

# TABLE DES MATIÈRES

LISTE DES FIGURES . . . . .	XV
LISTE DES TABLEAUX . . . . .	XXV
LISTE DES SIGLES . . . . .	XXXI
INTRODUCTION . . . . .	1
CHAPITRE 1	
L'analyse statistique des données textuelles: champs et objets d'étude . . . . .	7
1. Le champ de l'analyse statistique des données textuelles (ADT) . . . . .	8
1.1. La statistique lexicale et l'analyse multidimensionnelle lexicale . . . . .	8
1.2. La démarche et les parcours méthodologiques . . . . .	11
2. Les disciplines voisines . . . . .	15
2.1. La linguistique de corpus . . . . .	16
2.2. Le traitement automatique des langues . . . . .	19
2.3. Fouille de données et de textes . . . . .	23
2.4. Les méthodes qualitatives . . . . .	26
3. Les corpus et les enquêtes . . . . .	28
3.1. Les corpus textuels . . . . .	29
3.2. Des textes spécifiques: les réponses aux questions ouvertes . . . . .	32

CHAPITRE 2	
Les unités d'analyse et les observations . . . . .	39
1. La segmentation du texte en unités . . . . .	43
1.1. La segmentation du texte en mots ( <i>tokenisation</i> ) . . . . .	43
1.2. Un point de vue statistique sur les unités segmentées . . . . .	46
1.3. Un point de vue linguistique sur la nature des unités segmentées . . . . .	52
1.4. Deux exemples d'impact de la lemmatisation sur les analyses . . . . .	56
2. L'annotation de surface automatique . . . . .	61
2.1. L'analyse morphosyntaxique ( <i>tagging</i> ) . . . . .	62
2.2. L'annotation syntaxique ( <i>parsing</i> ) . . . . .	67
3. Les unités séquentielles . . . . .	72
3.1. Les segments répétés . . . . .	72
3.2. Les unités séquentielles complexes . . . . .	76
Conclusion et perspectives . . . . .	80
Annexe A. Annexe Python du chapitre 2 . . . . .	83
CHAPITRE 3	
Les unités en contexte . . . . .	93
1. La concordance . . . . .	95
1.1. Une définition . . . . .	95
1.2. Le pivot . . . . .	97
1.3. La superposition et le tri . . . . .	98
1.4. Les localisations . . . . .	101
1.5. La taille du contexte . . . . .	102
2. Une typologie des formes de retour au texte . . . . .	104
2.1. Un système de désignations mnémorique . . . . .	104
2.2. Le relevé de termes . . . . .	105
2.3. La concordance . . . . .	107
2.4. Le relevé d'extraits . . . . .	108
2.5. Le retour au texte intégral . . . . .	110
3. La cooccurrence, une synthèse statistique des contextes . . . . .	113
3.1. Le principe . . . . .	113
3.2. Les paramètres . . . . .	115
3.3. Les graphes et les polycooccurrences . . . . .	118
3.4. Le calcul de cooccurrences basé sur les spécificités . . . . .	120
3.5. D'autres mesures . . . . .	121
4. Le calcul des spécificités, un outil pour la caractérisation contrastive des contextes locaux et globaux . . . . .	122
4.1. Le principe . . . . .	122

4.2. Les usages . . . . .	125
4.3. Quelques précautions interprétatives . . . . .	132
Annexe B. Annexe Python du chapitre 3 : concordances élémentaires . . . . .	135
CHAPITRE 4	
Exploration, visualisation, validation et inférence :	
les principes de base . . . . .	141
1. Les approches exploratoires et confirmatoires . . . . .	141
1.1. Les principes de base . . . . .	143
2. Les méthodes d'analyse en axes principaux . . . . .	144
2.1. Les principaux types de tableaux de données . . . . .	144
2.2. Représentation des tableaux et des méthodes . . . . .	147
2.3. Le noyau théorique de base: décomposition aux valeurs singulières (ou analyse générale) . . . . .	149
2.4. Les éléments supplémentaires (ou illustratifs) . . . . .	155
3. Les méthodes de classification. . . . .	157
4. La validation par rééchantillonnage . . . . .	160
4.1. Les techniques de rééchantillonnage avec remise ( <i>Bootstrap</i> ) . . . . .	160
4.2. Le principe du <i>rééchantillonnage avec remise</i> . . . . .	161
4.3. La mise en œuvre et le calcul des zones de confiance. . . . .	162
Annexe C. Annexe R du chapitre 4 . . . . .	165
CHAPITRE 5	
L'analyse en composantes principales (ACP) . . . . .	
1. Les interprétations géométriques . . . . .	170
2. Le problème des échelles de mesure et la transformation des données . . . . .	173
2.1. La distance entre les individus . . . . .	173
2.2. La matrice C à diagonaliser . . . . .	175
2.3. Les distances entre points-variables . . . . .	175
3. La représentation des mots et des répondants . . . . .	176
4. L'analyse du nuage des $p$ variables (colonnes) . . . . .	177
4.1. L'interprétation statistique des axes. . . . .	177
4.2. La distance à l'origine. . . . .	178
5. Observations et variables supplémentaires . . . . .	179
5.1. Observations supplémentaires . . . . .	180
5.2. Variables supplémentaires . . . . .	180
5.3. Catégories (de variables nominales) supplémentaires : . . . . .	181

6.	L'analyse factorielle en facteurs communs et spécifiques . . . . .	181
6.1.	Le modèle de l'analyse factorielle . . . . .	181
7.	La validation par rééchantillonnage ( <i>bootstrap</i> ) . . . . .	183
7.1.	La nécessité d'une validation adaptée . . . . .	183
7.2.	Le choix de la validation partielle par rééchantillonnage ( <i>partial bootstrap</i> ) . . . . .	184
8.	Deux exemples d'application . . . . .	185
8.1.	Exemple 1 : enquêtes sémiométriques . . . . .	185
8.2.	Exemple 2 : caractérisation morphosyntaxique de typologies textuelles . . . . .	192
	Annexe D. Annexe technique : analyse canonique, régression et variables supplémentaires . . . . .	201
	Annexe E. Annexe R du chapitre 5 . . . . .	207
	CHAPITRE 6	
	L'analyse des correspondances (AC) . . . . .	211
1.	La démarche d'après un exemple . . . . .	212
1.1.	Les données initiales : la table de contingence . . . . .	212
1.2.	La présentation du corpus et du tableau de données . . . . .	212
1.3.	Les transformations de la table de contingence . . . . .	214
1.4.	La visualisation sous forme de plan factoriel . . . . .	215
2.	La représentation simultanée des lignes et des colonnes . . . . .	219
2.1.	Les relations de transition (ou quasi barycentriques) . . . . .	219
2.2.	Le principe et l'interprétation de la représentation simultanée . . . . .	220
3.	Les éléments supplémentaires . . . . .	222
4.	Les aides à l'interprétation . . . . .	223
4.1.	La variance totale et le test d'indépendance . . . . .	223
4.2.	Les contributions absolues et cosinus carrés . . . . .	225
5.	La validation par rééchantillonnage . . . . .	227
5.1.	Le rééchantillonnage pour l'analyse des correspondances . . . . .	227
5.2.	Le principe des répliques . . . . .	228
5.3.	Les zones de confiance . . . . .	228
6.	L'analyse des correspondances multiples (ACM) . . . . .	230
6.1.	Une simple extension de l'analyse des correspondances simples . . . . .	230
6.2.	La structure de base d'un échantillon d'enquête . . . . .	233
6.3.	Le positionnement des variables illustratives . . . . .	236
7.	D'autres méthodes . . . . .	237
7.1.	L'analyse logarithmique . . . . .	238
7.2.	L'analyse sémantique latente . . . . .	238

Annexe F. Annexe technique du chapitre 6 . . . . .	241
Annexe G. Annexe R du chapitre 6 . . . . .	251
CHAPITRE 7	
La classification des mots et des textes . . . . .	255
1. La classification ascendante hiérarchique (CAH) d'après un exemple . . . . .	256
1.1. Le principe . . . . .	257
1.2. Un exemple de base . . . . .	257
1.3. Le dendrogramme . . . . .	258
1.4. Le détail des étapes de l'agrégation . . . . .	260
1.5. La confrontation avec le plan principal de la figure 6.1 . . . . .	261
2. Les méthodes de classification hiérarchique, les représentations arborées . . . . .	262
2.1. Les distances entre éléments et entre groupes . . . . .	263
2.2. L'algorithme basique de classification ascendante hiérarchique (CAH) . . . . .	263
2.3. L'arbre de longueur minimale (ALM) . . . . .	264
2.4. Le critère d'agrégation selon la variance . . . . .	268
2.5. L'analyse arborée, les arbres additifs (AA) . . . . .	269
2.6. La classification descendante hiérarchique (ou classification divisive) . . . . .	271
2.7. Exemple 2: les discours sur l'état de l'Union de 19 présidents des États-Unis (1900-2009) . . . . .	273
3. Les méthodes de partitionnement . . . . .	276
3.1. L'agrégation autour des centres mobiles . . . . .	277
3.2. Les cartes auto-organisées (SOM) . . . . .	279
4. La classification mixte et autres modèles . . . . .	283
4.1. La stratégie de classification mixte . . . . .	283
4.2. Le choix du nombre de classes par coupure de l'arbre hiérarchique . . . . .	284
4.3. La factorisation non négative de matrices et la recherche de thèmes . . . . .	285
5. La sériation . . . . .	288
5.1. Le principe général . . . . .	288
5.2. L'application au corpus STATE OF THE UNION . . . . .	289
6. La validation des classifications . . . . .	291
6.1. Le cadre général . . . . .	291
6.2. Le nombre de classes à retenir . . . . .	292
Annexe H. Annexe R du chapitre 7: calcul et tracé de l'arbre de longueur minimale (ALM) . . . . .	295

## CHAPITRE 8

Les stratégies d'analyse et la complémentarité entre analyse en axes principaux et classification . . . . .	301
1. Les forces et les faiblesses des méthodes en axes principaux . . . . .	302
1.1. Le caractère partiel des visualisations . . . . .	302
1.2. Les difficultés d'interprétation . . . . .	303
1.3. Le manque de robustesse . . . . .	303
1.4. Des graphiques factoriels inextricables . . . . .	303
1.5. Les limites de la compression structurale par SVD (décomposition aux valeurs singulières) . . . . .	303
1.6. Cas des axes peu différenciés: le recours aux arbres additifs . . . . .	305
1.7. Les analyses en axes principaux, malgré tout . . . . .	307
2. L'utilisation conjointe des axes principaux et de la classification . . . . .	307
2.1. La classification comme remède et complément . . . . .	308
2.2. D'autres travaux sur la complémentarité . . . . .	309
2.3. Mise en œuvre pratique . . . . .	309
3. La description statistique des classes ou des catégories: valeurs-test et spécificités . . . . .	311
3.1. Les variables caractéristiques d'une classe . . . . .	311
3.2. Les valeurs-test pour les variables continues . . . . .	312
3.3. Les valeurs-test pour les variables nominales (catégories, mots) . . . . .	313
3.4. Exemple de mise en œuvre des valeurs-test . . . . .	315
3.5. Le problème des tests multiples . . . . .	317
4. Les fragments caractéristiques (ou réponses modales) . . . . .	318
4.1. La sélection à partir des distances . . . . .	318
4.2. La sélection à partir des mots caractéristiques . . . . .	320
4.3. Exemple de mise en œuvre . . . . .	320
5. Les stratégies d'analyse: le cas des corpus de réponses libres (ou de textes courts, nombreux, qualifiés) . . . . .	322
5.1. Les prétraitements: seuil de fréquence, lemmatisation . . . . .	323
5.2. L'analyse directe des réponses ou des documents . . . . .	325
5.3. AC à partir d'une juxtaposition de tableaux lexicaux . . . . .	336
5.4. L'analyse des correspondances sur tableau lexical agrégé . . . . .	339
5.5. Conclusion sur les stratégies d'analyse . . . . .	343
6. La fragmentation d'un corpus en « unités de contexte » . . . . .	344
6.1. Hypothèse sous-jacente . . . . .	345
6.2. Les avantages de la fragmentation du corpus . . . . .	346

6.3.	Une application. Variations autour de huit présidents : le corpus . . . . .	347
6.4.	L'analyse (supervisée) de la table présidents × lemmes . . . . .	347
6.5.	L'analyse (non supervisée) des lignes et des blocs de lignes . . . . .	348
6.6.	Un retour sur l'analyse supervisée des huit discours intégraux des présidents. . . . .	358
	Conclusion . . . . .	359
CHAPITRE 9		
	L'articulation entre les analyses exploratoires et confirmatoires . . . . .	361
1.	Explorer, valider, prévoir... . . . . .	361
1.1.	Des instruments d'observation ou des modèles? . . . . .	362
1.2.	Les difficultés de l'articulation exploration-inférence statistique . . . . .	363
1.3.	Plan du chapitre . . . . .	363
2.	La stylométrie et la discrimination globale. . . . .	364
2.1.	Discrimination à partir de la forme: la stylométrie. . . . .	365
2.2.	Discrimination globale: recherche de documents, codification. . . . .	365
3.	Les unités statistiques de la stylométrie . . . . .	366
3.1.	« Mots-outils », parties du discours. . . . .	366
3.2.	D'autres mesures discriminantes . . . . .	367
4.	Un exemple de modèle statistique en stylométrie . . . . .	368
4.1.	La gamme des fréquences. . . . .	368
4.2.	Les textes utilisés pour le problème d'attribution et les résultats du test . . . . .	369
4.3.	Une approche multidimensionnelle du problème . . . . .	371
5.	Les analyses discriminantes globales. . . . .	375
5.1.	Le principe de l'analyse discriminante . . . . .	375
5.2.	Les unités statistiques de la discrimination . . . . .	377
5.3.	La discrimination et les réponses modales. . . . .	377
5.4.	La discrimination régularisée par l'AC. . . . .	378
5.5.	La validation d'une discrimination . . . . .	379
6.	Discrimination et validation: un exemple . . . . .	380
6.1.	L'exemple et le problème . . . . .	380
6.2.	Phase descriptive et validation. . . . .	382
6.3.	Réalité des configurations . . . . .	385
6.4.	L'analyse discriminante et les matrices de confusion . . . . .	387
6.5.	Conclusion de la section 6. . . . .	391

7.	La discrimination et les réseaux de neurones . . . . .	391
7.1.	Un perceptron multicouche particulier . . . . .	393
7.2.	Un perceptron multicouche non supervisé . . . . .	396
7.3.	Deux réseaux : deux approches dans le cas de l'AC . . . . .	397
8.	Les recherches de thèmes ( <i>Topic Modeling</i> ) : un point de vue . . . . .	399
8.1.	Un aperçu du contenu des sonnets de Shakespeare . . . . .	400
8.2.	Six méthodes sélectionnées pour la recherche de thèmes . . . . .	403
8.3.	Des extraits de la liste des thèmes (extraits limités à deux « thèmes » par méthode) . . . . .	405
8.4.	Une synthèse des <i>thèmes</i> produits . . . . .	406
8.5.	Un rapprochement avec les thèmes a priori . . . . .	408
8.6.	Conclusion de la section 8 . . . . .	411
	Conclusion . . . . .	412
	Annexe I. Annexe technique du chapitre 9 . . . . .	415
	Annexe J. Corpus . . . . .	419
	Annexe K. Logiciels d'analyse des données textuelles . . . . .	425
	BIBLIOGRAPHIE . . . . .	441
	INDEX . . . . .	467