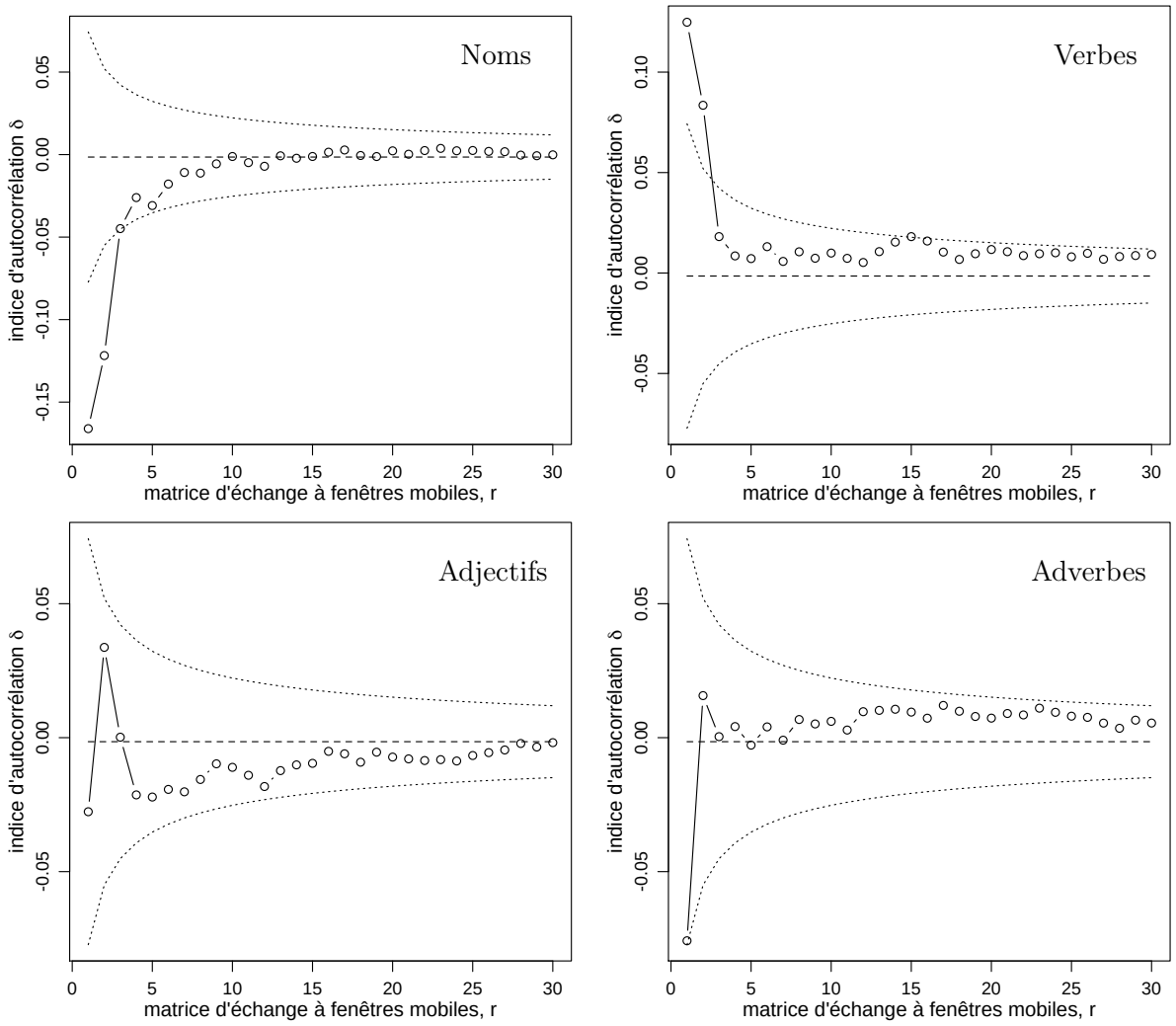


FIGURE 6.4 – Parties du discours : δ en fonction de r avec un matrice d'échange à fenêtres mobiles et r variant de 1 à 30. En haut à gauche : noms ; en haut à droite : verbes ; en bas à gauche : adjectifs ; en bas à droite : adverbes.

6.3 Sens des mots selon WordNet

6.3.1 Dissimilarités sémantiques

WordNet (Miller, 1995; Fellbaum, 1998) regroupe les mots (les nom, les verbes, les adjectifs ou les adverbes) en *synsets*, soit un ensemble de synonymes, et chaque « synset » représente un *concept*. Pour les noms et les verbes, ces concepts sont reliés, entre autres choses, sous forme d'ontologie, selon des relations sémantiques. Soit c_1 et c_2 deux concepts d'une ontologie, si le « concept c_1 est un (genre de) c_2 », noté $c_1 \leq c_2$, alors c_1 est *l'hyponyme* (*troponyme* dans le cas des verbes) de c_2 ; et c_2, *l'hyperonyme* de c_1. Le plus petit hyperonyme commun à ces deux concepts s'écrit $c_1 \vee c_2$. Par exemple, avec l'ontologie simplifiée présentée dans la figure 6.5 pour les noms : `bicyclette` $\leq$ `véhicule`, `voiture` $\leq$ `véhicule`, `véhicule` $\leq$ `entité`, `animal` $\leq$ `entité`, `bicyclette` $\vee$ `voiture` = `véhicule` et `bicyclette` $\vee$ `animal` = `entité`.

La probabilité $p(c)$ d'un concept c peut être estimée, en se basant sur un corpus de référence, comme la proportion d'occurrences $n(w)$ du mot w dont le sens $C(w)$ est une occurrence du concept c :

$$p(c) := \frac{\sum_w n(w)\, \mathbb{1}(C(w) \leq c)}{\sum_w n(w)} \tag{6.1}$$

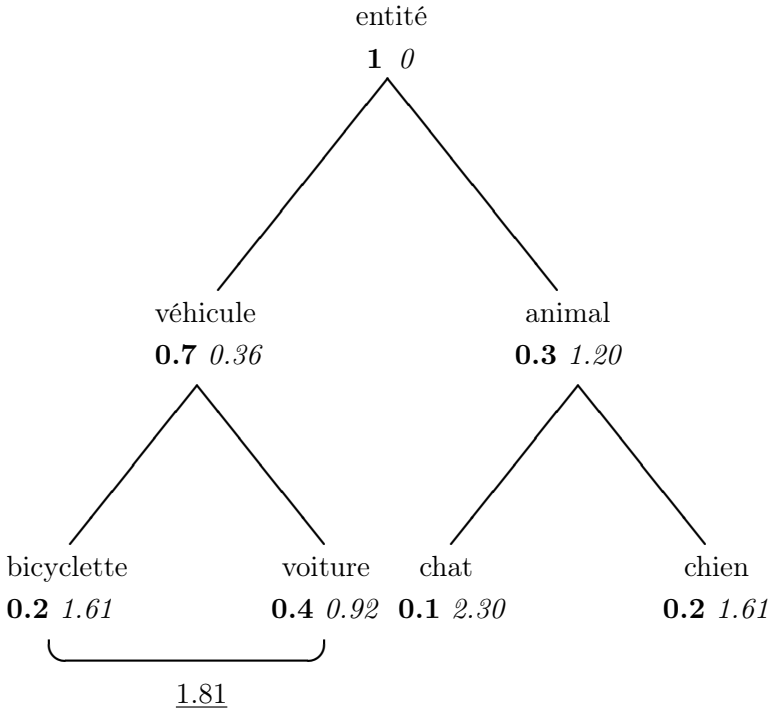


FIGURE 6.5 – Exemple d'une ontologie simplifiée composée de 7 concepts. Pour chacun des concepts, le nombre en gras représente la probabilité de ce concept (6.1) ; et le nombre en italique, la similarité d'un concept avec lui-même (6.2). Le nombre sous-ligné est la dissimilarité (6.3) entre les concepts `bicyclette` et `voiture`.

Avec cette probabilité, Resnik (1995, 1999) propose une mesure de *similarité* entre deux concepts :

$$s(c_1, c_2) := -\log p(c_1 \vee c_2) \geq 0 \tag{6.2}$$

Ces similarités permettent de définir la dissimilarité entre deux concepts comme :

$$D(c_1, c_2) := s(c_1, c_1) + s(c_2, c_2) - 2s(c_1, c_2) \tag{6.3}$$

qui est une dissimilarité d'arbre et donc, par suite, une dissimilarité euclidienne carrée (Bavaud *et al.*, 2012).

En se basant sur les probabilités données dans l'exemple de la figure 6.5, la dissimilarité entre les concepts `bicyclette` et `voiture` vaut :

$$\begin{aligned} D(\texttt{bicyclette}, \texttt{voiture}) &= s(\texttt{bicyclette}, \texttt{bicyclette}) + s(\texttt{voiture}, \texttt{voiture}) \\ &\quad - 2s(\texttt{bicyclette}, \texttt{voiture}) \\ &= -\log(0.2) - \log(0.4) + 2\log(0.7) \\ &= 1.81 \end{aligned}$$

6.3.1.1 Module Perl et implémentation des dissimilarités

L'implémentation a été faite pour les noms et les verbes qui sont liés, comme déjà mentionné, sous forme d'ontologie dans WordNet. Le traitement des noms et des verbes a été effectué séparément, mais en utilisant la même procédure.

Pour commencer, les noms (respectivement les verbes) sont extraits et lemmatisés à l'aide de TreeTagger, en se servant du même module Perl que celui présenté dans la section 4.1.3. Puis, la matrice des dissimilarités (6.3) pour les noms (respectivement les verbes) est créée en utilisant

les similarités (6.2) calculées avec le module Perl `WordNet::Similarity` (Pedersen, Patwardhan et Michelizzi, 2004). En particulier, ce module se base sur la version de WordNet 3.0. De plus, l'option « resnik » a été utilisée et pour chaque mot, le premier sens a été sélectionné dans WordNet, soit le plus probable. Concernant les probabilités des concepts (6.1), elles se basent sur le fichier `ic-brown-resnik-add1.dat` que propose le module Perl. Ce fichier considère le corpus de Brown [1] (Francis et Kučera, 1967, 1982) avec un « resnik counting », ce qui signifie que chaque concept associé à une forme graphique (*type*) obtient une proportion égale de chaque unité de décompte ; et avec l'option « add1 », ce qui signifie qu'avant de commencer le décompte, chaque concept prend une valeur de 1 afin d'éviter qu'un concept qui n'apparaît pas dans le corpus ait une valeur de 0. Finalement, l'option d'ajouter un noeud racine (*root node*) à l'ensemble des concepts, s'il n'existe pas, est choisie (soit l'hyperonyme commun le plus général qui serait `entité` dans la figure 6.5).

Pour terminer, si un mot n'est pas reconnu par WordNet, car il est mal lemmatisé ou n'existe pas dans cette base, il est soit modifié manuellement, soit sa position est supprimée.

6.3.2 Autocorrélation sémantique

Deux textes sont traités dans cette application. Le premier, l'*Atlantic Charter*, est une déclaration de principe rendue publique en août 1941 et rédigée par le président des États-Unis Franklin Delano Roosevelt et le Premier ministre britannique Winston Churchill. Le second, *The Masque of the Red Death*, est une nouvelle d'Edgar Allan Poe publiée pour la première fois en 1842. Comme expliqué dans la section 6.3.1.1, deux matrices de dissimilarités sémantiques sont créées pour chaque texte, une pour les noms et l'autre pour les verbes. En combinant chacune de ces quatre matrices de dissimilarités avec la matrice d'échange à fenêtres mobiles, l'indice d'autocorrélation a pu être calculé.

6.3.2.1 Atlantic Charter

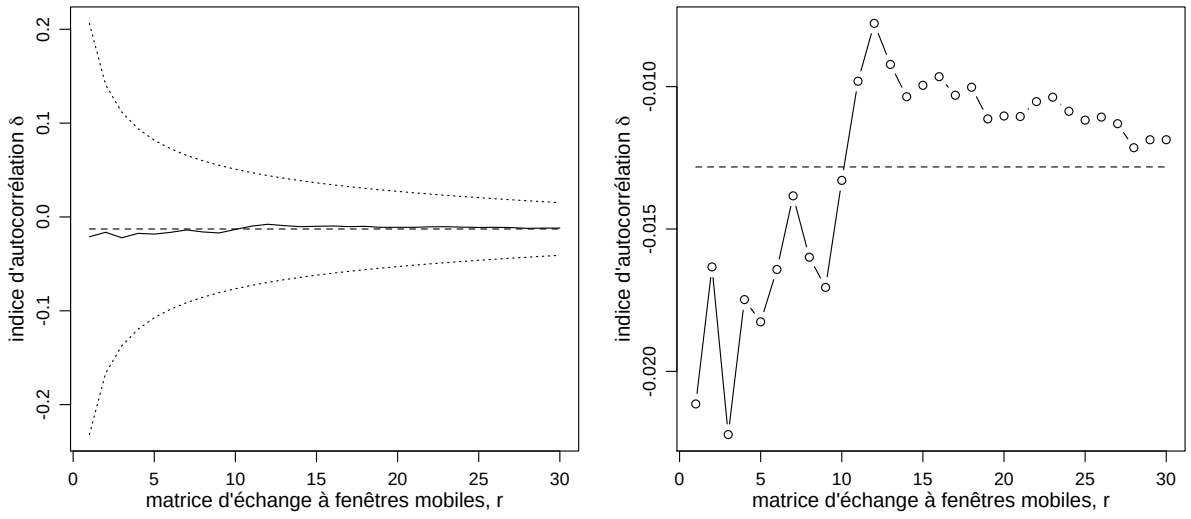


FIGURE 6.6 – Sens des noms de l'*Atlantic Charter* : δ en fonction de r avec un matrice d'échange à fenêtres mobiles. Gauche : r variant de 1 à 30 ; droite : zoom avec r variant aussi de 1 à 30.

Concernant le texte de l'*Atlantic Charter*, il y a $n = 79$ positions de noms et $n = 57$ positions de verbes, dont une, non reconnue par WordNet, a été supprimée, ramenant le total à $n = 56$

1. Plus précisément, les auteurs ont utilisé le corpus de Brown de l'ICAME Collection of English Language Corpora, Second Edition, 1999, `http://www.hit.uib.no/icame/cd`.

occurences de verbes. En effet TreeTagger a considéré le mot *pending* comme un verbe qu'il a lemmatisé comme *pend*, alors qu'il s'agit d'une préposition signifiant « en attendant ».

La figure 6.6 présente l'autocorrélation obtenue pour les noms. Bien qu'elle ne soit pas significative (graphique de gauche), on remarque qu'elle est négative pour $r \leq 10$ (graphique de droite), laissant penser qu'il y a répulsion sémantique pour les noms dans un voisinage restreint.

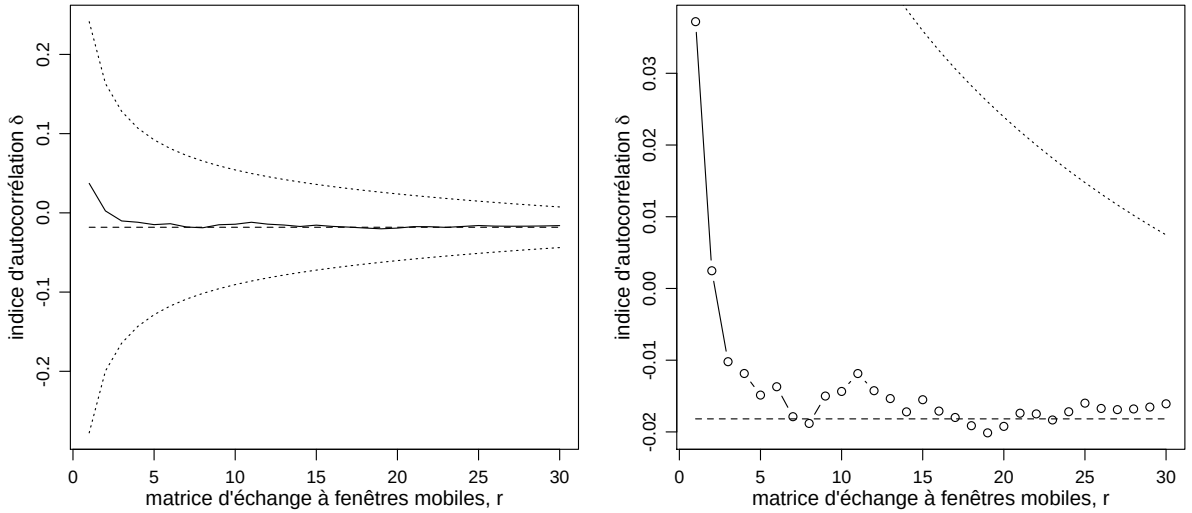


FIGURE 6.7 – Sens des verbes de l'*Atlantic Charter* : δ en fonction de r avec un matrice d'échange à fenêtres mobiles. Gauche : r variant de 1 à 30 ; droite : zoom avec r variant aussi de 1 à 30.

A propos des verbes (figure 6.7), bien que plus élevée que pour les noms, l'autocorrélation n'est à nouveau jamais significative. Cependant, elle est systématiquement positive pour $r \leq 7$ et en général proche de $E_0(\delta)$. Il y a donc une attraction sémantique pour les verbes dans un proche voisinage et peu de variabilité dans l'ensemble du texte.

6.3.2.2 The Masque of the Red Death

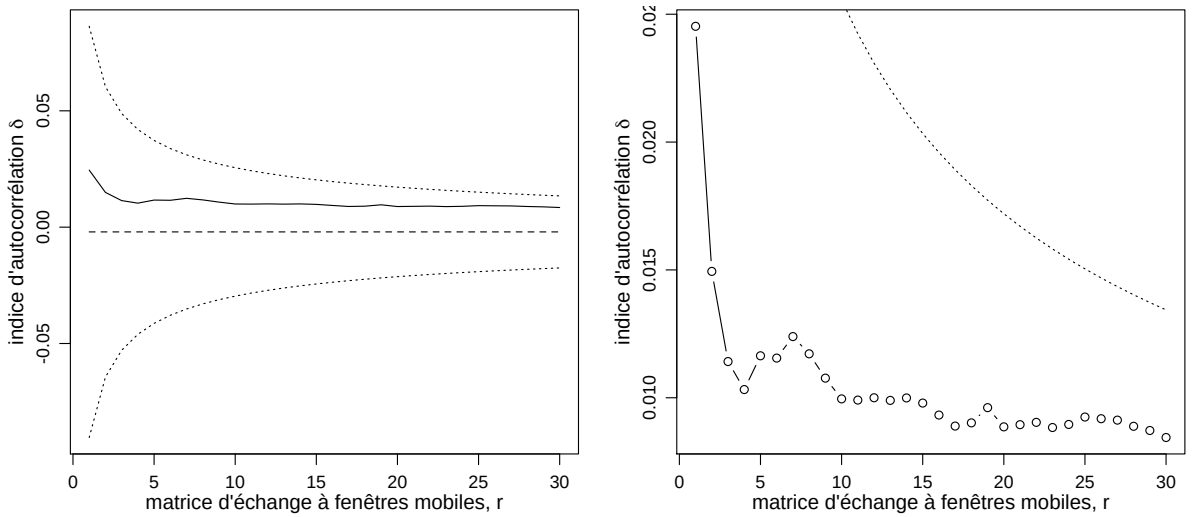


FIGURE 6.8 – Sens des noms de *The Masque of the Red Death* : δ en fonction de r avec un matrice d'échange à fenêtres mobiles. Gauche : r variant de 1 à 30 ; droite : zoom avec r variant aussi de 1 à 30.

Pour ce second texte, le nombre total de noms reconnus par TreeTagger s'élève à 497. Pour pouvoir être lemmatisés par TreeTagger et/ou reconnus par WordNet, quatre noms ont été

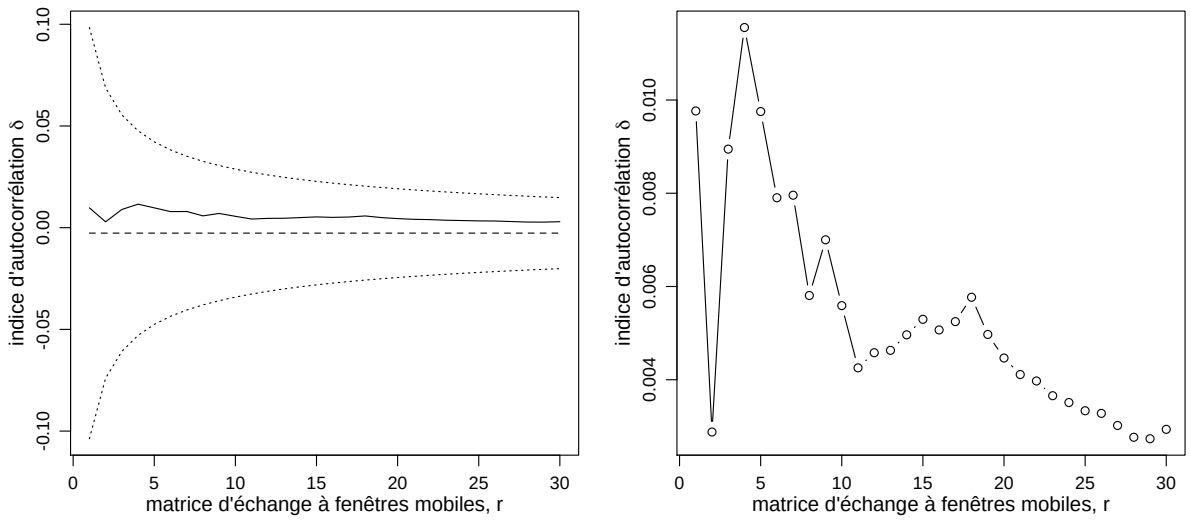


FIGURE 6.9 – Sens des verbes de *The Masque of the Red Death* : δ en fonction de r avec un matrice d'échange à fenêtres mobiles. Gauche : r variant de 1 à 30 ; droite : zoom avec r variant aussi de 1 à 30.

modifiés : *dominions* a été transformé en *dominion*, *fire-light* en *firelight*, *minute-hand* en *minute_ hand* et *ballet-dancers* en *ballet_ dancer*. De plus, six mots ont dû être supprimés, car ils ne sont pas répertoriés dans WordNet, à savoir : *fellow-men*, *improvisatori*, *tremulousness*, *decora*, *something* et *grave-cerements*, conduisant à un total de $n = 491$ positions pour les noms. En ce qui concerne les verbes, 379 verbes ont été identifiés par TreeTagger, dont deux ont été supprimés, car non reconnus par WordNet : *stiff-frozen* et *untenanted*. Au final, il reste $n = 375$ positions pour les verbes.

Comme pour le texte de l'*Atlantic Charter*, les autocorrélations ne sont jamais significatives, aussi bien pour les noms (figure 6.8) que pour les verbes (figure 6.9). Cependant, $\delta > E_0(\delta)$ pour $r \leq 226$ avec les noms (respectivement $r \leq 370$ avec les verbes), ce qui laisse supposer qu'il y a une attraction sémantique entre les noms (respectivement les verbes) dans un large voisinage. Plus particulièrement, on remarque que pour les noms, le résultat est opposé à celui obtenu pour le texte de l'*Atlantic Charter*, avec une attraction sémantique relativement élevée dans un proche voisinage, ce qui signifie peut-être qu'avec ce texte sous forme de nouvelle, le champ lexical est plus similaire dans un proche voisinage.

6.3.3 MDS et autocorrélation sur les premiers facteurs

Les dissimilarités entre concepts (6.3) étant euclidiennes carrées, il est possible d'appliquer un MDS, une approche originale à notre connaissance dans le cas sémantique. Les mots, que ce soit les noms ou les verbes, n'ont pas été pondérés, ainsi le MDS ordinaire est utilisé, équivalent à la version pondérée (1.25) en prenant des poids uniformes.

Dans un second temps, les coordonnées $x_{j\alpha}$ (1.25b) des deux premiers facteurs ($\alpha = 1, 2$) ont été extraites et une nouvelle dissimilarité euclidienne carrée a été calculée, telle que $D^{\alpha}_{ij} = (x_{i\alpha} - x_{j\alpha})^2$. Avec ces dissimilarités et une matrice d'échange à fenêtre mobiles, l'autocorrélation est à nouveau mesurée pour ces deux premières dimensions sémantiques.

6.3.3.1 Atlantic Charter

Le résultat obtenu en appliquant un MDS sur les dissimilarités sémantiques entre les noms est exposé dans la figure 6.10. Malgré un pourcentage faible d'inertie expliquée par les deux premiers facteurs (18.4%), on remarque trois groupes de noms clairement distincts. Dans le

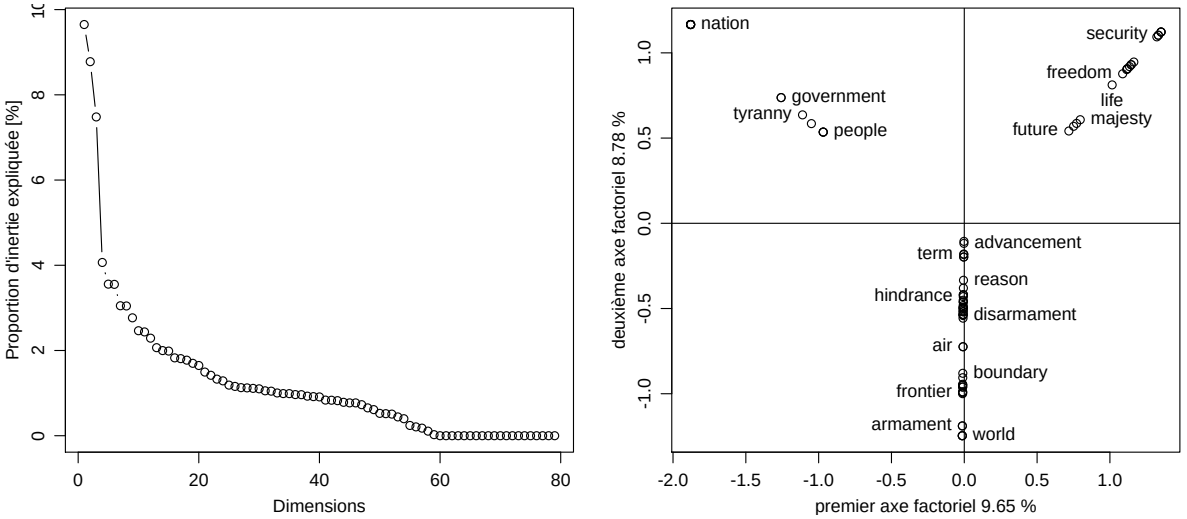


FIGURE 6.10 – MDS sur le sens des noms de l'*Atlantic Charter*. Gauche : valeurs propres ; droite : coordonnées pour les deux premiers axes factoriels.

quadrant nord-ouest, on trouve des noms tels que *nation*, *government*, *country* ou *people*, dont le plus petit hyperonyme commun est le concept « group, grouping », défini dans WordNet comme « any number of entities (members) considered as a unit ». Dans le quadrant nord-est se trouvent des noms, tels que *freedom*, *security* ou *majesty*, qui sont englobés dans le concept « attribute » qui est défini comme étant « an abstraction belonging to or characteristic of an entity ». Finalement, le troisième groupe, dans la zone sud, est composé de tous les autres noms, donc ceux qui ne sont pas englobés dans les concepts « group, grouping » ou « attribute ». Ainsi, le premier axe différencie les noms concernant le « group, grouping » de ceux concernant « attribute », et le deuxième oppose les noms englobés dans ces deux concepts aux autres.

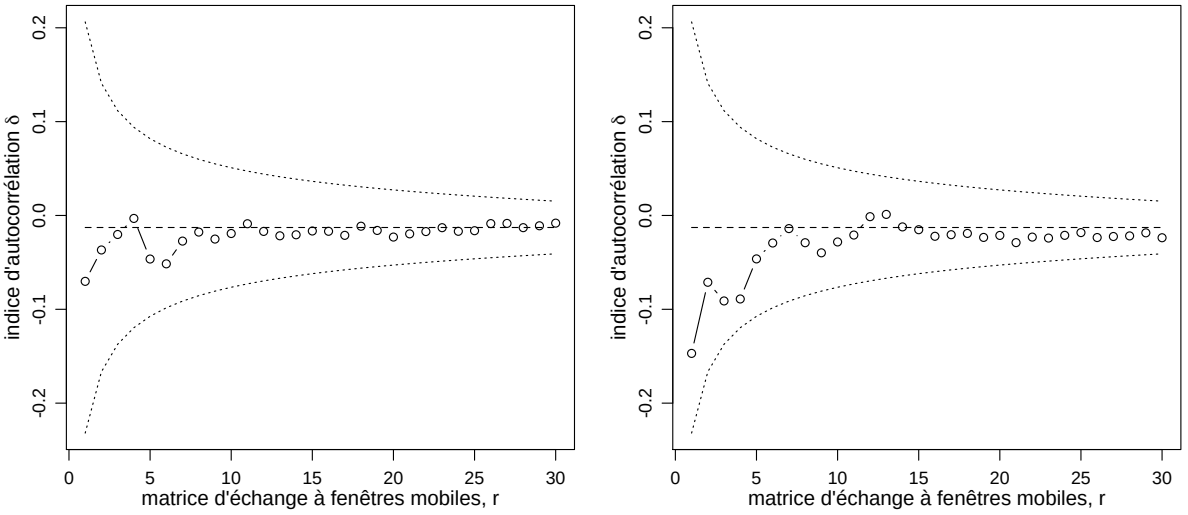


FIGURE 6.11 – Sens des noms de l'*Atlantic Charter* : δ pour la première (gauche) et la deuxième (droite) dimension sémantique en fonction de $r = 1, \ldots, 30$, avec un matrice d'échange à fenêtres mobiles.

L'autocorrélation mesurée sur la première dimension sémantique des noms (figure 6.11 gauche) n'est jamais significative et inférieure à $E_0(\delta)$ pour $r \leq 3$. Ceci laisse penser que dans un voisinage restreint, il peut y avoir alternance entre les noms relatifs au concept « group, grouping » et ceux relatifs au concept « attribute ». Quant à la deuxième dimension sémantique (figure 6.11

gauche), on remarque que l'autocorrélation n'est à nouveau pas significative, mais que les valeurs négatives sont plus élevées, en valeur absolue, pour r petit, ce qui semble indiquer une alternance entre les noms en relation avec les concepts « group, grouping » et « attribute » et tous les autres noms.

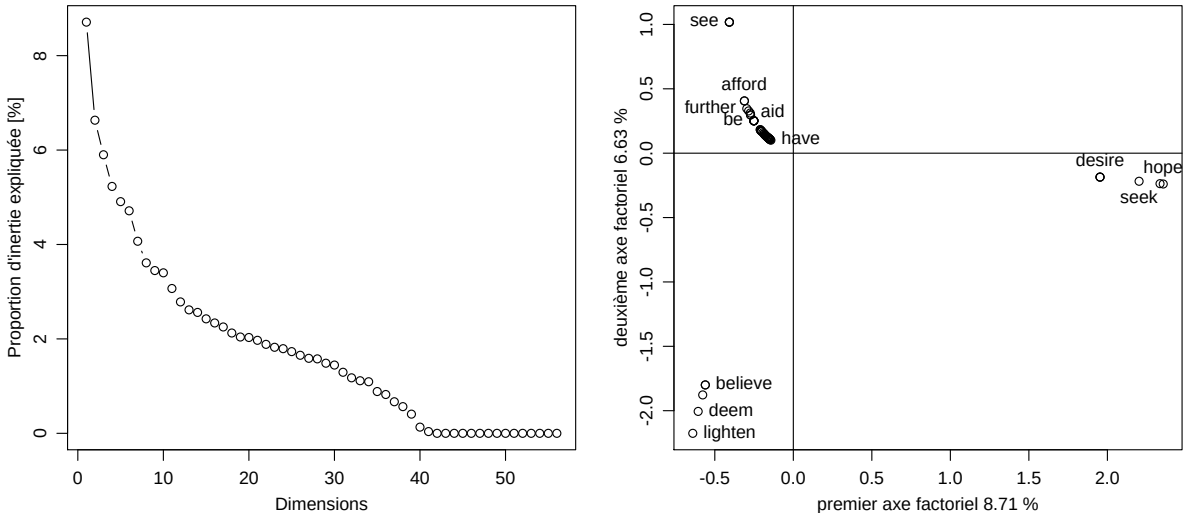


FIGURE 6.12 – MDS sur le sens des verbes de l'*Atlantic Charter*. Gauche : valeurs propres ; droite : coordonnées pour les deux premiers axes factoriels.

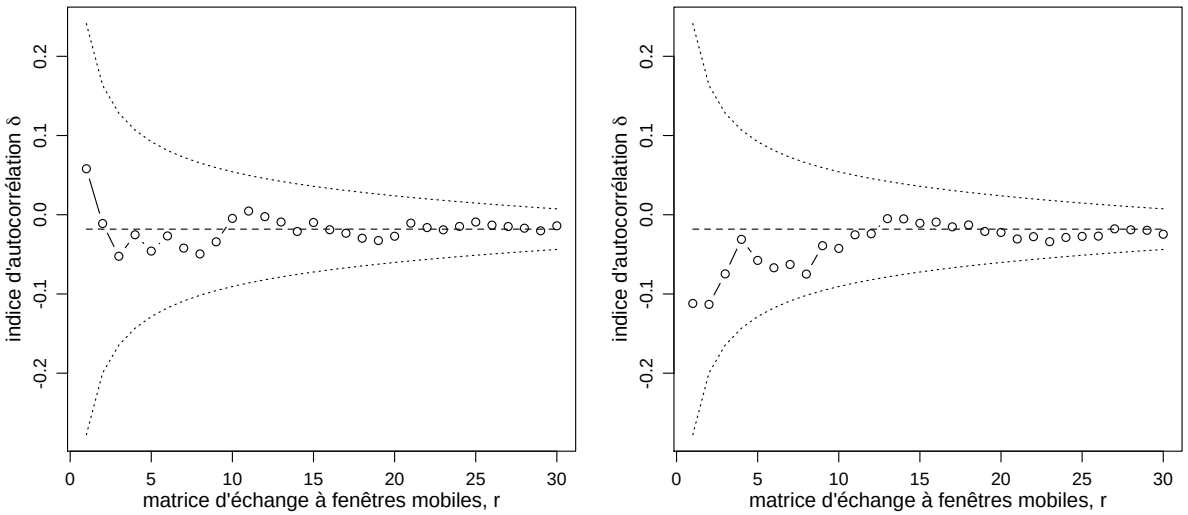


FIGURE 6.13 – Sens des verbes de l'*Atlantic Charter* : δ pour la première (gauche) et la deuxième (droite) dimension sémantique en fonction de $r = 1, \ldots, 30$, avec un matrice d'échange à fenêtres mobiles.

Comme pour les noms, avec un pourcentage d'inertie expliquée un peu plus faible pour les deux premiers facteurs (15.3%), trois groupes apparaissent en appliquant le MDS sur les dissimilarités sémantiques entre les verbes (figure 6.12). Le plus petit hyperonyme commun des verbes présents dans le quadrant sud-est est le concept « desire, want », défini comme « feel or have a desire for ; want strongly ». Dans le quadrant sud-ouest se trouvent des verbes tels que *lighten*, *deem* ou *respect*, tous englobés dans le concept « think, cogitate, cerebrate », défini comme « use or exercise the mind or one's power of reason in order to make inferences, decisions, or arrive at a solution or judgments ». Enfin, le troisième groupe, dans le quadrant nord-ouest, est composé de verbes sémantiquement hétérogènes et ne possédant pas un hyperonyme com-

mun. Il semble donc que le premier facteur oppose les verbes englobés dans le concept « desire, want » à ceux qui ne le sont pas. Pareillement, le second facteur différencie les verbes relatifs au concept « think, cogitate, cerebrate » des autres.

Comme pour les noms, on remarque qu'en mesurant l'autocorrélation sur la première et la deuxième dimension sémantique, elle n'est jamais significative (figure 6.13) (sauf pour $r = 50$ dans la seconde dimension). Cependant, elle est supérieure à $E_0(\delta)$ pour $r \leq 2$ avec la première dimension, et inférieure à $E_0(\delta)$ pour $r \leq 12$. Il semblerait donc que, dans un voisinage restreint, il y a peu d'alternance entre les verbes reliés au concept « desire, want » et les autres, et qu'il y en a plus entre les verbes englobés dans le concept « think, cogitate, cerebrate » et les autres.

6.3.3.2 The Masque of the Red Death

Pour ce texte, plus long que le précédent, l'inertie expliquée par les deux premiers axes factoriels dans le cas des noms est plus faible (14.9%) et les groupes sont plus nombreux (figure 6.14).

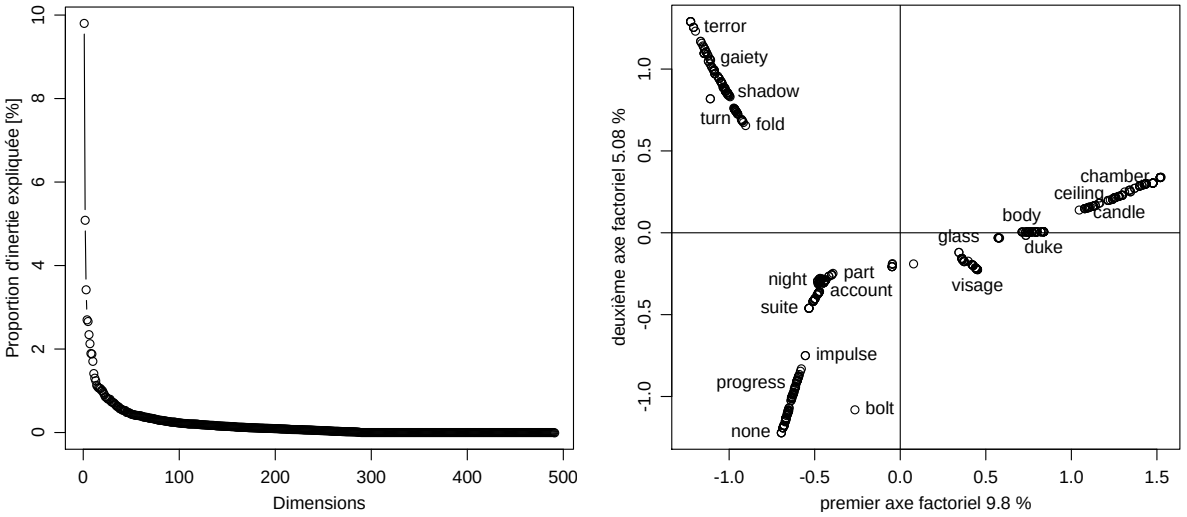


FIGURE 6.14 – MDS sur le sens des noms de *The Masque of the Red Death*. Gauche : valeurs propres ; droite : coordonnées pour les deux premiers axes factoriels.

Le groupe dans le quadrant nord-ouest est composé de noms (*terror*, *gaiety*, *courage*, *magnificence*, etc.) englobés dans le concept « attribute » qui est « an abstraction belonging to or characteristic of an entity », lui-même englobé dans le concept plus général de « abstraction, abstract entity ». Dans le quadrant sud-ouest, on distingue deux groupes. Celui qui est plus au sud et qui contient des noms tels que *progress*, *impulse*, *sympathy* ou *creation*, a pour plus petit hyperonyme commun le concept « psychological feature », défini comme « a feature of the mental life of a living organism » et est à nouveau hyponyme de « abstraction, abstract entity ». L'autre groupe de ce quadrant contient tous les noms qui sont englobés dans le concept « abstraction, abstract entity », soit « a general concept formed by extracting common features from specific examples », mais qui ne sont pas des hyponymes de « attribute » ou de « psychological feature », comme par exemple : *night*, *part*, *music* ou *orchestra*.

Les noms *chamber*, *candle*, *minute hand*, *structure*, etc. forment un groupe dans le quadrant nord-est et ont comme plus petit hyperonyme commun le concept « artifact, artefact », soit « a man-made object taken as a whole ». A l'est, entre les quadrants nord-est et sud-est, se trouve un groupe composé de noms, tels que *body*, *duke*, *mummer* ou *violet*, englobés dans le concept « whole, unit » qui désigne « an assemblage of parts that is regarded as a single entity », mais qui ne sont pas un « artifact, artefact ». En effet, il faut préciser que le concept « artifact, artefact » est un hyponyme de « whole, unit », lui-même hyponyme indirect de « physical entity ». Enfin,

le groupe dans le quadrant sud-est, proche du centre, contient tous les noms, tels que *glass*, *visage*, *flame* ou *stream*, englobés dans le concept de « physical entity », mais qui ne sont pas des hyponymes de « whole, unit ».

En conclusion, le premier axe s'interprète comme l'oppositions entre « abstraction, abstract entity » et « physical entity », mais le second axe reste difficile à interpréter de manière univoque.

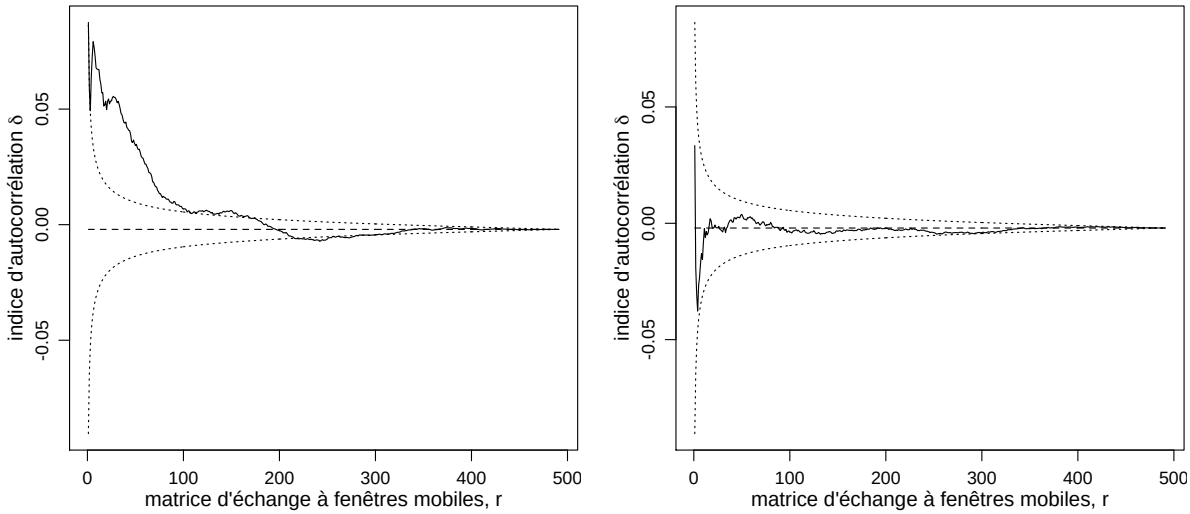


FIGURE 6.15 – Sens des noms de *The Masque of the Red Death* : δ pour la première (gauche) et la deuxième (droite) dimension sémantique en fonction de r, qui varie de 1 à $n = 491$, avec un matrice d'échange à fenêtres mobiles.

La mesure de δ sur le premier axe factoriel (figure 6.15 gauche) est positive et clairement significative lorsque $r \leq 107$ (sauf pour $r = 2$). Ainsi, de longs segments de textes doivent contenir une majorité de noms relatifs à un seul des deux concepts : « abstraction, abstract entity » ou « physical entity ». Le second axe factoriel étant difficile à interpréter, l'autocorrélation mesurée sur celui-ci (figure 6.15 droite) l'est tout autant. On peut simplement constater que δ n'est presque jamais significatif , qu'il est positif pour $r = 1$, puis négatif avant de rapidement s'approcher de $E_0(\delta)$.

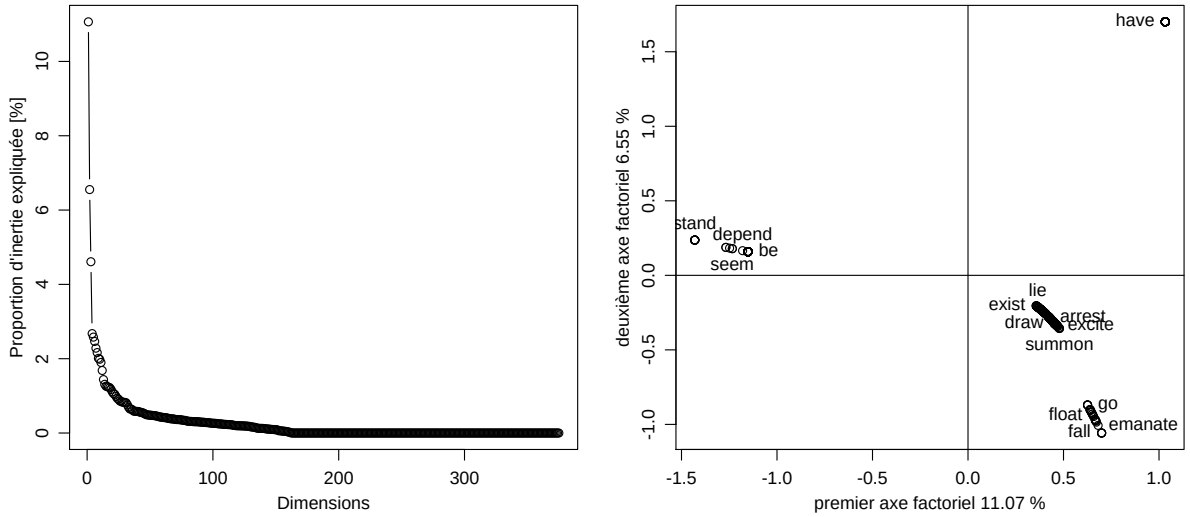


FIGURE 6.16 – MDS sur le sens des verbes de *The Masque of the Red Death*. Gauche : valeurs propres ; droite : coordonnées pour les deux premiers axes factoriels.

Concernant les verbes (figure 6.16), l'inertie expliquée par les deux premiers facteurs est de

17.6%. On peut distinguer quatre groupes. Dans le quadrant nord-ouest se trouvent les verbes d'état (*seem*, *gleam*, *sound*, etc.), tous englobés dans le concept « be », défini comme « have the quality of being ; (copula, used with an adjective or a predicate noun) ». Seul le verbe *have* se situe dans le quadrant nord-ouest. Finalement, on observe deux groupes dans le quadrant sud-est : celui plus au sud contient des verbes, tels que *go*, *fall*, *approach* ou *rush*, dont le plus petit hyperonyme commun est le concept de « travel, go, move, locomote » défini comme « change location ; move, travel, or proceed, also metaphorically » ; quant à celui plus proche du centre, il contient tous les autres verbes qui n'ont pas d'hypernoyme commun. En résumé, le premier axe factoriel oppose les verbes d'états aux autres ; et le second, le verbe *have* aux autres.

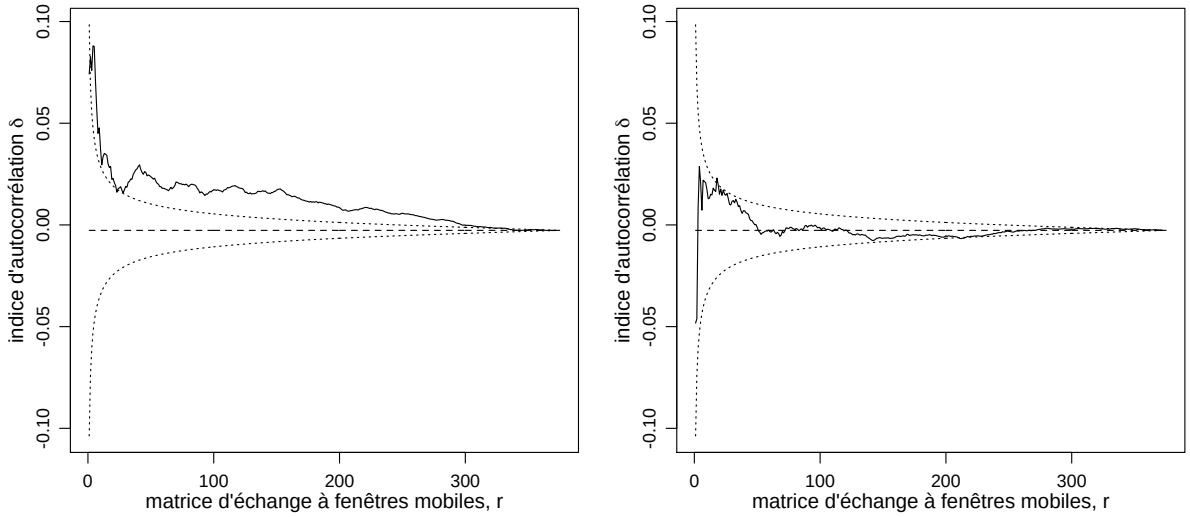


FIGURE 6.17 – Sens des verbes de *The Masque of the Red Death* : δ pour la première (gauche) et la deuxième (droite) dimension sémantique en fonction de r, qui varie de 1 à $n = 375$, avec un matrice d'échange à fenêtres mobiles.

L'autocorrélation mesurée sur la première dimension sémantique (figure 6.17 gauche) est clairement positive pour l'ensemble des voisinages, même s'elle n'est pas significative pour $r = 1$ et quelques autres r, ce qui laisse supposer qu'il y a peu d'alternances entre les verbes d'état et les verbes d'action. L'explication la plus plausible de ce phénomène est que les verbes d'état sont peu nombreux et donc que, généralement, les verbes d'actions se suivent. Concernant la deuxième dimension sémantique (figure 6.17 droite), bien que δ ne soit pas significatif, il est clairement inférieur à $E_0(\delta)$ pour $r = 1, 2$, ce qui est peut-être dû à l'utilisation de *have* comme auxiliaire des temps composés.

6.4 Discussion

Ce chapitre a présenté l'application de l'indice d'autocorrélation, exposé dans la section 3.2, à différents traitements textuels. Calculé sur la base de deux éléments, une matrice d'échange E et une matrice de dissimilarités euclidiennes carrées D, cet indice, d'abord développé pour l'analyse spatiale, dont les séries temporelles sont un cas particulier, permet de modéliser différentes navigations séquentielles dans un texte, grâce à E, et de mesurer la variation de caractéristiques mesurées sur les unités textuelles dans ces navigations, grâce à D.

Le premier exemple (section 6.1), assez simple, concernant la longueur des mots, a permis de comparer les différentes matrices d'échange et de retrouver le résultat, présupposé, d'alternance entre mots longs et mots courts. Puis, le second exemple (section 6.2), sur les parties du discours, a mis en lumière certaines structures syntaxiques. Finalement, le troisième exemple (section 6.3), sur l'autocorrélation sémantique, a montré qu'il est possible de mesurer une sorte de

variabilité sémantique dans un voisinage donné. De plus, les dissimilarités sémantiques, qui sont euclidiennes carrées, ont pu être, par le biais du MDS, visualisées et décomposées en dimensions factorielles, sur lesquelles l'autocorrélation a pu être à nouveau mesurée. Pour une approche comparable, quoique distincte, sur l'autocorrélation sémantique, voir Samsonovich (2014).

Seul un petit aperçu des applications textuelles possibles ont été présentées ici. Signalons que l'on peut également mesurer, pour un texte, l'autocorrélation de la présence et l'absence de termes. Concernant un dialogue ou une pièce de théâtre, il est possible de calculer l'autocorrélation de la longueur d'une réplique, du sexe de l'interlocuteur ou du profil de catégories morpho-syntaxiques d'une réplique par l'intermédiaire d'une table de contingence et de dissimilarités du khi2. De surcroît, en plus de modéliser la navigation à l'intérieur d'un document, on peut aussi modéliser la navigation hypertextuelle dans un réseau textuel (voir Bavaud *et al.*, 2012 ; Bavaud *et al.*, accepté pour publication). Cet indice permet aussi de mesurer les variations présentes dans les séquences musicales (voir section 8.2). En conclusion, cet indice, δ, permet d'explorer une large palette de données textuelles en résumant l'information concernant une dissimilarité et un voisinage à un seul indicateur.

Partie III

APPLICATIONS MUSICALES

CHAPITRE 7

Formats symboliques de données musicales

La musique se transmet principalement de deux manières : par le son ou par l'écriture (pour une revue des sources de données musicales et de leur historique, voir par exemple Vatolkin, 2013, section 2.1.2). Si l'on compare cela à la linguistique, le son représente la parole ; et l'écriture, le texte qui retranscrit cette parole. Concernant le son, des fichiers audio sont utilisés (voir par exemple Kriesel, 2013, section 2.2). Dans le cas de l'écriture, ce qui nous intéresse ici, on utilise le plus souvent des partitions (section 7.1).

Cependant, pour traiter les partitions avec un ordinateur, il faudra les numériser et leur donner un aspect « textuel ». Ceci est à peu près équivalent à utiliser, pour l'analyse textuelle, un fichier en format .txt et non un .pdf. On parlera alors de **partitions numériques** (*digital scores*) ou de **formats symboliques** (*symbolic formats*) [1]. Pour rappel, les données musicales symboliques sont définies comme « La description détaillée de toutes les informations nécessaires à l'affichage (ou gravure) précis d'une partition. » (Faget, 2011, p. 12).

Un format symbolique très connu pour la musique est le MIDI qui, de plus, produit du son (section 7.2). Bien que ce ne soit pas fait dans ce travail, il est possible d'extraire l'information d'un fichier MIDI pour pouvoir l'analyser de manière « textuelle ». Cependant, il existe aussi d'autres formats qui reproduisent les partitions sous forme de texte et qui sont souvent accompagnés de logiciels permettant de transformer des fichiers MIDI dans ce format et inversément. Parmi les nombreux formats existants, uniquement trois seront présentés dans ce qui suit (section 7.3) : Melisma, ABC et Humdrum.

7.1 Partitions

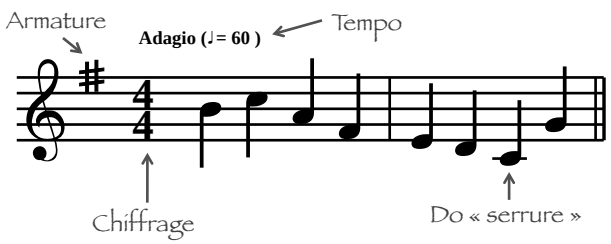


FIGURE 7.1 – Principales informations transmises par une partition.

Les principales informations transmises par une partition (figure 7.1) sont :

1. Il faut éviter de parler de formats numériques, car ces derniers incluent aussi les fichiers audio.

- **les informations générales**, telles que le titre, le nom du compositeur, etc. ;
- **le tempo**, indiqué par un mot ou un groupe des mots, comme par exemple *lento*, *adagio*, *allegretto*, *presto* ou *andante non troppo e con molta espressione* ; ou indiqué par une pulsation pas minute pour une durée ou une valeur de note donnée ;
- **les instruments**, lorsque la partition concerne plusieurs instruments ;
- **l'armure** ou **l'armature**, qui est l'ensemble d'altérations indiquant la tonalité du morceau de musique ;
- **le chiffrage** ou **la mesure**, qui donne une information sur la rythmique ;
- **les répétitions** ;
- **les notes**, et en particulier :
  - **leur hauteur** (do (*C* en anglais), ré (*D*), ..., la (*A*) et si (*B*), et au milieu d'un clavier de piano, le do « serrure ») et
  - **leur durée** ou **leur valeur** (croche (♪), noire (♩), blanche (𝅗𝅥), ronde (𝅝), etc.) ;
- **les silences** ;
- **les nuances** (*ppp, pp, p, mp, mf, f, ff, fff, crescendo, diminuendo, appassionato, pesante*, etc.) ;
- etc.

Deux extraits de partitions, qui seront utilisés pour les exemples concernant les formats symboliques dans la suite de ce chapitre, sont présentés dans les figures 7.2 et 7.3.

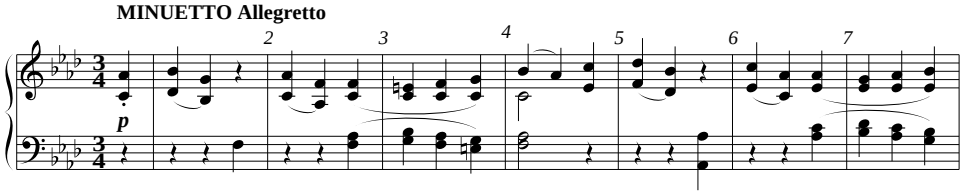


FIGURE 7.2 – Extrait de la « sonate pour piano n°1 en fa mineur, op. 2 n°1, 3ème mouvement » de Beethoven.

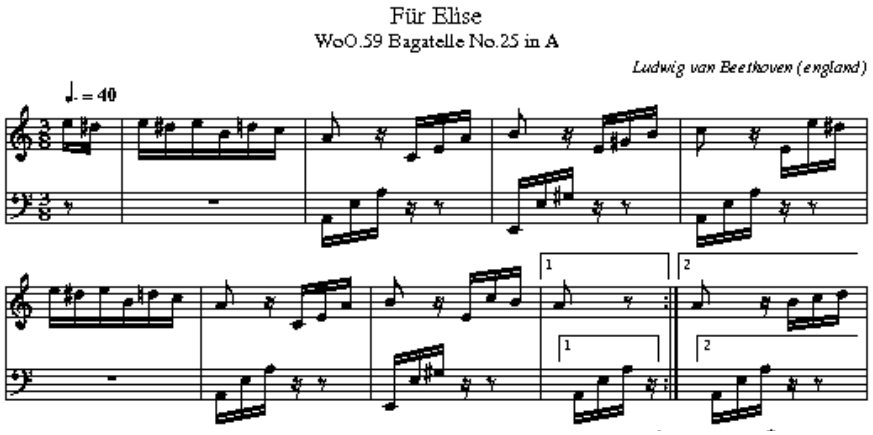


FIGURE 7.3 – Extrait de « Für Elise » de Beethoven, en Angleterre [*sic*], résultant du code au format ABC de la figure 7.7. Source : `http://abcnotation.com/getResource/downloads/image/fur-elise.png?a=trillian.mit.edu/~jc/music/abc/mirror/home.quicknet.nl/england/1837`.

7.2 Format MIDI en bref

Le M.I.D.I (*Musical Instrument Digital Interface* = Interface numérique pour instrument de musiques) est apparu en 1982-1983 (`http://www.midi.org`). À la base, c'est à la fois une interface et un protocole qui permettent aux instruments de musique numérique ou électronique de communiquer entre eux. Ce qui nous intéresse ici est le format de fichier MIDI qui est une structure de données permettant de transcrire de la musique. Ce fichier ne contient pas des sons, mais des instructions individuelles correspondant à des notes de musique pour chaque instrument. Les principales instructions qu'on trouve dans un fichier MIDI sont :

— les informations générales, telles que le rythme, le chiffrage, la tonalité, etc. ; et
— des pistes contenant le début et la fin des notes, leur hauteur, leur volume, etc.

En particulier, la hauteur des notes est codée par des nombres compris entre 0 et 127, avec le nombre 60 pour le do « serrure ».

7.3 Formats « textuels »

7.3.1 Le format Melisma

Le système *The Melisma Music Analyzer* (`http://www.link.cs.cmu.edu/music-analysis/`) permet d'analyser de la musique et prend, en entrée, des fichiers que l'on appellera « fichiers au format Melisma ». Ce système contient un programme, « mftext », qui permet de convertir des fichiers MIDI en fichiers Melisma. L'extension de ces derniers est .notes. Ils contiennent, dans l'ordre chronologique, les notes jouées avec leur hauteur en nombres, identiques à ceux des fichiers MIDI, ainsi que les temps de début et de fin de ces notes en millisecondes. Il existe deux types de formats Melisma : un dont chaque ligne représente une note (figures 7.4 et 7.5) ; et un autre dont chaque note est écrite sur deux lignes, soit une ligne pour le début de la note et l'autre pour la fin de la même note (figure 7.6).

```
Note 750 937 68
Note 1125 1500 61
Note 1125 1500 70
Note 1500 1875 58
Note 1500 1875 67
Note 2250 2625 60
Note 2250 2625 68
Note 2625 2906 56
Note 2625 2906 65
Note 3000 3375 60
Note 3000 3375 65
Note 3375 3750 60
Note 3375 3750 64
```

```
Note 3750 4125 60
Note 3750 4125 65
Note 4125 4500 60
Note 4125 4406 67
Note 4500 4875 70
Note 4500 5250 60
Note 4875 5156 68
Note 5250 5437 63
Note 5250 5437 72
Note 5625 6000 65
Note 5625 6000 73
Note 6000 6281 61
Note 6000 6281 70
```

```
Note 6750 7125 63
Note 6750 7125 72
Note 7125 7500 60
Note 7125 7500 68
Note 7500 7781 63
Note 7500 7781 68
Note 7875 8250 63
Note 7875 8250 67
Note 8250 8625 63
Note 8250 8625 68
Note 8625 9000 63
Note 8625 8906 70
```

FIGURE 7.4 – Format Melisma (une note par ligne) pour l'extrait de la partition de la figure 7.2. Ce fichier est une retranscription de la partition, mais il semblerait qu'il ne contienne que la première des deux portées, qu'il manque le premier do et que la durée d'une noire soit environ égale à 375 ms. Source : `http://www.link.cs.cmu.edu/link/ftp-site/music-analysis/notefiles/misc/beet.fmison.III.q.notes`. Il faut remarquer que dans cette figure, ainsi que dans les suivantes de ce chapitre, les encadrés sont « empilés », c'est-à-dire qu'ils constituent les parties successives d'un seul et même fichier.

```
Reference COM: Beethoven, Ludwig van
Reference CDT: 1770///-1827///
Reference OTL: Piano Sonata no. 1, mvmt. 3
Reference OMD: Minuet: Allegretto
Reference OPS: Op. 2
Reference ONM: No. 1
Reference OMV: No. 3
Reference AGN: Minuet and Trio
Comment Minuetto: Allegretto
Info keysig 4 sharps
Info key F Minor
Info Tempo 118 MM per quarter note
Note 0 508 60
Note 0 508 68
Note 508 1017 61
Note 508 1017 70
Note 1017 1525 58
Note 1017 1525 67
Note 1525 2034 53
Note 2034 2542 60
Note 2034 2542 68
Note 2542 3051 56
Note 2542 3051 65
Note 3051 3559 53
```

```
Note 3051 3559 56
Note 3051 3559 60
Note 3051 3559 65
Note 3559 4068 55
Note 3559 4068 58
Note 3559 4068 60
Note 3559 4068 64
Note 4068 4576 53
Note 4068 4576 56
Note 4068 4576 60
Note 4068 4576 65
Note 4576 5085 52
Note 4576 5085 55
Note 4576 5085 60
Note 4576 5085 67
Note 5085 6102 53
Note 5085 6102 56
Note 5085 5593 70
Note 5085 6102 60
Note 5593 6102 68
Note 6102 6610 63
Note 6102 6610 72
Note 6610 7119 65
Note 6610 7119 73
```

```
Note 7119 7627 61
Note 7119 7627 70
Note 7627 8136 44
Note 7627 8136 56
Note 8136 8644 63
Note 8136 8644 72
Note 8644 9153 60
Note 8644 9153 68
Note 9153 9661 56
Note 9153 9661 60
Note 9153 9661 63
Note 9153 9661 68
Note 9661 10169 58
Note 9661 10169 61
Note 9661 10169 63
Note 9661 10169 67
Note 10169 10678 56
Note 10169 10678 60
Note 10169 10678 63
Note 10169 10678 68
Note 10678 11186 55
Note 10678 11186 58
Note 10678 11186 63
Note 10678 11186 70
```

FIGURE 7.5 – Format Melisma (une note par ligne) pour l'extrait de la partition de la figure 7.2, avec le même principe « d'empilement » des encadrés que dans la figure 7.4. Ce fichier a été créé par une conversion automatique d'un fichier Humdrum (cf. section 7.3.3), c'est pourquoi les premières lignes contiennent des informations générales supplémentaires. De plus, le tempo indique 118 pulsations par minutes pour une noire, soit une durée d'environ 508.5 ms pour une noire. Source : `http://kern.ccarh.org/cgi-bin/ksdata?l=users/craig/classical/beethoven/piano/sonata&file=sonata01-3.krn&f=melisma`.

```
Note-on 2326 60
Note-on 2327 68
Note-off 2395 68
Note-off 2462 60
Note-on 2687 70
Note-on 2704 61
Note-on 3020 58
Note-off 3025 61
Note-on 3037 67
Note-off 3038 70
Note-off 3158 67
Note-off 3213 58
Note-on 3384 53
Note-off 3517 53
Note-on 3739 68
Note-on 3747 60
Note-off 4027 68
Note-on 4055 56
Note-off 4063 60
Note-on 4078 65
Note-off 4157 56
Note-off 4172 65
Note-on 4422 65
Note-on 4443 60
Note-on 4446 56
Note-on 4477 53
Note-off 4698 60
Note-off 4707 53
Note-off 4739 56
Note-on 4772 55
```

```
Note-on 4779 64
Note-on 4780 58
Note-on 4784 60
Note-off 4833 65
Note-off 4990 60
Note-off 5037 58
Note-off 5097 55
Note-on 5134 60
Note-on 5135 56
Note-on 5138 53
Note-on 5139 65
Note-off 5178 64
Note-off 5333 60
Note-off 5408 56
Note-off 5437 53
Note-on 5460 67
Note-on 5468 55
Note-on 5470 60
Note-on 5476 52
Note-off 5493 65
Note-off 5718 60
Note-off 5795 55
Note-on 5812 53
Note-off 5817 52
Note-on 5827 70
Note-on 5828 60
Note-on 5839 56
Note-off 5840 67
Note-off 6108 70
Note-on 6156 68
```

```
Note-off 6169 53
Note-off 6199 60
Note-off 6221 56
Note-off 6235 68
Note-on 6521 63
Note-on 6538 72
Note-off 6628 72
Note-off 6672 63
Note-on 6888 65
Note-on 6900 73
Note-off 7221 73
Note-on 7231 61
Note-on 7237 70
Note-off 7281 65
Note-off 7356 70
Note-off 7400 61
Note-on 7579 44
Note-on 7598 56
Note-off 7697 44
Note-off 7769 56
Note-on 7941 63
Note-on 7951 72
Note-off 8244 72
Note-off 8253 63
Note-on 8264 60
Note-on 8277 68
Note-off 8370 60
Note-off 8416 68
Note-on 8604 63
Note-on 8614 68
```

```
Note-on 8617 56
Note-on 8628 60
Note-off 8847 63
Note-off 8907 56
Note-on 8946 58
Note-on 8962 63
Note-on 8966 61
Note-on 8967 67
Note-off 8969 60
Note-off 8995 68
Note-off 9167 63
Note-off 9243 61
Note-on 9295 63
Note-on 9304 60
Note-on 9310 68
Note-on 9317 56
Note-off 9352 67
Note-off 9355 58
Note-off 9504 63
Note-off 9565 60
Note-off 9590 56
Note-on 9640 58
Note-on 9649 63
Note-on 9652 70
Note-on 9672 55
Note-off 9695 68
Note-off 9878 63
Note-off 9982 55
Note-off 10004 70
```

FIGURE 7.6 – Format Melisma (deux lignes pour une note) pour l'extrait de la partition de la figure 7.2, avec le même principe « d'empilement » des encadrés que dans la figure 7.4. Ce fichier a été produit par la conversion automatique d'un fichier MIDI (cf. section 7.2), ainsi la durée d'une noire peut varier selon l'interprétation du musicien. Source : `http://www.link.cs.cmu.edu/link/ftp-site/music-analysis/notefiles/misc/beet.fmison.III.p.notes`.

7.3.2 Le format ABC

```
X:1838
T:F\"ur Elise
T:Bagatelle No.25 in A, WoO.59
O:england
C:Ludwig van Beethoven
%http://www.musicaviva.com/beethoven-ludwig-van.abc
V:1 Program 1 0 %Piano
V:2 Program 1 0 bass %Piano
M:3/8
L:1/16
Q:3/8=40
K:Am
V:1
e^d|e^deB=dc|A2 z CEA|B2 z E^GB|c2 z Ee^d|
V:2
z2|z6|A,,E,A, z z2|E,,E,^G, z z2|A,,E,A, z z2|
%
V:1
e^deB=dc|A2 z CEA|B2 z EcB|[1A2 z2:|[2A2z Bcd|
V:2
z6|A,,E,A, z z2|E,,E,^G, z z2|[1A,,E,A, z :|[2A,,E,A, z z2|
```

FIGURE 7.7 – Format ABC pour l'extrait de la partition de la figure 7.3. Source : `http://abcnotation.com/tunePage?a=trillian.mit.edu/~jc/music/abc/mirror/home.quicknet.nl/england/1837`.

Un fichier au format ABC (`http://abcnotation.com/`) a comme extension : .abc. [2] Comme pour le format Melisma, il existe un programme permettant de transformer un fichier MIDI en fichier ABC. Il se compose d'un préambule et d'un corps (figures 7.7 et 7.8). Les principales informations du préambule sont :

- un numéro de référence (X) ;
- un titre (T) ;
- le nom du compositeur (C) ;
- la durée de référence des notes (L), qui va servir de base pour indiquer la durée de chaque note dans le corps du fichier, où 1/4 correspond à une noire, 1/8, à une croche, 1/16, à une double croche, etc. ;
- le chiffrage (M) ;
- le tempo (Q), indiqué, comme pour les partitions (cf. section 7.1), avec un mot ou des pulsations par minute pour une durée de note donnée ;
- la tonalité (K) ;
- etc.

L'ordre du préambule est strict concernant le numéro de référence et le titre, qui doivent toujours être au début, et la tonalité, qui doit toujours être à la fin du préambule, contrairement aux autres éléments.

Dans le corps, chaque ligne représente une portée telle qu'elle apparaît sur la partition. Lorsqu'il s'agit d'un système de portées, toutes les portées du système sont représentées à la suite et indiquées par « V » suivi d'un nombre. Les principales notations utilisées dans le corps sont les suivantes :

2. Il existe une variante de ce format, très similaire, nommée « ABC Plus » (`http://abcplus.sourceforge.net/`), dont un exemple est présenté dans la figure 7.8.

— des lettres pour la hauteur des notes, correspondant aux noms des notes en anglais, avec « C » pour le do « serrure », « C, » pour le do une octave en-dessous, et pour chaque octave plus basse, une virgule est ajoutée ; « c » représente le do une octave en-dessus du do « serrure », « c' », le do encore une octave au-dessus, et des apostrophes sont ajoutées pour chaque octave plus haute ;
— la lettre « z » pour les silences ;
— des nombres, pour la durée des notes, relatifs à la durée de référence indiquée dans le préambule (L) et précédés d'un « / » lorsque la durée est plus courte que celle de référence ;
— d'autres symboles pour les altérations : _ pour ♭ , = pour ♮ et ^ pour ♯ ;
— des guillemets pour les accords écrits explicitement sur une partition, par exemple "Gm7" ;
— des crochets pour les notes jouées simultanément, ou en d'autres termes, les accords écrits note par note sur une portée, par exemple [CEGc] ;
— divers symboles pour représenter les différentes barres de mesure, tels que | pour une barre de mesure simple, || pour une barre de mesure double marquant une partie du morceau, |] pour la barre de mesure indiquant la fin d'un morceau, :| pour la barre de mesure qui indique une répétition, etc. ;
— etc.

En conclusion, ce format est particulièrement adapté pour la création de partitions.

```
X: 1
T: Piano Sonata no. 1, mvmt. 3
C: Ludwig van Beethoven
%%abc-version 2.0
%%abcx-abcm2ps-target-version 5.9.1 (29 Sep 2008)
%%abc-creator hum2abc beta
%%abcx-conversion-date 2012/04/13 12:40:19
%%abc-edited-by Craig Stuart Sapp
%%abcx-initial-encoding-date 2004/04/06/
%%gracespace 0 6 6
%%notespacingfactor 1.85
%%humdrum-veritas 3897117643
%%humdrum-veritas-data 871200473
%%continueall 1
%%barnumbers 0
F: http://kern.ccarh.org/cgi-bin/ksdata?l=users/craig/
classical/beethoven/piano/
sonata&file=sonata01-3.krn&f=abcplus
L: 1/4
M: 3/4
Q: "Minuet: Allegretto" 1/4=116
```

```
%%staves {1 2}
V: 1 clef=treble
V: 2 clef=bass
K: Ab
[V:1] .[CA] [I:setbarnb 1]|
[V:2] z |
[V:1] ([DB][B,G])z |
[V:2] zzF, |
[V:1] ([CA][A,F])([CF] |
[V:2] zz([F,A,] |
[V:1] [C=E][CF][CG]) |
[V:2] [G,B,][F,A,][=E,G,]) |
[V:1] (BA)[Ec] & C2z |
[V:2] [F,2A,2]z |
[V:1] ([Fd][DB])z |
[V:2] zz[A,,A,] |
[V:1] ([Ec][CA])([EA] |
[V:2] zz([A,C] |
[V:1] [EG][EA][EB]) |
[V:2] [B,D][A,C][G,B,]) |
```

FIGURE 7.8 – Format ABC Plus pour l'extrait de la partition de la figure 7.2, avec le même principe « d'empilement » des encadrés que dans la figure 7.4. Ce fichier a été créé par une conversion automatique d'un fichier Humdrum (cf. section 7.3.3). Source : `http://kern.ccarh.org/cgi-bin/ksdata?l=users/craig/classical/beethoven/piano/sonata&file=sonata01-3.krn&f=abcplus`.

7.3.3 Le format Humdrum

Les fichiers au format Humdrum (ou format **kern), disponibles sur le site `http://kern.ccarh.org/`, ont été créés, le plus souvent avec un programme de reconnaissance optique de musique, pour être traités avec le *Humdrum Toolkit for Music Research*[3] (Sapp, 2005). Ce logiciel a été conçu pour assister les chercheurs en musique et offre de nombreuses possibilités

3. `http://humdrum.org/Humdrum/install.html`.

(Huron, 1994, 1998). En plus de ce logiciel, il existe une série de programmes (*Humdrum extras*, `http://extra.humdrum.org/`) qui permettent, comme le logiciel, la transposition de partitions ou la sélection de différentes parties, mais aussi de de convertir les fichiers Humdrum en d'autres formats, tels ceux présentés ci-dessus (figures 7.5 et 7.8). Cette série de programmes sera utilisée pour les manipulations des fichiers dans le chapitre 8.

```
!!!COM: Beethoven, Ludwig van
!!!CDT: 1770///-1827///
!!!OTL: Piano Sonata no. 1, mvmt. 3
!!!OMD: Minuet: Allegretto
!!!OPS: Op. 2
!!!ONM: No. 1
!!!OMV: No. 3
!!!AGN: Minuet and Trio
**kern **dynam **kern **dynam
*Ipiano *Ipiano *Ipiano *Ipiano
*>[A,A,B,B,C,C,D,D,A,B] *>[A,A,B,B,C,C,D,D,A,B]
*>[A,A,B,B,C,C,D,D,A,B] *>[A,A,B,B,C,C,D,D,A,B]
*>norep[A,B,C,D,A,B] *>norep[A,B,C,D,A,B]
*>norep[A,B,C,D,A,B] *>norep[A,B,C,D,A,B]
!! Minuetto: Allegretto
*>A *>A *>A *>A
*clefF4 *clefF4 *clefG2 *clefG2
*k[b-e-a-d-] *k[b-e-a-d-] *k[b-e-a-d-] *k[b-e-a-d-]
*f: *f: *f: *f:
*M3/4 *M3/4 *M3/4 *M3/4
*MM118 *MM118 *MM118 *MM118
4r . 4c'/ 4a-'/ p
=1 =1 =1 =1
4r . (4d-/ 4b-/ .
4r . 4B-/) 4g/ .
4F\ . 4r .
=2 =2 =2 =2
4r . (4c/ 4a-/ .
4r . 4A-/) 4f/ .
4F\ (4A-\ . (4c/ 4f/ .
=3 =3 =3 =3
4G\ 4B-\ . 4c/ 4e/ .
```

```
4F\ 4A-\ . 4c/ 4f/ .
4E\ 4G\) . 4c/) 4g/ .
=4 =4 =4 =4
* * *^ *
2F\ 2A-\ . (4b-/ 2c\ .
. . 4a-/) . .
4r . 4e-/ 4cc/ 4r .
* * *v *v *
=5 =5 =5 =5
4r . (4f/ 4dd-/ .
4r . 4d-/) 4b-/ .
4AA-\ 4A-\ . 4r .
=6 =6 =6 =6
4r . (4e-/ 4cc/ .
4r . 4c/) 4a-/ .
4A-\ (4c\ . (4e-/ 4a-/ .
=7 =7 =7 =7
4B-\ 4d-\ . 4e-/ 4g/ .
4A-\ 4c\ . 4e-/ 4a-/ .
4G\ 4B-\) . 4e-/) 4b-/ .
=8 =8 =8 =8

etc.

==:|! ==:|! ==:|! ==:|!
*-*-*-*-
!!!ENC: Craig Stuart Sapp
!!!END: 2004/04/06/
!!!ONB: preliminary proof
reading done on 2008/10/20/
!!!hum2abc: --spacing 1.85
```

FIGURE 7.9 – Format Humdrum pour l'extrait de la partition de la figure 7.2, avec le même principe « d'empilement » des encadrés que dans la figure 7.4. Source : `http://kern.ccarh.org/cgi-bin/ksdata?l=users/craig/classical/beethoven/piano/sonata&file=sonata01-3.krn&f=kern`.

L'extension utilisée pour ce format est : .krn. Comme pour le format ABC, il est composé d'un préambule et d'un corps (figure 7.9). En plus, une série de commentaires est généralement présente au début du fichier. La structure de ces fichiers est très différente de celle des fichiers ABC, car ici chaque colonne représente une voix de la partition.

Les informations générales (comme le titre et le compositeur) se trouvent dans les commentaires au début du fichier [4]. Le préambule est divisé en colonnes comme le corps. Chaque colonne contient :

— le début, indiqué par l'expression : **kern ;
— une indication de la portée ou de l'instrument ;

4. Selon la littérature, les commentaires généraux sont indiqués par deux point d'exclamation (!!) et les commentaires concernant une seule des voix, par un point d'exclamation (!). Cependant, il semblerait que dans les fichiers disponibles sur `http://kern.ccarh.org/`, un point d'exclamation ait été ajouté à cette convention pour les premiers et les derniers commentaires du fichier.

— la clé ;
— l'armature (k[...]) ;
— la tonalité ;
— le chiffrage (M) ;
— parfois, le tempo (MM) ;
— etc.

et chaque information commence par une étoile.

Dans le corps du fichier, chaque ligne représente un moment, apparaissant dans l'ordre chronologique. Les principales notations utilisées sont les suivantes :

— des lettres pour la hauteur des notes, correspondant aux noms des notes en anglais, avec « c » pour le = do « serrure », puis « cc » pour le do une octave en-dessus, puis « ccc », etc., et le même principe est appliqué avec des lettre majuscules pour les notes plus graves, soit « C » pour le do une octave en-dessous de « c », puis « CC », etc. ;
— la lettre « r »pour les silences ;
— des nombres fixes pour la durée des notes, avec, par exemple, 1 pour la ronde, 4 pour la noire et 2. pour la blanche pointée ;
— d'autres symboles pour les altérations : - pour ♭, n pour ♮ et # pour ♯[5] ;
— des signes d'égalité pour indiquer les barres de mesures, éventuellement suivis du numéro de la mesure dans la partition.

7.3.4 Comparaison de ces trois formats

Le format ABC est particulièrement bien conçu pour conserver un maximum d'informations et donc pour écrire des partitions (après transposition, ou autre changement), mais moins pour le traitement informatique, ne serait-ce qu'en raison de sa flexibilité. Le format Melisma est beaucoup plus simple, certainement le plus pratique pour l'analyse informatique, mais perd beaucoup d'information. Entre les deux, le format **kern est suffisamment structuré et fixe pour être traité informatiquement et conserve la grande majorité des informations contenues sur la partition. De plus, comme il a déjà été mentionné, il existe une base de données dédiée à ce format et il a l'avantage d'être lié à de nombreux programmes permettant, d'une part, d'obtenir les autres formats les plus utilisés, dont les trois présentés dans ce chapitre et, d'autre part, de transposer les partitions, d'extraire certaines informations, etc.

5. Contrairement au format ABC, ici les altérations sont notifiées pour chaque note, même lorsqu'elles sont déjà mentionnées dans le préambule.

CHAPITRE 8

Analyse de données musicales

Ce chapitre, qui reprend la structure, les méthodes, une partie du texte traduite et les résultats présentés dans l'article Cocco et Bavaud (accepté pour publication) et ajoute de nombreux résultats, présente une analyse exploratoire de données de musique polyphonique en format symbolique.

À cet effet (section 8.1), on divise la partition en durées égales, puis on transforme des partitions numériques (cf. chapitre 7) en tables de contingence qui comptent la durée de chaque note pour chaque intervalle de temps. Cette représentation, très proche de la représentation sur rouleau de piano pneumatique (*piano-roll representation*) et, pour les fichiers audio, de la représentation Chroma (voir par exemple Ellis et Poliner, 2007; Müller et Ewert, 2011; Kriesel, 2013, section 2.4), a l'avantage de représenter de la musique polyphonique dans un format compatible avec des méthodes d'analyse de données courantes, telles que l'AFC, et d'être invariante sous agrégation (cf. section 8.1.1).

Pour commencer (section 8.2), des morceaux de musique complets sont analysés, par l'intermédiaire de l'AFC et de l'indice d'autocorrélation. Ces deux méthodes permettent de découvrir des structures intrinsèques dans des partitions de musique, ainsi que d'en visualiser les patterns. Elles sont illustrées par un exemple monophonique et par plusieurs exemples polyphoniques.

Ensuite, dans la section 8.3, les différentes voix d'une même partition, ainsi que les liens qui existent entre elles, sont analysés par l'intermédiaire d'une analyse des correspondances multiples (ACM) floue et de l'indice d'autocorrélation croisée. Ces deux méthodes sont appliquées à deux partitions polyphoniques composées pour plusieurs instruments.

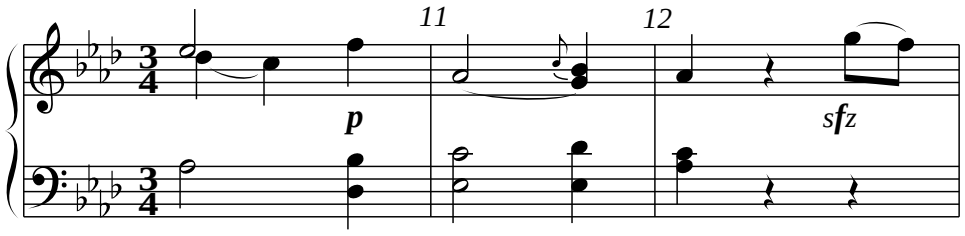
Finalement, une mesure de similarité entre deux partitions, basée sur la représentation des partitions de musique par des tables de contingence, est présentée dans la section 8.4. À partir de cette mesure de similarité, des partitions écrites par plusieurs compositeurs sont regroupées par une classification ascendante hiérarchique.

8.1 Représentation des données

8.1.1 Formalisme

Une partition musicale peut être représentée par une table de contingence *brute* $X = (x_{tj})$ qui croise les *intervalles de temps* ($t = 1, \ldots, n$) et la *hauteur des notes* ($j = 0, \ldots, m$). Cette table compte la durée de chaque hauteur de note dans chaque intervalle de temps. Ainsi, la répétition de notes de même hauteur dans un intervalle de temps n'est pas codée.

Aussi, toutes les hauteurs de note sont rapportées à l'octave et l'on attribue la valeur de 0 à do ; de 1 à do♯ ou ré♭ ; de 2 à ré, etc. [1] Ensuite, un *vrai silence*, z, qui correspond à un moment durant lequel aucune note n'est jouée, est ajouté. Au final, j peut prendre 13 valeurs différentes : 0 à 11 et z. Concernant les intervalles de temps, ils ont une durée constante qui vaut τ. Cette durée peut prendre n'importe quelle valeur, telle qu'un nombre de doubles croches, de mesures ou de millisecondes. Par conséquent, la durée totale d'une partition (ou d'un extrait) vaut $\tau_{\text{tot}} = n\tau$. Les figures 8.1 et 8.2 présentent deux exemples de la table de contingence transposée, un pour l'extrait d'une partition de piano et un autre pour l'extrait d'une partition pour un quatuor à cordes, chacune avec deux valeurs de τ.



	Intervalles de temps t avec $\tau =$ ♩								
notes j	29	30	31	32	33	34	35	36	37
0	0	4	0	4	4	0	4	0	0
1	4	0	4	0	0	4	0	0	0
3	4	4	0	4	4	4	0	0	0
5	0	0	4	0	0	0	0	0	2
7	0	0	0	0	0	4	0	0	2
8	4	4	0	4	4	0	8	0	0
10	0	0	4	0	0	4	0	0	0
z	0	0	0	0	0	0	0	4	0

	Intervalles de temps t avec $\tau =$ 𝅗𝅥.		
notes j	10	11	12
0	4	8	4
1	8	4	0
3	8	12	0
5	4	0	2
7	0	4	2
8	8	8	8
10	4	4	0
z	0	0	4

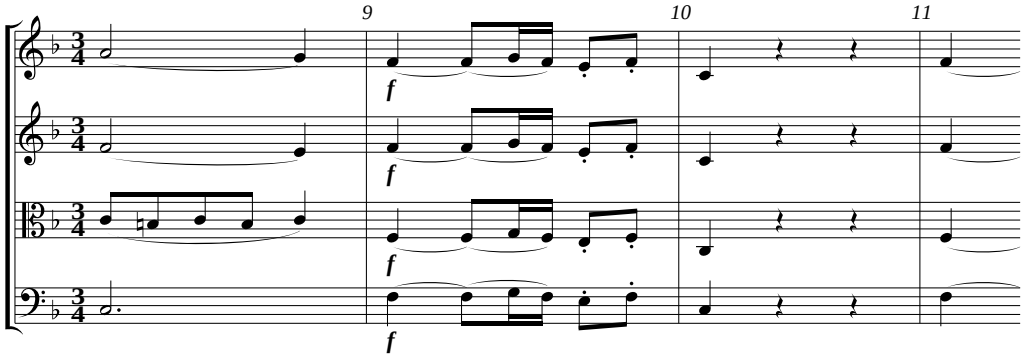
FIGURE 8.1 – Extrait de la « sonate pour piano n°1 en fa mineur, op. 2 n°1, 3ème mouvement » de Beethoven. Table de contingence **transposée** $X = (x_{tj})$, qui donne la durée de chaque hauteur de note en nombre de double-croches pour τ égal à une noire (haut) et à une blanche pointée (bas).

En plus de permettre de traiter de la musique polyphonique, cette représentation a l'avantage d'être invariante sous agrégation : ainsi doubler la valeur de τ revient à sommer les effectifs de deux intervalles de temps successifs. Donc, si T représente un intervalle de temps composé d'intervalles de temps plus petits t, alors les nouveaux effectifs deviennent :

$$\tilde{x}_{Tj} := \sum_{t \in T} x_{tj}$$

comme l'illustrent les figures 8.1 et 8.2. Lavrenko et Pickens (2003) et Morando (1981) utilisent des représentations relativement similaires, à ceci près que les premiers ne considèrent ni la durée des notes, ni celle entre les notes, et que le second base sa représentation sur la succession des accords. De plus, ces représentations ne sont pas invariantes sous agrégation, à l'inverse de celle présentée ici.

1. Si l'on décidait de ne pas reporter les notes à l'octave, le formalisme serait strictement identique, il suffirait d'augmenter le nombre de modalités j.



	Intervalles de temps t avec $\tau = \text{♩}$									
notes j	22	23	24	25	26	27	28	29	30	31
0	6	6	8	0	0	0	16	0	0	0
4	0	0	4	0	0	8	0	0	0	0
5	4	4	0	16	12	8	0	0	0	16
7	0	0	4	0	4	0	0	0	0	0
9	4	4	0	0	0	0	0	0	0	0
11	2	2	0	0	0	0	0	0	0	0
z	0	0	0	0	0	0	0	4	4	0

	Intervalles de temps t avec $\tau = \text{𝅗𝅥.}$			
notes j	8	9	10	11
0	20	0	16	0
4	4	8	0	0
5	8	36	0	16
7	4	4	0	0
9	8	0	0	0
11	4	0	0	0
z	0	0	8	0

FIGURE 8.2 – Extrait du « 1er mouvement *« Allegro con brio »* du Quatuor à cordes en fa majeur, op. 18 no 1 » de Beethoven. Table de contingence **transposée** $X = (x_{tj})$, qui donne la durée de chaque hauteur de note en nombre de double-croches pour τ égal à une noire (haut) et à une blanche pointée (bas).

Dans un second temps, la table brute $X = (x_{tj})$ est normalisée à $\Xi = (\xi_{tj})$, de façon à ce que la somme de chaque ligne $\xi_{t\bullet}$ soit égale à 1 :

$$\xi_{tj} = \frac{x_{tj}}{x_{t\bullet}} \tag{8.1}$$

Par conséquent, la même importance est donnée à chaque intervalle de temps ($f_t = 1/n$), quels que soient la durée et le nombre de notes qu'il contient, ce qui implique que $\xi_{\bullet\bullet} = n$.

Comme pour la table brute, il est possible d'agréger les intervalles de temps de la table normalisée. La table normalisée agrégée $\tilde{\Xi}$ s'obtient soit par des moyennes pondérées :

$$\tilde{\xi}_{Tj} = \frac{\sum_{t\in T} x_{t\bullet}\xi_{tj}}{\sum_{t\in T} x_{t\bullet}}$$

soit directement à partir de la table brute :

$$\tilde{\xi}_{Tj} = \frac{\sum_{t\in T} x_{tj}}{\sum_{t\in T} x_{t\bullet}} = \frac{\tilde{x}_{Tj}}{\tilde{x}_{T\bullet}}$$

8.1.2 Pré-traitement

Pour obtenir la représentation des données exposée ci-dessus, on commence par utiliser des fichiers au format Humdrum, qui sont bien structurés, indépendants de l'interprétation d'un

musicien et disponibles sur Internet (cf. section 7.3.3). En particulier, pour conserver l'œuvre dans sa version complète, on utilise les fichiers comportant toutes les répétitions, telles qu'elles sont indiquées sur la partition. Lorsque des modifications sont nécessaires, telle qu'une transposition ou l'extraction d'une voix, par exemple, la série de programmes *Humdrum extras* est utilisée. Ensuite, ils sont transformés en format Melisma (cf. section 7.3.1), plus simple pour le traitement informatique, à l'aide du programme « kern2melisma » de la série *Humdrum extras*. Avant cette transformation, le tempo des fichiers Humdrum est fixé, arbitrairement et sans conséquences sur les applications, à 100 pulsations par minute pour une noire (*MM100). Ainsi, lors de la transformation en fichier Melisma, une noire aura une durée exacte de 600 ms. Finalement, les fichiers Melisma sont transformés en tables de contingence brutes par l'intermédiaire d'un programme en Perl.

Soit un fichier Melisma dont chaque ligne, $l = 1, \ldots, L$, représente une note j, avec t_{deb} le temps de début de la note et t_{fin} le temps de fin de cette note. On choisit une durée τ, en millisecondes [2], puis, pour chaque ligne l, on obtient, pour les s entiers compris entre $\lfloor \frac{t_{\text{deb}}}{\tau} \rfloor$ (inclus) et $\lceil \frac{t_{\text{fin}}}{\tau} \rceil$ (non inclus), les éléments de la table de contingence (temporaire) comme :

$$X_{sj}^{\text{temp}} = \min\left(t_{\text{fin}}, \tau\left(s+1\right)\right) - \max\left(t_{\text{deb}}, \tau s\right) \tag{8.2}$$

On procède de la même manière pour les silences ($j = z$), soit lorsque le temps de début de la note sur la ligne l, $t_{\text{deb}}\left(l\right)$, est plus grand que les temps de fin des notes précédentes, soit des lignes $1, \ldots, l-1$, $t_{\text{fin}}\left(l-1\right)$, en posant, dans (8.2), $t_{\text{deb}} = t_{\text{fin}}\left(l-1\right)$ et $t_{\text{fin}} = t_{\text{deb}}\left(l\right)$. Pour terminer, tous les effectifs de ces tables temporaires sont additionnés pour obtenir la table de contingence brute X.

Lors de cette procédure, les silences présents sur la partition à la fin ou au début du morceau de musique sont perdus. Ils sont alors ajoutés « manuellement » pour conserver toutes les informations de la partition.

Dans un second temps, l'agrégation est exécutée dans R. Dans toutes les applications de ce chapitre, à l'exception des figures 8.21 et 8.22, lorsque la durée τ est plus grande ou égale à une mesure et que le morceau de musique commence avec une *anacrouse* (ou levée), cette dernière est ajoutée à la première mesure lors de l'agrégation. Aussi, si le choix de la durée τ ne permet pas d'obtenir des diviseurs entiers de τ_{tot}, alors le dernier intervalle de temps, n, sera plus court lors de l'agrégation. Finalement, toujours dans R, les tables de contingence brutes sont normalisées pour obtenir la table Ξ (8.1).

8.2 Analyses d'une partition

8.2.1 Traitements

Comme expliqué dans la section 1.4 et mis en œuvre dans la section 4.2, il est possible de pratiquer une AFC pour visualiser des données représentées sous la forme d'une table de contingence. Pour ce faire, on utilise le MDS (cf. section 1.4.1). En premier lieu, les dissimilarités du khi2 entre les intervalles de temps $\hat{D}_{st}$ (respectivement entre les hauteurs de notes $\check{D}_{ij}$) sont calculées par (1.6) (resp. (1.7)) sur la table de contingence normalisée (8.1). Ensuite, par (1.24), on obtient la matrice des produits scalaires pondérés entre les intervalles de temps $\hat{K}$ (resp. entre les hauteurs de note $\check{K}$), dont la décomposition spectrale va permettre de calculer les coordonnées factorielles (1.25).

D'autre part, les intervalles de temps, ordonnés chronologiquement, peuvent s'interpréter comme des positions. Ainsi, il est possible de mesurer la différence entre la variabilité de l'ensemble des dissimilarités du khi2 entre les intervalles de temps ($\hat{D}_{st}$) et la variabilité locale

2. Pour ne procéder qu'une fois à la transformation des fichiers Melisma en tables de contingence avant les éventuelles agrégations, on choisit une valeur assez faible de τ, par exemple une croche.

de ces dissimilarités dans un voisinage défini par E, grâce à l'indice d'autocorrélation δ (3.4), comme il a été fait pour les textes dans le chapitre 6. Concernant la matrice d'échange, seule la matrice périodique (3.2), déjà utilisée pour les textes dans la figure 6.2, sera adoptée. Pour rappel, cette dernière, contrairement aux autres matrices d'échange, a l'avantage de permettre de considérer deux positions (une à gauche et une à droite) à une distance r d'une position donnée, sans considérer les positions qui les séparent. De plus, le voisinage est périodique, ce qui correspond au cas d'un morceau de musique joué en continu.

8.2.2 Partition monophonique

Afin de mieux appréhender les résultats obtenus avec ces méthodes, le premier exemple traite une chanson enfantine, dont la mélodie est connue et qui, en plus, a l'avantage d'être monophonique.

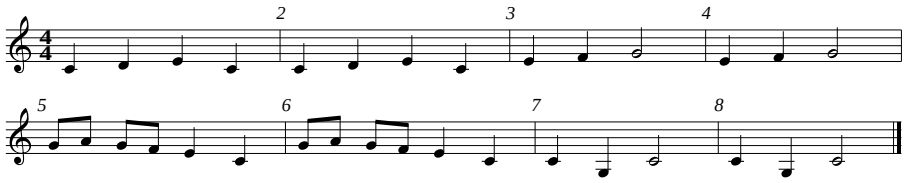


FIGURE 8.3 – Partition de « Frère Jacques » en do majeur.

La figure 8.3 présente la partition de Frère Jacques transposée en do majeur (le fichier Humdrum original était en mi♭ majeur) ; et la figure 8.4, l'AFC appliquée sur cette partition. Dans cette dernière, lorsque τ est égal à une croche (graphiques du haut), alors une note, au maximum, est jouée durant chaque intervalle de temps, ce qui signifie que la représentation est totalement monophonique. Dans ce cas, les dissimilarités euclidiennes carrées entre les intervalles de temps sont des *dissimilarités en étoile*, donc de la forme $\hat{D}_{st} = a_s + a_t$ (voir par exemple Critchley et Fichet, 1994). Par conséquent, toutes les valeurs propres sont identiques et il est difficile de compresser les données par l'intermédiaire d'une analyse factorielle. Aussi, sur le biplot, les coordonnées des intervalles de temps coïncident exactement avec les coordonnées des hauteurs de notes, il est donc possible de suivre visuellement la partition. En augmentant la valeur de τ à une noire (graphique en bas, à gauche), le nombre d'intervalles de temps diminue et, ainsi, l'inertie expliquée par les deux premières dimensions augmente. On remarque aussi que les coordonnées factorielles dans les deux premières dimensions sont identiques pour trois notes, à savoir fa (5), sol (7) et la (9). Finalement, avec τ égal à une mesure (graphique en bas, à droite), la structure du morceau de musique apparaît, avec chaque mesure jouée deux fois. On remarque aussi l'alignement de la succession des intervalles de temps en forme de « fer à cheval ». Cet alignement est typique d'un *effet de Guttman* (*arch* ou *horseshoe effect*) se produisant lorsque les modalités sont ordonnées, ce qui est le cas ici selon l'ordre chronologique (voir par exemple Gauch, Whittaker et Wentworth, 1977; Camiz, 2005).

De plus, on observe sur la figure 8.4, comme déjà évoqué, que l'inertie expliquée par les deux premiers facteurs varie en fonction de τ, car le nombre d'intervalles de temps diminue lorsque la durée τ augmente et, par conséquent, le nombre de dimensions α (cf. section 1.4.1) décroît aussi. Ainsi, l'inertie expliquée par les premiers facteurs augmente (graphique de gauche de la figure 8.5) et l'inertie totale Δ (1.17) diminue (graphique de droite de la figure 8.5) avec τ. En particulier, dans ces deux figures, l'inertie reste constante lorsque τ est plus petit ou égal à une croche, soit la plus petite durée d'une note dans la partition, et lorsque τ est compris entre une ronde, donc une mesure, et deux rondes, car chaque mesure est répétée une fois. On observe aussi, qu'à l'inverse des résultats obtenus avec les diviseurs entiers de τ_{tot} qui évoluent régulièrement, les résultats calculés avec toutes les valeurs de τ sont plus fluctuants. En fait, lors

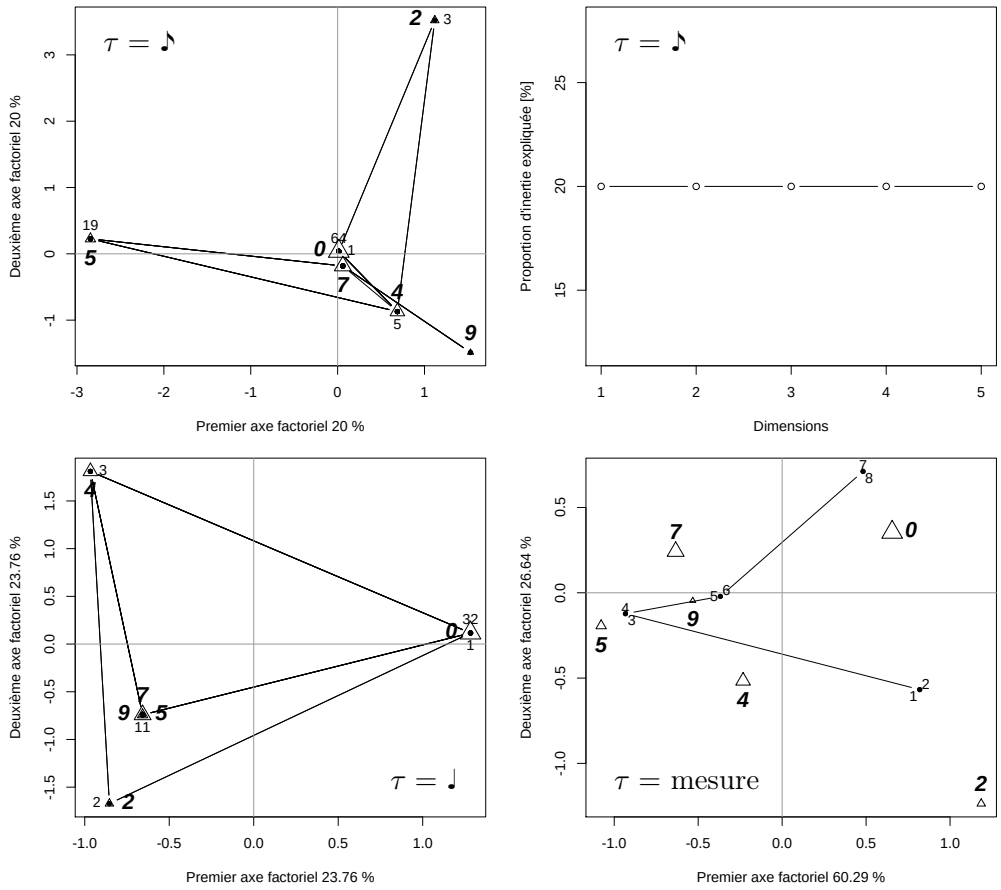


FIGURE 8.4 – AFC sur « Frère Jacques ». En haut, à gauche : biplot avec τ égal à une croche. Sur ce graphique et sur les suivants dans ce chapitre, les triangles, avec des nombres en italique de grande taille, représentent les hauteurs de notes, la taille des triangles étant proportionnelle au nombre de notes dans le morceau de musique ; et les cercles pleins, parfois étiquetés avec des nombres de petite taille, les intervalles de temps. Ces derniers sont reliés dans l'ordre chronologique selon la progression du temps. En haut, à droite : valeurs propres pour le biplot de gauche. En bas à gauche : biplot avec τ égal à une noire. En bas à droite : biplot avec τ égal à une mesure.

de l'agrégation des effectifs des tables de contingence dans ce second cas, comme déjà mentionné, la durée du dernier intervalle de temps est plus courte et, par conséquent, le partitionnement du morceau de musique n'est pas régulier. Finalement, on constate que la courbe de l'inertie totale décroît de façon convexe, comme une hyperbole ou une exponentielle à exposant négatif.

La figure 8.6 présente l'indice d'autocorrélation calculé sur le morceau de musique « Frère Jacques » avec deux valeurs de τ différentes. En premier lieu, comme déjà expliqué pour la figure 6.2, on remarque que $\delta = 1$ lorsque $r = 0$ et que le graphique est symétrique. Sur le graphique de gauche, soit pour une valeur de τ égal à une noire, un pic significatif ($\delta = 0.495$) apparaît lorsque $r = 4$, soit pour une distance correspondant à une mesure. En fait, en raison de la répétition systématique de chaque mesure, à chaque moment t, les mêmes notes sont jouées à une distance $r = 4$, parfois à gauche, parfois à droite de t. Ce pic correspond à la durée τ d'une mesure, soit celle qui permet d'obtenir la meilleure visualisation de la partition par l'AFC dans cet exemple (graphique en bas à droite de la figure 8.4).

En posant τ égal à une mesure (graphique gauche de la figure 8.6), aucune valeur n'est significative et aucun pic n'apparaît. Il semble donc que cette durée soit trop élevée et que, par conséquent, trop d'information soit perdue. Cependant, il est tout de même possible d'observer que l'autocorrélation est positive ($\delta = 0.382$) lorsque $r = 1$, soit pour la répétition de chaque

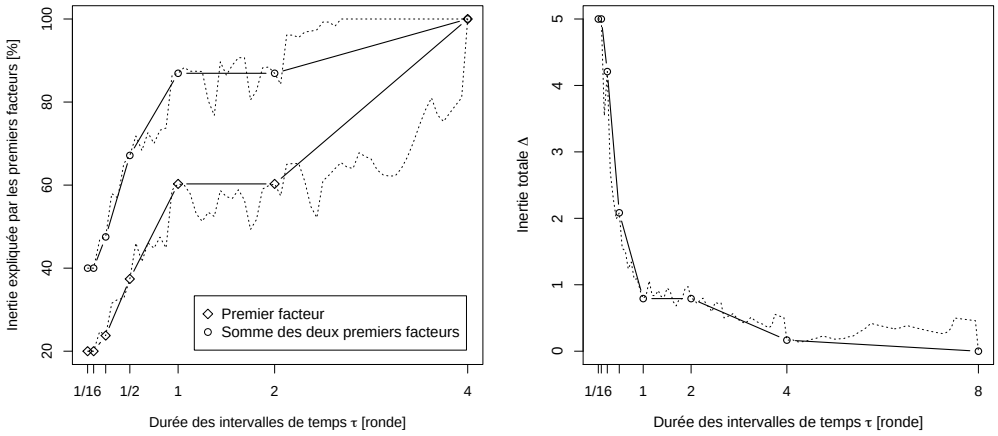


FIGURE 8.5 – AFC sur « Frère Jacques ». Proportion d'inertie expliquée par les premiers facteurs (gauche) et inertie totale (droite) en fonction de la valeur de τ. Dans ces deux graphiques, la ligne pointillée représente les résultats pour toutes les durées et la ligne continue, les résultats pour les diviseurs entiers de τ_{tot}.

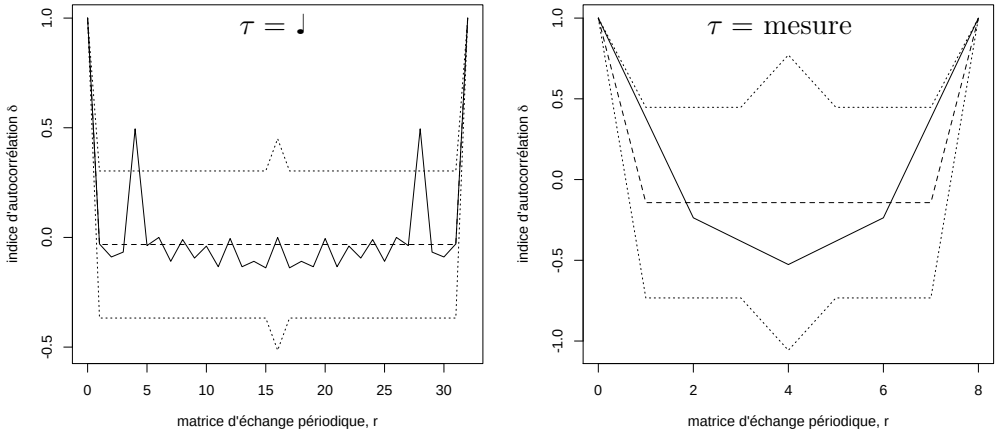


FIGURE 8.6 – Indice d'autocorrélation en fonction du décalage r variant entre 0 et n pour « Frère Jacques », avec τ égal à une noire (gauche) et à une mesure (droite). Dans cette figure, ainsi que dans les suivantes de ce chapitre, la ligne continue représente l'indice d'autocorrélation ; la ligne traitillée, la valeur attendue $E_0(\delta)$ (3.5) ; et les lignes pointillées, l'intervalle de confiance à 95% (3.6).

mesure, et qu'elle est négative pour r compris entre 2 et 4, soit quand on compare des mesures qui sont différentes, ce qui semble cohérent.

8.2.3 Partitions polyphoniques avec un seul instrument

Dans cette section, quatre partitions polyphoniques pour piano sont étudiées :
- la « Mazurka en fa♯ mineur, Op. 6, N°1 » de Chopin :
  - avec un chiffrage 3/4 et 112 mesures, passages répétés inclus ;
- le « Prélude N°1 en do majeur, BWV 846 » de J. S. Bach :
  - avec un chiffrage 4/4 et 35 mesures ;
- la « Sonate en ré majeur, Andante cantabile, L. 12 (K. 478) » de Scarlatti :
  - avec un chiffrage 3/4 et 230 mesures, passages répétés inclus ;
- le 3e mouvement, « Minuetto e Trio », de la « Sonate pour piano N°1 en fa mineur, Op. 2, N°1 » de Beethoven :
  - avec un chiffrage 3/4 et 186 mesures, passages répétés inclus.

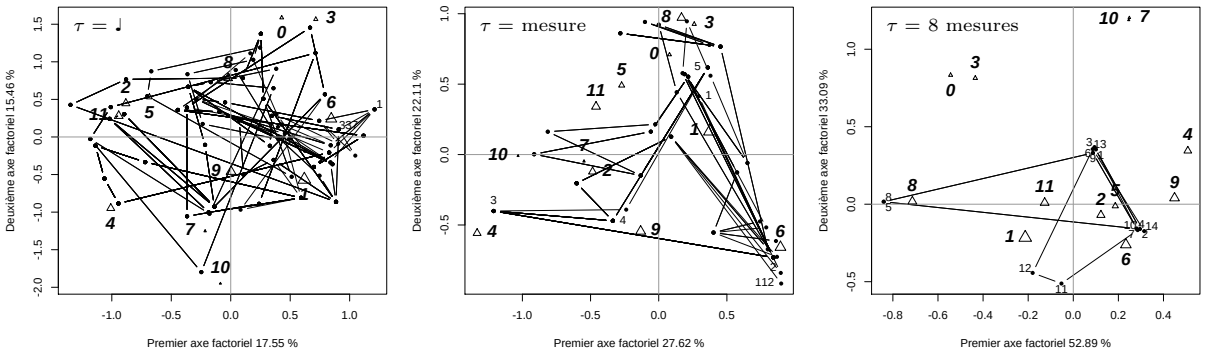


FIGURE 8.7 – AFC sur la « Mazurka en fa♯ mineur, Op. 6, N°1 » de Chopin. Biplots avec τ égal à une noire (gauche), à une mesure (centre) et à huit mesures (droite).

La figure 8.7 présente les résultats de l'AFC appliquée sur la Mazurka de Chopin, avec trois valeurs différentes de τ. La structure de la partition de musique apparaît plus clairement pour des valeurs de τ élevées. En particulier, le graphique de droite, lorsque τ est égal à huit mesures, révèle les passages similaires (1, 3, 6, 9 et 13 d'une part ; 2, 4, 7, 10 et 14 d'autre part ; ainsi que 5 et 8) et les passages différents (12 par rapport à 13 par exemple).

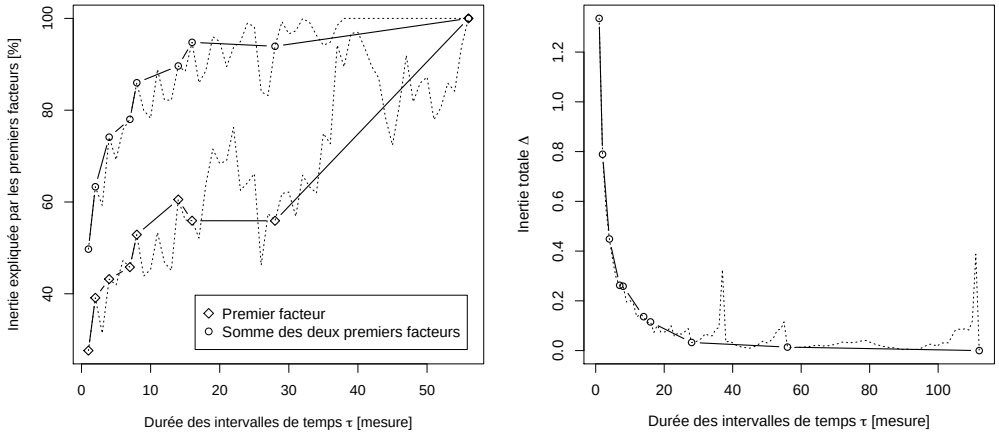


FIGURE 8.8 – AFC sur la « Mazurka en fa♯ mineur, Op. 6, N°1 » de Chopin. Proportion d'inertie expliquée par les premiers facteurs (gauche) et inertie totale (droite) en fonction de la valeur de τ.

Comme pour la partition musicale de « Frère Jacques », augmenter la valeur de τ implique une augmentation de l'inertie expliquée par les premiers facteurs et une diminution de l'inertie totale Δ (figure 8.8). En particulier, pour l'inertie expliquée par le premier facteur, on constate qu'elle est plus élevée lorsque τ est égal à 14 mesures, puis qu'elle ne varie que très peu entre τ égal à 16 mesures et τ égal à 28 mesures, l'explication de ce phénomène restant à établir. Concernant le graphique τ - Δ, il possède, à nouveau la même structure que celui pour « Frère Jacques » (graphique de droite de la figure 8.5). Ceci se produisant pour toutes les partitions, ce graphique sera donc omis dans les prochains exemples.

Le choix consistant à sélectionner τ égal à huit mesures dans le graphique de droite de la figure 8.7, résulte, d'une part, de l'étude de la partition, et d'autre part, des résultats obtenus pour l'indice d'autocorrélation (figure 8.9). En effet, on observe que des pics significatifs se produisent toutes les 24 noires (graphique de gauche) ou toutes les 8 mesures (graphique de droite), ce qui est équivalent. Les deux graphiques apportent donc une information semblable, si ce n'est que dans le premier cas, les résultats sont plus détaillés. Ainsi, pour l'étude des trois autres partitions de piano, on choisira systématiquement τ égal à une mesure pour les indices

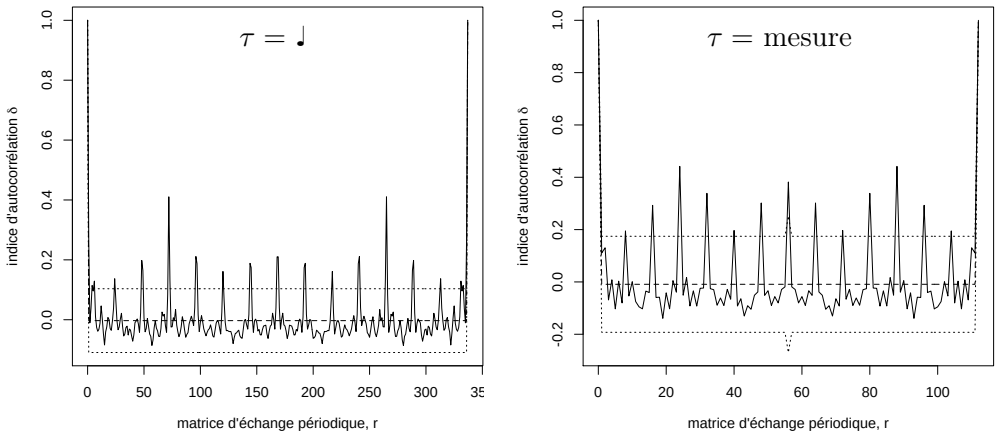


FIGURE 8.9 – Indice d'autocorrélation en fonction du décalage r variant entre 0 et n pour la « Mazurka en fa♯ mineur, Op. 6, N°1 » de Chopin, avec τ égal à une noire (gauche) et à une mesure (droite).

d'autocorrélation, car il semble être plus adapté à la mise en évidence de la structure globale de ces partitions polyphoniques. On remarque aussi un pic plus élevé lorsque $r = 72$ avec τ égal à une noire (respectivement $r = 24$ avec τ égal à une mesure), ce qui s'explique certainement par le fait qu'un passage composé de 24 mesures se répète, donc que les proportions de notes des intervalles de temps $t = 33, \ldots, 56$ sont identiques aux proportions de notes des intervalles $t = 57, \ldots, 80$.

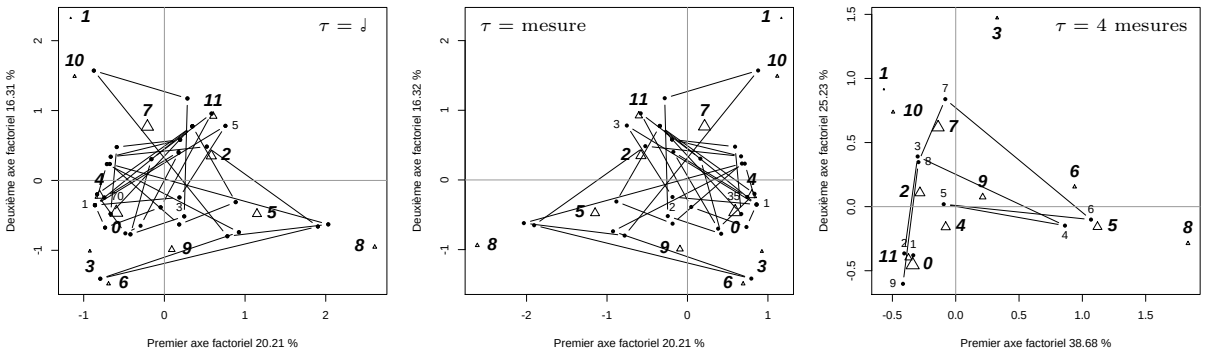


FIGURE 8.10 – AFC sur le « Prélude N°1 en do majeur, BWV 846 » de J. S. Bach. Biplots avec τ égal à une blanche (gauche), à une mesure (centre) et à quatre mesures (droite).

Concernant le prélude de Bach (figure 8.10), aucune structure claire n'apparaît dans les biplots obtenus par l'AFC, excepté lorsque la durée τ est égale à quatre mesures. En observant l'indice d'autocorrélation pour cette même partition (graphique de gauche de la figure 8.13), on remarque un pic lorsque $r = 4$ mesures qui, bien qu'il ne soit pas significatif, semble donc constituer une division intéressante de la partition. Il faut préciser que pour l'AFC, le morceau comportant 35 mesures, le temps $t = 9$ n'est composé que des 3 dernières mesures. Aussi, en observant les graphiques de gauche et du centre de la figure 8.10, on constate que les graphiques pour τ égal à une blanche et τ égal à une mesure sont quasiment identiques (au signe du premier facteur près). Cela s'explique par le fait que dans tout ce morceau de musique, à l'exception des mesures 33 et 34, les deux derniers temps d'une mesure sont identiques aux deux premiers. De plus, dans les mesures 33 et 34, il existe des différences entre les deux premiers et les deux derniers temps, mais les hauteurs de note rapportées à l'octave sont identiques, bien que de durées différentes. Évidemment, le même phénomène aurait pu être observé sur la partition de « Frère Jacques » : exactement le même résultat (aux signes des facteurs près) aurait été obtenu

avec τ égal à deux mesures que celui qui est obtenu avec τ égal à une mesure (figure 8.4).

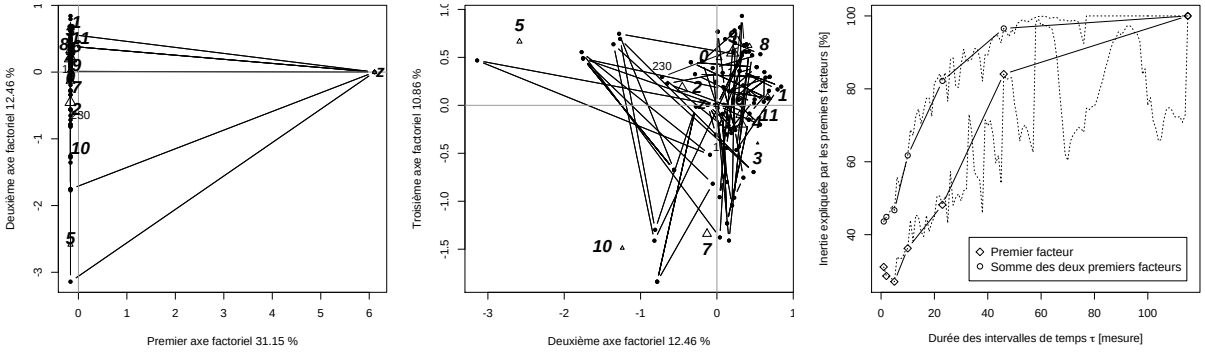


FIGURE 8.11 – AFC sur la « Sonate en ré majeur, Andante cantabile, L. 12 (K. 478) » de Scarlatti. Biplots avec τ égal à une mesure, 1re et 2e dimensions (gauche) et 2e et 3e dimensions (centre). Proportion d'inertie expliquée par les premiers facteurs (droite).

Les résultats obtenus pour l'AFC sur la sonate de Scarlatti, avec τ égal à une mesure (graphique de gauche et du centre de la figure 8.11) sont considérablement différents de ceux obtenus pour les autres partitions de musique, en raison de la présence de vrais silences z. En fait, le profil de z est opposé au profil des autres hauteurs de note et cette opposition est capturée par le premier facteur. Par construction, le même phénomène se produit lorsque τ est plus petit ou égal à une mesure. Ainsi, pour ce morceau de musique, on remarque que l'inertie expliquée par le premier facteur (graphique de droite de la figure 8.11) n'augmente pas systématiquement avec τ, mais qu'elle diminue lorsque τ est plus petit ou égal à cinq mesures.

Aussi, il n'a pas été trouvé de valeur de τ permettant de mettre clairement en évidence la structure de la partition. En observant l'indice d'autocorrélation (graphique du centre de la figure 8.13), deux pics significatifs apparaissent ($\delta = 0.251$ et $\delta = 0.208$) lorsque $r = 54$ et $r = 61$ mesures. Cela s'explique par le fait que les 61 premières mesures sont répétées une fois, puis les 54 mesures suivantes sont aussi répétées une fois et que ces deux parties constituent le morceau entier. Évidemment, il aurait été possible de choisir τ égal à 54 mesures pour obtenir un biplot plus simple à lire avec l'AFC, car dans ce cas, il n'y aurait eu que cinq intervalles de temps qui auraient, comme pour la partition de « Frère Jacques », manifesté un effet de Guttman.

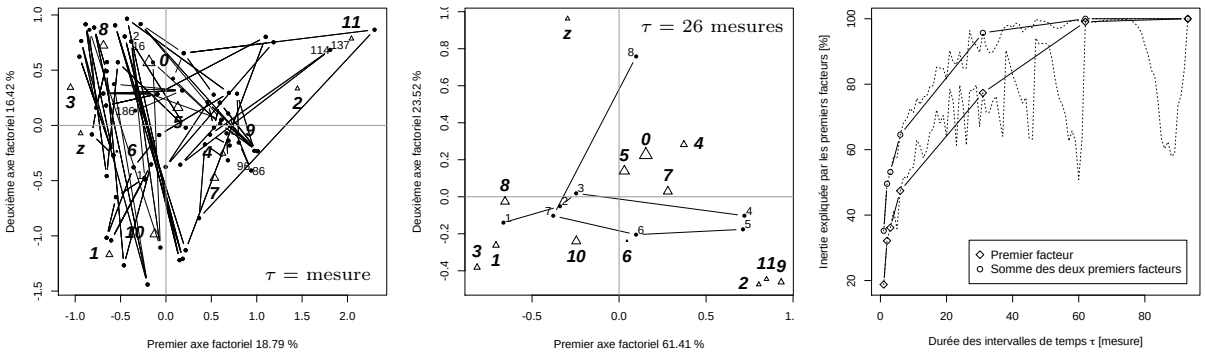


FIGURE 8.12 – AFC sur le 3e mouvement, « Minuetto e Trio », de la « Sonate pour piano N°1 en fa mineur, Op. 2, N°1 » de Beethoven. Biplots avec τ égal à une mesure (gauche) et à vingt-six mesures (centre). Proportion d'inertie expliquée par les premiers facteurs (droite).

Le biplot obtenu avec l'AFC sur le 3e mouvement de la sonate de Beethoven, avec τ égal à une mesure (graphique de gauche de la figure 8.12) montre que de nombreux intervalles de temps sont superposés (par exemple le 114 et le 137), ce qui est dû aux multiples répétitions présentes dans ce morceau, mais aussi au fait que certaines mesures se composent exactement des mêmes

hauteurs de note rapportées à l'octave (elles sont parfois sur des octaves différentes), avec les mêmes durées. Si l'on avait produit le biplot pour τ égal à une noire, on aurait obtenu, comme pour la partition de Scarlatti avec τ égal à une mesure, un premier axe différenciant les vrais silences des autres notes, car la durée la plus longue d'un vrai silence dans ce morceau vaut une noire.

Pour sélectionner une valeur de τ permettant de visualiser la structure de la partition, l'indice d'autocorrélation, avec τ égal à une mesure, est examiné (graphique de droite de la figure 8.13). Trois pics apparaissent clairement, soit lorsque $r = 26$, $r = 54$ et $r = 80$ mesures. Alors que le premier pic s'explique certainement par le fait que la deuxième partie du morceau, qui est la plus longue, s'étend sur 26 mesures qui sont répétées une fois, le troisième pic s'explique peut-être par la présence des deux autres pics, puisque $26 + 54 = 80$. Cependant, la signification du deuxième pic reste encore à établir. Ainsi, le premier pic, soit le seul à être significatif, est sélectionné comme valeur de τ pour le biplot présenté sur le graphique du centre de la figure 8.12. Bien que peu d'intervalles de temps soient représentés sur ce graphique, il reste tout de même difficile de l'interpréter. À noter que le dernier intervalle de temps $t = 8$ n'est composé que de 4 mesures, car la division de τ_{tot} par 26 mesures ne donnait pas un nombre entier.

Finalement, on observe que bien que l'inertie expliquée par les premiers facteurs augmente avec la durée τ, comme pour les autres partitions de musique, la courbe croît de manière concave, sans paliers, ni pics (graphique de droite de la figure 8.12).

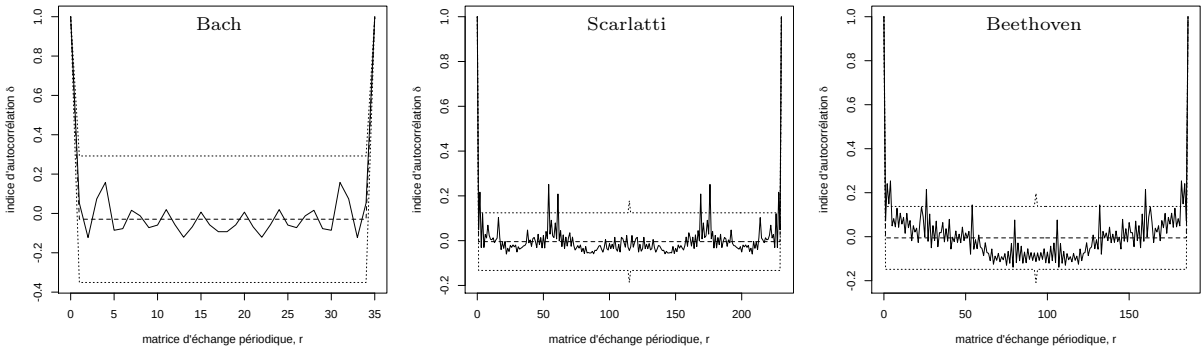


FIGURE 8.13 – Indice d'autocorrélation en fonction du décalage r variant entre 0 et n, avec τ égal à une mesure, pour le « Prélude N°1 en do majeur, BWV 846 » de J. S. Bach (gauche), pour la « Sonate en ré majeur, Andante cantabile, L. 12 (K. 478) » de Scarlatti (centre) et pour le 3e mouvement, « Minuetto e Trio », de la « Sonate pour piano N°1 en fa mineur, Op. 2, N°1 » de Beethoven (droite).

8.2.4 Partition polyphonique avec plusieurs instruments

Pour terminer cette analyse de partitions complètes, un morceau polyphonique composé pour quatre instruments est étudié, à savoir le « Canon en ré majeur » de Pachelbel, qui comporte 57 mesures, avec un chiffrage 4/4.

La figure 8.14 présente les résultats obtenus avec l'AFC. Lorsque τ est égal à une noire (graphique en haut à gauche), une structure du morceau de musique apparaît clairement, bien qu'elle soit difficile à comprendre. En retirant les lignes qui relient les intervalles de temps et en attribuant le même symbole aux intervalles de temps avec un même décalage de $t \bmod 8$ (graphique en haut à droite), on observe que la position d'un intervalle de temps chaque huit noires ne varie que peu sur les deux premiers axes factoriels. En fait, le canon est joué par quatre instruments : trois violons et un clavecin. Alors que le clavecin joue continuellement, le premier violon commence à jouer la mélodie à la 3e mesure, puis le second violon reprend cette mélodie à la 5e mesure et finalement, le troisième violon recommence la même mélodie à partir de la 7e mesure. Ainsi, la structure de base de ce morceau de musique semble se baser sur deux

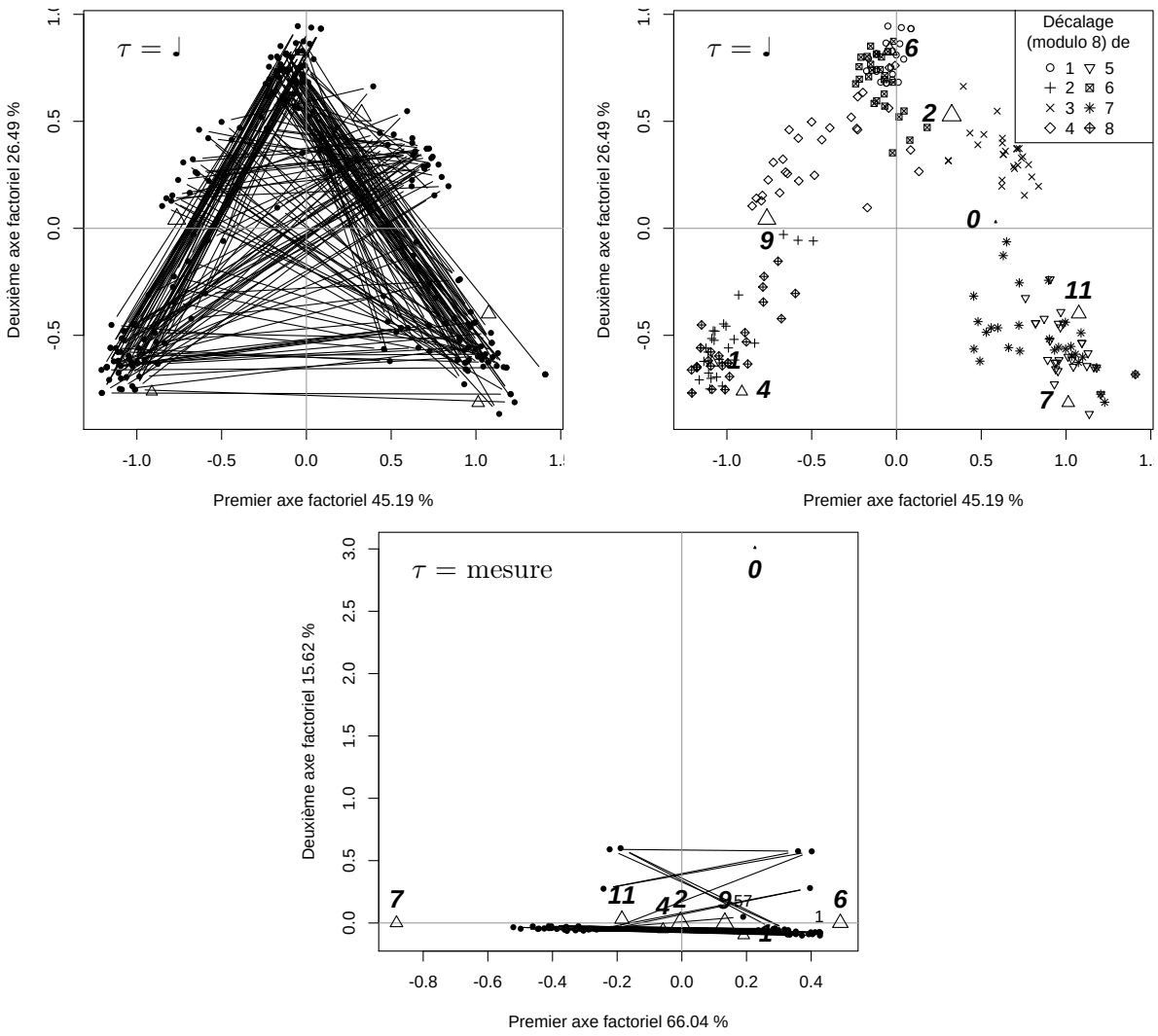


FIGURE 8.14 – AFC sur le « Canon en ré majeur » de Pachelbel. Biplots avec τ égal à une noire (haut) et à une mesure (bas).

mesures, soit huit noires. En particulier, le clavecin, qui constitue la basse du morceau, joue plus de notes simultanément et influence donc fortement le résultat obtenu.

Concernant le biplot obtenu lorsque τ est égal à une mesure (graphique du bas), il est plus difficile de visualiser la structure du morceau de musique, car le second axe ne différencie que le do naturel (0) des autres notes. Cela s'explique par le fait que ce do naturel n'apparaît que dans quelques mesures.

Contrairement aux partitions de musique polyphoniques pour un seul instrument étudiées dans la section 8.2.3, l'indice d'autocorrélation a été calculé avec τ égal à une noire (graphique de gauche de la figure 8.15), car le résultat obtenu apporte des informations supplémentaires à celles que l'on peut observer lorsque τ est égal à une mesure (graphique de droite de la figure 8.15). En effet, lorsque τ vaut une noire, δ exhibe de nombreuses fluctuations régulières. En particulier, des pics significativement positifs et plus élevés apparaissent toutes les huit noires, ce qui semble cohérent avec l'AFC produite pour τ égal à une noire. Aussi, certaines valeurs de δ, toujours à intervalles réguliers, sont significativement négatives, ce qui n'a jamais été observé pour les autres partitions de musique étudiées. De plus, bien qu'ils soient moins élevés que les premiers, d'autres pics significativement positifs apparaissent pour $r = 8c + 2$ et pour $r = 8c + 6$ noires, où $c \in \mathbb{N}$. Ces derniers correspondent probablement, en se basant sur une structure de huit noires, aux distances entre les intervalles de temps similaires, soit le cinq

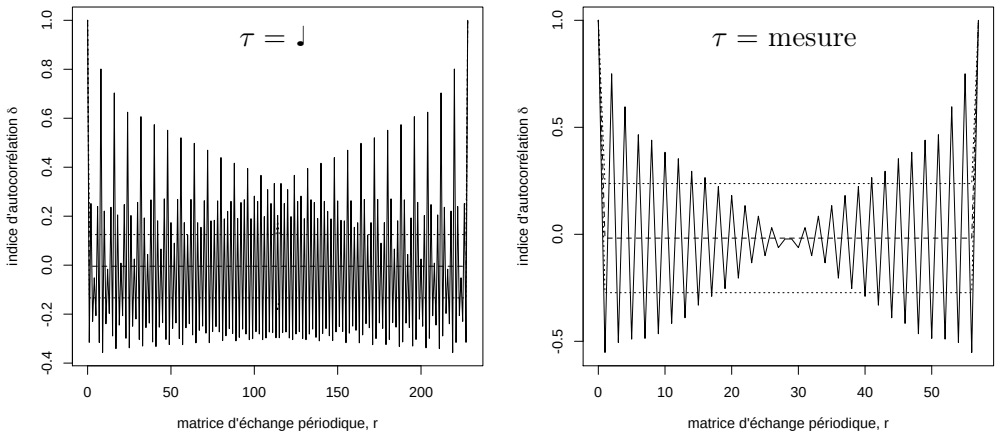


FIGURE 8.15 – Indice d'autocorrélation en fonction du décalage r variant entre 0 et n pour le « Canon en ré majeur » de Pachelbel, avec τ égal à une noire (gauche) et à une mesure (droite).

et le sept, pour le premier, et le deux et le huit, pour le second, selon le graphique en haut à droite de la figure 8.14. En se basant sur ce même graphique, on pourrait s'attendre à trouver des valeurs positives de δ aussi lorsque $r = 8c + 5$, en raison de la similarité des intervalles de temps un et six. Cependant, ces valeurs sont négatives, car plusieurs positions s'opposent selon le premier axe factoriel, dont l'inertie expliquée est élevée, comme par exemple les intervalles deux et sept.

Lorsque τ est égal à une mesure, on constate que δ forme une courbe en dents de scie, oscillant entre des valeurs positives et négatives. Ainsi, lorsque r est paire, δ est positif, et inversement, ce qui, à nouveau, semble cohérent avec une structure de la partition de musique basée sur deux mesures.

8.3 Analyses inter-voix

8.3.1 Traitements

Soit Ξ^v, la table de contingence normalisée pour une des *voix* $v = 1, \ldots, V$ d'une partition de musique. Alors, la table de contingence *complète* pour une partition s'obtient comme la matrice concaténée $\Xi^{\text{COMP}} = (\Xi^1|\Xi^2|...|\Xi^V)$. Une AFC est appliquée sur cette table de contingence, de manière identique à celle expliquée dans la section 8.2.1. Alors qu'une analyse des correspondances multiples (AMC) se pratique sur une table disjonctive (voir par exemple Lebart *et al.*, 1995, section 1.4 ; Saporta, 2006, chapitre 10 ; Le Roux et Rouanet, 2010), la procédure est appliquée ici sur des lignes qui, en raison de la normalisation (8.1), contiennent les *proportions* des hauteurs de note de chaque voix pour un t donné, ce qui constitue une variante « floue » de l'AMC.

D'autre part, afin d'étendre l'indice d'autocorrélation à deux voix (α et β), l'indice d'autocorrélation croisée, mesurant la similarité entre la distribution de la hauteur des notes de la voix α et la distribution de la hauteur des notes de la voix β dans un voisinage fixé, est utilisé (cf. section 3.3). Pour ce faire, les coordonnées de haute dimensionalité des lignes, ${}^*\xi^v_{tj}$, sont obtenues par (1.10), puis ces dernières permettent de calculer $\delta({}^*\Xi^\alpha, {}^*\Xi^\beta)$ (3.7). Comme pour l'indice d'autocorrélation, on utilise la matrice d'échange périodique (3.2).

Pour rappel, plusieurs conditions sont nécessaires à l'application de l'indice d'autocorrélation croisée, à savoir que les deux tables comparées comportent *1)* le même nombre de positions, ici les intervalles de temps t, *2)* le même nombre de caractéristiques, ici les hauteurs de notes j et que *3)* les poids des lignes des deux tables, f_t, soient identiques. La condition *1)* est systématiquement remplie, car les partitions sont de même longueur pour toutes les voix ; et

la condition *3)*, car la table Ξ^v est normalisée. Quant à la condition *2)*, elle n'est pas toujours remplie, car une hauteur de note peut être présente dans une voix et non dans une autre. Le cas échéant, la note absente dans une des voix est ajoutée avec une faible valeur (10^{-30}) pour chaque t (cf. section 8.3.2).

8.3.2 Un canon

Pour commencer, les méthodes décrites ci-dessus sont appliqués au Canon de Pachelbel déjà traité lors de l'étude des partitions complètes (cf. section 8.2.4). En premier lieu, il faut préciser que le fichier Humdrum comportait cinq voix, dont deux pour le clavecin. Néanmoins, pour ce travail, on a choisi de ne considérer que quatre voix, soit une pour chaque instrument.

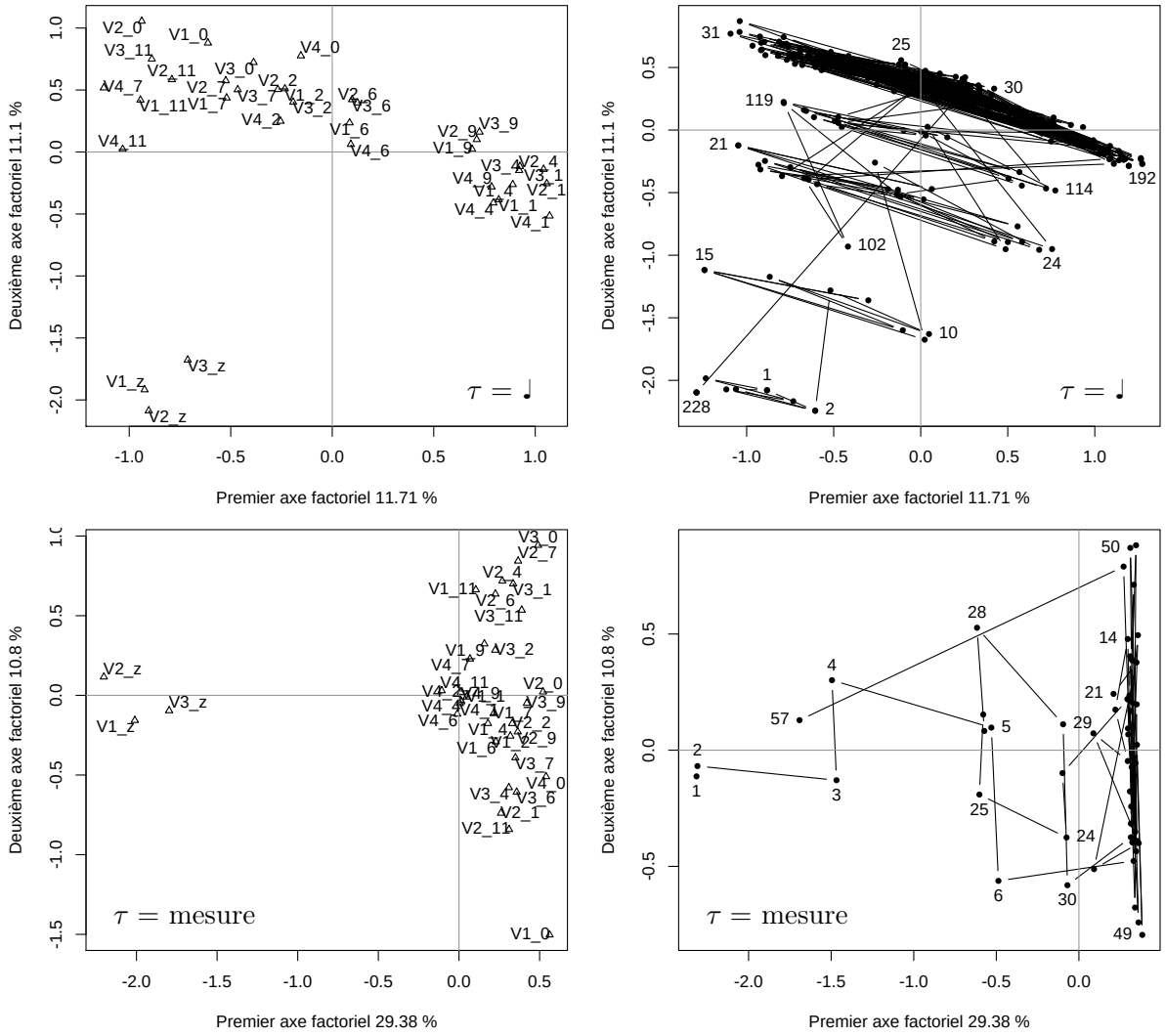


FIGURE 8.16 – ACM floue sur le « Canon en ré majeur » de Pachelbel, avec τ égal à une noire (haut) et à une mesure (bas). Gauche : coordonnées factorielles des hauteurs de note, dont les noms sont précédés par V1 pour le violon I, V2 pour le violon II, V3 pour le violon III et V4 pour le clavecin. Droite : coordonnées factorielles des intervalles de temps.

En appliquant l'ACM sur le canon de Pachelbel (figure 8.16), on constate que lorsque τ est égal à une noire (graphiques du haut), le premier et le second axe factoriel différencient les silences des trois violons de toutes les autres notes, pour la même raison évoquée lorsque l'AFC a été appliquée à la sonate de Scarlatti (cf. figure 8.11). Cela permet aussi de remarquer que des

vrais silences existent pour les trois violons, mais non pour le clavecin [3]. Cette opposition entre les vrais silences et les autres notes met en évidence la structure de la partition de musique par l'intermédiaire de la représentation des intervalles de temps (graphique en haut à droite). En effet, les huit premiers intervalles de temps sont regroupés dans l'extrémité sud-ouest du quadrant sud-ouest, correspondant au début du morceau de musique, lorsque seul le clavecin joue. Ensuite, on observe un regroupement des intervalles de temps neuf à seize, soit la durée pendant laquelle le violon I a rejoint le clavecin. Puis, durant les intervalles de temps dix-sept à vingt-quatre, les violons I et II jouent avec le clavecin. Et finalement, le plus grand groupe au nord est constitué de la majorité des intervalles de temps pendant lesquels tous les instruments jouent. On constate aussi qu'il existe un autre groupe, contenant, par exemple, les intervalles de temps 114 ou 119, et qui correspond à des moments durant lesquels des silences, qui durent une croche, se produisent pour l'un des violons.

Les graphiques du bas de la figure 8.16, obtenus avec τ égal à une mesure, ont une interprétation similaire. En effet, dans ce cas, le premier axe factoriel (graphique de gauche) oppose les silences aux autres notes et on retrouve (graphique de droite) les deux premières mesures dans cette zone, puis les mesures trois et quatre plus proche du centre, etc. Ces graphiques comportant moins de points que les précédents (graphiques du haut de la figure 8.16), il est aussi possible de mieux observer les mesures contenant des silences lorsque tous les instruments jouent, comme par exemple, les mesures vingt-quatre ou vingt-cinq. La principale différence entre les résultats obtenus avec τ égal à une noire ou égal à une mesure réside dans le fait que les mêmes notes jouées par des instruments différents sont regroupées dans le premier cas et non dans le second (graphiques de gauche).

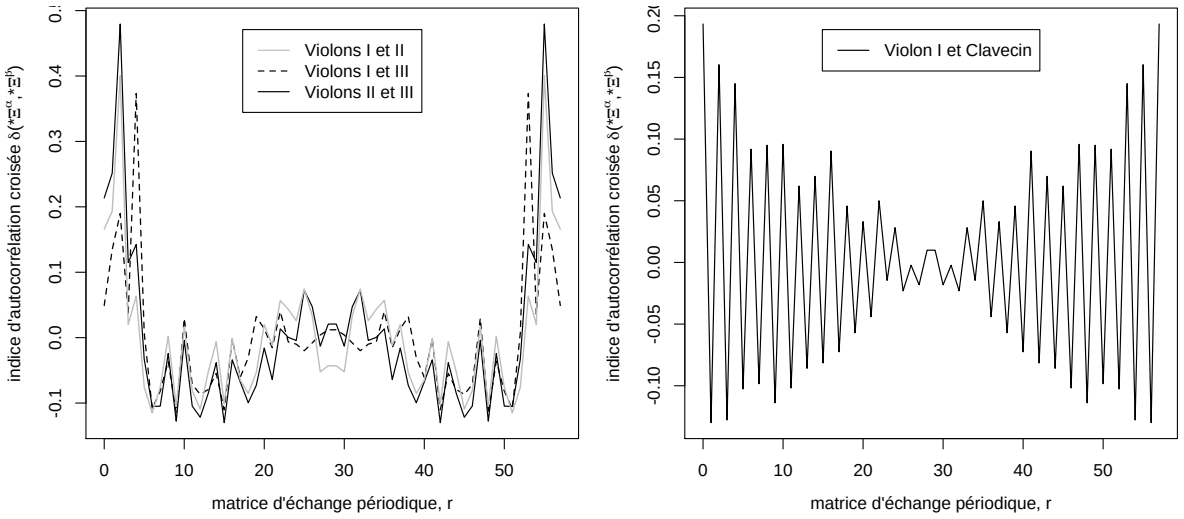


FIGURE 8.17 – Indice d'autocorrélation croisée en fonction de la distance r variant entre 0 et n pour le « Canon en ré majeur » de Pachelbel, avec τ égal à une mesure.

Sur le graphique de gauche de la figure 8.17, représentant l'indice d'autocorrélation entre les trois violons avec τ égal à une mesure, on observe trois pics plus importants : le premier, entre les violons I et II lorsque $r = 2$, le second entre les violons II et III aussi lorsque $r = 2$ et le troisième entre les violons I et III lorsque $r = 4$; ce qui correspond bien aux décalages de deux ou quatre mesures entre les départs de chaque violon.

Concernant l'autocorrélation croisée entre le violon I et le clavecin avec τ égal à une mesure, on observe un comportement très similaire à celui de l'autocorrélation pour l'ensemble des instruments avec la même durée τ (cf. figure 8.15), soit des valeurs positives lorsque r est paire et

3. Ainsi, pour calculer l'indice d'autocorrélation croisée entre le clavecin et un autre instrument, il faudra ajouter le silence au premier avec de faibles valeurs, comme il est expliqué dans la section 8.3.1.

inversement. En fait, l'indice d'autocorrélation entre le clavecin et n'importe quel autre violon suit toujours cette même alternance. Aussi, en prenant τ égal à une noire, l'autocorrélation croisée entre l'un des violons et le clavecin est très similaire à l'indice d'autocorrélation obtenu avec la même durée τ. Il semble donc que le clavecin comportant plus de notes influence totalement l'indice d'autocorrélation croisée, à l'inverse de chacun des violons.

8.3.3 Un quatuor à cordes

Le second et dernier exemple étudié pour l'analyse inter-voix d'une partition est le 1er mouvement « Allegro con brio » du « Quatuor à cordes N°1 en fa majeur, Op. 18 N°1 » de Beethoven, avec un chiffrage 3/4 et 427 mesures, répétitions incluses.

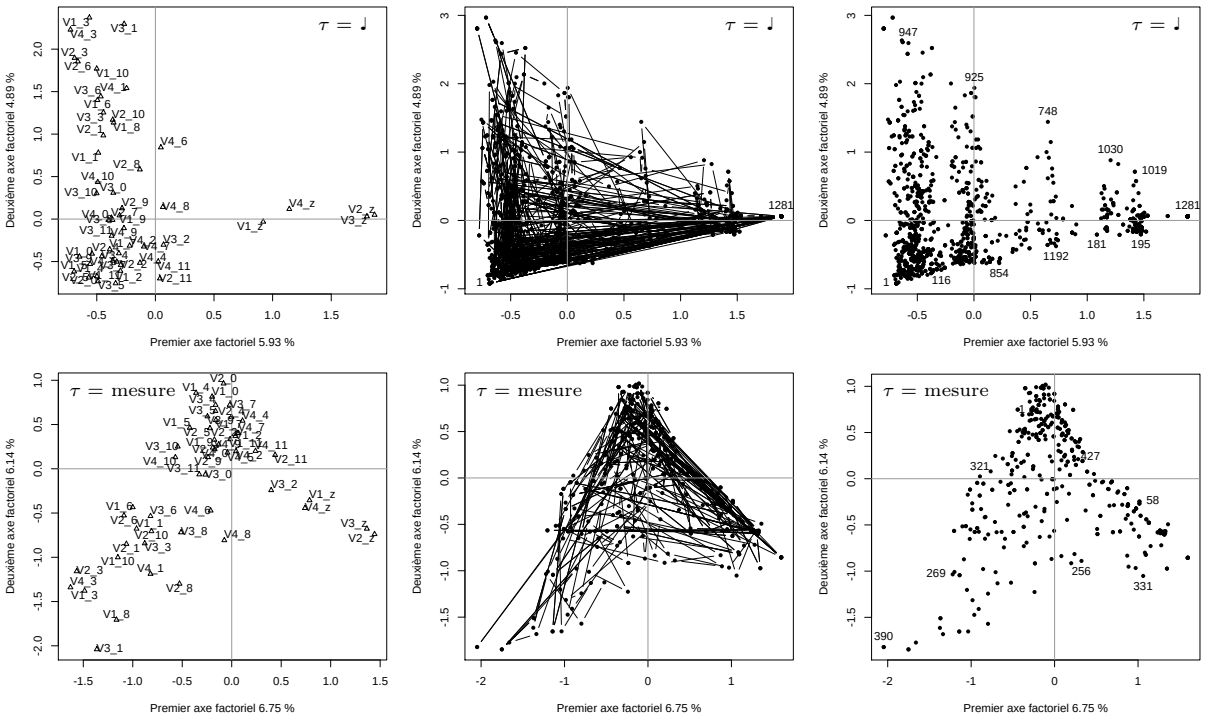


FIGURE 8.18 – ACM floue sur le 1er mouvement du « Quatuor à cordes N°1 en fa majeur, Op. 18 N°1 » de Beethoven, avec τ égal à une noire (haut) et à une mesure (bas). Gauche : coordonnées factorielles des hauteurs de note, dont les noms sont précédés par V1 pour le violoncelle, V2 pour l'alto, V3 pour le violon II et V4 pour le violon I. Centre : coordonnées factorielles des intervalles de temps reliées dans l'ordre chronologique. Droite : coordonnées factorielles des intervalles de temps non reliées.

Les résultats obtenus avec l'ACM « floue » sont présentés dans la figure 8.18. Lorsque τ est égal à une noire (graphiques du haut), le premier axe oppose, comme pour le Canon de Pachelbel, les silences aux autres hauteurs de note. Ne s'agissant pas d'un canon, il semble difficile de déterminer des zones pour les intervalles de temps du graphique du centre. Cependant, en supprimant les lignes qui relient les intervalles de temps (graphique de droite), plusieurs zones distinctes apparaissent. À l'extrême est se trouvent les intervalles de temps durant lesquels aucun instrument ne joue et à l'extrême ouest, ceux durant lesquels tous les instruments jouent.

Concernant les résultats obtenus en posant que τ est égal à une mesure (graphiques du bas), les vrais silences des quatre instruments à cordes sont regroupés dans le quadrant sud-est, mais il est difficile de visualiser la structure du morceau de musique.

La figure 8.19 présente l'indice d'autocorrélation croisée entre les différentes paires d'instruments. On remarque diverses oscillations pour toutes les courbes, difficiles à interpréter. Cependant, on retrouve un pic plus important, pour plusieurs des courbes, lorsque $r = 114$

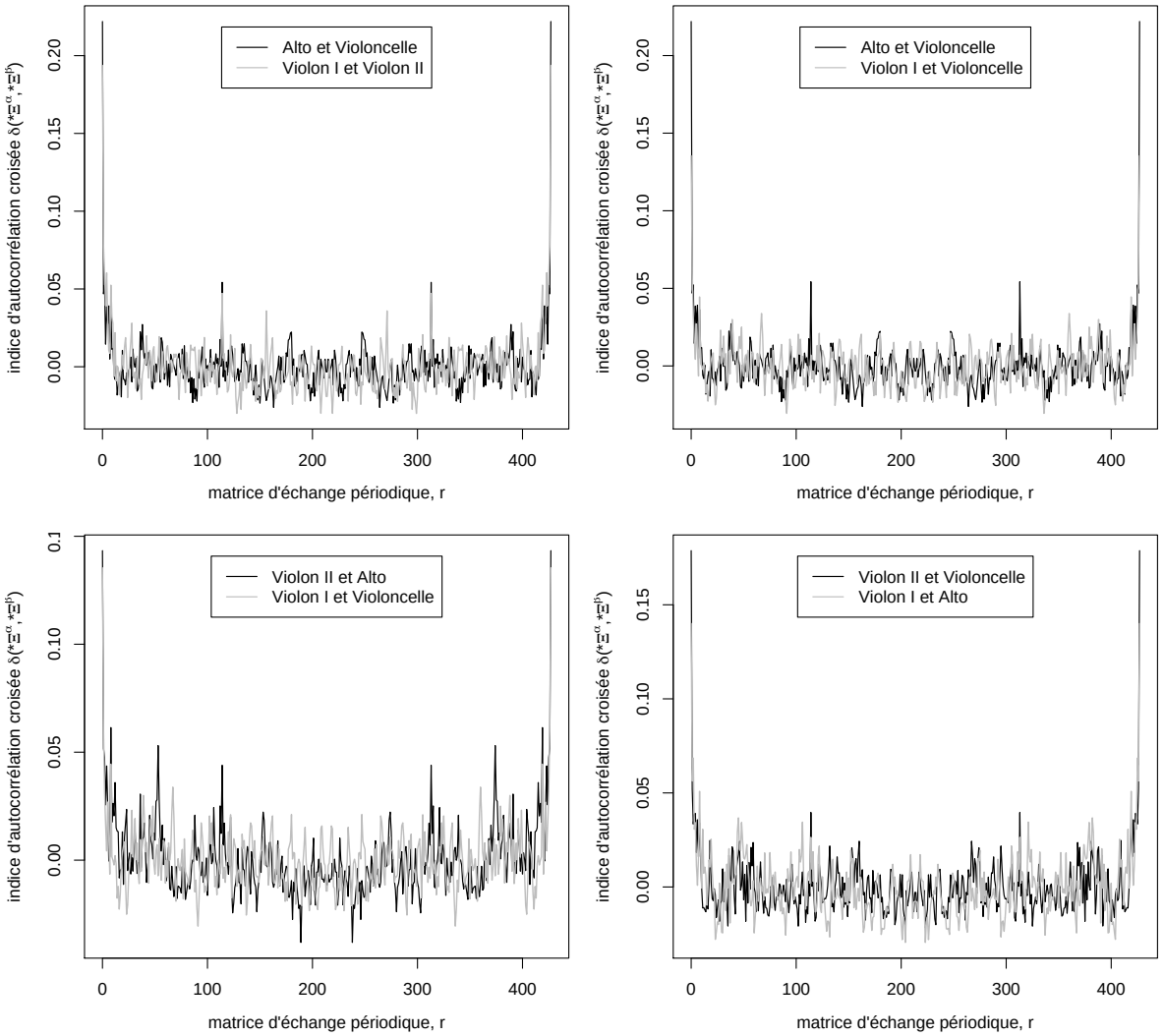


FIGURE 8.19 – Indice d'autocorrélation croisée en fonction de la distance r variant entre 0 et n pour le 1er mouvement du « Quatuor à cordes N°1 en fa majeur, Op. 18 N°1 » de Beethoven, avec τ égal à une mesure.

mesures, correspondant à la répétition de la première partie de la partition de musique. Deux autres pics ($r = 8$ et $r = 53$ mesures) apparaissent pour l'autocorrélation croisée entre le violon II et l'alto, probablement dus à des passages joués une première fois par l'un des instruments et repris par l'autre, ou simplement des hauteurs de notes similaires ; le recours à l'interprétation d'un expert serait ici nécessaire.

Aussi, lorsque $r = 0$, il n'existe pas de décalage entre les deux voix α et β, et l'indice d'autocorrélation croisée $\delta({}^*\Xi^\alpha, {}^*\Xi^\beta)^{(r=0)}$ s'interprète alors comme une mesure de similarité entre ces deux voix. Sur la figure 8.19, on constate que certaines paires de voix sont plus similaires que d'autres.

La dissimilarité entre les deux voix s'obtient comme $D_{\alpha\beta} = 1 - \delta({}^*\Xi^\alpha, {}^*\Xi^\beta)^{(r=0)}$ qui se trouve être une dissimilarité euclidienne carrée. Ainsi, il est possible d'utiliser la classification ascendante hiérarchique, avec le critère de Ward, pour classifier les différents instruments (cf. section 2.1.1). Le résultat obtenu, avec la fonction « hclust » de R, est présenté dans la figure 8.20. Il en ressort que l'alto et le violoncelle, d'une part, et que le violon I et le violon II, d'autre part, partagent plus de similarités mélodiques que les autres paires d'instruments.

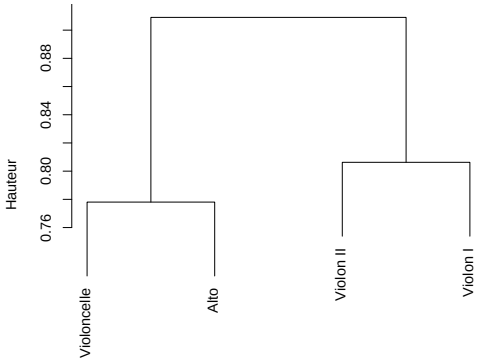


FIGURE 8.20 – Classification ascendante hiérarchique des quatre voix du 1er mouvement du « Quatuor à cordes N°1 en fa majeur, Op. 18 N°1 » de Beethoven selon $\delta({}^*\Xi^{\alpha}, {}^*\Xi^{\beta})^{(r=0)}$, avec τ égal à une mesure.

8.4 Analyses inter-partitions

Dans cette dernière section, on cherche à déterminer les similarités qui existent entre les partitions de musique, de manière à pouvoir les classer selon leur compositeur. Pour ce faire, un échantillon de vingt partitions de musique est constitué (section 8.4.1), puis, en se basant sur une mesure de similarité, une classification ascendante hiérarchique de ces morceaux de musique est effectuée (section 8.4.2).

8.4.1 Données

Le jeu de données comprend 20 morceaux de musique écrits par 4 compositeurs, à savoir :
- cinq sonates de Domenico Scarlatti (1685 - 1757) ayant toutes un chiffrage 2/2, soit :
  - la « Sonate en do majeur, Allegro, L. 1 (K. 514) »,
  - la « Sonate en mi bémol majeur, Allegro, L. 16 (K. 306) »,
  - la « Sonate en sol mineur, Allegro, L. 336 (K. 93) »,
  - la « Sonate en la majeur, Allegrissimo, L. 345 (K. 113) », et
  - la « Sonate en si mineur, Allegro, L. 346 (K. 408) » ;
- le premier mouvement de cinq sonates pour piano de Wolfgang Amadeus Mozart (1756 - 1791), soit :
  - la « Sonate pour piano N°1 en do majeur, K^1 279 / K^6 189d, 1. Allegro »,
  - la « Sonate pour piano N°2 en fa majeur, K^1 280 / K^6 189e, 1. Allegro assai »,
  - la « Sonate pour piano N°3 en si bémol majeur, K^1 281 / K^6 189f, 1. Allegro »,
  - la « Sonate pour piano N°4 en mi bémol majeur, K^1 282 / K^6 189g, 1. Adagio », et
  - la « Sonate pour piano N°5 en sol majeur, K^1 283 / K^6 189h, 1. Allegro » ;
- le premier mouvement de cinq sonates pour piano de Ludwig van Beethoven (1770 - 1827), soit :
  - la « Sonate pour piano N°1 en fa mineur, Op. 2, N°1, 1. Allegro »,
  - la « Sonate pour piano N°2 en la majeur, Op. 2, N°2, 1. Allegro vivace »,
  - la « Sonate pour piano N°3 en do majeur, Op. 2, N°3, 1. Allegro con brio »,
  - la « Sonate pour piano N°4 en mi bémol majeur, Op. 7, 1. Allegro molto con brio », et
  - la « Sonate pour piano N°5 en do mineur, Op. 10, N°1, 1. Allegro molto e con brio » ; et
- cinq mazurkas de Frédéric François Chopin (1810 - 1849), soit :
  - la « Mazurka en fa dièse mineur, Op. 6, N°1 »,
  - la « Mazurka en si bémol majeur, Op. 7, N°1 »,
  - la « Mazurka en si bémol majeur, Op. 17, N°1 »,

— la « Mazurka en sol mineur, Op. 24, N°1 »,
— la « Mazurka en do mineur, Op. 30, N°1 ».

8.4.2 Traitement et résultat

Pour mesurer la similarité de la configuration (*configuration similarity*) entre deux partitions a et b, on utilise une version duale pondérée du coefficient RV proposé par Robert et Escoufier (1976), à savoir :

$$\mathrm{CS}_{ab} = \frac{\mathrm{Tr}(\check{K}^a \check{K}^b)}{\sqrt{\mathrm{Tr}((\check{K}^a)^2)\mathrm{Tr}((\check{K}^b)^2)}}$$

où $\check{K}^a$ (respectivement $\check{K}^b$) sont les produits scalaires pondérés entre les hauteurs de notes de la partition de musique a (resp. b), identiques à ceux calculés dans la section 8.2.1 par (1.24). Cela implique que les deux partitions possèdent les mêmes hauteurs de note. Cependant, si une note est présente dans une des partitions de musique et non dans l'autre, les composantes $\check{K}^a$ (ou $\check{K}^b$) sont nulles par définition. Ainsi, des composantes nulles ont simplement été ajoutées dans les matrices le cas échéant. De plus, pour rendre les partitions comparables, elles ont toutes été transposées en do.

Ensuite, on définit la dissimilarité entre deux partitions comme $D_{ab} = 1 - \mathrm{CS}_{ab}$. Cette dissimilarité, tout comme la dissimilarité entre deux voix $D_{\alpha\beta}$, peut s'interpréter comme une généralisation de la distance du cosinus (voir par exemple Weihs, Ligges, Mörchen et Müllensiefen, 2007) et se trouve être une dissimilarité euclidienne carrée. Ainsi, les méthodes de classification usuelles (cf. chapitre 2) peuvent être utilisées sur les dissimilarités D_{ab}, et on utilise, à nouveau, la classification ascendante hiérarchique avec le critère de Ward, par l'intermédiaire de la fonction « hclust ».

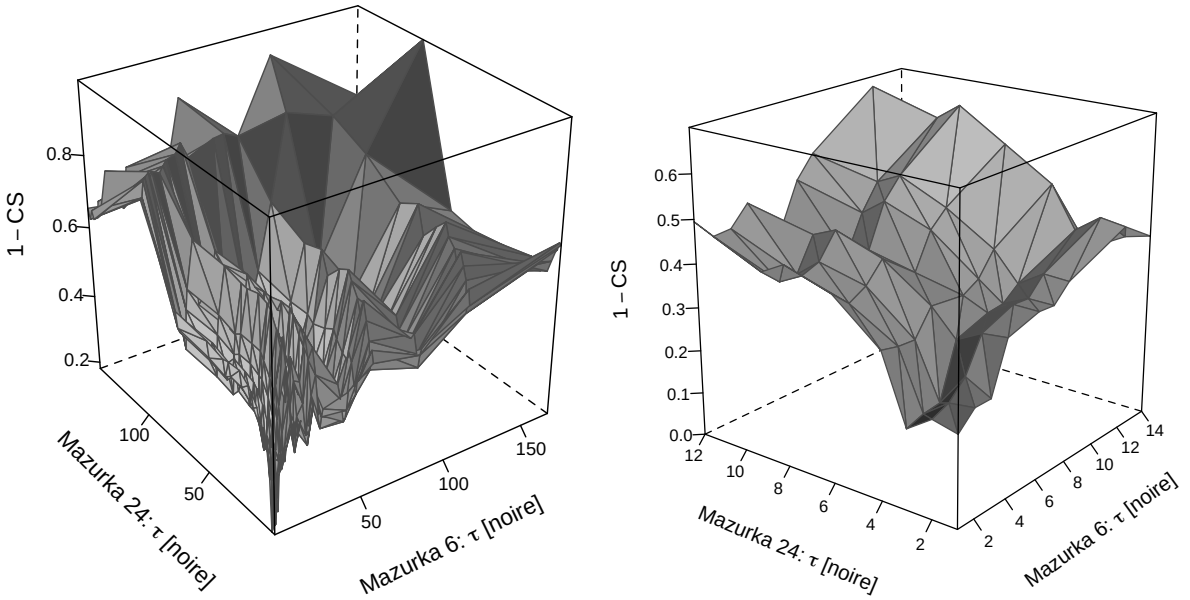


FIGURE 8.21 – Dissimilarité entre la « Mazurka en fa dièse mineur, Op. 6, N°1 » et la « Mazurka en sol mineur, Op. 24, N°1 » de Chopin en fonction de τ.

Avant de procéder à une classification, il faut noter qu'étant donné que $\check{K}^a$ et $\check{K}^b$ dépendent de la durée τ, il en sera de même pour la dissimilarité D_{ab}. On observe, sur les deux exemples présentés dans les figures 8.21 et 8.22, que la dissimilarité entre les deux partitions D_{ab} augmente de façon irrégulière lorsque la durée τ augmente [4].

4. Il faut noter que pour créer ces figures, comme déjà mentionné dans la section 8.1.2, les éventuelles anacrouses ont été supprimées pour pouvoir agréger les intervalles de temps de manière complètement automatique.

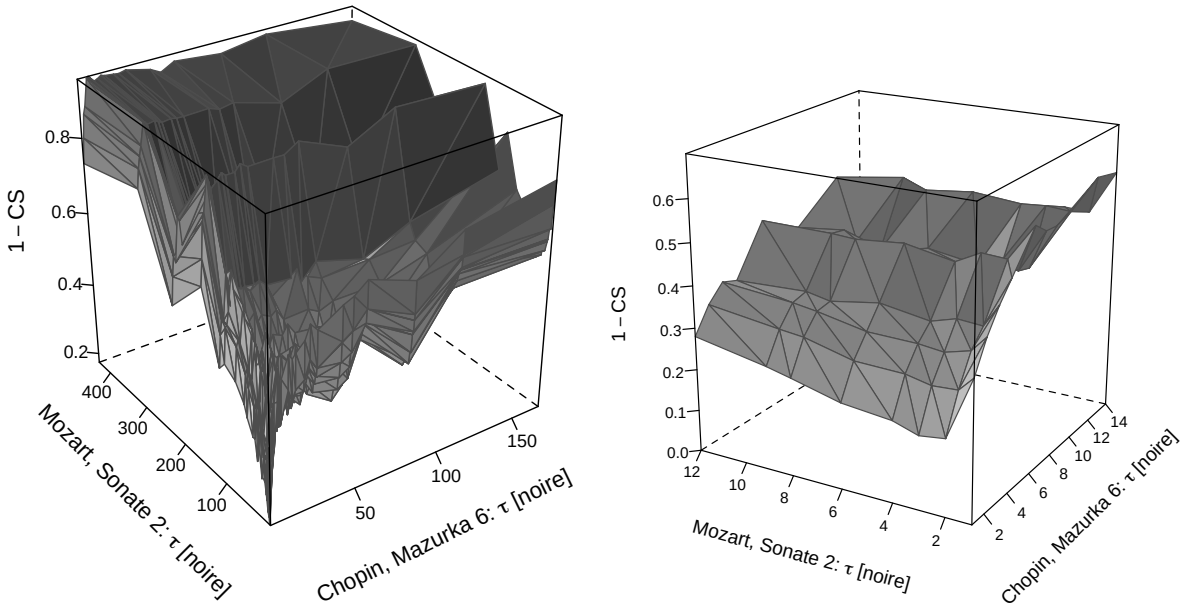


FIGURE 8.22 – Dissimilarité entre la « Mazurka en fa dièse mineur, Op. 6, N°1 » de Chopin et le 1er mouvement de la « Sonate pour piano N°2 en fa majeur, K^1 280 / K^6 189e » de Mozart en fonction de τ.

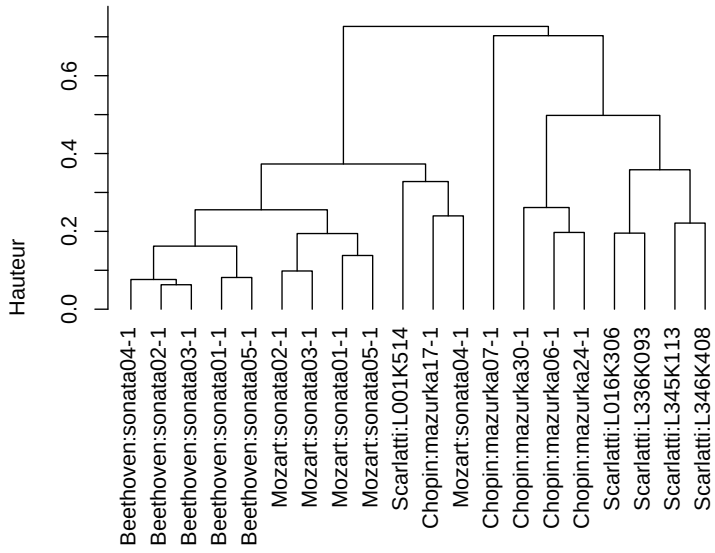


FIGURE 8.23 – Classification ascendante hiérarchique des 20 partitions de musique avec le critère d'agrégation de Ward.

Ainsi, afin d'avoir une unité de durée commune à toutes les partitions de musique lors du calcul des distances D_{ab}, on pose τ égal à une mesure. Le résultat obtenu est présenté dans la figure 8.23. Bien que le jeu de données soit restreint, ce premier résultat est encourageant, car les morceaux de musique sont bien regroupés selon leur compositeur, en particulier en ce qui concerne les partitions de Beethoven.

8.5 Discussion

Pour commencer, il faut se rappeler que seul l'aspect catégoriel des notes a été exploité, et non les valeurs numériques des hauteurs. Ainsi, une transposition de l'ensemble de la partition ne changerait pas les résultats et il en serait de même si deux hauteurs de notes étaient systématiquement échangées. À l'inverse, un partie répétée *mais transposée* aura une représentation différente de l'originale. En d'autres termes, la représentation choisie implique que, à l'intérieur d'un intervalle de temps donné t, les notes forment « un sac de notes ». Néanmoins, l'ordre temporel des notes est pris en compte lorsqu'elles n'apparaissent pas durant le même intervalle de temps.

Concernant les analyses pratiquées sur les partitions complètes (section 8.2), l'AFC et l'autocorrélation ont été utilisées, et ces deux méthodes ont permis de visualiser certains aspects de la structure des partitions. En particulier, l'AFC a mis en évidence la structure du morceau lorsqu'une valeur adéquate de τ était utilisée et que la partition contenait des motifs récurrents. Les résultats sont plus difficiles à interpréter lorsqu'un motif est transposé ou lorsqu'un vrai silence apparaît, car comme on l'a vu (figure 8.11), dans ce second cas, le premier facteur n'exprime que l'opposition entre le silence et le son. Concernant l'indice d'autocorrélation, il permet principalement de détecter les répétitions, qu'elles soient exactes ou approximatives, mais à la condition qu'elles ne soient pas transposées. De plus, il est souvent un bon indicateur des valeurs de τ pouvant donner lieu à des AFC intéressantes.

Au sujet des analyses inter-voix des partitions (section 8.3), l'ACM floue, tout comme l'AFC dans le cas des partitions de musique complètes, a permis de visualiser des éléments structurels des partitions, de manière plus ou moins évidente selon le choix de la valeur de τ. Quant à l'indice d'autocorrélation croisée, il a permis de comparer les différentes voix d'une même partition en mesurant leur similarité selon une distance r. Il est particulièrement adapté pour révéler les passages similaires, mais dans deux voix distinctes. Cet indice pourrait aussi être utilisé pour comparer deux variantes d'un même morceau de musique. Par exemple, Ellis et Poliner (2007) utilisent l'auto-corrélation croisée pour comparer des variantes d'un même morceau dans des fichiers audio. Finalement, l'analyse inter-partitions (section 8.4) a montré des premiers résultats encourageants.

En conclusion, la représentation de la musique polyphonique en tables de contingence a permis de visualiser certaines structures inhérentes des partitions, ainsi que d'obtenir une classification non supervisée avec de bons résultats. Évidemment, de nombreuses pistes restent encore à explorer.

En premier lieu, seul un petit nombre de partitions a été étudié dans l'ensemble de ce chapitre et il serait assurément intéressant d'en analyser un plus grand nombre, afin de déterminer si certains résultats sont systématiques. En particulier, il faudrait découvrir s'il est possible de déterminer la valeur de τ idéale pour les analyses factorielles, et selon quel critère.

Concernant les différents choix opérés lors de la représentation des partitions, d'autres possibilités pourraient être envisagées. Par exemple, les parties répétées et explicitement indiquées comme telles sur la partition pourraient être omises (étape facile à réaliser à partir du format Humdrum). Ainsi, ces parties répétées ne seraient plus détectées, ce qui permettrait peut-être de voir émerger d'autres structures. Aussi, les anacrouses pourraient être retirées, ce qui permettrait d'automatiser davantage la procédure pour la suite des opérations.

Pour terminer, concernant la classification des partitions de musique selon les compositeurs, une prochaine étape pourrait consister à augmenter le jeu de données, puis à utiliser des méthodes de classification supervisée, telles que l'analyse discriminante (cf. section 2.2.1).

Conclusion et discussion

Comme il a été expliqué dans l'introduction, la première visée de ce travail était de pratiquer une analyse exploratoire de données textuelles et musicales au moyen d'un formalisme et de méthodes bien contrôlés et compatibles avec des unités de poids possiblement non-uniformes. En particulier, le formalisme s'appuyait sur trois concepts fondamentaux : (i) une table de contingence, (ii) une matrice de dissimilarités *euclidiennes carrées* et (iii) une matrice d'échange. Grâce à ce formalisme, plusieurs méthodes ont pu être exprimées, à savoir : l'AFC, basée sur (i) ou sur (ii), à condition que ces dernières soient produites sur (i) ; la classification supervisée ou non, parfois combinée aux transformations de Schoenberg, de nouveau basée sur (ii) ; et les indices d'autocorrélation et d'autocorrélation croisée, basés sur (ii) et (iii). Les dissimilarités euclidiennes carrées sont donc au cœur de ces méthodes.

Ainsi la première question qu'on est en droit de se poser est « Quelles structures ont pu être découvertes sur les données textuelles et musicales choisies par l'intermédiaire de ces méthodes ? », ainsi que « Quelles conclusions peut-on en retirer le cas échéant ? ». Comme déjà mentionné, on ne se positionne pas ici comme spécialiste de l'un ou l'autre des domaines spécifiques aux données traitées, mais comme un observateur, ou même un explorateur, espérant que l'une de ses découvertes puisse être utile et offrir un nouveau point de vue à des spécialistes.

Concernant la classification automatique de propositions énoncées en types de discours (chapitre 4), plusieurs conclusions émergent. En premier lieu, il faut se demander si le choix de ne représenter les propositions énoncées que par les uni-, bi- et trigrammes de CMS qu'elles contiennent constituait une bonne approche [1]. En d'autres termes, est-ce que le choix d'utiliser des représentations si simples était suffisant pour un problème si complexe ? La littérature relative à ce type de problèmes ne semblait pas aller à l'encontre de ce choix. Ensuite, une première analyse inférentielle (test du khi2) et descriptive (quotient d'indépendance) sur les liens existant entre les CMS et les types de discours, basés sur l'annotation des propositions énoncées par un expert humain, a confirmé ce choix. De plus, les visualisations obtenues par l'AFC semblaient encourageantes, bien que parfois atténuées par la validation par bootstrap.

Les résultats de la classification non supervisée pour chacun des quatre contes de Maupassant, par l'intermédiaire de la méthode K-means, combinée à la transformation de puissance de Schoenberg, ainsi que par l'algorithme K-means flou, n'ont finalement pas été aussi concluants que l'on aurait pu l'espérer. Pour commencer, comme il avait déjà été remarqué lors de l'analyse descriptive et de l'AFC, des tendances différentes sont obtenues pour les quatre contes. Qui plus est, les deux indices d'accord entre partitions utilisés ont des comportements très différents et il était donc difficile de parvenir à une conclusion générale pour un texte donné. Néanmoins,

1. On ne reviendra pas ici sur la capacité de TreeTagger à associer correctement, ou non, les CMS à chaque mot (voir par exemple Dejean, Fortun, Massot, Pottier, Poulard et Vernier, 2010, pour le français).

sans pouvoir en élucider vraiment la raison, il est clair que cette représentation des textes par les CMS a été plus performante pour retrouver les types de discours du conte « Le Voleur » que pour les autres textes. Aussi, au regard des résultats obtenus avec la méthode K-means flou avec laquelle on a fait varier le nombre de groupes construit par l'algorithme, il semble que parfois la génération de plus de six groupes permette un meilleur accord avec les six types de discours annotés. Sans pouvoir l'assurer, faute d'analyses à cet effet, on peut imaginer qu'un autre niveau de structure entre en jeu, peut-être en lien avec la structure hiérarchique des types de discours.

On pourrait aussi se demander s'il n'aurait pas été plus pertinent de faire usage de classification supervisée, sortant ainsi du cadre strict de l'analyse exploratoire. Cependant, étant donné qu'une classification supervisée nécessite un ensemble d'apprentissage, le nombre de propositions de certains types de discours semblait trop faible lorsque l'on considère chaque texte séparément. Une alternative pourrait consister à modifier la méthodologie de ce chapitre et concaténer les quatre textes. Finalement, comme on aurait peut-être pu s'y attendre, n'utiliser que les CMS fut un choix un peu trop drastique au vu de la complexité de la tâche à effectuer. Comme il est vrai pour chacune des applications, le sujet reste ouvert et les pistes, nombreuses.

Dans le chapitre 5 qui traitait de la classification supervisée multi-étiquette des tours de parole des pages de discussion de *Simple English Wikipedia* en actes de dialogue, la démarche était clairement différente s'agissant d'un problème supervisé. À nouveau, la représentation des données se voulait simple, intelligible et uniquement axée sur des caractéristiques linguistiques. Pour rappel, les tours de parole étaient représentés par trois caractéristiques considérées séparément : les lemmes, les CMS et le sens des verbes selon WordNet. Ces trois caractéristiques, qui semblaient en accord avec la littérature, ont été sélectionnées pour diverses raisons. Concernant les lemmes, ils ont été pris en compte car il s'agit d'une unité souvent considérée dans ce type d'approches, simple et généralement efficace. Quant aux CMS, elles ont été utilisées au vu des liens qui avaient été déterminés entre ces dernières et les types de discours dans le chapitre précédent. En effet, bien qu'un acte de dialogue ne soit pas un type de discours, il nous semblait, en tant que néophyte dans ces domaines, que ces deux concepts devaient avoir des points communs, choix qui s'avéra judicieux au vu des résultats obtenus. Finalement, l'idée d'utiliser les hyperonymes les plus généraux naquit après avoir travaillé sur les dissimilarités sémantiques présentées dans le chapitre 6. Il nous semblait en effet que certains verbes devaient souvent se retrouver dans certains actes de dialogue et qu'il faudrait donc considérer des classes de verbes. Ainsi, plutôt que de constituer manuellement des classes de verbes comme il avait déjà été fait dans la littérature, on a décidé de les constituer automatiquement par l'intermédiaire de WordNet [2].

Concernant la méthode de classification multi-étiquette, il fallait commencer par décider si les données devaient être transformées pour aboutir à une série de classifications binaires pour chaque étiquette ; ou si au contraire il fallait opter pour une transformation de l'algorithme permettant de traiter en une fois toutes les étiquettes pour chaque individu. Ainsi, une première analyse consista à déterminer l'existence, ou non, de liens entre les étiquettes. Ces liens étant faibles, le choix s'est porté sur la série de classifications binaires. Elle a été effectuée grâce à l'analyse discriminante, encore une fois combinée à la transformation de puissance de Schoenberg. Plusieurs conclusions émergent. En premier lieu, le critère du plus proche centroïde est souvent plus efficace que le critère des plus proches voisins sur ces données. Deuxièmement, la transformation de puissance améliore les résultats obtenus avec la classification selon la F-mesure. Troisièmement, le meilleur résultat sur l'ensemble des étiquettes est obtenu en utilisant les CMS comme caractéristiques. Finalement, bien que les CMS permettent une meilleure classification de l'ensemble des étiquettes, certaines étiquettes sont mieux discriminées par une des deux autres caractéristiques. Au final, il nous semble que l'intérêt de chacune des caractéristiques

2. Comme pour TreeTagger, on ne s'étendra pas ici sur les limitations de WordNet et sur le fait de sélectionner systématiquement le premier sens des mots rencontrés (voir par exemple Hawker et Honnibal, 2006).

linguistiques a été mis en évidence.

Le chapitre 6 concernait la dernière application sur des textes, à savoir l'autocorrélation textuelle. Dans ce chapitre, on s'est intéressé à différentes caractéristiques concernant les mots d'un texte, pensé comme une séquence d'unités ordonnées. Pour rappel, les textes ont été choisis pour des raisons techniques, l'accent étant davantage mis sur les propriétés génériques que spécifiques à un texte. Pour commencer, l'indice d'autocorrélation a été mesuré en considérant la longueur des mots, avec des voisinages définis par les trois matrices d'échange présentées dans le chapitre 3, en fonction de r (itérations, distance ou largeur). Comme on l'espérait pour le français, l'indice a correctement détecté une alternance entre des mots courts et des mots longs, correspondant certainement aux mots outils et aux mots pleins lorsque l'on considère un voisinage restreint.

Ensuite, l'apparition, ou non, de quatre parties du discours a été analysée avec la matrice d'échange à fenêtres mobiles. Concernant les noms, on a observé qu'ils sont rarement suivis (ou précédés) par un autre nom, ce qui est cohérent avec l'utilisation du français. Inversement, les verbes sont souvent suivis par d'autres verbes, ce que l'on a attribué à l'emploi de temps composés. Concernant les adjectifs et les adverbes, l'indice d'autocorrélation n'était jamais significatif, et seule une tendance à ne pas avoir deux adjectifs (respectivement deux adverbes) qui se suivent, mais à avoir deux adjectifs (respectivement deux adverbes) séparés par une autre CMS, a été observée.

Pour terminer, des dissimilarités sémantiques, basées sur WordNet, ont été étudiées avec la matrice d'échange à fenêtres mobiles sur deux textes différents, pour deux parties du discours : les noms et les verbes. Bien qu'en général le choix des textes ait été arbitraire, ici le second texte a été sélectionné avec l'espoir qu'une nouvelle contienne un matériel plus intéressant et hétérogène du point de vue sémantique. En premier lieu, l'indice d'autocorrélation a été mesuré pour chacun des textes et pour chacune des parties du discours, sur la matrice des dissimilarités sémantiques. Dans les deux textes et pour les deux parties du discours, il n'était jamais significatif et se comportait de façon assez régulière en fonction de la taille du voisinage envisagé. On remarque cependant que pour la nouvelle, il est clairement positif et ce pour une grande gamme de voisinages. On peut donc supposer l'existence d'un champ lexical cohérent dans un proche voisinage. Après avoir représenté graphiquement ces dissimilarités au moyen du *Multidimensional Scaling* (MDS), l'autocorrélation a été mesurée sur les deux premiers facteurs extraits de ce MDS. Finalement, l'interprétation des résultats obtenus pour les deux textes était différente, car les structures produites par le MDS étaient différentes. Premièrement, alors que l'indice d'autoccorélation n'était jamais significatif pour le premier texte, il l'était pour le second texte plus littéraire lorsque le premier facteur était considéré. En particulier, on a constaté que de longs segments de textes contenaient des noms appartenant au même concept parmi les deux concepts observés. Concernant les verbes, on a pu constater la coprésence des verbes d'action, plus nombreux que les verbes d'état.

En résumé, même en se limitant à quelques applications, on peut déjà observer les possibilités de cet indice dans une analyse exploratoire de textes. En plus des autres applications qui sont présentées dans les articles cités dans ce chapitre, il en existe de nombreuses autres.

Comme dernière application, le chapitre 8 s'intéressait à l'analyse purement exploratoire de partitions musicales. Dans une première partie, les partitions, représentées à l'aide de tables de contingence selon différentes durées d'intervalles de temps, ont été analysées dans leur ensemble. Pour ce faire, l'AFC et l'indice d'autocorrélation ont été utilisés. Notre espoir était ici, d'une part, de pouvoir visualiser la structure d'une partition avec l'AFC et de percevoir des groupements de notes selon les accords ; et d'autre part, de détecter des motifs avec l'indice d'autocorrélation. L'analyse a d'abord été effectuée sur une partition monophonique, ce qui nous a permis de mieux appréhender le genre de résultats que l'on était en droit d'attendre avec ces analyses. Il a rapidement semblé évident que certaines structures pouvaient être visualisées par l'AFC et mises en évidence par l'indice d'autocorrélation, mais qu'il serait difficile de détecter

des accords. L'analyse effectuée ensuite sur des partitions polyphoniques a confirmé qu'il était possible de visualiser des structures présentes dans une partition, à condition de sélectionner la « bonne » durée d'intervalles de temps. En revanche, l'indice d'autocorrélation a repéré les structures partiellement répétées, à condition qu'il n'y ait pas de modulation, donc de transposition à l'intérieur d'une partition. Finalement, on comprend qu'il est plus simple de synthétiser l'information de morceaux de musique contenant des formes de répétition et qui correspondent probablement à ceux que l'on retient facilement.

L'analyse s'est ensuite portée sur les différentes voix que comprend une partition. Dans ce second cas, il n'y avait pas d'attentes particulières, l'idée étant plutôt d'étendre les analyses effectuées dans la première partie de ce chapitre. Ainsi, une variante flou de l'analyse multiple des correspondances, ainsi qu'un indice d'autocorrélation croisée, ont été utilisés. À nouveau, il a été possible de visualiser certaines structures présentes dans une partition au moyen de l'analyse factorielle et de repérer des répétitions (partielles ou complètes), cette fois entre les différentes voix, grâce à l'indice d'autocorrélation croisée. Malgré ces résultats intéressants, on est derechef confronté au problème de la détermination de la « bonne » durée d'intervalles de temps, question qu'il reste à élucider.

Finalement, au vu des caractérisations particulières obtenues pour chaque partition dans la première partie de l'analyse grâce à la représentation choisie, il semblait cohérent de comparer les représentations de chacune des partitions pour examiner si des similitudes se dégageaient entre certaines d'entre elles. Pour cela, une adaptation du coefficient RV permettant de mesurer la similarité entre deux tables de contingence a été proposée. Après avoir transformé les similarités entre 20 partitions en dissimilarités, une classification ascendante hiérarchique a permis de constater qu'en général, les partitions d'un même compositeur étaient regroupées.

S'agissant d'une thèse, la recherche ne s'achève pas avec cette conclusion : plusieurs questions restent ouvertes et un grand nombre de pistes concernant les suites possibles de ce travail peuvent être explorées. Ainsi, la deuxième question sera : « Quelles pistes de recherche semblent les plus pertinentes pour continuer cette recherche ? »

Pour reprendre la structure de la thèse, on se penchera d'abord sur les nouvelles possibilités à envisager concernant les méthodes. La méthode de visualisation (AFC), ainsi que les méthodes classiques de classification, supervisée ou non, ont été choisies pour leur compatibilité avec des dissimilarités euclidiennes carrées. Ainsi, il a été possible de les utiliser avec des transformations de Schoenberg. Dans cette thèse, seule la transformation de puissance a été envisagée. Bien qu'elle permette des améliorations dans certaines des applications de classification proposées, il serait sans nul doute judicieux d'en expérimenter d'autres, telles que, pour n'en citer qu'une, la transformation gaussienne :

$$\varphi(D) = 1 - \exp(-qD) \qquad \text{où} \qquad q \geq 0$$

Aussi, comme il a été fait avec la classification non supervisée par l'algorithme K-means, et celle supervisée, avec l'analyse discriminante, il serait possible de combiner facilement les transformations de Schoenberg avec la méthode K-means flou. De plus, il serait également imaginable de visualiser des dissimilarités du khi2 alliées aux transformations de Schoenberg par le MDS. Finalement, une piste certainement pertinente serait, comme il a déjà été proposé dans la discussion du chapitre 5 pour un cas restreint (5.1), de combiner différentes matrices de dissimilarités euclidiennes carrées (correspondant à différentes caractéristiques mesurées sur des données, éventuellement associées à des transformations de Schoenberg), avec des poids non-négatifs β et normalisés, ce qui permettrait d'obtenir une matrice de dissimilarités euclidiennes carrées globale :

$$D_{\text{tot}} = \beta_1 D_1 + \beta_2 D_2 + \ldots + \beta_p D_p \qquad \text{où} \qquad \beta_k \geq 0 \qquad \text{et} \qquad \beta_1 + \beta_2 + \ldots + \beta_p = 1$$

Cette nouvelle matrice pourrait alors être utilisée de manière analogue aux autres matrices de dissimilarités avec les différentes méthodes proposées dans cette thèse. À l'extrême, il serait toujours possible de l'associer à de nouvelles transformations de Schoenberg.

Pour revenir sur le choix des dissimilarités, seules les dissimilarités du khi2, ou les dissimilarités euclidiennes carrées en général ont été utilisées, car ces dernières possédaient la forme adéquate pour l'ensemble des méthodes expérimentées dans cette thèse. Cela étant, d'autres distances auraient pu être utilisées pour faire de la classification. En particulier, il serait intéressant de considérer des dissimilarités adaptées, pour autant qu'elles existent, aux matrices creuses rencontrées dans certaines de nos applications, en particulier lors de la classification non supervisée des types de discours.

À propos de la classification non supervisée : seule une partie des possibilités de la méthode K-means flou a été exploitée dans ce travail. Il serait possible, en particulier, d'y implémenter le principe du recuit-simulé. En bref, ce dernier consiste à démarrer l'itération de l'algorithme, comme dans la version proposée ici, à partir d'une température élevée, puis, à abaisser lentement la température à chaque itération, *i.e.* suffisamment lentement pour que la convergence soit assurée à chaque étape. L'algorithme convergera alors vers une solution dure, dont on peut espérer que la variance intra-groupe résultante sera plus faible que celle résultante de la méthode K-means dur. Les itérations initiales à température élevée visent, en effet, à permettre au système d'explorer plus efficacement l'ensemble des partitions floues possibles et d'éviter ainsi d'être emprisonné dans une configuration locale dont la variance intra-groupe serait trop élevée (voir par exemple Rose *et al.*, 1990).

Au sujet des applications, divers axes de recherche qui nous semblent prometteurs pourraient encore être exploités. Ils seront simplement résumés ici et en partie étendus, car ils ont déjà été largement exposés dans la dernière section de chaque chapitre concernant les applications.

Concernant le chapitre 4, la question de la structure hiérarchique des types de discours a simplement été éludée, bien qu'elle puisse, à première vue, déboucher sur différentes nouvelles pistes d'étude. Par exemple, on pourrait analyser si certains types de discours tendent à être systématiquement inclus dans d'autres. On le sait déjà pour le type injonctif qui est systématiquement inclus dans le type dialogal dans les textes traités ici, mais il existe peut-être d'autres inclusions significatives. Aussi, il semblerait que certains passages aient un type de discours dominant. Il serait alors utile qu'un expert humain indique ces dominances dans la hiérarchie, ce qui permettrait de travailler sur des segments plus longs et donc susceptibles de contenir plus d'information à classer. Finalement, il serait certainement précieux de pouvoir disposer d'un plus grand nombre de textes annotés.

Dans le chapitre 4, tout comme dans le chapitre 5, il pourrait aussi être intéressant de se pencher sur une vision plus « ingénieure » du problème, c'est-à-dire clairement axée sur la performance, en contraste avec le propos principal de la thèse, axé sur l'analyse exploratoire des données. Il faudrait alors combiner un grand nombre de caractéristiques, y appliquer une sélection de ces caractéristiques, puis combiner différentes méthodes. De plus, il faudrait disposer de données plus étendues : la prise en compte d'un grand nombre de caractéristiques sur des jeux de données relativement réduits étant particulièrement susceptible de produire des règles surparamétrées.

Spécifiquement, au sujet du chapitre 5, il a été mis en évidence que certaines des caractéristiques linguistiques utilisées étaient plus efficaces pour discriminer certains actes de dialogue et que les transformations de Schoenberg s'avéraient utiles. En plus de combiner ces caractéristiques et ces transformations comme proposé dans le chapitre ou ci-dessus, on pourrait étudier l'intérêt des différentes caractéristiques pour chaque acte de dialogue. En d'autres termes, il faudrait analyser le rapport entre ces actes et ces caractéristiques. Aussi, comme déjà mentionné dans le chapitre, il pourrait s'avérer intéressant d'utiliser un algorithme qui permette d'attribuer directement toutes les étiquettes à un tour de parole donné, plutôt que de procéder à une série de classifications binaires pour chaque étiquette, malgré la faiblesse des liens statistiques entre

les différentes étiquettes.

Concernant le chapitre 6, on ne reviendra pas sur l'étendue des applications qui pourraient être effectuée avec cet indice sur des textes, telles que la navigation hypertextuelle. Bien qu'il ne s'agisse pas d'un point strictement central dans ce chapitre, on peut se rappeler que les dissimilarités sémantiques ont pu être représentées grâce à un MDS sur des dissimilarités adéquates, permettant l'émergence de différents groupes de mots. Bien que ces derniers étaient concentrés, effectuer une validation expérimentale de ces groupes par le *bootstrap* aurait pu certifier la présence desdits groupes. Aussi, ces dissimilarités sémantiques nous semblent particulièrement fascinantes et il serait assurément profitable de les utiliser sur d'autres textes sur lesquels les analyses proposées dans ce chapitre pourraient être effectuées. Un autre axe de recherche pourrait se concentrer sur la définition de nouvelles matrices d'échange, alternatives aux trois familles proposées dans cette thèse, et susceptibles de modéliser différents modes de lecture.

Finalement, le chapitre 8 reposait sur une représentation originale de la musique, exploitée en partie seulement. Pour rappel, il resterait encore à définir systématiquement la durée de l'intervalle de temps qui serait la mieux à même de faire émerger des structures intéressantes. Il faudrait aussi étudier dans quelle mesure le problème des passages répétés *transposés* pourrait être contourné. Effectuer les mêmes analyses que celles présentées dans cette thèse, mais sur un plus grand nombre de partitions, semble une voie toute tracée pour favoriser l'apparition de régularités robustes et de classifications susceptibles d'être interprétées de façon plus stable. Dans le même esprit, il serait également possible d'appliquer le *bootstrap* pour vérifier la significativité des représentations obtenues à l'aide de l'AFC et de l'analyse des correspondances multiples, comme il a été fait dans le chapitre 4 pour les données textuelles.

En guise de conclusion ouverte à ce travail, on peut proposer quelques perspectives supplémentaires, dont certaines se situent au-delà des théories ou des données considérées dans cette thèse.

La musique et le texte ont clairement été étudiés séparément dans cette thèse, se voyant même dédier deux parties différentes. Cependant, on aurait pu, par exemple, envisager le texte comme une séquence d'unités ordonnées, tel qu'il l'a été fait pour la musique. On pourrait typiquement s'intéresser à la suite des propositions de l'un des textes étudiés au chapitre 4 et, grâce à la table de contingence propositions – CMS à disposition, on pourrait bien évidemment utiliser l'indice d'autocorrélation pour étudier la similarité entre ces propositions (des applications de ce type ont été produites dans les articles cités dans le chapitre 6). Cependant, il serait aussi envisageable de représenter graphiquement ces propositions reliées selon le déroulement du texte avec l'AFC, comme il a été fait pour la musique. On pourrait aussi imaginer de comparer des textes représentés ainsi en mesurant la similarité des configurations avec la version pondérée du coefficient RV utilisé pour la musique, à condition que les textes comportent le même nombre de positions, ce qui est typiquement le cas de corpus parallèles.

Aussi, un indice d'autocorrélation croisée a été proposé et appliqué uniquement à la musique. En particulier, il a servi à mesurer les similarités entre les différentes voix d'une même partition. Dans ce contexte particulier, toutes les conditions d'utilisation de l'indice étaient à peu près remplies, à savoir : le même nombre de positions (les intervalles de temps), le même nombre de caractéristiques (les hauteurs de notes) et les mêmes poids en ligne (poids uniformes). Cependant, on imagine l'intérêt que pourrait avoir cet indice pour les textes. Il serait par exemple possible, selon certaines caractéristiques qu'il reste à préciser, de comparer un texte et sa traduction dans une autre langue ou deux versions d'un même texte. Il faudrait alors soit étudier s'il est possible de remplir les différentes conditions d'utilisation, soit généraliser l'indice d'autocorrélation afin qu'il puisse être utilisé dans d'autres conditions.

Pour terminer, la musique est souvent accompagnée de texte. Il peut s'agir de métadonnées, mais aussi des paroles d'une chanson ou d'un opéra. Il apparaît alors qu'un grand nombre d'analyses, combinant les deux types de données, pourraient être effectuées. Pour n'en citer

que quelques unes : le rapport entre le texte et la musique, à un instant donné, pourrait être examiné ; les textes associés à une partition pourraient constituer, en des termes qu'il resterait à préciser, une caractéristique supplémentaire pour une classification ; ou encore il serait possible de compléter les dissimilarités concernant la musique par celles caractérisant le texte, produisant finalement de nouvelles mesures d'autocorrélation.

ANNEXES

ANNEXE A

Textes de Maupassant annotés

Cette annexe présente les textes de Maupassant annotés en types du discours, avec des balises XML, par Raphaël Pittier, étudiant de master en sciences du langage et de la communication, ainsi qu'en français moderne (orientation linguistique française), en 2011. Ces textes ont été utilisés pour les analyses du chapitre 4. La définition des balises employées pour l'annotation, ainsi que la description de ce corpus, se trouvent dans la section 4.1.2.

Les quatre contes annotés sont :

— « L'Orient » (section A.1),
— « Le Voleur » (section A.2),
— « Un Fou ? » (section A.3) et
— « Un Fou » (section A.4).

A.1 L'Orient

```
1 <?xml version="1.0" encoding="ISO-8859-1" ?>
2 <text source="http://un2sg4.unige.ch/athena/selva/maupassant/
    textes/orient.html" date="2011.03.05">
3   <title>L'Orient</title>
4   <div type="narratif">
5     <e>Voici l'automne !</e>
6     <e>Je ne puis sentir ce premier frisson d'hiver sans songer à
        l'ami</e>
7     <e>qui vit là-bas sur la frontière de l'Asie.</e><cr/>
8     <e>La dernière fois que j'entrai chez lui,</e>
9     <e>je compris</e>
10    <e>que je ne le reverrais plus.</e>
11    <div type="descriptif">
12      <e>C'était vers la fin de septembre, voici trois ans.</e>
13    </div>
14    <e>Je le trouvai tantôt couché sur un divan, en plein rêve d'
        opium.</e>
15    <e>Il me tendit la main sans remuer le corps,</e>
16    <e>et me dit :</e><cr/>
17    <div type="dialogal">
```

```
18      <div type="injonctif">
19        <e>Reste là, parle,</e>
20      </div>
21      <div type="argumentatif">
22        <e>je te répondrai de temps en temps,</e>
23        <div type="explicatif">
24          <e>mais je ne bougerai point,</e>
25          <e>car tu sais qu'une fois la drogue avalée</e>
26          <e>il faut demeurer sur le dos.</e><cr/>
27        </div>
28      </div>
29    </div>
30    <e>Je m'assis</e>
31    <e>et je lui racontai mille choses, des choses de Paris et du
         boulevard.</e><cr/>
32    <e>Il me dit :</e><cr/>
33    <div type="dialogal">
34      <e>- Tu ne m'intéresses pas ;</e>
35      <e>je ne songe plus qu'aux pays clairs.</e>
36      <e>Oh ! comme ce pauvre Gautier devait souffrir, toujours
           habité par le désir de l'Orient.</e>
37      <e>Tu ne sais pas</e>
38      <e>ce que c'est,</e>
39      <e>comme il vous prend, ce pays,</e>
40      <e>vous captive,</e>
41      <e>vous pénètre jusqu'au coeur,</e>
42      <e>et ne vous lâche plus.</e>
43      <e>Il entre en vous par l'oeil, par la peau, par toutes ses
           séductions invincibles,</e>
44      <e>et il vous tient par un invisible fil</e>
45      <e>qui vous tire sans cesse, en quelque lieu du monde</e>
46      <e>que le hasard vous ait jeté.</e>
47      <div type="explicatif">
48        <e>Je prends la drogue</e>
49        <e>pour y penser dans la délicieuse torpeur de l'opium.</e
             ><cr/>
50      </div>
51    </div>
52    <e>Il se tut</e>
53    <e>et ferma les yeux.</e>
54    <e>Je demandai :</e><cr/>
55    <div type="dialogal">
56      <div type="explicatif">
57        <e>- Qu'éprouves-tu de si agréable à prendre ce poison ?</
             e>
58        <e>Quel bonheur physique donne-t-il donc,</e>
59        <e>qu'on en absorbe jusqu'à la mort ?</e><cr/>
60      </div>
61    </div>
62    <e>Il répondit :</e><cr/>
63    <div type="dialogal">
```

```
64      <div type="explicatif">
65        <div type="descriptif">
66          <e>- Ce n'est point un bonheur physique ;</e>
67          <e>c'est mieux,</e>
68          <e>c'est plus.</e>
69          <e>Je suis souvent triste ;</e>
70        </div>
71        <e>je déteste la vie,</e>
72        <e>qui me blesse chaque jour par tous ses angles, par
             toutes ses duretés.</e>
73        <e>L'opium console de tout,</e>
74        <e>fait prendre son parti de tout.</e>
75        <e>Connais-tu cet état de l'âme</e>
76        <e>que je pourrais appeler l'irritation harcelante ?</e>
77        <e>Je vis ordinairement dans cet état.</e>
78        <e>Deux choses m'en peuvent guérir : l'opium, ou l'Orient
             .</e>
79        <div type="narratif">
80          <e>A peine ai-je pris l'opium</e>
81          <e>que je me couche,</e>
82          <e>et j'attends.</e>
83          <e>J'attends une heure, deux heures parfois.</e>
84          <e>Puis, je sens d'abord de légers frémissements dans
               les mains et dans les pieds, non pas une crampe, mais
                un engourdissement vibrant.</e>
85          <e>Puis peu à peu j'ai l'étrange et délicieuse sensation
                de la disparition de mes membres.</e>
86          <e>Il me semble</e>
87          <e>qu'on me les ôte.</e>
88          <e>Cela gagne,</e>
89          <e>monte,</e>
90          <e>m'envahit entièrement.</e>
91          <e>Je n'ai plus de corps.</e>
92          <e>Je n'en garde plus qu'une sorte de souvenir agréable.
               </e>
93          <e>Ma tête seule est là,</e>
94          <e>et travaille.</e>
95          <e>Je pense.</e>
96          <e>Je pense avec une joie matérielle infinie, avec une
               lucidité sans égale, avec une pénétration surprenante
               .</e>
97          <e>Je raisonne,</e>
98          <e>je déduis,</e>
99          <e>je comprends tout,</e>
100         <e>je découvre des idées</e>
101         <e>qui ne m'avaient jamais effleuré ;</e>
102         <e>je descends en des profondeurs nouvelles,</e>
103         <e>je monte à des hauteurs merveilleuses ;</e>
104         <e>je flotte dans un océan de pensées,</e>
105         <e>et je savoure l'incomparable bonheur, l'idéale
               jouissance de cette pure et sereine ivresse de la
```

```
            seule intelligence.</e><cr/>
106       </div>
107     </div>
108   </div>
109   <e>Il se tut encore</e>
110   <e>et ferma de nouveau les yeux.</e>
111   <e>Je repris :</e><cr/>
112   <div type="dialogal">
113     <div type="explicatif">
114       <e>- Ton désir de l'Orient ne vient que de cette constante
              ivresse.</e>
115       <e>Tu vis dans une hallucination.</e>
116       <e>Comment désirer ce pays barbare</e>
117       <e>où l'Esprit est mort,</e>
118       <e>où la Pensée stérile ne sort point des étroites limites
              de la vie,</e>
119       <e>ne fait aucun effort pour s'élancer, grandir et conqué
             rir ?</e><cr/>
120     </div>
121   </div>
122   <e>Il répondit :</e><cr/>
123   <div type="dialogal">
124     <div type="explicatif">
125       <e>- Qu'importe la pensée pratique !</e>
126       <e>Je n'aime que le rêve.</e>
127       <e>Lui seul est bon,</e>
128       <e>lui seul est doux.</e>
129       <e>La réalité implacable me conduirait au suicide</e>
130       <e>si le rêve ne me permettait d'attendre.</e><cr/>
131     </div>
132     <div type="argumentatif">
133       <e>"Mais tu as dit</e>
134       <div type="descriptif">
135         <e>que l'Orient était la terre des barbares ;</e>
136       </div>
137       <div type="injonctif">
138         <e>tais-toi, malheureux</e>
139       </div>
140       <div type="descriptif">
141         <e>c'est la terre des sages, la terre chaude</e>
142         <e>où on laisse couler la vie,</e>
143         <e>où on arrondit les angles.</e><cr/>
144       </div>
145       <div type="descriptif">
146         <e>Nous sommes les barbares, nous autres gens de l'
              Occident</e>
147         <e>qui nous disons civilisés ;</e>
148         <e>nous sommes d'odieux barbares</e>
149         <e>qui vivons durement, comme des brutes.</e><cr/>
150       </div>
151       <div type="injonctif">
```

```
152         <e>"Regarde nos villes de pierres, nos meubles de bois
               anguleux et durs.</e>
153       </div>
154       <div type="explicatif">
155         <e>Nous montons en haletant des escaliers étroits et
               rapides</e>
156         <e>pour entrer en des appartements étranglés,</e>
157         <e>où le vent glacé pénètre en sifflant pour s'enfuir
               aussitôt par un tuyau de cheminée en forme de pompe,<
               /e>
158         <e>qui établit des courants d'air mortels, forts à faire
                tourner des moulins.</e>
159       </div>
160       <div type="descriptif">
161         <e>Nos chaises sont dures,</e>
162         <e>nos murs froids, couverts d'un odieux papier ;</e>
163         <e>partout des angles nous blessent.</e>
164         <e>Angles des tables, des cheminées, des portes, des
               lits.</e>
165       </div>
166       <div type="explicatif">
167         <e>Nous vivons debout ou assis, jamais couchés, sauf
               pour dormir,</e>
168         <e>ce qui est absurde,</e>
169         <e>car on ne perçoit plus dans le sommeil le bonheur d'ê
               tre étendu.</e><cr/>
170       </div>
171       <div type="injonctif">
172         <e>"Mais songe aussi à notre vie intellectuelle.</e>
173       </div>
174       <div type="descriptif">
175         <e>C'est la lutte, la bataille incessante.</e>
176       </div>
177       <e>Le souci plane sur nous,</e>
178       <e>les préoccupations nous harcèlent ;</e>
179       <e>nous n'avons plus le temps de chercher et de poursuivre
              les deux ou trois bonnes choses à portée de nos mains
             .</e><cr/>
180       <div type="descriptif">
181         <e>"C'est le combat à outrance.</e>
182       </div>
183       <e>Plus que nos meubles encore, notre caractère a des
             angles, toujours des angles !</e><cr/>
184       <e>"A peine levés, nous courons au travail par la pluie ou
              la gelée.</e>
185       <e>Nous luttons contre les rivalités, les compétitions,
             les hostilités.</e>
186       <div type="descriptif">
187         <e>Chaque homme est un ennemi</e>
188         <e>qu'il faut craindre et terrasser,</e>
189         <e>avec qui il faut ruser.</e>
```

```
190         </div>
191         <div type="descriptif">
192           <e>L'amour même a, chez nous, des aspects de victoire et
                  de défaite :</e>
193           <e>c'est encore une lutte."</e><cr/>
194         </div>
195       </div>
196     </div>
197     <e>Il songea quelques secondes et reprit :</e><cr/>
198     <div type="dialogal">
199       <div type="descriptif">
200         <e>- La maison que je vais acheter,</e>
201         <e>je la connais.</e>
202         <e>Elle est carrée, avec un toit plat et des découpures de
              bois à la mode orientale.</e>
203         <e>De la terrasse, on voit la mer,</e>
204         <e>où passent ces voiles blanches, en forme d'ailes
              pointues, des bateaux grecs ou musulmans.</e>
205         <e>Les murs du dehors sont presque sans ouvertures.</e>
206         <e>Un grand jardin,</e>
207         <e>où l'air est lourd sous le parasol des palmiers,</e>
208         <e>forme le milieu de cette demeure.</e>
209         <e>Un jet d'eau monte sous les arbres</e>
210         <e>et s'émiette en retombant dans un large bassin de
              marbre</e>
211         <e>dont le fond est sablé de poudre d'or.</e>
212         <e>Je m'y baignerai à tout moment, entre deux pipes, deux
              rêves ou deux baisers.</e><cr/>
213       </div>
214       <e>"Je n'aurai point la servante, la hideuse bonne au
            tablier gras,</e>
215       <e>et qui relève en s'en allant, d'un coup de sa savate usée
            , le bas fangeux de sa jupe.</e>
216       <e>Oh ! ce coup de talon</e>
217       <e>qui montre la cheville jaune,</e>
218       <e>il me remue le coeur de dégoût,</e>
219       <e>et je ne le puis éviter.</e>
220       <e>Elles l'ont toutes, les misérables !</e><cr/>
221       <e>"Je n'entendrai plus le claquement de la semelle sur le
            parquet, le battement des portes lancées à toute volée,
            le fracas de la vaisselle</e>
222       <e>qui tombe.</e><cr/>
223       <e>"J'aurai des esclaves noirs et beaux, drapés dans un
            voile blanc</e>
224       <e>et qui courent, nu-pieds, sur les tapis sourds.</e><cr/>
225       <e>"Mes murs seront moelleux et rebondissants comme des
            poitrines de femmes,</e>
226       <e>et, sur mes divans en cercle autour de chaque appartement
            , toutes les formes de coussins me permettront de me
            coucher dans toutes les postures</e>
227       <e>qu'on peut prendre.</e><cr/>
```

```
228       <e>"Puis,</e>
229       <e>quand je serai las du repos délicieux, las de jouir de l'
             immobilité de mon rêve éternel, las du calme plaisir d'ê
             tre bien,</e>
230       <e>je ferai amener devant ma porte un cheval blanc ou noir</
             e>
231       <e>qui courra très vite.</e><cr/>
232       <e>"Et je partirai sur son dos, en buvant l'air</e>
233       <e>qui fouette</e>
234       <e>et grise,</e>
235       <e>l'air sifflant des galops furieux.</e><cr/>
236       <e>"Et j'irai comme une flèche sur cette terre colorée</e>
237       <e>qui enivre le regard,</e>
238       <e>dont la vue est savoureuse comme un vin.</e><cr/>
239       <e>"A l'heure calme du soir, j'irai, d'une course affolée,
             vers le large horizon</e>
240       <e>que le soleil couchant teinte en rose.</e>
241       <div type="descriptif">
242         <e>Tout devient rose, là-bas, au crépuscule : les
               montagnes brûlées, le sable, les vêtements des Arabes,
               la robe blanche des chevaux.</e><cr/>
243       </div>
244       <e>"Les flamants roses s'envoleront des marais sur le ciel
             rose ;</e>
245       <e>et je pousserai des cris de délire, noyé dans la roseur
             illimitée du monde.</e><cr/>
246       <e>"Je ne verrai plus, le long des trottoirs, assourdis par
             le bruit dur des fiacres sur les pavés, des hommes vêtus
             de noir, assis sur des chaises incommodes, boire l'
             absinthe en parlant d'affaires.</e><cr/>
247       <e>"J'ignorerai le cours de la Bourse, les fluctuations des
             valeurs, toutes les inutiles bêtises</e>
248       <e>où nous gaspillons notre courte, misérable et trompeuse
             existence.</e>
249       <div type="explicatif">
250         <e>Pourquoi ces peines, ces souffrances, ces luttes ?</e>
251       </div>
252       <e>Je me reposerai à l'abri du vent dans ma somptueuse et
             claire demeure.</e><cr/>
253       <e>"Et j'aurai quatre ou cinq épouses en des appartements
             moelleux, cinq épouses venues des cinq parties du monde
             ,</e>
254       <e>et qui m'apporteront la saveur de la beauté féminine é
             panouie dans toutes les races."</e><cr/>
255     </div>
256     <e>Il se tut encore,</e>
257     <e>puis prononça doucement :</e><cr/>
258     <div type="dialogal">
259       <div type="injonctif">
260         <e>- Laisse-moi.</e><cr/>
261       </div>
```

```
262     </div>
263     <e>Je m'en allai.</e>
264     <e>Je ne le revis plus.</e><cr/>
265     <e>Deux mois plus tard, il m'écrivit ces trois mots seuls :</e
          >
266     <div type="dialogal">
267       <e>"Je suis heureux."</e>
268     </div>
269     <e>Sa lettre sentait l'encens et d'autres parfums très doux.</
          e><cr/>
270   </div>
271 </text>
```

A.2 Le Voleur

```
1 <?xml version="1.0" encoding="ISO-8859-1" ?>
2 <text source="http://un2sg4.unige.ch/athena/selva/maupassant/
    textes/voleur.html" date="2011.07.06" date-origin="1882.06.21">
3   <title>LE VOLEUR</title>
4   <div type="dialogal">
5     <e>"Puisque je vous dis</e>
6     <e> qu'on ne la croira pas.</e><cr/>
7     <div type="injonctif">
8       <e>- Racontez tout de même.</e><cr/>
9     </div>
10    <div type="argumentatif">
11      <e>- Je le veux bien.</e>
12      <e> Mais j'éprouve d'abord le besoin de vous affirmer</e>
13      <e> que mon histoire est vraie en tous points, quelque
          invraisemblable qu'elle paraisse.</e>
14      <e> Les peintres seuls ne s'étonneront point, surtout les
          vieux</e>
15      <e> qui ont connu cette époque</e>
16      <e> où l'esprit farceur sévissait si bien</e>
17      <e> qu'il nous hantait encore dans les circonstances les
          plus graves."</e><cr/>
18    </div>
19  </div>
20  <div type="narratif">
21    <e>Et le vieil artiste se mit à cheval sur une chaise.</e><cr/
        >
22    <div type="descriptif">
23      <e>Ceci se passait dans la salle à manger d'un hôtel de
          Barbizon.</e><cr/>
24    </div>
25    <e>Il reprit :</e>
26    <div type="dialogal">
27      <div type="descriptif">
28        <e> "Donc nous avions dîné ce soir-là chez le pauvre
            Sorieul, aujourd'hui mort, le plus enragé de nous.</e>
29        <e> Nous étions trois seulement : Sorieul, moi et Le
```

```
          Poittevin, je crois ;</e>
30        <e> mais je n'oserais affirmer</e>
31        <e> que c'était lui.</e>
32        <e> Je parle, bien entendu, du peintre de marine Eugène Le
              Poittevin, mort aussi, et non du paysagiste, bien
          vivant et plein de talent.</e><cr/>
33        <e>Dire que nous avions dîné chez Sorieul, cela signifie</
          e>
34        <e> que nous étions gris.</e>
35        <e> Le Poittevin seul avait gardé sa raison, un peu noyée
          il est vrai, mais claire encore.</e>
36        <e> Nous étions jeunes, en ce temps-là.</e>
37        <e> Etendus sur des tapis, nous discourions extravagamment
              dans la petite chambre qui touchait à l'atelier.</e>
38        <e> Sorieul, le dos à terre, les jambes sur une chaise,
          parlait bataille,</e>
39        <e> discourait sur les uniformes de l'Empire,</e>
40      </div>
41      <div type="narratif">
42        <e> et soudain se levant, il prit dans sa grande armoire
          aux accessoires une tunique complète de hussard,</e>
43        <e> et s'en revêtit.</e>
44        <e> Après quoi il contraignit Le Poittevin à se costumer
          en grenadier.</e>
45        <e> Et comme celui-ci résistait,</e>
46        <e> nous l'empoignâmes,</e>
47        <e> et, après l'avoir déshabillé, nous l'introduisîmes
          dans un uniforme immense</e>
48        <e> où il fut englouti.</e><cr/>
49        <e>Je me déguisai moi-même en cuirassier.</e>
50        <e> Et Sorieul nous fit exécuter un mouvement compliqué.</
          e>
51        <e> Puis il s'écria :</e>
52        <div type="dialogal">
53          <e> "Puisque nous sommes ce soir des soudards,</e>
54          <div type="injonctif">
55            <e> buvons comme des soudards."</e><cr/>
56          </div>
57        </div>
58        <e>Un punch fut allumé, avalé,</e>
59        <e> puis une seconde fois la flamme s'éleva sur le bol
          rempli de rhum.</e>
60        <e> Et nous chantions à pleine gueule des chansons
          anciennes, des chansons</e>
61        <e> que braillaient jadis les vieux troupiers de la grande
              armée.</e><cr/>
62        <e>Tout à coup Le Poittevin,</e>
63        <e> qui restait, malgré tout, presque maître de lui,</e>
64        <e> nous fit taire,</e>
65        <e> puis, après un silence de quelques secondes, il dit à
          mi-voix :</e>
```

```
66        <div type="dialogal">
67          <e> "Je suis sûr qu'on a marché dans l'atelier."</e>
68        </div>
69        <e> Sorieul se leva comme il put,</e>
70        <e> et s'écria :</e>
71        <div type="dialogal">
72          <e> "Un voleur ! quelle chance !"</e>
73        </div>
74        <e> Puis, soudain, il entonna la Marseillaise :</e><cr/>
75        <div type="dialogal">
76          <e>Aux armes, citoyens !</e><cr/>
77        </div>
78        <e>Et, se précipitant sur une panoplie, il nous équipa,
             selon nos uniformes.</e>
79        <e> J'eus une sorte de mousquet et un sabre ;</e>
80        <e> Le Poittevin, un gigantesque fusil à baïonnette,</e>
81        <e> et Sorieul, ne trouvant pas ce qu'il fallait, </e>
82        <e>s'empara d'un pistolet d'arçon</e>
83        <e> qu'il glissa dans sa ceinture, et d'une hache d'
             abordage</e>
84        <e> qu'il brandit.</e>
85        <e> Puis il ouvrit avec précaution la porte de l'atelier
             ,</e>
86        <e> et l'armée entra sur le territoire suspect.</e><cr/>
87        <e>Quand nous fûmes au milieu de la vaste pièce encombrée
             de toiles immenses, de meubles, d'objets singuliers et
             inattendus,</e>
88        <e> Sorieul nous dit :</e>
89        <div type="dialogal">
90          <e> "Je me nomme général.</e>
91          <e> Tenons un conseil de guerre.</e>
92          <e> Toi, les cuirassiers, tu vas couper la retraite à l'
               ennemi, c'est-à-dire donner un tour de clef à la
               porte.</e>
93          <e> Toi, les grenadiers, tu seras mon escorte."</e><cr/>
94        </div>
95        <e>J'exécutai le mouvement commandé,</e>
96        <e> puis je rejoignis le gros des troupes</e>
97        <e> qui opérait une reconnaissance.</e><cr/>
98        <e>Au moment où j'allais le rattraper derrière un grand
             paravent,</e>
99        <e> un bruit furieux éclata.</e>
100       <e> Je m'élançai, portant toujours une bougie à la main.</
             e>
101       <e> Le Poittevin venait de traverser d'un coup de baï
             onnette la poitrine d'un mannequin </e>
102       <e>dont Sorieul fendait la tête à coups de hache.</e>
103       <e> L'erreur reconnue, le général commanda :</e>
104       <div type="dialogal">
105         <div type="injonctif">
106           <e> "Soyons prudents",</e>
```

```
107         </div>
108       </div>
109       <e> et les opérations recommencèrent.</e><cr/>
110       <div type="descriptif">
111         <e>Depuis vingt minutes au moins on fouillait tous les
              coins et recoins de l'atelier, sans succès,</e>
112         <e> quand Le Poittevin eut l'idée d'ouvrir un immense
              placard.</e>
113         <e> Il était sombre et profond,</e>
114         <e> j'avançai mon bras</e>
115         <e> qui tenait la lumière,</e>
116         <e> et je reculai stupéfait ;</e>
117         <e> un homme était là, un homme vivant,</e>
118         <e> qui m'avait regardé.</e><cr/>
119       </div>
120       <e>Immédiatement, je refermai le placard à deux tours de
            clef,</e>
121       <e> et on tint de nouveau conseil.</e><cr/>
122       <div type="descriptif">
123         <e>Les avis étaient très partagés.</e>
124         <e> Sorieul voulait enfumer le voleur.</e>
125         <e> Le Poittevin parlait de le prendre par la famine.</e
              >
126         <e> Je proposai de faire sauter le placard avec de la
              poudre.</e><cr/>
127       </div>
128       <e>L'avis de Le Poittevin prévalut ;</e>
129       <e> et, pendant qu'il montait la garde avec son grand
            fusil,</e>
130       <e> nous allâmes chercher le reste du punch et nos pipes ;
            </e>
131       <e> puis on s'installa devant la porte fermée,</e>
132       <e> et on but au prisonnier.</e><cr/>
133       <e>Au bout d'une demi-heure, Sorieul dit :</e>
134       <div type="dialogal">
135         <e> "C'est égal,</e>
136         <e> je voudrais bien le voir de près.</e>
137         <e> Si nous nous emparions de lui par la force ?"</e><cr
              />
138       </div>
139       <e>Je criai :</e>
140       <div type="dialogal">
141         <e> "Bravo !"</e>
142       </div>
143       <e> Chacun s'élança sur ses armes ;</e>
144       <e> la porte du placard fut ouverte,</e>
145       <e> et Sorieul, armant son pistolet</e>
146       <e> qui n'était pas chargé,</e>
147       <e> se précipita le premier.</e><cr/>
148       <e>Nous le suivîmes en hurlant.</e>
149       <e> Ce fut une bousculade effroyable dans l'ombre ;</e>
```

```
150       <e> et après cinq minutes d'une lutte invraisemblable,
             nous ramenâmes au jour une sorte de vieux bandit à
             cheveux blancs, sordide et déguenillé.</e><cr/>
151       <e>On lui lia les pieds et les mains,</e>
152       <e> puis on l'assit dans un fauteuil.</e>
153       <e> Il ne prononça pas une parole.</e><cr/>
154       <e>Alors Sorieul, pénétré d'une ivresse solennelle, se
             tourna vers nous :</e><cr/>
155       <div type="dialogal">
156         <e>"Maintenant nous allons juger ce misérable."</e><cr/>
157       </div>
158       <e>J'étais tellement gris</e>
159       <e> que cette proposition me parut toute naturelle.</e><cr
             />
160       <e>Le Poittevin fut chargé de présenter la défense</e>
161       <e> et moi de soutenir l'accusation.</e><cr/>
162       <e>Il fut condamné à mort à l'unanimité moins une voix,
             celle de son défenseur.</e><cr/>
163       <div type="dialogal">
164         <e>"Nous allons l'exécuter"</e>
165       </div>
166       <e>, dit Sorieul.</e>
167       <e> Mais un scrupule lui vint :</e>
168       <div type="dialogal">
169         <e> "Cet homme ne doit pas mourir privé des secours de
               la religion.</e>
170         <e> Si on allait chercher un prêtre ?"</e>
171       </div>
172       <e> J'objectai</e>
173       <e> qu'il était tard.</e>
174       <div type="argumentatif">
175         <e> Alors Sorieul me proposa de remplir cet office ;</e>
176         <e> et il exhorta le criminel à se confesser dans mon
               sein.</e><cr/>
177       </div>
178       <e>L'homme, depuis cinq minutes, roulait des yeux épouvant
             és,</e>
179       <e> se demandant à quel genre d'êtres il avait affaire.</e
             >
180       <e> Alors il articula d'une voix creuse, brûlée par l'
             alcool </e>
181       <div type="dialogal">
182         <e>"Vous voulez rire, sans doute."</e>
183       </div>
184       <e> Mais Sorieul l'agenouilla de force,</e>
185       <e> et, de crainte que ses parents eussent omis de le
             faire baptiser,</e>
186       <e> il lui versa sur le crâne un verre de rhum.</e><cr/>
187       <e>Puis il dit :</e><cr/>
188       <div type="dialogal">
189         <div type="injonctif">
```

```
190           <e>"Confesse-toi à monsieur ;</e>
191         </div>
192         <e> ta dernière heure a sonné."</e><cr/>
193       </div>
194       <e>Eperdu, le vieux gredin se mit à crier :</e><cr/>
195       <div type="dialogal">
196         <e>"Au secours !"</e>
197       </div>
198       <e> avec une telle force qu'on fut contraint de le bâ
             illonner pour ne pas réveiller tous les voisins.</e>
199       <e> Alors il se roula par terre, ruant et se tordant,
             renversant les meubles, crevant les toiles.</e>
200       <e> A la fin, Sorieul, impatienté, cria :</e>
201       <div type="dialogal">
202         <div type="injonctif">
203           <e> "Finissons-en."</e>
204         </div>
205       </div>
206       <e> Et visant le misérable étendu par terre, il pressa la
             détente de son pistolet.</e>
207       <e> Le chien tomba avec un bruit sec.</e>
208       <e> Emporté par l'exemple, je tirai à mon tour.</e>
209       <e> Mon fusil, qui était à pierre, lança une étincelle</e>
210       <e> dont je fus surpris.</e><cr/>
211       <e>Alors Le Poittevin prononça gravement ces paroles :</e>
212       <div type="dialogal">
213         <e> "Avons-nous bien le droit de tuer cet homme ?"</e><
               cr/>
214       </div>
215       <e>Sorieul, stupéfait, répondit :</e>
216       <div type="dialogal">
217         <div type="explicatif">
218           <e> "Puisque nous l'avons condamné à mort !"</e><cr/>
219         </div>
220       </div>
221       <div type="argumentatif">
222         <e>Mais Le Poittevin reprit :</e>
223         <div type="dialogal">
224           <e> "On ne fusille pas les civils,</e>
225           <e> celui-ci doit être livré au bourreau.</e>
226           <e> Il faut le conduire au poste."</e><cr/>
227         </div>
228       </div>
229       <e>L'argument nous parut concluant.</e>
230       <e> On ramassa l'homme,</e>
231       <div type="explicatif">
232         <e> et comme il ne pouvait marcher,</e>
233         <e> il fut placé sur une planche de table à modèle,
               solidement attaché,</e>
234         <e> et je l'emportai avec Le Poittevin,</e>
235         <e> tandis que Sorieul, armé jusqu'aux dents, fermait la
```

```
          marche.</e><cr/>
      </div>
      <e>Devant le poste, la sentinelle nous arrêta.</e>
      <e> Le chef de poste, mandé, nous reconnut,</e>
      <div type="explicatif">
        <e> et, comme chaque jour il était témoin de nos farces,
            de nos scies, de nos inventions invraisemblables,</e
           >
        <e> il se contenta de rire</e>
        <e> et refusa notre prisonnier.</e><cr/>
      </div>
      <e>Sorieul insista :</e>
      <e> alors le soldat nous invita sévèrement à retourner
         chez nous sans faire de bruit.</e><cr/>
      <e>La troupe se remit en route </e>
      <e>et rentra dans l'atelier. </e>
      <e>Je demandai : </e>
      <div type="dialogal">
        <e>"Qu'allons-nous faire du voleur ?"</e><cr/>
      </div>
      <e>Le Poittevin, attendri, affirma</e>
      <e> qu'il devait être bien fatigué, cet homme.</e>
      <e> En effet, il avait l'air agonisant, ainsi ficelé, bâ
         illonné, ligaturé sur sa planche.</e><cr/>
      <e>Je fus pris à mon tour d'une pitié violente, une pitié
         d'ivrogne,</e>
      <e> et, enlevant son bâillon, je lui demandai :</e>
      <div type="dialogal">
        <e> "Eh bien, mon pauv'vieux, comment ça va-t-il ?"</e><
           cr/>
      </div>
      <e>Il gémit :</e>
      <div type="dialogal">
        <e> "J'en ai assez, nom d'un chien !"</e>
      </div>
      <e> Alors Sorieul devint paternel.</e>
      <e> Il le délivra de tous ses liens,</e>
      <e> le fit asseoir,</e>
      <e> le tutoya,</e>
      <e> et, pour le réconforter, nous nous mîmes tous trois à
         préparer bien vite un nouveau punch.</e>
      <e> Le voleur, tranquille dans son fauteuil, nous
         regardait.</e>
      <e> Quand la boisson fut prête,</e>
      <e> on lui tendit un verre-</e>
      <e> nous lui aurions volontiers soutenu la tête,</e>
      <e> et on trinqua.</e><cr/>
      <e>Le prisonnier but autant qu'un régiment.</e>
      <e> Mais, comme le jour commençait à paraître,</e>
      <e> il se leva, et, d'un air fort calme :</e>
      <div type="dialogal">
```

```
278           <div type="explicatif">
279             <e> "Je vais être obligé de vous quitter,</e>
280             <e> parce qu'il faut que je rentre chez moi."</e><cr/>
281           </div>
282         </div>
283         <e>Nous fûmes désolés ;</e>
284         <e> on voulut le retenir,</e>
285         <e> mais il se refusa à rester plus longtemps.</e><cr/>
286         <e>Alors on se serra la main,</e>
287         <e> et Sorieul, avec sa bougie, l'éclaira dans le
              vestibule. en criant :</e>
288         <div type="dialogal">
289           <div type="injonctif">
290             <e> "Prenez garde à la marche sous la porte cochère
                  ."</e><cr/>
291           </div>
292         </div>
293       </div>
294     </div>
295     <e>On riait franchement autour du conteur.</e>
296     <e> Il se leva, alluma sa pipe,</e>
297     <e> et il ajouta, en se campant en face de nous .</e><cr/>
298     <div type="dialogal">
299       <e>"Mais le plus drôle de mon histoire c'est qu'elle est
            vraie."</e><cr/>
300     </div>
301   </div>
302 </text>
```

A.3 Un Fou ?

```
1 <?xml version="1.0" encoding="ISO-8859-1" ?>
2 <text source="http://un2sg4.unige.ch/athena/maupassant/maup_fou.
    html" date="2011.02.07">
3   <title>Un fou ?</title>
4   <div type="explicatif">
5     <div type="narratif">
6       <e>Quand on me dit:</e>
7       <div type="dialogal">
8         <e>"Vous savez</e>
9         <e>que Jacques Parent est mort fou dans une maison de sant
            é",</e>
10      </div>
11      <e>un frisson douloureux, un frisson de peur et d'angoisse
          me courut le long des os;</e>
12      <e>et je le revis brusquement, ce grand garçon étrange, fou
          depuis longtemps peut-être, maniaque inquiétant,
          effrayant même.</e><cr/>
13      <div type="descriptif">
14        <e>C'était un homme de quarante ans, haut, maigre, un peu
            voûté, avec des yeux d'halluciné, des yeux noirs, si
```

```
          noirs</e>
15        <e>qu'on ne distinguait pas la pupille,</e>
16        <e>des yeux mobiles, rôdeurs, malades, hantés.</e>
17        <e>Quel être singulier, troublant</e>
18        <div type="narratif">
19          <e>qui apportait, qui jetait un malaise autour de lui,
              un malaise vague, de l'âme, du corps, un de ces é
              nervements incompréhensibles</e>
20          <e>qui font croire à des influences surnaturelles.</e><
              cr/>
21        </div>
22        <e>Il avait un tic gênant: la manie de cacher ses mains.</
            e>
23        <div type="narratif">
24          <e>Presque jamais il ne les laissait errer,</e>
25          <e>comme nous faisons tous sur les objets, sur les
              tables.</e>
26          <e>Jamais il ne maniait les choses traînantes avec ce
              geste familier</e>
27          <e>qu'ont presque tous les hommes.</e>
28          <e>Jamais il ne les laissait nues, ses longues mains
              osseuses, fines, un peu fébriles.</e><cr/>
29          <e>Il les enfonçait dans ses poches, sous les revers de
              ses aisselles en croisant les bras.</e>
30          <div type="explicatif">
31            <e>On eût dit</e>
32            <e>qu'il avait peur</e>
33            <e>qu'elles ne fissent, malgré lui, quelque besogne dé
                fendue,</e>
34            <e>qu'elles n'accomplissent quelque action honteuse ou
                 ridicule</e>
35            <e>s'il les laissait libres et maîtresses de leurs
                mouvements.</e><cr/>
36          </div>
37          <e>Quand il était obligé de s'en servir pour tous les
              usages ordinaires de la vie,</e>
38          <e>il le faisait par saccades brusques, par élans
              rapides du bras</e>
39          <div type="explicatif">
40            <e>comme s'il n'eût pas voulu leur laisser le temps d'
                agir par elles-mêmes, de se refuser à sa volonté, d
                'exécuter autre chose.</e>
41          </div>
42          <e>A table, il saisissait son verre, sa fourchette ou
              son couteau si vivement</e>
43          <e>qu'on n'avait jamais le temps de prévoir</e>
44          <e>ce qu'il voulait faire</e>
45          <e>avant qu'il ne l'eût accompli.</e><cr/>
46        </div>
47      </div>
48      <div type="argumentatif">
```

```
49        <e>Or, j'eus un soir l'explication de la surprenante
             maladie de son âme.</e><cr/>
50      </div>
51      <e>II venait passer de temps en temps quelques jours chez
           moi, à la campagne,</e>
52      <div type="descriptif">
53        <e>et ce soir-là il me paraissait particulièrement agité
             !</e><cr/>
54      </div>
55      <div type="descriptif">
56        <e>Un orage montait dans le ciel, étouffant et noir, après
             une journée d'atroce chaleur.</e>
57        <e>Aucun souffle d'air ne remuait les feuilles.</e>
58        <e>Une vapeur chaude de four passait sur les visages,</e>
59        <e>faisait haleter les poitrines.</e>
60      </div>
61      <div type="descriptif">
62        <e>Je me sentais mal à l'aise, agité,</e>
63      </div>
64      <e>et je voulus gagner mon lit.</e><cr/>
65      <e>Quand il me vit me lever pour partir,</e>
66      <e>Jacques Parent me saisit le bras d'un geste effaré.</e><
           cr/>
67      <div type="dialogal">
68        <e>- Oh! non,</e>
69        <div type="injonctif">
70          <e>reste encore un peu,</e>
71        </div>
72      </div>
73      <e>me dit-il.</e><cr/>
74      <e>Je le regardai avec surprise en murmurant:</e><cr/>
75      <div type="dialogal">
76        <e>- C'est que cet orage me secoue les nerfs.</e><cr/>
77      </div>
78      <e>Il gémit,</e>
79      <e>ou plutôt il cria:</e><cr/>
80      <div type="dialogal">
81        <e>- Et moi donc! Oh!</e>
82        <div type="injonctif">
83          <e>reste,</e>
84          <e>je te prie ;</e>
85        </div>
86        <e>je ne voudrais pas demeurer seul.</e><cr/>
87      </div>
88      <div type="descriptif">
89        <e>Il avait l'air affolé.</e><cr/>
90      </div>
91      <e>Je prononçai:</e><cr/>
92      <div type="dialogal">
93        <e>- Qu'est-ce que tu as?</e>
94        <e>Perds-tu la tête?</e><cr/>
```

```
95      </div>
96      <e>Et il balbutia:</e><cr/>
97      <div type="dialogal">
98        <div type="explicatif">
99          <e>- Oui, par moments, dans les soirs comme celui-ci,
               dans les soirs d'électricité... j'ai... j'ai... j'ai
               peur... j'ai peur de moi...</e>
100         <e>tu ne me comprends pas?</e>
101         <e>C'est que je suis doué d'un pouvoir... non... d'une
               puissance... non... d'une force...</e>
102         <e>Enfin je ne sais pas dire</e>
103         <e>ce que c'est,</e>
104         <div type="argumentatif">
105           <e>mais j'ai en moi une action magnétique si
                 extraordinaire</e>
106           <e>que j'ai peur, oui, j'ai peur de moi,</e>
107           <e>comme je te le disais tout à l'heure!</e><cr/>
108         </div>
109       </div>
110     </div>
111     <e>Et il cachait, avec des frissons éperdus, ses mains
          vibrantes sous les revers de sa jaquette.</e>
112     <div type="descriptif">
113       <e>Et moi-même je me sentis soudain tout tremblant d'une
            crainte confuse, puissante, horrible.</e>
114     </div>
115     <e>J'avais envie de partir, de me sauver, de ne plus le voir
          , de ne plus voir son oeil errant passer sur moi, puis s'
          enfuir, tourner autour du plafond, chercher quelque coin
          sombre de la pièce pour s'y fixer,</e>
116     <div type="explicatif">
117       <e>comme s'il eût voulu cacher aussi son regard redoutable
            .</e><cr/>
118     </div>
119     <e>Je balbutiai:</e><cr/>
120     <div type="dialogal">
121       <e>- Tu ne m'avais jamais dit ça!</e><cr/>
122     </div>
123     <e>Il reprit:</e><cr/>
124     <div type="dialogal">
125       <e>- Est-ce que j'en parle à personne?</e>
126       <div type="injonctif">
127         <e>Tiens,</e>
128         <e>écoute,</e>
129       </div>
130       <e>ce soir je ne puis me taire.</e>
131       <e>Et j'aime mieux</e>
132       <e>que tu saches tout;</e>
133       <e>d'ailleurs, tu pourras me secourir.</e><cr/>
134       <div type="explicatif">
135         <div type="argumentatif">
```

```
136             <e>Le magnétisme!</e>
137             <e>Sais-tu ce que c'est?</e>
138             <e>Non.</e>
139             <e>Personne ne sait.</e>
140             <e>On le constate pourtant.</e>
141             <e>On le reconnaît,</e>
142             <e>les médecins eux-mêmes le pratiquent;</e>
143             <e>un des plus illustres, M. Charcot, le professe;</e>
144             <e>donc, pas de doute, cela existe.</e><cr/>
145             <e>Un homme, un être a le pouvoir, effrayant et
                   incompréhensible, d'endormir, par la force de sa
                   volonté, un autre être, et,</e>
146             <e>pendant qu'il dort,</e>
147             <e>de lui voler sa pensée</e>
148             <e>comme on volerait une bourse.</e>
149             <e>Il lui vole sa pensée, c'est-à-dire son âme, l'âme,
                    ce sanctuaire, ce secret du Moi, l'âme, ce fond de
                    l'homme</e>
150             <e>qu'on croyait impénétrable,</e>
151             <e>l'âme, cet asile des inavouables idées,</e>
152             <e>de tout ce qu'on cache,</e>
153             <e>de tout ce qu'on aime,</e>
154             <e>de tout ce qu'on veut celer à tous les humains,</e>
155             <e>il l'ouvre,</e>
156             <e>la viole,</e>
157             <e>l'étale,</e>
158             <e>la jette au public!</e>
159           </div>
160           <e>N'est-ce pas atroce, criminel, infâme?</e><cr/>
161           <e>Pourquoi, comment cela se fait-il?</e>
162           <e>Le sait-on?</e>
163           <e>Mais que sait-on?</e><cr/>
164           <e>Tout est mystère.</e>
165           <e>Nous ne communiquons avec les choses que par nos misé
                 rables sens, incomplets, infirmes, si faibles</e>
166           <e>qu'ils ont à peine la puissance de constater</e>
167           <e>ce qui nous entoure.</e>
168           <e>Tout est mystère.</e>
169           <div type="argumentatif">
170             <div type="injonctif">
171               <e>Songe à la musique, cet art divin, cet art</e>
172               <e>qui bouleverse l'âme,</e>
173               <e>l'emporte,</e>
174               <e>la grise,</e>
175               <e>l'affole,</e>
176             </div>
177             <e>qu'est-ce donc?</e>
178             <e>Rien.</e><cr/>
179             <e>Tu ne me comprends pas?</e>
180             <div type="injonctif">
181               <e>Ecoute.</e>
```

```
182           </div>
183           <e>Deux corps se heurtent.</e>
184           <e>L'air vibre.</e>
185           <e>Ces vibrations sont plus ou moins nombreuses, plus
                 ou moins rapides, plus ou moins fortes, selon la
                 nature du choc.</e>
186           <e>Or nous avons dans l'oreille une petite peau</e>
187           <e>qui reçoit ces vibrations de l'air</e>
188           <e>et les transmet au cerveau sous forme de son.</e>
189           <div type="injonctif">
190             <e>Imagine qu'un verre d'eau se change en vin dans
                   ta bouche.</e>
191           </div>
192           <e>Le tympan accomplit cette incroyable métamorphose,
                 ce surprenant miracle de changer le mouvement en
                 son.</e>
193           <e>Voilà.</e><cr/>
194           <e>La musique, cet art complexe et mystérieux, précis
                 comme l'algèbre et vague comme un rêve, cet art
                 fait de mathématiques et de brise, ne vient donc
                 que de la propriété étrange d'une petite peau.</e>
195           <e>Elle n'existerait point, cette peau,</e>
196           <e>que le son non plus n'existerait pas,</e>
197           <e>puisque par lui-même il n'est qu'une vibration.</e>
198           <e>Sans l'oreille, devinerait-on la musique?</e>
199           <e>Non.</e>
200           <div type="explicatif">
201             <e>Eh bien ! nous sommes entourés de choses</e>
202             <e>que nous ne soupçonnerons jamais,</e>
203             <e>parce que les organes nous manquent</e>
204             <e>qui nous les révéleraient.</e><cr/>
205           </div>
206         </div>
207         <e>Le magnétisme est de celles-là peut-être.</e>
208         <e>Nous ne pouvons que pressentir cette puissance,</e>
209         <e>que tenter en tremblant ce voisinage des esprits,</e>
210         <e>qu'entrevoir ce nouveau secret de la nature,</e>
211         <e>parce que nous n'avons point en nous l'instrument rév
              élateur.</e><cr/>
212       </div>
213       <e>Quant à moi... Quant à moi, je suis doué d'une
            puissance affreuse.</e>
214       <e>On dirait un autre être enfermé en moi,</e>
215       <e>qui veut sans cesse s'échapper,</e>
216       <e>agir malgré moi,</e>
217       <e>qui s'agite,</e>
218       <e>me ronge,</e>
219       <e>m'épuise.</e>
220       <e>Quel est-il?</e>
221       <e>Je ne sais pas,</e>
222       <div type="argumentatif">
```

```
223           <e>mais nous sommes deux dans mon pauvre corps,</e>
224           <e>et c'est lui, l'autre, qui est souvent le plus fort,
                 comme ce soir.</e><cr/>
225         </div>
226         <e>Je n'ai qu'à regarder les gens pour les engourdir</e>
227         <e>comme si je leur avais versé de l'opium.</e>
228         <e>Je n'ai qu'à étendre les mains pour produire des choses
               ... des choses... terribles.</e>
229         <e>Si tu savais?</e>
230         <e>Oui.</e>
231         <e>Si tu savais?</e>
232         <div type="argumentatif">
233           <e>Mon pouvoir ne s'étend pas seulement sur les hommes,
                 mais aussi sur les animaux et même... sur les objets
                 ...</e><cr/>
234         </div>
235         <e>Cela me torture</e>
236         <e>et m'épouvante.</e>
237         <e>J'ai eu envie souvent de me crever les yeux et de me
               couper les poignets.</e>
238         <e>Mais je vais...</e>
239         <e>je veux que tu saches tout.</e>
240         <div type="injonctif">
241           <e>Tiens.</e>
242         </div>
243         <div type="argumentatif">
244           <e>Je vais te montrer cela... non pas sur des créatures
                 humaines,</e>
245           <e>c'est ce qu'on fait partout,</e>
246           <e>mais sur... sur... des bêtes.</e><cr/>
247         </div>
248         <div type="injonctif">
249           <e>Appelle Mirza.</e><cr/>
250         </div>
251       </div>
252       <e>Il marchait à grands pas avec des airs d'halluciné,</e>
253       <e>et il sortit ses mains cachées dans sa poitrine.</e>
254       <div type="descriptif">
255         <e>Elles me semblèrent effrayantes</e>
256         <e>comme s'il eût mis à nu deux épées.</e><cr/>
257       </div>
258       <e>Et je lui obéis machinalement, subjugué, vibrant de
             terreur et dévoré d'une sorte de désir impétueux de voir
             .</e>
259       <e>J'ouvris la porte</e>
260       <e>et je sifflai ma chienne</e>
261       <e>qui couchait dans le vestibule.</e>
262       <e>J'entendis aussitôt le bruit précipité de ses ongles sur
             les marches de l'escalier,</e>
263       <e>et elle apparut, joyeuse, remuant la queue.</e><cr/>
264       <e>Puis je lui fis signe de se coucher sur un fauteuil;</e>
```

```
<e>elle y sauta,</e>
<e>et Jacques se mit à la caresser en la regardant.</e><cr/>
<div type="descriptif">
  <e>D'abord, elle sembla inquiète;</e>
  <e>elle frissonnait,</e>
  <e>tournait la tête pour éviter l'oeil fixe de l'homme,</e>
  <e>semblait agitée d'une crainte grandissante.</e>
</div>
<e>Tout à coup, elle commença à trembler,</e>
<e>comme tremblent les chiens.</e>
<e>Tout son corps palpitait, secoué de longs frissons,</e>
<e>et elle voulut s'enfuir.</e>
<div type="argumentatif">
  <e>Mais il posa sa main sur le crâne de l'animal</e>
  <e>qui poussa, sous ce toucher, un de ces longs hurlements</e>
  <e>qu'on entend, la nuit, dans la campagne.</e><cr/>
</div>
<div type="descriptif">
  <e>Je me sentais moi-même engourdi, étourdi,</e>
  <e>ainsi qu'on l'est</e>
  <e>lorsqu'on monte en barque.</e>
</div>
<e>Je voyais se pencher les meubles, remuer les murs.</e>
<e>Je balbutiai:</e>
<div type="dialogal">
  <div type="injonctif">
    <e>"Assez, Jacques, assez."</e>
  </div>
</div>
<div type="argumentatif">
  <e>Mais il ne m'écoutait plus,</e>
</div>
<e>il regardait Mirza d'une façon continue, effrayante.</e>
<e>Elle fermait les yeux maintenant</e>
<e>et laissait tomber sa tête</e>
<e>comme on fait en s'endormant.</e>
<e>Il se tourna vers moi.</e><cr/>
<div type="dialogal">
  <e>- C'est fait,</e>
</div>
<e>dit-il,</e>
<div type="dialogal">
  <div type="injonctif">
    <e>vois maintenant.</e><cr/>
  </div>
</div>
<e>Et jetant son mouchoir de l'autre côté de l'appartement,</e>
  il cria:</e>
<div type="dialogal">
```

```
313         <div type="injonctif">
314           <e>"Apporte!"</e>
315         </div>
316       </div>
317       <e>La bête alors se souleva</e>
318       <e>et chancelant, trébuchant</e>
319       <e>comme si elle eût été aveugle, remuant ses pattes</e>
320       <e>comme les paralytiques remuent leurs jambes,</e>
321       <e>elle s'en alla vers le linge</e>
322       <e>qui faisait une tache blanche contre le mur.</e>
323       <e>Elle essaya plusieurs fois de le prendre dans sa gueule
             ,</e>
324       <e>mais elle mordait à côté</e>
325       <div type="explicatif">
326         <e>comme si elle ne l'eût pas vu.</e>
327       </div>
328       <e>Elle le saisit enfin,</e>
329       <e>et revint de la même allure ballottée de chien somnambule
             .</e><cr/>
330       <div type="descriptif">
331         <e>C'était une chose terrifiante à voir.</e>
332       </div>
333       <e>Il commanda:</e>
334       <div type="dialogal">
335         <div type="injonctif">
336           <e>"Couche-toi."</e>
337         </div>
338       </div>
339       <e>Elle se coucha.</e>
340       <e>Alors, lui touchant le front, il dit:</e>
341       <div type="dialogal">
342         <div type="injonctif">
343           <e>"Un lièvre, pille,</e>
344           <e>pille."</e>
345         </div>
346       </div>
347       <e>Et la bête, toujours sur le flanc, essaya de courir,</e>
348       <e>s'agita</e>
349       <e>comme font les chiens</e>
350       <e>qui rêvent,</e>
351       <e>et poussa, sans ouvrir la gueule, des petits aboiements é
             tranges, des aboiements de ventriloque.</e><cr/>
352       <div type="descriptif">
353         <e>Jacques semblait devenu fou.</e>
354       </div>
355       <e>La sueur coulait de son front.</e>
356       <e>Il cria:</e>
357       <div type="dialogal">
358         <div type="injonctif">
359           <e>"Mords-le,</e>
360           <e>mords ton maître."</e>
```

```
361         </div>
362       </div>
363       <e>Elle eut deux ou trois soubresauts terribles.</e>
364       <div type="explicatif">
365         <e>On eût juré</e>
366         <e>qu'elle résistait,</e>
367         <e>qu'elle luttait.</e>
368       </div>
369       <e>Il répéta:</e>
370       <div type="dialogal">
371         <div type="injonctif">
372           <e>"Mords-le."</e>
373         </div>
374       </div>
375       <e>Alors, se levant, ma chienne s'en vint vers moi,</e>
376       <e>et moi je reculais vers la muraille, frémissant d'é
             pouvante, le pied levé pour la frapper, pour la repousser
             .</e><cr/>
377       <e>Mais Jacques ordonna:</e>
378       <div type="dialogal">
379         <div type="injonctif">
380           <e>"Ici, tout de suite."</e>
381         </div>
382       </div>
383       <e>Elle se retourna vers lui.</e>
384       <e>Alors, de ses deux grandes mains, il se mit à lui frotter
             la tête</e>
385       <div type="explicatif">
386         <e>comme s'il l'eût débarrassée de liens invisibles.</e><
              cr/>
387       </div>
388       <e>Mirza rouvrit les yeux:</e>
389       <div type="dialogal">
390         <e>"C'est fini",</e>
391       </div>
392       <e>dit-il.</e><cr/>
393       <e>Je n'osais point la toucher</e>
394       <e>et je poussai la porte</e>
395       <div type="explicatif">
396         <e>pour qu'elle s'en allât.</e>
397       </div>
398       <e>Elle partit lentement, tremblante, épuisée,</e>
399       <e>et j'entendis de nouveau ses griffes frapper les marches.
             </e><cr/>
400       <e>Mais Jacques revint vers moi:</e>
401       <div type="dialogal">
402         <e>"Ce n'est pas tout.</e>
403         <e>Ce qui m'effraie le plus,</e>
404         <e>c'est ceci,</e>
405         <div type="injonctif">
406           <e>tiens.</e>
```

```
407         </div>
408         <e>Les objets m'obéissent."</e><cr/>
409       </div>
410       <div type="descriptif">
411         <e>Il y avait sur ma table une sorte de couteau-poignard</
              e>
412         <e>dont je me servais pour couper les feuillets des livres
              .</e>
413       </div>
414       <e>Il allongea sa main vers lui.</e>
415       <div type="descriptif">
416         <e>Elle semblait ramper,</e>
417         <e>s'approchait lentement;</e>
418       </div>
419       <e>et tout d'un coup je vis, oui, je vis le couteau lui-même
            tressaillir,</e>
420       <e>puis il remua,</e>
421       <e>puis il glissa doucement, tout seul, sur le bois vers la
            main arrêtée</e>
422       <e>qui l'attendait,</e>
423       <e>et il vint se placer sous ses doigts.</e><cr/>
424       <e>Je me mis à crier de terreur.</e>
425       <div type="argumentatif">
426         <e>Je crus</e>
427         <e>que je devenais fou moi-même,</e>
428         <e>mais le son aigu de ma voix me calma soudain.</e><cr/>
429       </div>
430       <e>Jacques reprit:</e><cr/>
431       <div type="dialogal">
432         <e>- Tous les objets viennent ainsi vers moi.</e>
433         <div type="explicatif">
434           <e>C'est pour cela que je cache mes mains.</e>
435         </div>
436         <e>Qu'est cela?</e>
437         <e>Du magnétisme, de l'électricité, de l'aimant?</e>
438         <e>Je ne sais pas,</e>
439         <div type="argumentatif">
440           <div type="descriptif">
441             <e>mais c'est horrible.</e><cr/>
442           </div>
443         </div>
444         <div type="explicatif">
445           <e>Et comprends-tu</e>
446           <e>pourquoi c'est horrible?</e>
447           <e>Quand je suis seul,</e>
448           <e>aussitôt que je suis seul,</e>
449           <e>je ne puis m'empêcher d'attirer tout</e>
450           <e>ce qui m'entoure.</e><cr/>
451           <e>Et je passe des jours entiers à changer des choses de
                place, ne me lassant jamais d'essayer ce pouvoir
              abominable,</e>
```

```
452           <e>comme pour voir</e>
453           <e>s'il ne m'a pas quitté.</e><cr/>
454         </div>
455       </div>
456       <e>Il avait enfoui ses grandes mains dans ses poches</e>
457       <e>et il regardait dans la nuit.</e>
458       <e>Un petit bruit, un frémissement léger semblait passer
             dans les arbres.</e><cr/>
459       <e>C'était la pluie qui commençait à tomber.</e><cr/>
460       <e>Je murmurai:</e>
461       <div type="dialogal">
462         <div type="descriptif">
463           <e>"C'est effrayant!"</e><cr/>
464         </div>
465       </div>
466       <e>Il répéta:</e>
467       <div type="dialogal">
468         <div type="descriptif">
469           <e>"C'est horrible."</e><cr/>
470         </div>
471       </div>
472       <div type="descriptif">
473         <e>Une rumeur accourut dans ce feuillage, comme un coup de
               vent.</e>
474         <e>C'était l'averse, l'ondée épaisse, torrentielle.</e><cr
               />
475       </div>
476       <e>Jacques se mit à respirer par grands souffles</e>
477       <e>qui soulevaient sa poitrine.</e><cr/>
478       <div type="dialogal">
479         <div type="injonctif">
480           <e>- Laisse-moi,</e>
481         </div>
482       </div>
483       <e>dit-il,</e>
484       <div type="dialogal">
485         <e>la pluie va me calmer.</e>
486         <e>Je désire être seul à présent.</e><cr/>
487       </div>
488     </div>
489   </div>
490 </text>
```

A.4 Un Fou

```
1 <?xml version="1.0" encoding="ISO-8859-1" ?>
2 <text source="http://un2sg4.unige.ch/athena/selva/maupassant/
    textes/unfou.html" date="2011.04.26">
3   <title>UN FOU</title>
4   <div type="narratif">
5     <div type="descriptif">
```

```
6       <e>Il était mort chef d'un haut tribunal, magistrat intègre
           </e>
7       <e> dont la vie irréprochable était citée dans toutes les
           cours de France.</e>
8       <e> Les avocats, les jeunes conseillers, les juges saluaient
            en s'inclinant très bas, par marque d'un profond respect
           , sa grande figure blanche et maigre</e>
9       <e> qu'éclairaient deux yeux brillants et profonds.</e><cr/>
10      <div type="argumentatif">
11        <e>Il avait passé sa vie à poursuivre le crime et à proté
             ger les faibles.</e>
12        <e> Les escrocs et les meurtriers n'avaient point eu d'
             ennemi plus redoutable,</e>
13        <e> car il semblait lire, au fond de leurs âmes, leurs
             pensées secrètes, et démêler, d'un coup d'oeil, tous
             les mystères de leurs intentions.</e><cr/>
14      </div>
15      <e>Il était donc mort, à l'âge de quatre-vingt-deux ans,
           entouré d'hommages et poursuivi par les regrets de tout
           un peuple.</e>
16      <e> Des soldats en culotte rouge l'avaient escorté jusqu'à
           sa tombe,</e>
17      <e> et des hommes en cravate blanche avaient répandu sur son
            cercueil des paroles désolées et des larmes</e>
18      <e> qui semblaient vraies.</e><cr/>
19    </div>
20    <e>Or, voici l'étrange papier que le notaire, éperdu, dé
         couvrit dans le secrétaire</e>
21    <e> où il avait coutume de serrer les dossiers des grands
         criminels.</e><cr/>
22    <e>Cela portait pour titre :</e><cr/><cr/>
23    <div type="explicatif">
24      <e>POURQUOI ?</e><cr/><cr/>
25      <div type="date">
26        <e>20 juin 1851.</e>
27      </div>
28      <e> - Je sors de la séance ?</e>
29      <e> J'ai fait condamner Blondel à mort !</e>
30      <e> Pourquoi donc cet homme avait-il tué ses cinq enfants ?<
           /e>
31      <e> Pourquoi ?</e>
32      <div type="argumentatif">
33        <e> Souvent, on rencontre de ces gens</e>
34        <e> chez qui détruire la vie est une volupté.</e>
35        <e> Oui, oui, ce doit être une volupté, la plus grande de
             toutes peut-être ;</e>
36        <e> car tuer n'est-il pas ce qui ressemble le plus à créer
             ?</e>
37        <e> Faire et détruire !</e>
38        <e> Ces deux mots enferment l'histoire des univers, toute
             l'histoire des mondes, tout ce qui est, tout !</e>
```

```
39      </div>
40      <e> Pourquoi est-ce enivrant de tuer ?</e><cr/>
41    </div>
42    <div type="date">
43      <e>25 juin.</e>
44    </div>
45    <e> - Songer qu'un être est là qui vit,</e>
46    <e> qui marche,</e>
47    <e> qui court...</e>
48    <e> Un être ?</e>
49    <e> Qu'est-ce qu'un être ?</e>
50    <e> Cette chose animée,</e>
51    <e> qui porte en elle le principe du mouvement et une volonté
        réglant ce mouvement !</e>
52    <e> Elle ne tient à rien cette chose.</e>
53    <e> Ses pieds ne communiquent pas au sol.</e>
54    <e> C'est un grain de vie</e>
55    <e> qui remue sur la terre ;</e>
56    <e> et ce grain de vie, venu je ne sais d'où, on peut le dé
        truire comme on veut.</e>
57    <e> Alors rien, plus rien.</e>
58    <e> Ça pourrit,</e>
59    <e> c'est fini.</e><cr/>
60    <div type="explicatif">
61      <div type="date">
62        <e>26 juin.</e>
63      </div>
64      <e> - Pourquoi donc est-ce un crime de tuer ?</e>
65      <e> oui, pourquoi ?</e>
66      <e> C'est, au contraire, la loi de la nature.</e>
67      <e> Tout être a pour mission de tuer :</e>
68      <e> il tue pour vivre</e>
69      <e> et il tue pour tuer.</e><cr/>
70    </div>
71    <div type="argumentatif">
72      <e>- Tuer est dans notre tempérament ;</e>
73      <e> il faut tuer !</e>
74      <e> La bête tue sans cesse, tout le jour, à tout instant de
          son existence.</e>
75      <e> - L'homme tue sans cesse pour se nourrir,</e>
76      <e> mais comme il a besoin de tuer aussi, par volupté,</e>
77      <e> il a inventé la chasse !</e>
78      <e> L'enfant tue les insectes</e>
79      <e> qu'il trouve,</e>
80      <e> les petits oiseaux, tous les petits animaux</e>
81      <e> qui lui tombent sous la main.</e>
82      <e> Mais cela ne suffisait pas à l'irrésistible besoin de
          massacre</e>
83      <e> qui est en nous.</e>
84      <e> Ce n'est point assez de tuer la bête ;</e>
85      <e> nous avons besoin aussi de tuer l'homme.</e>
```

```
86      <e> Autrefois, on satisfaisait ce besoin par des sacrifices
           humains.</e>
87      <e> Aujourd'hui la nécessité de vivre en société a fait du
           meurtre un crime.</e>
88      <e> On condamne</e>
89      <e> et on punit l'assassin !</e>
90      <e> Mais comme nous ne pouvons vivre</e>
91      <e> sans nous livrer à cet instinct naturel et impérieux de
           mort,</e>
92      <e> nous nous soulageons de temps en temps, par des guerres<
           /e>
93      <e> où un peuple entier égorge un autre peuple.</e>
94      <e> C'est alors une débauche de sang, une débauche</e>
95      <e> où s'affolent les armées</e>
96      <e> et dont se grisent encore les bourgeois, les femmes et
           les enfants</e>
97      <e> qui lisent, le soir, sous la lampe, le récit exalté des
           massacres.</e><cr/>
98      <e>Et on pourrait croire</e>
99      <e> qu'on méprise ceux destinés à accomplir ces boucheries d
           'hommes !</e>
100     <e> Non.</e>
101     <e> On les accable d'honneurs !</e>
102     <e> On les habille avec de l'or et des draps éclatants ;</e>
103     <e> ils portent des plumes sur la tête, des ornements sur la
            poitrine ;</e>
104     <e> et on leur donne des croix, des récompenses, des titres
           de toute nature.</e>
105     <e> Ils sont fiers, respectés, aimés des femmes, acclamés
           par la foule,</e>
106     <e> uniquement parce qu'ils ont pour mission de répandre le
           sang humain !</e>
107     <e> Ils traînent par les rues leurs instruments de mort</e>
108     <e> que le passant vêtu de noir regarde avec envie.</e>
109     <e> Car tuer est la grande loi jetée par la nature au coeur
           de l'être !</e>
110     <e> Il n'est rien de plus beau et de plus honorable que de
           tuer !</e><cr/>
111   </div>
112   <div type="explicatif">
113     <div type="date">
114       <e>30 juin.</e>
115     </div>
116     <e> - Tuer est la loi ;</e>
117     <e> parce que la nature aime l'éternelle jeunesse.</e>
118     <e> Elle semble crier par tous ses actes inconscients :</e>
119     <div type="dialogal">
120       <e> "Vite ! vite ! vite !"</e>
121     </div>
122     <e> Plus elle détruit,</e>
123     <e> plus elle se renouvelle.</e><cr/>
```

```
124     </div>
125     <div type="argumentatif">
126       <div type="date">
127         <e>2 juillet.</e>
128       </div>
129       <e> - L'être - qu'est-ce que l'être ?</e>
130       <e> Tout et rien.</e>
131       <e> Par la pensée, il est le reflet de tout.</e>
132       <e> Par la mémoire et la science, il est un abrégé du monde
            ,</e>
133       <e> dont il porte l'histoire en lui.</e>
134       <e> Miroir des choses et miroir des faits, chaque être
            humain devient un petit univers dans l'univers !</e><cr/>
135       <div type="injonctif">
136         <e>Mais voyagez ;</e>
137         <e> regardez grouiller les races,</e>
138         <e> et l'homme n'est plus rien ! plus rien, rien !</e>
139         <e> Montez en barque,</e>
140         <e> éloignez-vous du rivage couvert de foule,</e>
141         <e> et vous n'apercevrez bientôt plus rien que la côte.</e
              >
142         <e> L'être imperceptible disparaît,</e>
143         <e> tant il est petit, insignifiant.</e>
144         <e> Traverser l'Europe dans un train rapide,</e>
145         <e> et regardez par la portière.</e>
146         <e> Des hommes, des hommes, toujours des hommes,
              innombrables, inconnus,</e>
147         <e> qui grouillent dans les champs,</e>
148         <e> qui grouillent dans les rues ;</e>
149         <e> des paysans stupides sachant tout juste retourner la
              terre ;</e>
150         <e> des femmes hideuses sachant tout juste faire la soupe
              du mâle et enfanter.</e>
151         <e> Allez aux Indes,</e>
152         <e> allez en Chine,</e>
153         <e> et vous verrez encore s'agiter des milliards d'êtres
              qui naissent,</e>
154         <e> vivent</e>
155         <e> et meurent</e>
156         <e> sans laisser plus de trace</e>
157         <e> que la fourmi écrasée sur les routes.</e>
158         <e> Allez au pays des noirs, gîtés en des cases de boue ;
              au pays des Arabes blancs, abrités sous une toile brune
              </e>
159         <e> qui flotte au vent,</e>
160         <e> et vous comprendrez</e>
161         <e> que l'être isolé, déterminé, n'est rien, rien.</e>
162         <e> La race est tout !</e>
163         <e> Qu'est-ce que l'être, l'être quelconque d'une tribu
              errante du désert ?</e>
164         <e> Et ces gens,</e>
```

```
165       <e> qui sont des sages</e>
166       <e>, ne s'inquiètent pas de la mort.</e>
167       <e> L'homme ne compte point chez eux.</e>
168       <e> On tue son ennemi :</e>
169       <e> c'est la guerre.</e>
170       <e> Cela se faisait ainsi jadis, de manoir à manoir, de
              province à province.</e><cr/>
171       <div type="explicatif">
172         <e>Oui, traversez le monde</e>
173         <e> et regardez grouiller les humains innombrables et
              inconnus.</e>
174         <e> Inconnus ?</e>
175         <e> Ah ! voilà le mot du problème !</e>
176         <e> Tuer est un crime</e>
177         <e> parce que nous avons numéroté les êtres !</e>
178       </div>
179       <e> Quand ils naissent,</e>
180       <e> on les inscrit,</e>
181       <e> on les nomme,</e>
182       <e> on les baptise.</e>
183       <e> La loi les prend !</e>
184       <e> Voilà !</e>
185       <e> L'être qui n'est point enregistré ne compte pas :</e>
186       <e> tuez-le dans la lande ou dans le désert,</e>
187       <e> tuez-le dans la montagne ou dans la plaine, qu'importe
              !</e>
188       <e> La nature aime la mort ;</e>
189       <e> elle ne punit pas, elle !</e><cr/>
190     </div>
191     <div type="explicatif">
192       <e>Ce qui est sacré, par exemple,</e>
193       <e> c'est l'état civil !</e>
194       <e> Voilà !</e>
195       <e> C'est lui qui défend l'homme.</e>
196       <e> L'être est sacré</e>
197       <e> parce qu'il est inscrit à l'état civil !</e>
198       <e> Respect à l'état civil, le Dieu légal.</e>
199       <e> A genoux !</e><cr/>
200       <e>L'état peut tuer, lui,</e>
201       <e> parce qu'il a le droit de modifier l'état civil.</e>
202       <e> Quand il a fait égorger deux cent mille hommes dans
              une guerre,</e>
203       <e> il les raye sur son état civil,</e>
204       <e> il les supprime par la main de ses greffiers.</e>
205       <e> C'est fini.</e>
206       <e> Mais nous, qui ne pouvons point changer les écritures
              des mairies,</e>
207       <e> nous devons respecter la vie.</e>
208       <e> État civil, glorieuse Divinité</e>
209       <e> qui règnes dans les temples des municipalités,</e>
210       <e> je te salue.</e>
```

```
211         <e> Tu es plus fort que la Nature.</e>
212         <e> Ah ! Ah !</e><cr/>
213       </div>
214     </div>
215     <div type="descriptif">
216       <div type="date">
217         <e>3 juillet.</e>
218       </div>
219       <e> - Ce doit être un étrange et savoureux plaisir que de
             tuer,</e>
220       <e> d'avoir là, devant soi, l'être vivant, pensant ;</e>
221       <e> de faire devant un petit trou, rien qu'un petit trou,</e
             >
222       <e> de voir couler cette chose rouge</e>
223       <e> qui est le sang,</e>
224       <e> qui fait la vie,</e>
225       <e> et de n'avoir plus devant soi, qu'un tas de chair molle,
              froide, inerte, vide de pensée !</e><cr/>
226     </div>
227     <div type="date">
228       <e>5 août.</e>
229     </div>
230     <e> - Moi qui ai passé mon existence à juger, à condamner, à
           tuer par des paroles prononcées, à tuer par la guillotine
           ceux</e>
231     <e> qui avaient tué par le couteau, moi ! moi !</e>
232     <e> si je faisais comme tous les assassins</e>
233     <e> que j'ai frappés, moi ! moi !</e>
234     <e> qui le saurait ?</e><cr/>
235     <div type="date">
236        <e>10 août.</e>
237     </div>
238     <e> - Qui le saurait jamais</e>
239     <e> Me soupçonnerait-on, moi, moi,</e>
240     <e> surtout si je choisis un être</e>
241     <e> que je n'ai aucun intérêt à supprimer ?</e><cr/>
242     <div type="descriptif">
243       <div type="date">
244         <e>15 août.</e>
245       </div>
246       <e> - La tentation !</e>
247       <e> La tentation, elle est entrée en moi comme un ver</e>
248       <e> qui rampe.</e>
249       <e> Elle rampe,</e>
250       <e> elle va ;</e>
251       <e> elle se promène dans mon corps entier, dans mon esprit
             ,</e>
252       <e> qui ne pense plus qu'à ceci :</e>
253       <e> tuer ;</e>
254       <e> dans mes yeux,</e>
255       <e> qui ont besoin de regarder du sang,</e>
```

```
256       <e> de voir mourir ;</e>
257       <e> dans mes oreilles,</e>
258       <e> où passe sans cesse quelque chose d'inconnu, d'horrible,
              de déchirant et d'affolant, comme le dernier cri d'un ê
             tre ;</e>
259       <e> dans mes jambes,</e>
260       <e> où frissonne le désir d'aller, d'aller à l'endroit</e>
261       <e> où la chose aura lieu ;</e>
262       <e> dans mes mains</e>
263       <e> qui frémissent du besoin de tuer.</e>
264       <e> Comme cela doit être bon, rare, digne d'un homme libre,
             au-dessus des autres, maître de son coeur</e>
265       <e> et qui cherche des sensations raffinées !</e><cr/>
266     </div>
267     <div type="narratif">
268       <div type="date">
269         <e>22 août.</e>
270       </div>
271       <e> - Je ne pouvais plus résister.</e>
272       <e> J'ai tué une petite bête pour essayer, pour commencer.</
             e><cr/>
273       <e>Jean, mon domestique, avait un chardonneret dans une cage
              suspendue à la fenêtre de l'office.</e>
274       <e> Je l'ai envoyé faire une course,</e>
275       <e> et j'ai pris le petit oiseau dans ma main, dans ma main<
             /e>
276       <e> où je sentais battre son coeur.</e>
277       <e> Il avait chaud.</e>
278       <e> Je suis monté dans ma chambre.</e>
279       <e> De temps en temps, je le serrais plus fort ;</e>
280       <e> son coeur battait plus vite ;</e>
281       <e> c'était atroce et délicieux.</e>
282       <div type="argumentatif">
283         <e> J'ai failli l'étouffer.</e>
284         <e> Mais je n'aurais pas vu le sang.</e><cr/>
285         <e>Alors j'ai pris des ciseaux, de courts ciseaux à ongles
               ,</e>
286         <e> et je lui ai coupé la gorge en trois coups, tout
               doucement.</e>
287       </div>
288       <e> Il ouvrait le bec,</e>
289       <e> il s'efforçait de m'échapper,</e>
290       <e> mais je le tenais,</e>
291       <e> oh ! je le tenais ;</e>
292       <e> j'aurais tenu un dogue enragé</e>
293       <e> et j'ai vu le sang couler.</e>
294       <e> Comme c'est beau, rouge, luisant, clair, du sang !</e>
295       <e> J'avais envie de le boire.</e>
296       <e> J'y ai trempé le bout de ma langue !</e>
297       <e> C'est bon.</e>
298       <e> Mais il en avait si peu,</e>
```

```
299       <e> ce pauvre petit oiseau !</e>
300       <e> Je n'ai pas eu le temps de jouir de cette vue</e>
301       <e> comme j'aurais voulu.</e>
302       <e> Ce doit être superbe de voir saigner un taureau.</e><cr
             />
303       <e>Et puis j'ai fait comme les assassins, comme les vrais.</
             e>
304       <e> J'ai lavé les ciseaux,</e>
305       <e> je me suis lavé les mains ;</e>
306       <e> j'ai jeté l'eau</e>
307       <e> et j'ai porté le corps, le cadavre, dans le jardin pour
             l'enterrer.</e>
308       <e> Je l'ai enfoui sous un fraisier.</e>
309       <e> On ne le trouvera jamais.</e>
310       <e> Je mangerai tous les jours une fraise à cette plante.</e
             >
311       <e> Vraiment, comme on peut jouir de la vie,</e>
312       <e> quand on sait !</e><cr/>
313       <e>Mon domestique a pleuré ;</e>
314       <e> il croit son oiseau parti.</e>
315       <e> Comment me soupçonnerait-il ?</e>
316       <e> Ah ! Ah !</e><cr/>
317     </div>
318     <div type="date">
319       <e>25 août.</e>
320     </div>
321     <e> - Il faut que je tue un homme !</e>
322     <e> Il le faut.</e><cr/>
323     <div type="date">
324       <e>30 août.</e>
325     </div>
326     <e> - C'est fait.</e>
327     <e> Comme c'est peu de chose !</e><cr/>
328     <div type="narratif">
329       <div type="descriptif">
330         <e>J'étais allé me promener dans le bois de Vernes.</e>
331         <e> Je ne pensais à rien, non, à rien.</e>
332         <e> Voilà un enfant dans le chemin, un petit garçon</e>
333         <e> qui mangeait une tartine de beurre.</e><cr/>
334       </div>
335       <e>Il s'arrête pour me voir passer</e>
336       <e> et dit :</e><cr/>
337       <div type="dialogal">
338         <e>- Bonjour, m'sieu le président.</e><cr/>
339       </div>
340       <e>Et la pensée m'entre dans la tête :</e>
341       <div type="dialogal">
342         <e> "Si je le tuais ?"</e><cr/>
343       </div>
344       <e>Je réponds :</e><cr/>
345       <div type="dialogal">
```

```
346         <e>- Tu es tout seul, mon garçon ?</e><cr/>
347         <e>- Oui, m'sieu.</e><cr/>
348         <e>- Tout seul dans le bois ?</e><cr/>
349         <e>- Oui, msieur.</e><cr/>
350       </div>
351       <div type="descriptif">
352         <e>L'envie de le tuer me grisait comme de l'alcool.</e>
353       </div>
354       <e> Je m'approchai tout doucement, persuadé</e>
355       <e> qu'il allait s'enfuir.</e>
356       <e> Et voilà que je le saisis à la gorge...</e>
357       <e> Je le serre,</e>
358       <e> je le serre de toute ma force !</e>
359       <e> Il m'a regardé avec des yeux effrayants !</e>
360       <div type="descriptif">
361         <e> Quels yeux !</e>
362         <e> Tout ronds, profonds, limpides, terribles !</e>
363         <e> Je n'ai jamais éprouvé une émotion si brutale...</e>
364         <e> mais si courte !</e>
365         <e> Il tenait mes poignets dans ses petites mains,</e>
366         <e> et son corps se tordait ainsi qu'une plume sur le feu
              .</e>
367         <e> Puis il n'a plus remué.</e><cr/>
368       </div>
369       <e>Mon coeur battait,</e>
370       <e> ah ! le coeur de l'oiseau !</e>
371       <e> J'ai jeté le corps dans le fossé, puis de l'herbe par-
            dessus.</e><cr/>
372       <e>Je suis rentré,</e>
373       <e> j'ai bien dîné.</e>
374       <e> Comme c'est peu de chose !</e>
375       <e> Le soir, j'étais très gai, léger, rajeuni,</e>
376       <e> j'ai passé la soirée chez le préfet.</e>
377       <e> On m'a trouvé spirituel.</e><cr/>
378       <div type="argumentatif">
379         <e>Mais je n'ai pas vu le sang !</e>
380         <e> Je suis tranquille.</e><cr/>
381       </div>
382       <div type="date">
383        <e>30 août.</e>
384       </div>
385       <e> - On a découvert le cadavre.</e>
386       <e> On cherche l'assassin.</e>
387       <e> Ah ! ah !</e><cr/>
388       <div type="date">
389         <e>1er septembre.</e>
390       </div>
391       <e> - On a arrêté deux rôdeurs.</e>
392       <e> Les preuves manquent.</e><cr/>
393       <div type="date">
394         <e>2 septembre.</e>
```

```
395       </div>
396       <e> - Les parents sont venus me voir.</e>
397       <e> Ils ont pleuré !</e>
398       <e> Ah ! ah !</e><cr/>
399       <div type="date">
400         <e>6 octobre.</e>
401       </div>
402       <e> - On n'a rien découvert.</e>
403       <e> Quelque vagabond errant aura fait le coup.</e>
404       <e> Ah ! ah !</e>
405       <div type="argumentatif">
406         <e> Si j'avais vu le sang couler,</e>
407         <e> il me semble</e>
408         <e> que je serais tranquille à présent !</e><cr/>
409       </div>
410     </div>
411     <div type="date">
412       <e>10 octobre.</e>
413     </div>
414     <e> - L'envie de tuer me court dans les moelles.</e>
415     <e> Cela est comparable aux rages d'amour</e>
416     <e> qui vous torturent à vingt ans.</e><cr/>
417     <div type="narratif">
418       <div type="date">
419         <e>20 octobre.</e>
420       </div>
421       <e> - Encore un.</e>
422       <e> J'allais le long du fleuve, après déjeuner.</e>
423       <e> Et j'aperçus, sous un saule, un pêcheur endormi.</e>
424       <e> Il était midi.</e>
425       <div type="descriptif">
426         <e> Une bêche semblait, tout exprès, plantée dans un champ
              de pommes de terre voisin.</e><cr/>
427       </div>
428       <e>Je la pris,</e>
429       <e> je revins ;</e>
430       <e> je la levai comme une massue</e>
431       <e> et, d'un seul coup, par le tranchant, je fendis la tête
             du pêcheur.</e>
432       <e> Oh ! il a saigné, celui-là !</e>
433       <div type="descriptif">
434         <e> Du sang rose, plein de cervelle !</e>
435         <e> Cela coulait dans l'eau, tout doucement.</e>
436       </div>
437       <e> Et je suis parti d'un pas grave.</e>
438       <div type="argumentatif">
439         <e> Si on m'avait vu !</e>
440         <e> Ah ! ah ! j'aurais fait un excellent assassin.</e><cr
              />
441       </div>
442       <div type="date">
```

```
443         <e>25 octobre.</e>
444       </div>
445       <e> - L'affaire du pêcheur soulève un grand bruit.</e>
446       <e> On accuse du meurtre son neveu,</e>
447       <e> qui pêchait avec lui.</e><cr/>
448       <div type="date">
449         <e>26 octobre.</e>
450       </div>
451       <e> - Le juge d'instruction affirme</e>
452       <e> que le neveu est coupable.</e>
453       <e> Tout le monde le croit par la ville.</e>
454       <e> Ah ! ah !</e><cr/>
455       <div type="date">
456         <e>27 octobre.</e>
457       </div>
458       <e> - Le neveu se défend bien mal.</e>
459       <e> Il était parti au village acheter du pain et du fromage
              ,</e>
460       <e> affirme-t-il.</e>
461       <e> Il jure</e>
462       <e> qu'on a tué son oncle pendant son absence !</e>
463       <e> Qui le croirait ?</e><cr/>
464       <div type="date">
465         <e>28 octobre.</e>
466       </div>
467       <div type="argumentatif">
468         <e> - Le neveu a failli avouer,</e>
469         <e> tant on lui fait perdre la tête !</e>
470       </div>
471       <e> Ah ! ah !</e>
472       <e> La justice !</e><cr/>
473       <div type="date">
474         <e>15 novembre.</e>
475       </div>
476       <e> - On a des preuves accablantes contre le neveu,</e>
477       <e> qui devait hériter de son oncle.</e>
478       <e> Je présiderai les assises.</e><cr/>
479       <div type="date">
480         <e>15 janvier.</e>
481       </div>
482       <e> - A mort ! à mort ! à mort !</e>
483       <e> Je l'ai fait condamner à mort !</e>
484       <e> Ah ! ah !</e>
485       <e> L'avocat général a parlé comme un ange !</e>
486       <e> Ah ! ah !</e>
487       <e> Encore un.</e>
488       <e> J'irai le voir exécuter.</e><cr/>
489       <div type="date">
490         <e>10 mars.</e>
491       </div>
492       <e> - C'est fini.</e>
```

```
493       <e> On l'a guillotiné ce matin.</e>
494       <e> Il est très bien mort !</e>
495       <e> très bien !</e>
496       <e> Cela m'a fait plaisir !</e>
497       <e> Comme c'est beau de voir trancher la tête d'un homme !</
              e>
498       <e> Le sang a jailli comme un flot, comme un flot !</e>
499       <e> Oh ! si j'avais pu,</e>
500       <e> j'aurais voulu me baigner dedans.</e>
501       <e> Quelle ivresse de me coucher là-dessous,</e>
502       <e> de recevoir cela dans mes cheveux et sur mon visage,</e>
503       <e> et de me relever tout rouge, tout rouge !</e>
504       <e> Ah ! si on savait !</e><cr/>
505       <e>Maintenant j'attendrai,</e>
506       <e> je puis attendre.</e>
507       <e> Il faudrait si peu de chose</e>
508       <e> pour me laisser surprendre.</e><cr/>
509     </div>
510     <div type="descriptif">
511       <e>Le manuscrit contenait encore beaucoup de pages,</e>
512       <e> mais sans relater aucun crime nouveau.</e><cr/>
513     </div>
514     <e>Les médecins aliénistes,</e>
515     <e> à qui on l'a confié,</e>
516     <e> affirment</e>
517     <e> qu'il existe dans le monde beaucoup de fous ignorés, aussi
            adroits et aussi redoutables que ce monstrueux dément.</e
           ><cr/>
518   </div>
519 </text>
```

ANNEXE B

Liens entre types de discours et CMS

La section B.1 présente, pour les quatre textes de Maupassant exposés dans l'annexe A réunis et pour chacun de ces textes considéré séparément, le nombre de CMS contenu dans chacun des types de discours. La section B.2 montre, pour les mêmes textes, les valeurs obtenues pour le khi2 ponctuel (1.3) et celles qui sont significatives au niveau $\alpha = 0.1\%$. Pour rappel, section 1.2.2.1, $\chi^2_{1-0.001}[1] = 10.83$. Ces résultats sont utilisés dans la section 4.1.4 du chapitre 4.

Les définitions de toutes les abréviations de CMS utilisées par TreeTagger pour le français sont les suivantes [1] :

ABR	Abréviation
ADJ	Adjectif
ADV	Adverbe
DET:ART	Déterminant article
DET:POS	Déterminant possessif (mon, ma, ton, ta,...)
INT	Interjection
KON	Conjonction
NAM	Nom propre
NOM	Nom
NUM	Numéral
PRO	Pronom (un exemple est présenté dans la figure 4.2)
PRO:DEM	Pronom démonstratif
PRO:IND	Pronom indéfini
PRO:PER	Pronom personnel
PRO:POS	Pronom possessif (mien, tien,...)
PRO:REL	Pronom relatif
PRP	Préposition
PRP:det	Préposition et article (au,du,aux,des)
PUN	Ponctuation
PUN:cit	Ponctuation de citation

1. Il s'agit de traductions des définitions disponibles sur le site : `http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/french-tagset.html`.

SENT Balise de phrase
VER:cond Verbe au conditionnel
VER:futu Verbe au futur
VER:impe Verbe à l'impératif
VER:impf Verbe à l'imparfait
VER:infi Verbe à l'infinitif
VER:pper Verbe au participe passé
VER:ppre Verbe au participe présent
VER:pres Verbe au présent
VER:simp Verbe au passé simple
VER:subi Verbe à l'imparfait du subjonctif
VER:subp Verbe au présent du subjonctif

B.1 Tables des effectifs croisés

	nar	arg	expl	descr	dial	inj
ABR	1	0	0	0	0	0
ADJ	142	58	57	106	58	22
ADV	193	89	73	65	71	42
DET:ART	284	129	68	129	84	37
DET:POS	70	21	17	19	15	6
INT	22	8	7	0	10	3
KON	148	75	63	49	48	18
NAM	37	4	7	18	7	5
NOM	515	239	147	243	161	83
NUM	18	7	5	10	4	0
PRO	2	0	0	0	0	0
PRO:DEM	45	40	33	28	27	6
PRO:IND	20	16	13	11	10	0
PRO:PER	475	127	126	89	129	36
PRO:REL	51	28	26	30	22	10
PRP	323	123	99	142	91	37
PRP:det	42	30	14	33	33	18
PUN	369	131	91	155	101	67
PUN:cit	0	3	4	7	60	24
SENT	232	76	77	73	85	39
VER:cond	12	8	3	1	4	0
VER:futu	6	2	2	2	21	3
VER:impe	0	0	0	0	0	1
VER:impf	83	13	9	60	8	1
VER:infi	106	39	51	39	34	8
VER:pper	113	27	28	39	25	7
VER:ppre	34	2	4	7	5	2
VER:pres	144	97	96	58	82	64
VER:simp	169	7	4	8	1	0
VER:subi	3	0	8	1	0	0
VER:subp	1	2	2	0	4	1

TABLE B.1 – Effectifs des CMS pour chaque type de discours pour les quatre textes de Maupassant réunis.

	« L'Orient »						« Le Voleur »						« Un Fou ? »						« Un Fou »					
	nar	arg	expl	descr	dial	inj	nar	arg	expl	descr	dial	inj	nar	arg	expl	descr	dial	inj	nar	arg	expl	descr	dial	inj
ABR																			1	0	0	0	0	0
ADJ	23	2	21	21	42	4	41	4	1	13	7	3	43	29	21	27	6	1	35	23	14	45	3	14
ADV	32	5	19	12	24	2	51	10	3	17	16	2	46	38	36	11	26	7	64	36	15	25	5	31
DET:ART	27	9	26	35	50	0	107	8	3	29	16	2	64	48	12	26	16	6	86	64	27	39	2	29
DET:POS	3	3	4	2	10	3	19	2	4	2	4	0	29	9	5	2	0	2	19	7	4	13	1	1
INT	0	0	0	0	3	0	0	0	0	0	2	0	1	3	2	0	3	3	21	5	5	0	2	0
KON	17	5	12	7	27	2	46	5	10	12	7	0	43	32	30	10	13	1	42	33	11	20	1	15
NAM	2	0	4	3	3	0	25	2	2	11	2	0	9	2	0	1	2	2	1	0	1	3	0	3
NOM	51	19	49	73	98	5	175	12	12	37	35	5	131	86	40	45	23	16	158	122	46	88	5	57
NUM	5	2	1	4	4	0	5	0	0	2	0	0	5	3	0	2	0	0	3	2	4	2	0	0
PRO																			2	0	0	0	0	0
PRO:DEM	4	0	9	10	6	0	8	2	0	5	8	0	5	25	16	8	13	2	28	13	8	5	0	4
PRO:IND	1	0	5	3	9	0	6	2	1	1	0	0	7	4	5	2	1	0	6	10	2	5	0	0
PRO:PER	62	8	27	20	44	2	134	11	14	20	31	2	135	51	61	28	49	9	144	57	24	21	5	23
PRO:REL	4	0	10	10	12	0	17	3	1	3	2	0	11	11	8	2	8	1	19	14	7	15	0	9
PRP	32	10	35	28	56	3	113	6	12	26	13	4	98	42	32	23	21	5	80	65	20	65	1	25
PRP:det	5	3	6	16	23	0	13	0	1	4	7	1	9	8	3	4	3	0	15	19	4	9	0	17
PUN	33	8	23	39	49	5	135	5	11	28	13	3	103	64	35	30	30	19	98	54	22	58	9	40
PUN:cit	0	2	0	2	15	2	0	1	4	1	35	8	0	0	0	4	6	14	0	0	0	0	4	0
SENT	22	4	19	18	22	3	51	4	4	14	25	4	39	27	23	19	29	13	120	41	31	22	9	19
VER:cond	1	0	2	0	0	0	1	0	0	1	1	0	0	4	1	0	2	0	10	4	0	0	1	0
VER:futu	0	1	1	1	18	0	0	1	0	0	2	0	0	0	1	0	1	0	6	0	0	1	0	3
VER:impe							0	0	0	0	0	1												
VER:impf	2	0	1	2	1	0	21	2	3	20	2	0	33	4	4	21	4	0	27	7	1	17	1	1
VER:infi	3	2	17	5	8	0	23	3	4	8	12	0	39	5	15	7	14	0	41	29	15	19	0	8
VER:pper	7	2	5	5	13	0	32	1	5	11	4	0	14	1	10	10	8	0	60	23	8	13	0	7
VER:ppre	0	0	2	1	4	0	16	0	0	1	0	0	16	1	2	1	1	0	2	1	0	4	0	2
VER:pres	28	6	27	32	23	6	7	4	4	4	22	5	22	38	33	5	36	17	87	49	32	17	1	36
VER:simp	17	0	0	0	0	0	96	3	4	4	0	0	50	4	0	4	1	0	6	0	0	0	0	0
VER:subi							1	0	0	0	0	0	2	0	8	1	0	0						
VER:subp	0	0	0	0	2	0	0	1	0	0	0	0	0	0	0	0	2	0	1	1	2	0	0	1

TABLE B.2 – Effectifs des CMS pour chaque type de discours pour chacun des quatre textes de Maupassant, considéré séparément.

B.2 Khi2 ponctuel

	nar	arg	expl	descr	dial	inj
ABR	1.56	0.18	0.14	0.18	0.15	0.06
ADJ	9.74	1.29	0.24	27.50*	0.03	0.55
ADV	2.00	1.32	1.32	3.95	0.12	4.62
DET:ART	0.02	4.45	5.91	3.69	1.26	0.73
DET:POS	4.23	0.07	0.06	0.65	0.97	0.82
INT	0.50	0.04	0.17	9.01	2.31	0.00
KON	0.86	4.58	5.07	2.88	0.27	1.27
NAM	2.29	5.99	0.73	3.79	1.04	0.06
NOM	2.77	6.46	3.58	6.75	2.19	0.13
NUM	0.06	0.03	0.02	1.95	0.55	2.71
PRO	3.11	0.35	0.28	0.36	0.29	0.12
PRO:DEM	14.97*	7.79	6.84	0.03	0.83	1.96
PRO:IND	3.29	3.44	2.76	0.01	0.13	4.32
PRO:PER	39.47*	3.59	0.52	32.03*	0.10	8.94
PRO:REL	5.25	0.43	1.90	1.01	0.02	0.01
PRP	0.10	0.01	0.00	3.43	2.20	2.49
PRP:det	15.10*	0.97	2.45	2.39	6.72	7.39
PUN	0.67	0.33	4.45	2.44	2.85	4.53
PUN:cit	63.63*	11.04*	6.01	4.99	207.51*	63.81*
SENT	0.15	1.79	0.72	3.39	1.76	0.99
VER:cond	0.17	4.08	0.05	2.95	0.05	1.72
VER:futu	7.65	2.52	1.46	2.61	66.95*	0.44
VER:impe	0.64	0.18	0.14	0.18	0.15	16.33*
VER:impf	5.49	7.84	8.03	51.17*	10.73	8.80
VER:infi	0.09	0.18	10.61	0.28	0.08	4.36
VER:pper	6.87	2.60	0.04	0.24	1.23	3.64
VER:ppre	12.97*	5.42	1.13	0.21	0.62	0.43
VER:pres	37.66*	3.94	17.06*	8.93	2.79	38.76*
VER:simp	204.95*	19.24*	18.12*	17.99*	26.08*	11.81*
VER:subi	1.01	2.12	33.57*	0.44	1.77	0.74
VER:subp	3.56	0.20	0.58	1.79	6.61	0.33

TABLE B.3 – khi2 ponctuels entre les CMS et les types de discours pour les quatre textes de Maupassant réunis. Les valeurs significatives pour $\alpha = 0.1\%$ sont suivies d'une étoile.

	« L'Orient »						« Le Voleur »					
	nar	arg	expl	descr	dial	inj	nar	arg	expl	descr	dial	inj
ABR												
ADV	0.14	2.89	0.00	0.14	1.28	1.18	0.00	0.16	2.17	1.21	0.83	1.79
ADV	8.76	0.00	0.17	3.21	2.12	0.00	2.83	6.43	1.12	0.71	0.46	0.00
DET:ART	1.10	0.28	0.08	1.49	0.20	3.47	2.07	0.00	4.48	1.60	2.63	0.67
DET:POS	1.42	2.38	0.11	2.27	0.68	11.97*	0.04	0.19	3.52	1.58	0.02	0.67
INT	0.84	0.16	0.69	0.75	6.28	0.06	2.95	0.10	0.11	0.33	12.43*	0.04
KON	0.27	0.56	0.10	4.52	1.28	0.19	0.15	0.39	8.35	0.03	1.83	1.78
NAM	0.19	0.66	1.74	0.19	0.30	0.26	0.00	0.00	0.03	4.97	2.98	0.91
NOM	4.21	1.10	0.91	5.10	0.12	0.30	1.95	0.14	0.66	0.20	0.38	0.12
NUM	0.85	1.74	1.62	0.26	0.40	0.35	0.41	0.35	0.40	1.17	1.13	0.15
PRO												
PRO:DEM	1.11	1.62	3.02	3.90	1.84	0.64	5.95	0.77	1.32	1.06	8.52	0.50
PRO:IND	2.81	1.00	1.02	0.12	2.59	0.39	0.00	5.09	0.42	0.15	1.62	0.21
PRO:PER	27.87*	0.03	0.48	6.65	2.37	0.69	1.29	0.08	0.71	4.58	0.11	1.52
PRO:REL	2.46	2.02	2.05	1.41	0.02	0.79	0.37	2.62	0.12	0.16	0.84	0.56
PRP	0.55	0.29	0.91	0.94	0.26	0.07	2.27	0.76	0.88	0.07	6.56	0.04
PRP:det	4.89	0.02	1.90	3.58	3.04	1.18	1.01	1.33	0.12	0.03	3.76	0.40
PUN	0.06	0.00	1.76	2.58	0.10	0.95	8.37	2.37	0.03	0.00	9.43	0.32
PUN:cit	5.92	0.80	4.85	1.45	14.82*	5.63	74.16*	0.84	0.77	6.16	139.47*	49.94*
SENT	0.56	0.08	0.55	0.01	2.29	0.75	4.12	0.18	0.44	0.03	10.21	1.78
VER:cond	0.24	0.16	4.59	0.75	1.44	0.06	0.86	0.15	0.17	0.89	0.95	0.06
VER:futu	5.92	0.01	2.68	3.07	27.64*	0.46	4.43	5.36	0.17	0.50	7.01	0.06
VER:impe							1.48	0.05	0.06	0.17	0.16	46.97*
VER:impf	0.47	0.33	0.01	0.67	0.68	0.13	5.13	0.04	0.07	30.14*	3.88	1.05
VER:infi	3.66	0.02	21.23*	0.72	1.47	0.77	3.94	0.16	0.70	0.12	4.41	1.09
VER:pper	0.00	0.07	0.19	0.38	1.02	0.70	0.01	1.01	1.77	1.86	1.82	1.16
VER:ppre	1.96	0.39	0.46	0.14	1.97	0.15	8.49	0.86	0.97	0.99	2.76	0.37
VER:pres	0.10	0.02	1.09	3.23	10.93*	4.97	38.54*	1.57	1.03	1.20	45.50*	17.81*
VER:simp	61.64*	0.94	3.92	4.28	8.21	0.37	42.71*	0.99	0.59	10.30	18.25*	2.41
VER:subi							0.68	0.05	0.06	0.17	0.16	0.02
VER:subp	0.56	0.11	0.46	0.50	4.18	0.04	1.48	19.86*	0.06	0.17	0.16	0.02

	« Un Fou ? »						« Un Fou »					
	nar	arg	expl	descr	dial	inj	nar	arg	expl	descr	dial	inj
ABR							1.59	0.28	0.11	0.20	0.02	0.13
ADJ	0.36	0.43	0.14	13.72*	6.85	4.27	9.16	2.02	0.05	29.63*	0.32	0.09
ADV	5.20	0.75	5.86	3.50	2.30	0.02	0.38	0.31	0.39	0.72	1.70	7.56
DET:ART	0.06	6.13	9.94	2.90	1.37	0.43	1.59	2.20	0.33	0.10	1.13	0.06
DET:POS	13.30*	0.06	0.82	2.30	6.60	0.01	0.26	1.15	0.05	5.08	0.10	3.73
INT	4.09	0.15	0.02	1.51	1.88	11.81*	8.85	0.95	1.04	6.60	4.08	4.23
KON	0.53	1.52	6.52	1.59	0.53	4.37	0.92	1.77	0.11	0.00	0.52	0.14
NAM	2.76	0.64	2.92	0.39	0.00	2.40	2.30	2.28	0.06	2.56	0.13	5.53
NOM	0.73	5.28	3.96	1.64	10.61	0.04	6.86	3.95	0.03	1.60	1.18	0.29
NUM	0.81	0.55	1.82	0.79	1.38	0.47	0.59	0.10	8.69	0.02	0.18	1.40
PRO							3.19	0.57	0.22	0.40	0.03	0.25
PRO:DEM	25.93*	10.70	3.35	0.01	3.01	0.42	2.35	0.00	1.01	2.67	0.98	1.12
PRO:IND	0.00	0.00	1.77	0.01	0.84	0.90	1.52	6.11	0.04	0.46	0.38	2.94
PRO:PER	2.90	6.36	2.58	2.92	2.42	2.85	24.86*	0.32	0.43	17.08*	0.07	2.47
PRO:REL	1.63	1.01	0.55	1.66	2.14	0.41	2.17	0.00	0.08	2.28	1.08	0.52
PRP	6.68	0.35	0.14	0.14	1.55	2.80	6.31	1.70	1.35	15.99*	2.68	0.62
PRP:det	0.11	1.38	0.38	0.37	0.03	1.28	6.31	2.15	0.97	0.28	1.08	15.33*
PUN	0.01	0.97	2.03	0.07	0.61	3.76	1.78	1.55	1.47	3.83	4.76	2.75
PUN:cit	13.83*	6.26	4.39	0.74	3.78	163.53*	2.51	1.14	0.44	0.79	241.52*	0.51
SENT	7.36	0.63	0.00	0.36	7.79	6.45	13.47*	4.14	2.52	10.49	7.14	3.05
VER:cond	4.01	5.77	0.01	0.88	1.79	0.33	5.02	0.18	1.65	2.98	2.38	1.91
VER:futu	1.14	0.52	1.85	0.25	2.70	0.09	1.95	2.86	1.10	0.31	0.17	3.53
VER:impe												
VER:impf	5.46	8.69	4.50	29.13*	2.33	3.19	3.03	2.69	3.98	8.94	0.02	4.87
VER:infi	5.49	10.32	0.73	0.48	2.25	3.88	0.19	0.94	1.60	0.02	1.93	1.97
VER:pper	0.27	8.88	2.10	6.45	1.73	2.06	11.66*	0.14	0.93	1.92	1.91	2.82
VER:ppre	14.53*	3.23	0.55	0.87	1.07	1.00	1.02	0.64	0.99	5.11	0.15	1.09
VER:pres	32.83*	2.11	5.21	9.96	20.69*	17.07*	0.04	0.00	5.51	13.60*	2.08	5.89
VER:simp	61.12*	7.00	10.95*	1.17	6.15	2.84	9.58	1.71	0.66	1.19	0.10	0.76
VER:subi	1.57	2.85	27.98*	0.05	1.52	0.52						
VER:subp	1.14	0.52	0.36	0.25	14.52*	0.09	0.73	0.01	5.10	0.99	0.08	0.38

TABLE B.4 – khi2 ponctuels entre les CMS et les types de discours pour chaque texte de Maupassant considéré séparément. Les valeurs significatives pour $\alpha = 0.1\%$ sont suivies d'une étoile.

ANNEXE C

Classification non supervisée en types de discours

Les sections C.1 et C.2 présentent des résultats supplémentaires ou complémentaires à ceux de la section 4.3, concernant la classification non supervisée en types de discours des propositions des quatre contes de Maupassant, présentés dans l'annexe A.

C.1 K-means

Pour compléter les résultats présentés dans les figures 4.11 à 4.14 de la section 4.3.1.2, la section C.1.1 présente les mêmes résultats, obtenus avec l'algorithme K-means et les deux indices d'accord de partitions, Rand corrigé et Jaccard, mais en y ajoutant les écarts-types.

Dans la section C.1.2, les résultats, toujours pour la méthode K-means, sont présentés en utilisant le V de Cramer pour comparer les partitions, soit :

$$V := \sqrt{\frac{\text{khi2}}{\text{khi2}_{\max}}} = \sqrt{\frac{\text{khi2}}{n \min(m_1 - 1, m_2 - 1)}} \in [0, 1]$$

Le nombre de groupes à retrouver par l'algorithme étant imposé ($m = 6$) et le nombre de types de discours étant fixé aussi (à nouveau 6), le dénominateur du V de Cramer reste constant pour un texte et une longueur de n-gramme donnée, car dans ce cas le nombre de propositions, n, est constant aussi. Ainsi, le V de Cramer varie uniquement en fonction du khi2 pour chaque courbe.

C.1.1 Indices d'accord entre partitions

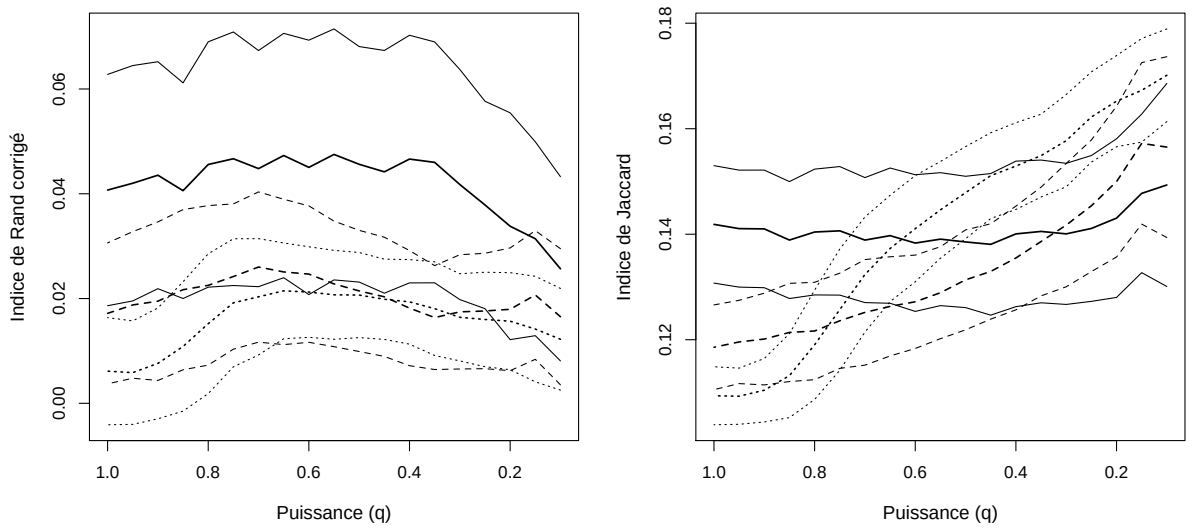


FIGURE C.1 – « L'Orient » avec l'algorithme K-means. Moyennes (lignes épaisses) et écarts-types (lignes fines) pour l'indice de Rand corrigé (gauche) et de Jaccard (droite) en fonction de la puissance q. (——— = unigrammes, − − − = bigrammes et ····· = trigrammes).

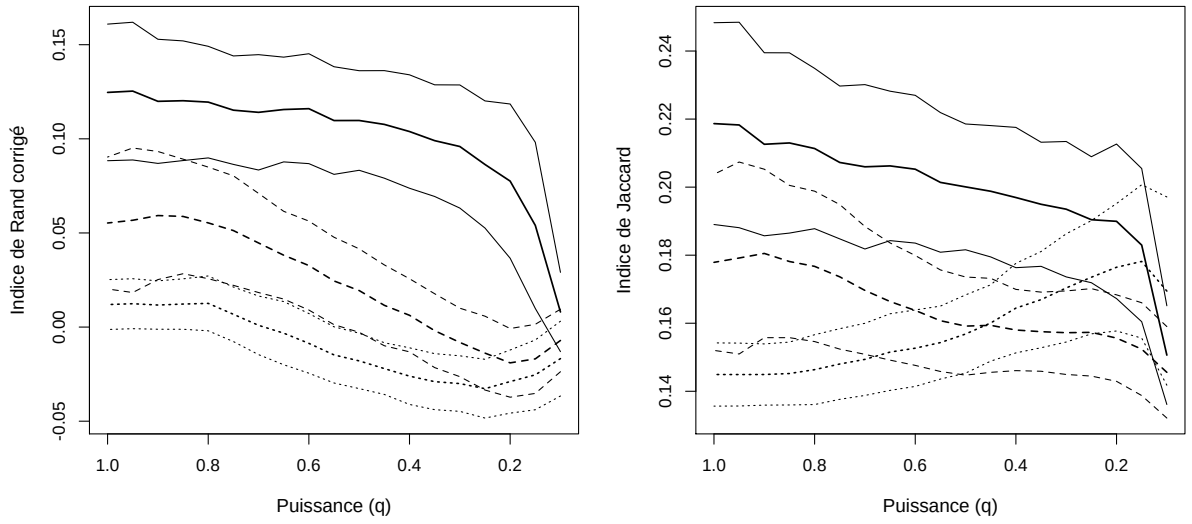


FIGURE C.2 – « Le Voleur » avec l'algorithme K-means. Moyennes (lignes épaisses) et écarts-types (lignes fines) pour l'indice de Rand corrigé (gauche) et de Jaccard (droite) en fonction de la puissance q. (——— = unigrammes, − − − = bigrammes et ····· = trigrammes).

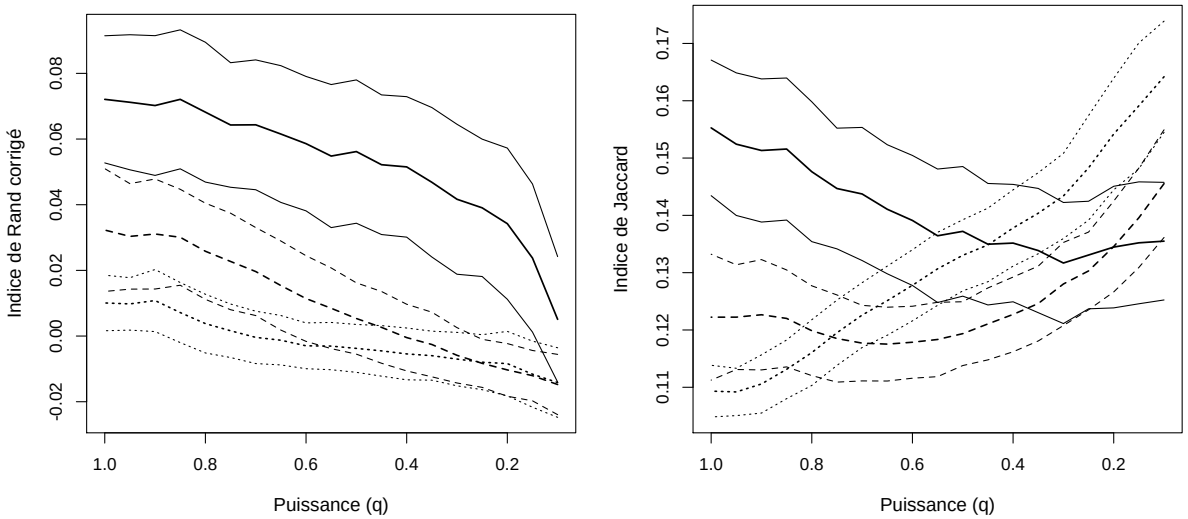


FIGURE C.3 – « Un Fou ? » avec l'algorithme K-means. Moyennes (lignes épaisses) et écarts-types (lignes fines) pour l'indice de Rand corrigé (gauche) et de Jaccard (droite) en fonction de la puissance q. (——— = unigrammes, − − − = bigrammes et ····· = trigrammes).

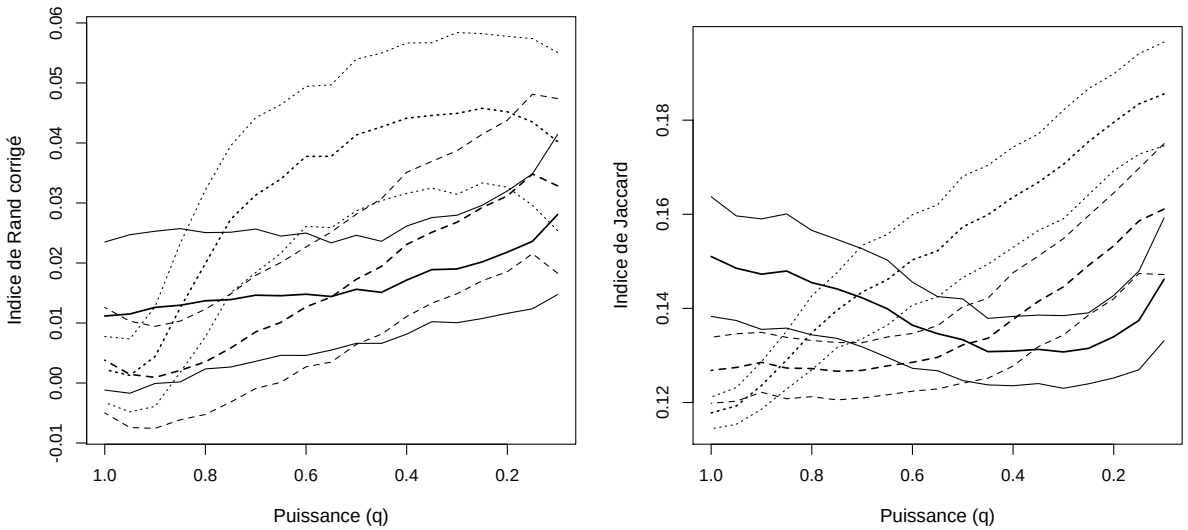


FIGURE C.4 – « Un Fou » avec l'algorithme K-means. Moyennes (lignes épaisses) et écarts-types (lignes fines) pour l'indice de Rand corrigé (gauche) et de Jaccard (droite) en fonction de la puissance q. (——— = unigrammes, − − − = bigrammes et ····· = trigrammes).

C.1.2 V de Cramer

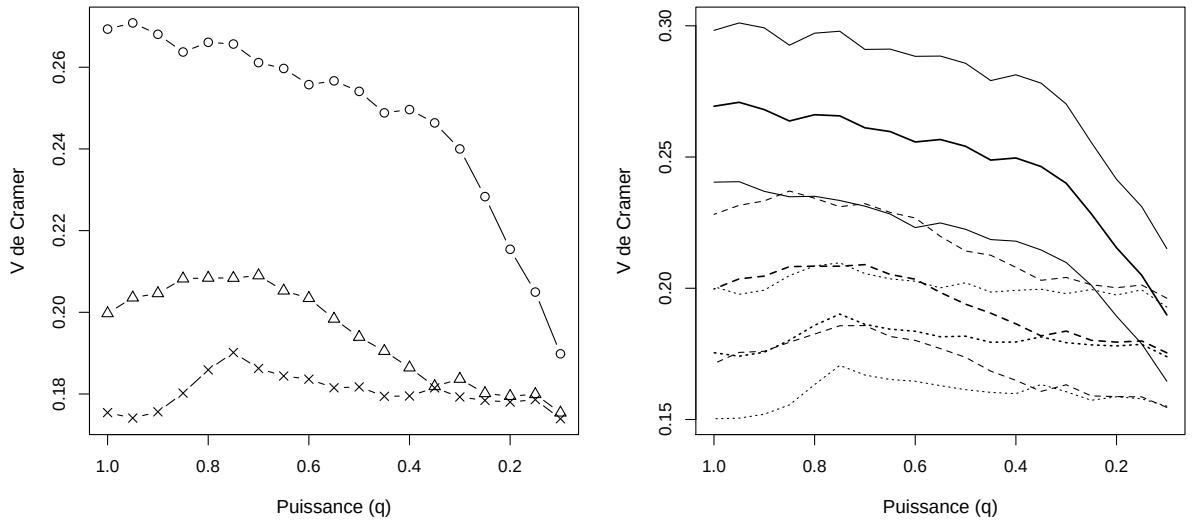


FIGURE C.5 – « L'Orient » avec l'algorithme K-means : V de Cramer en fonction de la puissance q. Gauche : moyenne des résultats obtenus pour chaque valeur de q, où $\circ$ = unigrammes, $\triangle$ = bigrammes et $\times$ = trigrammes. Droite : moyenne (lignes épaisses) et écarts-types (lignes fines), où ——— = unigrammes, − − − = bigrammes et ····· = trigrammes.

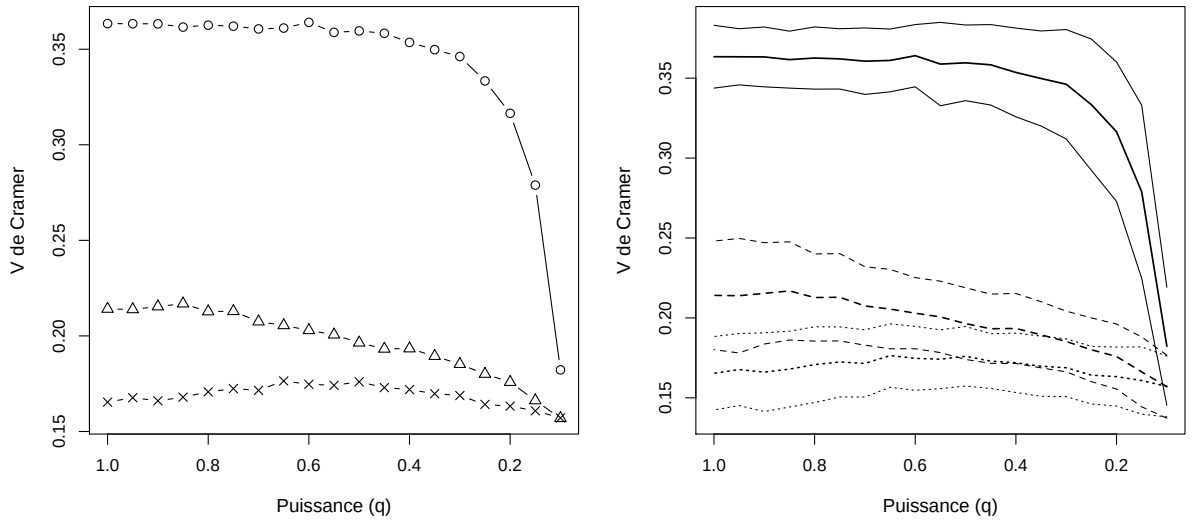


FIGURE C.6 – « Le Voleur » avec l'algorithme K-means : V de Cramer en fonction de la puissance q. Gauche : moyenne des résultats obtenus pour chaque valeur de q, où $\circ$ = unigrammes, $\triangle$ = bigrammes et $\times$ = trigrammes. Droite : moyenne (lignes épaisses) et écarts-types (lignes fines), où ——— = unigrammes, − − − = bigrammes et ····· = trigrammes.

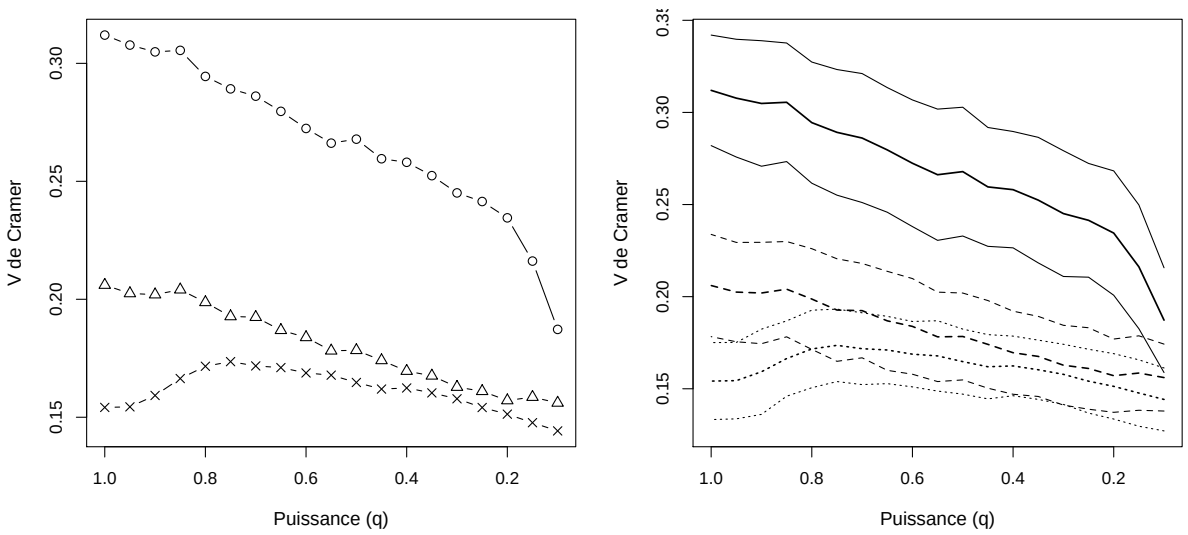


FIGURE C.7 – « Un Fou ? » avec l'algorithme K-means : V de Cramer en fonction de la puissance q. Gauche : moyenne des résultats obtenus pour chaque valeur de q, où $\circ$ = unigrammes, $\triangle$ = bigrammes et $\times$ = trigrammes. Droite : moyenne (lignes épaisses) et écarts-types (lignes fines), où ——— = unigrammes, − − − = bigrammes et ····· = trigrammes.

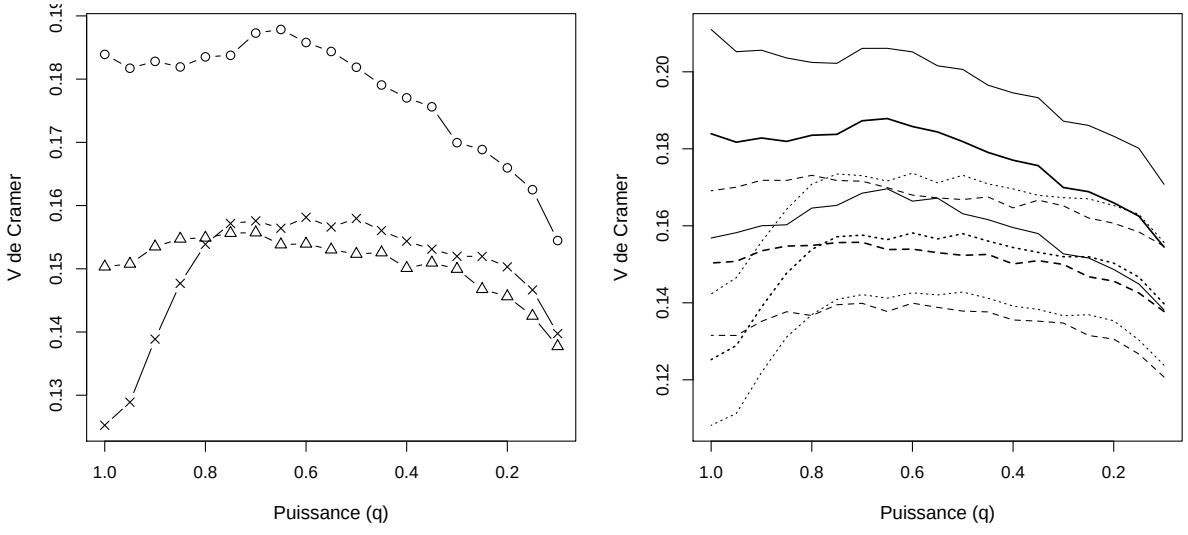


FIGURE C.8 – « Un Fou » avec l'algorithme K-means : V de Cramer en fonction de la puissance q. Gauche : moyenne des résultats obtenus pour chaque valeur de q, où $\circ$ = unigrammes, $\triangle$ = bigrammes et $\times$ = trigrammes. Droite : moyenne (lignes épaisses) et écarts-types (lignes fines), où ——— = unigrammes, − − − = bigrammes et ····· = trigrammes.

C.2 K-means flou

Les figures de cette section présentent, avec l'algorithme K-means flou, le nombre de groupes final M, l'indice de Rand corrigé et l'indice de Jaccard, en fonction de la température relative, t_{rel}. Les graphiques de droite des figures 4.15 à 4.22 de la section 4.3.2.2 sont des représentations paramétriques de ces résultats, en fonction de t_{rel}.

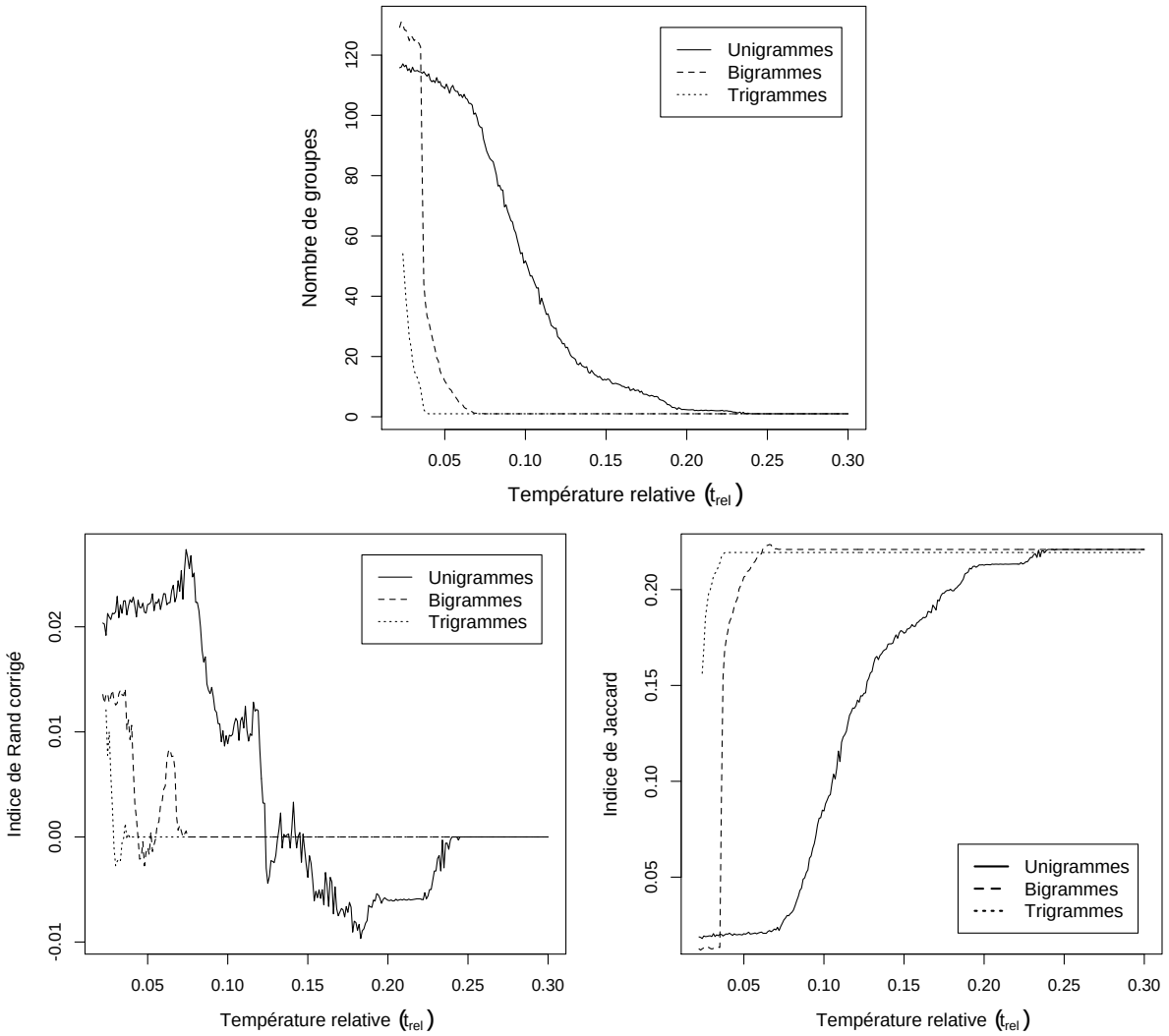


FIGURE C.9 – « L'Orient » avec l'algorithme K-means flou. En haut : nombre de groupes en fonction de la température relative. En bas, à gauche : indice de Rand corrigé en fonction de la température relative. En bas à droite : indice de Jaccard en fonction de la température relative.

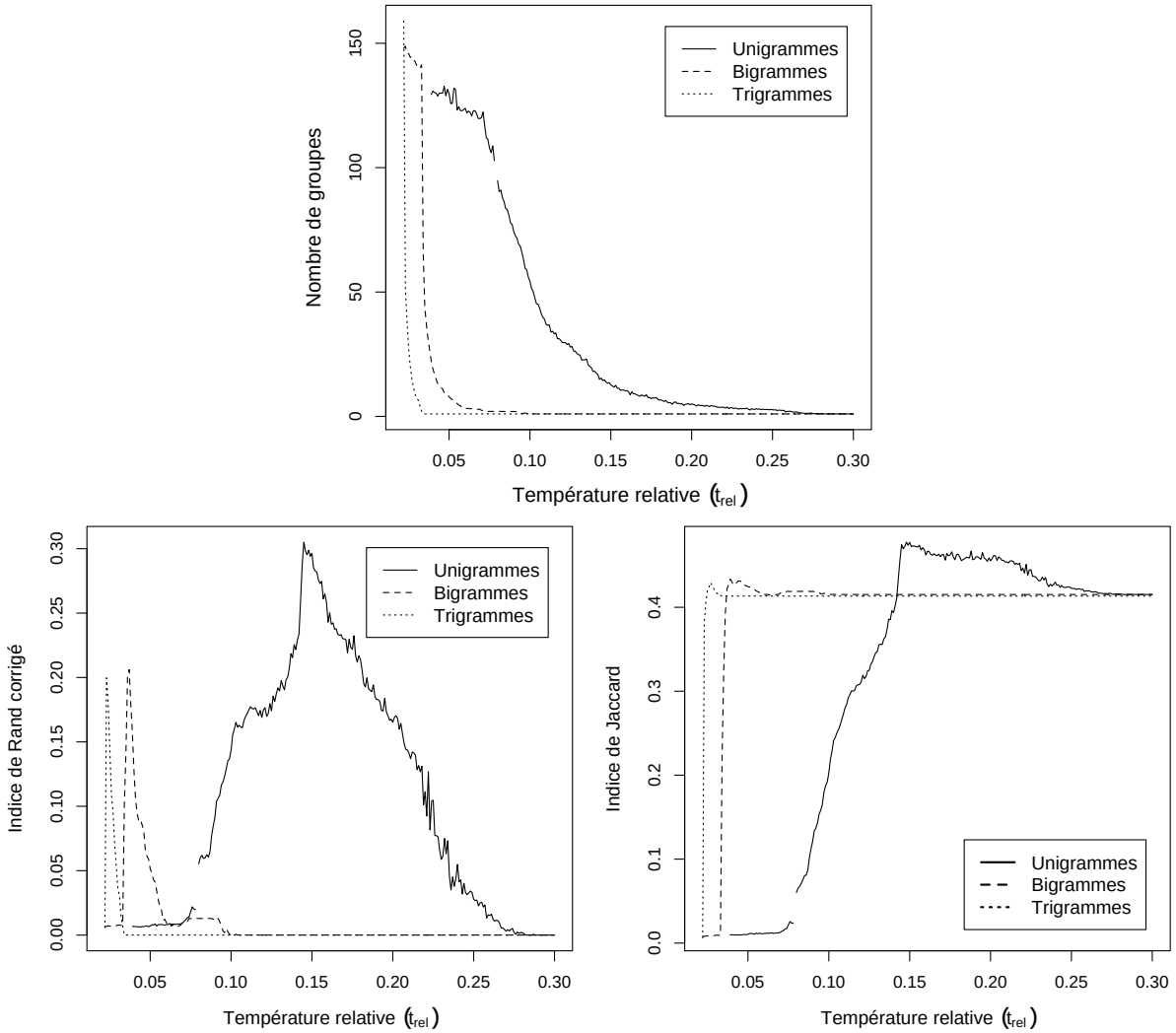


FIGURE C.10 – « Le Voleur » avec l'algorithme K-means flou. En haut : nombre de groupes en fonction de la température relative. En bas, à gauche : indice de Rand corrigé en fonction de la température relative. En bas à droite : indice de Jaccard en fonction de la température relative.

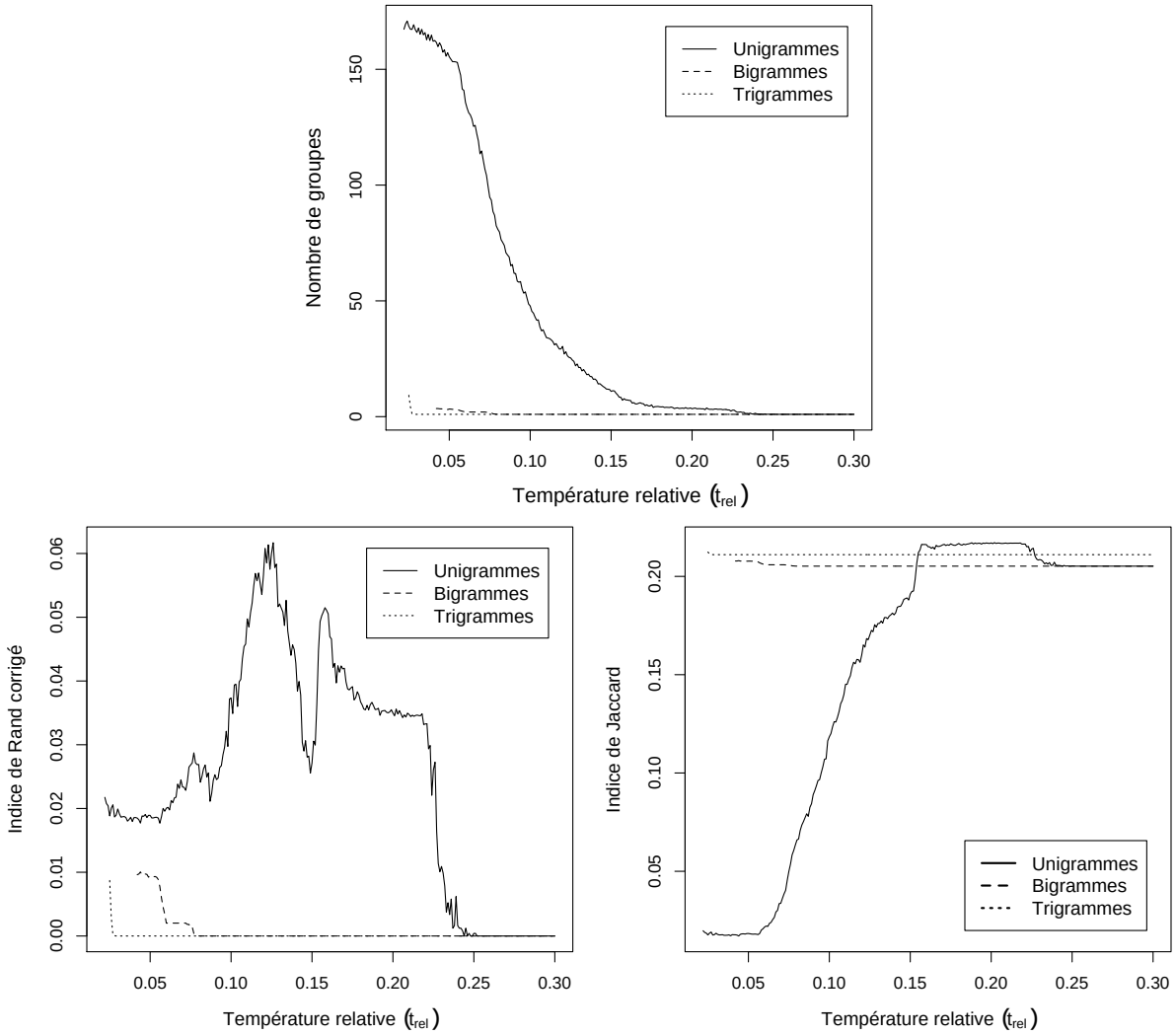


FIGURE C.11 – « Un Fou ? » avec l'algorithme K-means flou. En haut : nombre de groupes en fonction de la température relative. En bas, à gauche : indice de Rand corrigé en fonction de la température relative. En bas à droite : indice de Jaccard en fonction de la température relative.

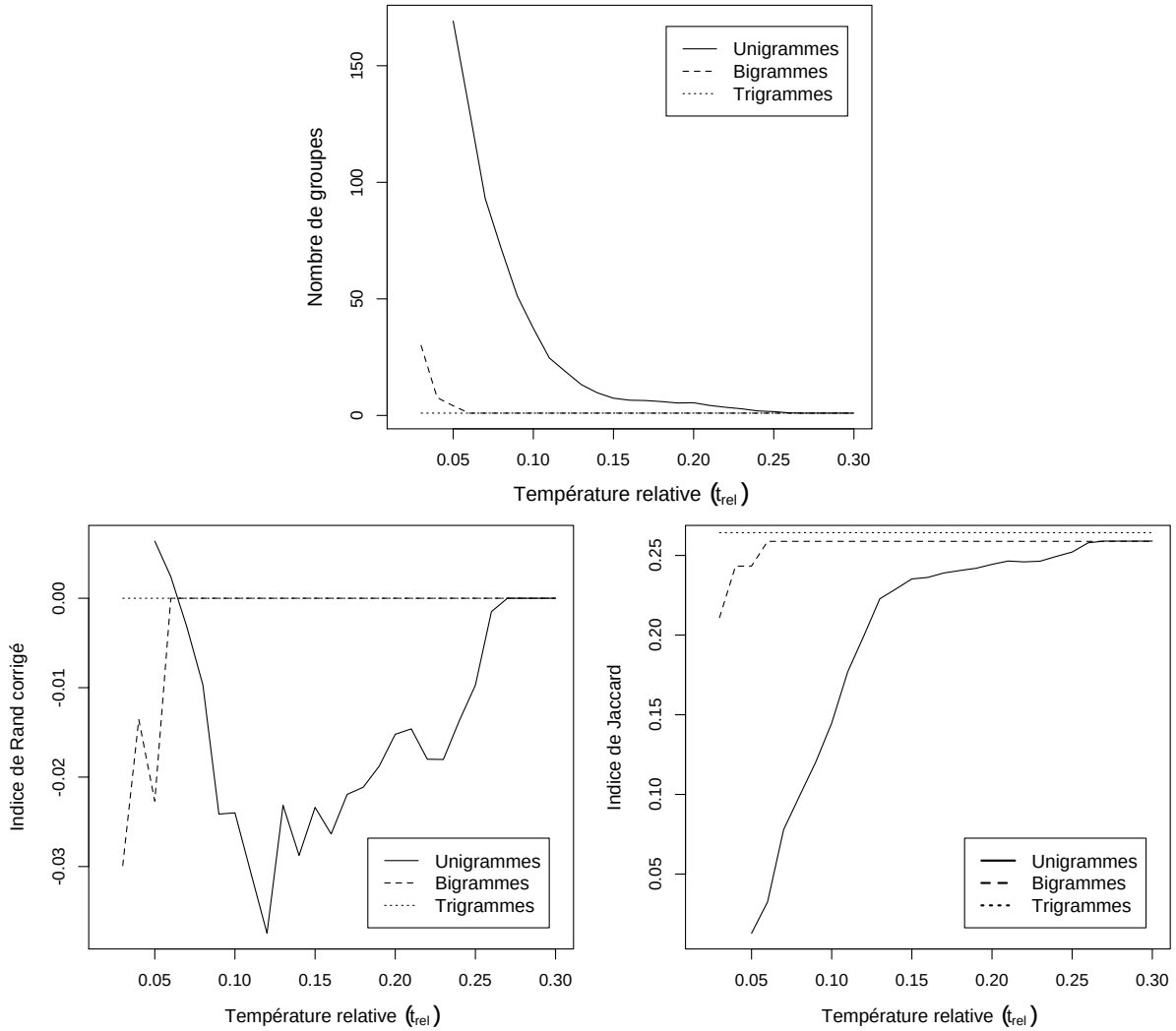


FIGURE C.12 – « Un Fou » avec l'algorithme K-means flou. En haut : nombre de groupes en fonction de la température relative. En bas, à gauche : indice de Rand corrigé en fonction de la température relative. En bas à droite : indice de Jaccard en fonction de la température relative.

Bibliographie

Adam, J.M. (2008*a*). *La linguistique textuelle : Introduction à l'analyse textuelle des discours*. 2ème éd. Paris : Armand Colin.

Adam, J.M. (2008*b*). *Les textes : types et prototypes*. 2ème éd. Paris : Armand Colin.

Austin, J.L. (1962). *How to do Things with Words*. London : Oxford University Press.

Bavaud, F. (2004). Generalized Factor Analyses for Contingency Tables. *In* D. Banks, F. McMorris, P. Arabie et W. Gaul (Eds.), *Classification, Clustering, and Data Mining Applications*, Studies in Classification, Data Analysis, and Knowledge Organisation, pp. 597–606. Berlin ; Heidelberg : Springer.

Bavaud, F. (2009). Aggregation invariance in general clustering approaches. *Advances in Data Analysis and Classification*, *3*(3) : 205–225.

Bavaud, F. (2010). Euclidean Distances, Soft and Spectral Clustering on Weighted Graphs. *In* J. Balcázar, F. Bonchi, A. Gionis et M. Sebag (Eds.), *Machine Learning and Knowledge Discovery in Databases*, t. 6321 de *Lecture Notes in Computer Science*, pp. 103–118. Berlin ; Heidelberg : Springer.

Bavaud, F. (2011). On the Schoenberg Transformations in Data Analysis : Theory and Illustrations. *Journal of Classification*, *28*(3) : 297–314.

Bavaud, F. (2013). Testing Spatial Autocorrelation in Weighted Networks : the Modes Permutation Test. *Journal of Geographical Systems*, *15*(3) : 233–247.

Bavaud, F. et Cocco, C. (accepté pour publication). Factor Analysis of Local Formalism. *In Data Analysis, Learning by Latent Structures, and Knowledge Discovery*, Studies in Classification, Data Analysis, and Knowledge Organization. Berlin ; Heidelberg : Springer.

Bavaud, F., Cocco, C. et Xanthos, A. (2012). Textual autocorrelation : formalism and illustrations. *In JADT 2012 : 11èmes Journées internationales d'Analyse statistique des Données Textuelles*, pp. 109–120.

Bavaud, F., Cocco, C. et Xanthos, A. (accepté pour publication). Textual navigation and autocorrelation. *In* G. Mikros et J. Mačutek (Eds.), *Sequences in Language and Text*, Quantitative Linguistics. Berlin : De Gruyter.

Benzécri, J.P. *et al.* (1973). *L'Analyse des Données : 1 La taxinomie*. Paris : Dunod.

Benzécri, J.P. *et al.* (1980). *L'Analyse des Données : 2 L'analyse des correspondances.* 3ème éd. Paris : Dunod.

Biber, D. (1988). *Variation across Speech and Writing.* Cambridge, UK : Cambridge University Press.

Box, G.E.P. et Jenkins, G.M. (1976). *Time series analysis : forecasting and control.* San Francisco, CA : Holden-Day.

Boyer, K., Ha, E.Y., Phillips, R., Wallis, M., Vouk, M. et Lester, J. (2010). Dialogue Act Modeling in a Complex Task-Oriented Domain. *In Proceedings of the SIGDIAL 2010 Conference*, pp. 297–305. Tokyo, Japan : Association for Computational Linguistics.

Bronckart, J.P. (1996). *Activité langagière, textes et discours : Pour un interactionisme socio-discursif.* Lausanne ; Paris : Delachaux et Niestlé.

Camiz, S. (2005). The Guttman Effect : its Interpretation and a New Redressing Method. *Tetradia Analushsq Dedomenwn (Data Analysis Bulletin)*, *5* : 7–34.

Celeux, G. et Govaert, G. (1992). A classification {EM} algorithm for clustering and two stochastic versions. *Computational Statistics & Data Analysis*, *14*(3) : 315 – 332.

Charaudeau, P. (1992). *Grammaire du sens et de l'expression.* Paris : Hachette.

Cliff, A.D. et Ord, J.K. (1981). *Spatial Processes : Models and Applications.* London : Pion.

Cocco, C. (2012*a*). Catégorisation automatique de propositions textuelles en types de discours. *In Lire demain : des manuscrits antiques à l'ère digitale = Reading tomorrow : from ancient manuscripts to the digital era*, pp. 689–707. Lausanne : PPUR.

Cocco, C. (2012*b*). Discourse Type Clustering using POS n-gram Profiles and High-Dimensional Embeddings. *In Proceedings of the Student Research Workshop at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 55–63. Avignon, France : Association for Computational Linguistics.

Cocco, C. (2014). Classification supervisée multi-étiquette en actes de dialogue : analyse discriminante et transformations de schoenberg. *In JADT 2014 : 12èmes Journées internationales d'Analyse statistique des Données Textuelles*, pp. 147–160.

Cocco, C. et Bavaud, F. (accepté pour publication). Correspondence Analysis, Cross-Autocorrelation and Clustering in Polyphonic Music. *In Data Analysis, Learning by Latent Structures, and Knowledge Discovery*, Studies in Classification, Data Analysis, and Knowledge Organization. Berlin ; Heidelberg : Springer.

Cocco, C., Pittier, R., Bavaud, F. et Xanthos, A. (2011). Segmentation and Clustering of Textual Sequences : a Typological Approach. *In Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*, pp. 427–433. Hissar, Bulgaria : RANLP 2011 Organising Committee.

Cohen, W.W., Carvalho, V.R. et Mitchell, T.M. (2004). Learning to Classify Email into "Speech Acts". *In* D. Lin et D. Wu (Eds.), *Proceedings of EMNLP 2004*, pp. 309–316. Barcelona, Spain : Association for Computational Linguistics.

Colineau, N. et Caelen, J. (1995). Étude de marqueurs dans les actes de dialogue dans un corpus de conception. *In 01Design'95 : Aspects communicatifs en conception, 4ème table ronde francophone sur la conception*, pp. 127–139.

Critchley, F. et Fichet, B. (1994). The partial order by inclusion of the principal classes of dissimilarity on a finite set, and some of their basic properties. *In* B.V. Cutsem (Ed.), *Classification and Dissimilarity Analysis*, n° 93 in Lecture Notes in Statistics, pp. 5–65. New York : Springer.

Cuadras, C.M. et Fortiana, J. (1996). Weighted continuous metric scaling. *In* A.K. Gupta et V.L. Girko (Eds.), *Multidimensional Statistical Analysis and Theory of Random Matrices*, pp. 27–40. Zeist, The Netherlands : VSP.

Daoust, F., Marcoux, Y. et Viprey, J.M. (2010). L'annotation structurelle. *In JADT 2010 : 10th International Conference on Statistical Analysis of Textual Data.*

de Maupassant, G. (1882). Le voleur. *Gil Blas.* `http://un2sg4.unige.ch/athena/selva/maupassant/textes/voleur.html`. Thierry Selva. Consulté le 6 juillet 2011.

de Maupassant, G. (1883). L'orient. *Le Gaulois.* `http://un2sg4.unige.ch/athena/selva/maupassant/textes/orient.html`. Thierry Selva. Consulté le 5 mars 2011.

de Maupassant, G. (1884). Un fou ? *Le Figaro.* `http://un2sg4.unige.ch/athena/maupassant/maup_fou.html`. Thierry Selva. Consulté le 7 février 2011.

de Maupassant, G. (1885). Un fou. *Le Gaulois.* `http://un2sg4.unige.ch/athena/selva/maupassant/textes/unfou.html`. Thierry Selva. Consulté le 26 avril 2011.

Dejean, C., Fortun, M., Massot, C., Pottier, V., Poulard, F. et Vernier, M. (2010). Un étiqueteur de rôles grammaticaux libre pour le français intégré à Apache UIMA. *In Actes de la 17e Conférence sur le Traitement Automatique des Langues Naturelles.* Montréal, Canada.

Denœud, L. et Guénoche, A. (2006). Comparison of Distance Indices Between Partitions. *In* V. Batagelj, H.H. Bock, A. Ferligoj et A. Žiberna (Eds.), *Data Science and Classification*, Studies in Classification, Data Analysis, and Knowledge Organization, pp. 21–28. Berlin ; Heidelberg : Springer.

Dupuis, F. et Lebart, L. (2009). Visualization, validation and seriation : Application to a corpus of medieval texts. *In* M. Dufresne, F. Dupuis et E. Vocaj (Eds.), *Historical Linguistics 2007*, n° 308 in Current Issues in Linguistic Theory, pp. 269–284. Amsterdam : John Benjamins Publishing Company.

Efron, B. et Tibshirani, R.J. (1993). *An Introduction to the Bootstrap.* N° 57 in Monographs on Statistics and Applied Probability. New York : Chapman & Hall.

Ellis, D. et Poliner, G.E. (2007). Identifying 'Cover Songs' with Chroma Features and Dynamic Programming Beat Tracking. *In IEEE International Conference on Acoustics, Speech and Signal Processing, 2007. ICASSP 2007*, t. 4, pp. IV–1429–IV–1432.

Estivill-Castro, V. (2002). Why So Many Clustering Algorithms : A Position Paper. *SIGKDD Explorations Newsletter*, *4*(1) : 65–75.

Faget, Z. (2011). *Un modèle pour la gestion des séquences temporelles synchronisées. Application aux données musicales symboliques.* Thèse de doctorat, Université Paris-Dauphine.

Fellbaum, C. (1998). *WordNet : An Electronic Lexical Database.* Cambridge, MA : MIT Press.

Ferschke, O., Gurevych, I. et Chebotar, Y. (2012). Behind the Article : Recognizing Dialog Acts in Wikipedia Talk Pages. *In Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 777–786. Avignon, France : Association for Computational Linguistics.

Filliettaz, L. (2001). Les types de discours. *Círculo de lingüistica aplicada a la communicación*, *8* : http://www.ucm.es/info/circulo/no8/filliettaz.htm.

Fisher, R.A. (1936). The Use of Multiple Measurements in Taxonomic Problems. *Annals of Eugenics*, *7*(2) : 179–188.

Francis, W.N. et Kučera, H. (1967). *Computational analysis of present-day American English*. Providence : Brown University Press.

Francis, W.N. et Kučera, H. (1982). *Frequency analysis of English usage : lexicon and grammar*. Boston : Houghton Mifflin.

Gauch, H.G., Whittaker, R.H. et Wentworth, T.R. (1977). A Comparative Study of Reciprocal Averaging and Other Ordination Techniques. *Journal of Ecology*, *65*(1) : 157–174.

Geary, R.C. (1954). The Contiguity Ratio and Statistical Mapping. *The Incorporated Statistician*, *5*(3) : 115–145.

Goldstein, J. et Sabin, R. (2006). Using Speech Acts to Categorize Email and Identify Email Genres. *In Proceedings of the 39th Annual Hawaii International Conference on System Sciences, 2006.*, t. 3, p. 50b.

Greenacre, M.J. (1984). *Theory and Applications of Correspondence Analysis*. London : Academic Press.

Halkidi, M., Batistakis, Y. et Vazirgiannis, M. (2002). Cluster Validity Methods : Part I. *SIGMOD Record*, *31*(2) : 40–45.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. et Witten, I.H. (2009). The WEKA Data Mining Software : An Update. *SIGKDD Explorations Newsletter*, *11*(1) : 10–18.

Hawker, T. et Honnibal, M. (2006). Improved default sense selection forword sense disambiguation. *In Proceedings of the Australasian Language Technology Workshop 2006*, pp. 11–17. Sydney, Australia.

Hildebrand, G.H. et Mace, A. (1950). The Employment Multiplier in an Expanding Industrial Market : Los Angeles County, 1940-47. *The Review of Economics and Statistics*, *32*(3) : 241–249.

Houle, M., Kriegel, H.P., Kröger, P., Schubert, E. et Zimek, A. (2010). Can Shared-Neighbor Distances Defeat the Curse of Dimensionality ? *In* M. Gertz et B. Ludäscher (Eds.), *Scientific and Statistical Database Management*, t. 6187 de *Lecture Notes in Computer Science*, pp. 482–500. Berlin ; Heidelberg : Springer.

Hubert, L. et Arabie, P. (1985). Comparing partitions. *Journal of Classification*, *2*(1) : 193–218.

Huron, D. (1994). *The Humdrum Toolkit : Reference Manual*. Menlo Park, CA : Center for Computer Assisted Research in the Humanities.

Huron, D. (1998). *Humdrum User's Guide*. http://humdrum.ccarh.org/. Consulté le 17 mars 2014.

Husson, F., Josse, J., Le, S. et Mazet, J. (2013). *FactoMineR : Multivariate Exploratory Data Analysis and Data Mining with R*. R package version 1.25.

Jain, A.K., Murty, M.N. et Flynn, P.J. (1999). Data Clustering : A Review. *ACM Computing Surveys*, *31*(3) : 264–323.

Karlgren, J. et Cutting, D. (1994). Recognizing Text Genres with Simple Metrics Using Discriminant Analysis. *In Proceedings of the 15th conference on Computational linguistics*, t. 2 de *COLING '94*, pp. 1071–1075. Stroudsburg, PA, USA : Association for Computational Linguistics.

Kim, S.N., Cavedon, L. et Baldwin, T. (2010). Classifying Dialogue Acts in One-on-One Live Chats. *In Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pp. 862–871. Cambridge, MA : Association for Computational Linguistics.

Koppel, M. et Schler, J. (2003). Exploiting Stylistic Idiosyncrasies for Authorship Attribution. *In IJCAI'03 Workshop on Computational Approaches to Style Analysis and Synthesis*, pp. 69–72.

Kriesel, V. (2013). *Music Synchronization, Audio Matching, Pattern Detection, and User Interfaces for a Digital Music Library System.* Thèse de doctorat, Universität Bonn.

Lavrenko, V. et Pickens, J. (2003). Polyphonic Music Modeling with Random Fields. *In Proceedings of the eleventh ACM international conference on Multimedia*, MULTIMEDIA '03, pp. 120–129. Berkeley, CA : ACM.

Lê, S., Josse, J. et Husson, F. (2008). FactoMineR : An R Package for Multivariate Analysis. *Journal of Statistical Software*, *25*(1) : 1–18.

Le Roux, B. et Rouanet, H. (2004). *Geometric Data Analysis : From Correspondence Analysis to Structured Data Analysis.* Dordrecht : Kluwer Academic Publishers.

Le Roux, B. et Rouanet, H. (2010). *Multiple Correspondence Analysis.* N° 163 in Quantitative Applications in the Social Sciences. Thousand Oaks, CA : Sage.

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

Lebart, L. (1969). Analyse Statistique de la Contiguïté. *Publications de l'Institut de Statistique des Universités de Paris*, *XVIII* : 81–112.

Lebart, L. (2007). Which Bootstrap for Principal Axes Methods ? *In* P. Brito, G. Cucumel, P. Bertrand et F. de Carvalho (Eds.), *Selected Contributions in Data Analysis and Classification*, Studies in Classification, Data Analysis, and Knowledge Organization, pp. 581–588. Berlin ; Heidelberg : Springer.

Lebart, L., Morineau, A. et Piron, M. (1995). *Statistique exploratoire multidimensionnelle.* Paris : Dunod.

Li, Y., Luo, C. et Chung, S.M. (2008). Text Clustering with Feature Selection by Using Statistical Data. *IEEE Transactions on Knowledge and Data Engineering*, *20* : 641–652.

Luaces, O., Dìez, J., Barranquero, J., del Coz, J.J. et Bahamonde, A. (2012). Binary relevance efficacy for multilabel classification. *Progress in Artificial Intelligence*, *1*(4) : 303–313.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. *In Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, t. 1, pp. 281–297. Berkeley : University of California Press.

Malrieu, D. et Rastier, F. (2001). Genres et variations morphosyntaxiques. *Traitement automatique des langues*, *42*(2) : 547–577.

Manning, C.D. et Schütze, H. (1999). *Foundations of Statistical Natural Language Processing.* 1ère éd. Cambridge, MA : The MIT Press.

Mardia, K.V., Kent, J.T. et Bibby, J.M. (1979). *Multivariate analysis*. London : Academic Press.

Matheron, G. (1965). *Les variables régionalisées et leur estimation : une application de la théorie des fonctions aléatoires aux sciences de la nature*. Paris : Masson.

McLachlan, G.J. et Krishnan, T. (1997). *The EM algorithm and extensions*. New York : John Wiley.

Meilă, M. (2003). Comparing Clusterings by the Variation of Information. *In* B. Schölkopf et M. Warmuth (Eds.), *Learning Theory and Kernel Machines*, t. 2777 de *Lecture Notes in Computer Science*, pp. 173–187. Berlin ; Heidelberg : Springer.

Miller, G.A. (1995). WordNet : A Lexical Database for English. *Communications of the ACM*, *38*(11) : 39–41.

Milligan, G.W. et Cooper, M.C. (1986). A Study of the Comparability of External Criteria for Hierarchical Cluster Analysis. *Multivariate Behavioral Research*, *21*(4) : 441–458.

Moran, P.A.P. (1950). Notes on continuous stochastic phenomena. *Biometrika*, *37* : 17–23.

Morando, B. (1981). L'analyse statistique des partitions de musique. *In* J.P. Benzécri *et al.* (Eds.), *Pratique de l'analyse des données, tome 3 : Linguistique et lexicologie*, pp. 507–522. Paris : Dunod.

Müller, M. et Ewert, S. (2011). Chroma Toolbox : Matlab Implementations for Extracting Variants of Chroma-based Audio Features. *In Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR)*, pp. 215–220.

Murtagh, F. et Legendre, P. (2011). Ward's hierarchical clustering method : Clustering criterion and agglomerative algorithm. *arXiv:1111.6285 [stat.ML]*.

Nenadic, O. et Greenacre, M. (2007). Correspondence Analysis in `R`, with Two- and Three-dimensional Graphics : The `ca` Package. *Journal of Statistical Software*, *20*(3) : 1–13.

Palmer, A., Ponvert, E., Baldridge, J. et Smith, C. (2007). A Sequencing Model for Situation Entity Classification. *In Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pp. 896–903. Prague, Czech Republic.

Park, C.H. et Lee, M. (2008). On applying linear discriminant analysis for multi-labeled problems. *Pattern Recognition Letters*, *29*(7) : 878 – 887.

Pedersen, T., Patwardhan, S. et Michelizzi, J. (2004). WordNet::Similarity - Measuring the Relatedness of Concepts. *In* D.M. Susan Dumais et S. Roukos (Eds.), *HLT-NAACL 2004 : Demonstration Papers*, pp. 38–41. Boston, Massachusetts, USA : Association for Computational Linguistics.

Pfitzner, D., Leibbrandt, R. et Powers, D. (2009). Characterization and evaluation of similarity measures for pairs of clusterings. *Knowledge and Information Systems*, *19*(3) : 361–394.

Qadir, A. et Riloff, E. (2011). Classifying Sentences as Speech Acts in Message Board Posts. *In Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pp. 748–758. Edinburgh, Scotland, UK. : Association for Computational Linguistics.

R Core Team (2013). *R : A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Read, J., Pfahringer, B., Holmes, G. et Frank, E. (2011). Classifier chains for multi-label classification. *Machine Learning*, *85*(3) : 333–359.

Rennie, J. (2000). WordNet::QueryData : a Perl module for accessing the WordNet database. `http://people.csail.mit.edu/~jrennie/WordNet`.

Resnik, P. (1995). Using Information Content to Evaluate Semantic Similarity in a Taxonomy. *In Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI-95)*, pp. 448–453.

Resnik, P. (1999). Semantic Similarity in a Taxonomy : An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language. *Journal of Artificial Intelligence Research*, *11* : 95–130.

Robert, P. et Escoufier, Y. (1976). A Unifying Tool for Linear Multivariate Statistical Methods : The *RV*-Coefficient. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, *25*(3) : pp. 257–265.

Rose, K., Gurewitz, E. et Fox, G.C. (1990). Statistical mechanics and phase transitions in clustering. *Physical Review Letters*, *65*(8) : 945–948.

Rukhin, A.L. et Vallejos, R. (2008). Codispersion coefficients for spatial and temporal series. *Statistics & Probability Letters*, *78*(11) : 1290 – 1300.

Salton, G. et McGill, M.J. (1983). *Introduction to Modern Information Retrieval*. McGraw-Hill computer science series. New York : McGraw-Hill Book Company.

Samsonovich, A.V. (2014). Semantic cross-correlation as a measure of social interaction. *Biologically Inspired Cognitive Architectures*, *7* : 1–8.

Saporta, G. (2006). *Probabilités, analyse des données et statistique*. 2ème éd. Paris : Editions Technip.

Sapp, C.S. (2005). Online Database of Scores in the Humdrum File Format. *In Proceedings of the 6th International Conference on Music Information Retrieval (ISMIR)*, pp. 664–665. London, UK.

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. *In Proceedings of the International Conference on New Methods in Language Processing*, pp. 44–49.

Schoenberg, I.J. (1937). On Certain Metric Spaces Arising From Euclidean Spaces by a Change of Metric and Their Imbedding in Hilbert Space. *Annals of Mathematics*, *38*(4) : 787–793.

Schoenberg, I.J. (1938). Metric Spaces and Positive Definite Functions. *Transactions of the American Mathematical Society*, *44*(3) : 522–536.

Searle, J.R. (1969). *Speech Acts : An Essay in the Philosophy of Language*. Cambridge, UK : Cambridge University Press.

Sebastiani, F. (2002). Machine Learning in Automated Text Categorization. *ACM Computing Surveys*, *34*(1) : 1–47.

Smith, C.S. (2003). *Modes of Discourse : The Local Structure of Texts*. N° 103 in Cambridge Studies in Linguistics. Cambridge, UK : Cambridge University Press.

Sokolova, M. et Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, *45*(4) : 427 – 437.

Stolcke, A., Ries, K., Coccaro, N., Shriberg, E., Bates, R., Jurafsky, D., Taylor, P., Martin, R., Ess-Dykema, C.V. et Meteer, M. (2000). Dialogue Act Modeling for Automatic Tagging and Recognition of Conversational Speech. *Computational Linguistics*, *26*(3) : 339–373.

Tsoumakas, G., Katakis, I. et Vlahavas, I. (2010). Mining Multi-label Data. *In* O. Maimon et L. Rokach (Eds.), *Data Mining and Knowledge Discovery Handbook*, pp. 667–685. 2ème éd. New York : Springer.

Van Asch, V. (2012). Macro- and micro-averaged evaluation measures. `http://www.clips.uantwerpen.be/~vincent/pdf/microaverage.pdf`, consulté le 04 février 2014.

van Rijsbergen, C.J. (1979). *Information retrieval*. 2ème éd. London : Butterworths.

Vatolkin, I. (2013). *Improving Supervised Music Classification by Means of Multi-Objective Evolutionary Feature Selection*. Thèse de doctorat, Technische Universität Dortmund.

Warrens, M.J. (2008). On Association Coefficients for 2×2 Tables and Properties That Do Not Depend on the Marginal Distributions. *Psychometrika*, *73*(4) : 777–789.

Weihs, C., Ligges, U., Mörchen, F. et Müllensiefen, D. (2007). Classification in Music Research. *Advances in Data Analysis and Classification*, *1*(3) : 255–291.

Yang, Y. (1999). An Evaluation of Statistical Approaches to Text Categorization. *Information Retrieval*, *1*(1-2) : 69–90.

Yang, Y. et Pedersen, J.O. (1997). A Comparative Study on Feature Selection in Text Categorization. *In Proceedings of the 14th International Conference on Machine Learning*, pp. 412–420.

Youness, G. et Saporta, G. (2004). Une Méthodologie pour la Comparaison de Partitions. *Revue de Statistique Appliquée*, *52*(1) : 97–120.

Young, G. et Householder, A. (1938). Discussion of a set of points in terms of their mutual distances. *Psychometrika*, *3*(1) : 19–22.

Yule, G.U. (1900). On the Association of Attributes in Statistics : With Illustrations from the Material of the Childhood Society, &c. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, *194*(252-261) : 257–319.

Yule, G.U. (1912). On the Methods of Measuring Association Between Two Attributes. *Journal of the Royal Statistical Society*, *75*(6) : 579–652.