Visualisation, validation et sériation. Application à un corpus de textes médiévaux.

Fernande Dupuis[1], Ludovic Lebart[2]

[1] UQAM, Montréal (dupuis.fernande@uqam.ca)

[2] ENST, Paris (lebart@enst.fr)

Abstract

Principal axes methods (such as correspondence analysis [CA]) provide useful visualizations of high-dimensional data sets. In the context of historical textual data, these techniques produce planar maps highlighting the associations between graphemes and texts (paragraphs, chapters, full texts, authors). In a first step, we remind the reader that a simple technique of seriation (re-ordering the rows and columns of a table) is readily derived from the first CA axis. In a second step, we stress the important role played by bootstrap techniques to allow for valid statistical inferences in a context in which classical analytical approach is both unrealistic and analytically complex. A series of medieval French texts (12^{th}-13^{th} centuries), rich in spelling variants, exemplify the proposed approaches. A free software is available.

Keywords: correspondence analysis, visualization, seriation of medieval corpora.

Résumé

Les méthodes d'analyse en axes principaux telles que l'analyse des correspondances fournissent des outils de visualisation précieux. Dans le cadre des études de textes anciens, ces méthodes permettent, par exemple, de représenter les associations entre graphèmes et textes (paragraphes, chapitres, textes complets, auteurs) sur des cartes planes. Dans un premier temps, on rappelle qu'une méthode élémentaire de sériation (ré-ordonnancement des lignes et des colonnes d'une table lexicale) est un simple sous-produit de l'analyse des correspondances. Puis on insiste sur le rôle important joué par les techniques de validation issues du bootstrap dans un contexte où les inférences statistiques classiques sont impossibles. L'exposé se fait à partir d'exemples qui concernent un corpus de textes médiévaux (XIIe-XIIIe siècles). Le logiciel utilisé est librement accessible.

Mots-clés : analyse des correspondances, visualisation, sériation corpus médiévaux.

1. Introduction

Les études de textes anciens se heurtent à des difficultés bien connues des spécialistes, dues à plusieurs facteurs étroitement interdépendants : variation des graphies[1], médiation fondamentale des copistes, effets régionaux et temporels marqués[2], absence de corpus de référence et de normes, enfin, conséquence des traits précédents, difficulté d'utilisation systématique d'outils de traitement automatique de la langue[3].

[1] L'ensemble des facteurs dont il est ici question est traité dans Dees (1987). La variation graphique est un phénomène connu de la littérature médiévale. Dees (1987, p. 535) dans son inventaire de forme, cite pour a carte no. 4, 36 formes en opposition pour le morphème ce.

[2] Voir Morin (à paraître) pour une étude récente de ce facteur.

[3] Voir Dupuis et Lemieux (2006) sur cette question.

Nous allons montrer, à partir d'un corpus de textes médiévaux (section 2) que l'application à une table lexicale basique croisant textes et formes de surface de trois techniques interdépendantes : analyse des correspondances (section 3) sériation (section 4), et zones de confiances (section 5) permet, en restant très proche des textes de base, d'obtenir des observations assez fines et de tester des hypothèses élaborées.

2. Le corpus de textes

Le corpus est formé de 15 textes en vers composés aux XIIe et XIIIe siècles[45]. Il comprend 383 193 occurrences de 27 459 formes graphiques. La liste et les caractéristiques des textes figurent ci-dessous.

- Identificateur de la Base de français médiéval : stbrend ; Auteur : Benedeit ; Titre : Voyage de saint Brendan ; Date : début XIIème ; Ed. Sc. : I. Short, B. Merrilees ; Manchester University Press ; 1979 ; Domaine : religieux ; Genre : hagiographie ; Dialecte : anglo-normand ; 10829 mots.

- Ident : roland ; Anonyme ; Titre : Chanson de Roland ; vers 1100 ; Ed. Sc. : G. Moignet ; Bordas ; Collection : n/a ; 1969 ; Domaine : littéraire ; Genre : épique ; Dialecte : anglo-normand ; 29338 mots.

- Ident : gormont ; Anonyme ; Titre : Gormont et Isembart ; vers 1130 ; Ed. Sc. : A. Bayot ; Champion; 1931 ; Domaine : littéraire ; Genre : épique ; Dialecte : non défini (ajout = centre ou sud-ouest de Paris) ; 3815 mots.

Ident : louis ; Anonyme ; Titre : Couronnement de Louis ; vers 1130 ; Ed. Sc. : E. Langlois ; Champion ; 1925 ; Domaine : littéraire ; Genre : épique ; Dialecte : non défini ; 19786 mots.

- Ident : thebes ; Anonyme ; Titre : Roman de Thèbes ; vers 1150 ; Ed. Sc. : G. Raynaud de Lage ; Champion ; 1968 ; Domaine : littéraire ; Genre : roman ; Dialecte : non défini ; 62698 ; mots.

- Ident : thomas ; Auteur : Guernes de Pont-Sainte-Maxence ; Titre : Vie de saint Thomas ; 1172 - 1174 ; Ed. Sc. : E. Walberg ; Champion ; 1936 ; Domaine : religieux ; Genre : hagiographie ; Dialecte : non déf. ; 53947 mots.

- Ident : eracle ; Auteur : Gautier d'Arras ; Titre : Eracle ; vers 1176 1184 ; Ed. Sc. : G. Raynaud de Lage ; Champion ; 1976 ; Domaine : littéraire ; Genre : roman ; Dialecte : non défini ; 40839 mots.

- Ident : beroul ; Auteur : Béroul ; Titre : Tristan ; entre 1165 et 1200 ; Ed. Sc. : L. M. Defourques, E. Muret ; Champion ; 1947 ; Domaine : littéraire ; Genre : roman ; Dialecte : franco-picard ; 27257 mots.

- Ident : amiamil ; Anonyme ; Titre : Ami et Amile ; vers 1200 ; Ed. Sc. : P.F. Dembowski ; Champion ; 1969 ; Domaine : littéraire ; Genre : épique ; Dialecte : non défini ; 25283 mots.

- Ident : belinc ; Auteur : Renaut de Beaujeu ; Titre : Bel Inconnu ; avant 1214 ; Ed. Sc. : P. Williams ; Champion, 1929 ; Domaine : littéraire ; Genre : roman ; Dialecte : non défini ; 36692 mots.

- Ident : renart10 ; Anonyme ; Titre : Roman de Renart (branche X) ; début XIIIème ; Ed. Sc. : M. Roques ; Champion ; 1948-1963 ; Domaine : littéraire ; Genre : récits brefs ; Dialecte : non défini ; 13472 mots.

- Ident : renart11 ; Anonyme ; Titre : Roman de Renart (branche XI) ; début XIIIème ; Ed. Sc. : M. Roques ; Champion ; 1948-1963 ; Domaine : littéraire ; Genre : récits brefs ; Dialecte : non défini ; 8563 mots.

- Ident : Escoufle ; Auteur : Jean Renart ; Titre : Escoufle ; entre 1200 et 1202 ; Ed. Sc. : F. P. Sweester ; Droz ; Collection : TLF ; 1974 ; Domaine : littéraire ; Genre : roman ; Dialecte : picard ; 57967 ; mots.

- Ident : dole ; Auteur : Jean Renart ; Titre : Roman de la Rose ou de Guillaume de Dole ; 1210 ou 1228 ; Ed. Sc. : F. Lecoy ; Champion ; 1962 ; Domaine : littéraire ; Genre : roman ; Dialecte : non vérifié ; 34555 mots.

- Ident : vergy ; Anonyme ; Titre : Châtelaine de Vergy ; mi XIIIème, avant 1288 ; Ed. Sc. : G. Raynaud, L. Foulet ; Champion ; 1921 ; Domaine : littéraire ; Genre : roman ; Dialecte : non défini ; 6117. mots.

[4] Les textes du corpus proviennent de la Base de français médiéval constituée par Christiane Marchello-Nizia de l'ENS-LSH, Lyon.

[5] On remarque que la date de composition est quelquefois fort imprécise, se situant de quelques années à plusieurs dizaines d'années.

Ce corpus bien que réduit atteste de la diversité des genres de l'époque médiévale : hagiographie, épique, romann et récit. On notera en outre que les textes sont de longueur inégale, variant de 62698 mots à 6117 mots ce qui constitue souvent un problème quand on s'intéresse aux phénomènes de faible fréquence.

3. Visualisation par AC

La première étape est une analyse des correspondances de tables lexicales. Dans un premier temps, le seuil de fréquence minimale pour les formes a été de 40, ce qui a conduit à garder 941 mots totalisant 290 769 occurrences. Le tableau 1 de la section 3 ci-dessous présente les premières lignes de cette table lexicale. C'est effectivement cette analyse qui sera utilisée pour l'étape de sériation (section 3) mais les graphiques correspondant ne sont pas publiables dans le format du présent document, et donc le graphique de la figure 1 ci-après, déjà fort encombré (nombreux points superposés dans la zone circulaire) correspond au seuil de fréquence minimale de 200, ce qui laisse 227 mots totalisant 233 697 occurrences.

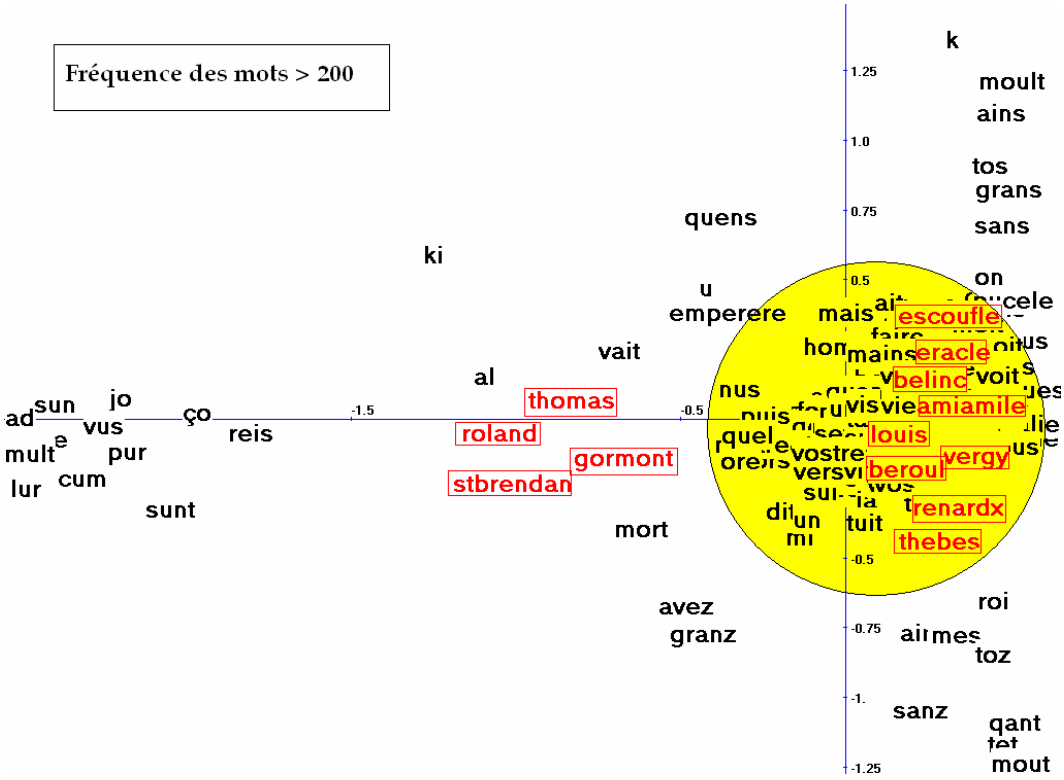

Figure 1. Analyse des correspondances de la table (227 x 15) correspondant au seuil 200.

Il est intéressant de noter que le pattern observé pour les textes est extrêmement voisin de celui obtenu avec la table (941 x 15) correspondant au seuil 40 Il en est de même pour la position des formes communes aux deux analyses. L'opposition constatée entre les graphies anglo-normandes (à gauche) et les autres nous permet de caractériser les copistes des quatre auteurs situés dans la partie gauche du premier plan factoriel[6].

[6] Cette procédure nous autorise à classer « gormont » sous le dialecte anglo-normand en conformité avec la localisation de Dees (1987) et de préciser la provenance dialectale de « thomas » pour lequel cette information manque dans les fiches descriptives de la BFM de Lyon.

4. La sériation

4.1. Principe général

Les techniques de sériation, comme les techniques de classification par bloc, sont largement utilisées par les praticiens. La référence la plus ancienne est probablement celle de l'égyptologue Petrie (1899). Les sériations sont fondées sur de simples permutations des lignes et des colonnes de la table étudiée et ont l'immense avantage pratique et cognitif de mettre l'utilisateur devant les données brutes, et donc de le dispenser d'utiliser des règles d'interprétation souvent délicates. Ces permutations peuvent faire apparaître des blocs homogènes des valeurs fortes, ou au contraire de valeurs faibles ou nulles. Elles peuvent aussi faire apparaître une évolution continue et progressive des profils. Il convient aussi de citer Bertin (1973) parmi les pionniers de ce type d'approche, avant les outils actuels de calcul électronique. Citons, parmi quelques références de base en classification automatique et en statistique les travaux de Hartigan (1972), de Arabie (et al., 1975), de Lerman (1972, 1981). Mais c'est dans un article de Hill (1974) que l'on trouve le résultat qui nous intéresse ici : les ordres des lignes et des colonnes suivant leurs coordonnées sur le premier axe d'une analyse des correspondances ont des propriétés optimales, pour réordonner simultanément les lignes et les colonnes de la table de nombre positifs soumise à l'analyse.

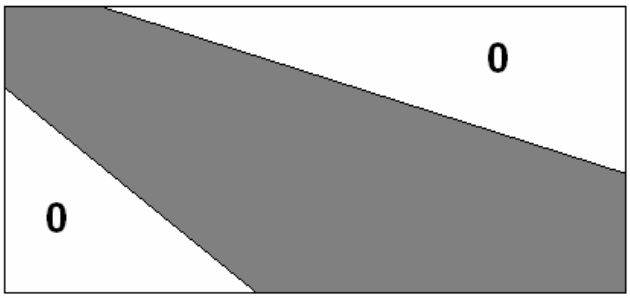

Figure 2. Une structure particulière du tableau de données
(zone grisée : éléments positifs ; zone blanche : éléments nuls)

Si le tableau de données initial **T**, après ré-ordonnancement des lignes et des colonnes, peut prendre la forme du tableau esquissé sur la figure 2 (en fait, les limites de la zone grisée ne sont pas forcément des droites) alors ce ré-ordonnancement est fourni par l'ordre des coordonnées des lignes et des colonnes selon le premier facteur (axe) de l'analyse des correspondances T.

[Cette propriété se démontre en utilisant le fait que la matrice à diagonaliser en AC est la matrice **L'C** produit de la matrice des profils-lignes **L'** par la matrice des profils-colonnes **C** (**L** et **C** ont la même structure que **T**). On montre alors que l'algorithme de la puissance itérée de Hotelling, qui fournit le premier vecteur propre, à partir d'un vecteur quelconque, en itérant des multiplications par **L'C**, permet de retrouver la structure cherchée].

4.2. Application au tableau lexical global

Tableau 1. 25 premières lignes (sur 941) de la table lexicale originale (15 textes en colonne)

	amia	beli	bero	dole	erac	esco	gorm	loui	rena	renx	rola	stbr	theb	thom	verg
a	719.	964.	727.	953.	1127.	1452.	69.	543.	514.	698.	392.	140.	830.	353.	180.
abat	2.	4.	1.	0.	1.	4.	4.	6.	1.	1.	25.	0.	5.	0.	0.
abes	0.	0.	0.	0.	0.	0.	0.	10.	0.	0.	0.	48.	0.	0.	0.
ad	0.	0.	0.	0.	0.	0.	26.	0.	0.	0.	442.	65.	0.	100.	0.
afaire	0.	6.	2.	1.	11.	47.	0.	0.	8.	4.	0.	0.	5.	0.	0.
ahi	5.	1.	5.	5.	7.	14.	3.	5.	2.	5.	1.	0.	0.	0.	0.
ai	44.	60.	70.	47.	71.	73.	6.	22.	39.	67.	47.	22.	26.	71.	23.

aidier	4.	17.	1.	3.	8.	7.	0.	31.	2.	4.	0.	0.	2.	0.	0.
aim	0.	6.	4.	7.	13.	9.	0.	0.	1.	4.	3.	0.	0.	5.	2.
aime	0.	1.	5.	17.	39.	17.	0.	2.	5.	2.	0.	0.	2.	9.	3.
ainc	0.	18.	0.	4.	35.	62.	0.	0.	0.	0.	0.	0.	0.	0.	3.
ains	1.	24.	0.	2.	79.	106.	0.	0.	0.	0.	0.	0.	0.	0.	0.
ainz	38.	2.	44.	55.	0.	0.	0.	30.	11.	32.	1.	20.	47.	2.	8.
ainçois	0.	1.	0.	18.	0.	0.	0.	0.	8.	5.	0.	0.	25.	0.	1.
aise	0.	0.	2.	3.	9.	17.	0.	1.	4.	7.	0.	1.	2.	4.	3.
ait	14.	20.	36.	23.	64.	51.	1.	21.	10.	9.	25.	4.	19.	30.	10.
al	0.	22.	0.	0.	63.	43.	29.	93.	0.	0.	92.	64.	6.	64.	0.
ala	6.	14.	6.	6.	3.	9.	1.	2.	3.	4.	0.	0.	2.	1.	2.
aler	22.	69.	7.	27.	7.	39.	0.	13.	10.	16.	15.	6.	24.	12.	4.
alez	0.	0.	13.	10.	0.	0.	0.	4.	8.	16.	10.	2.	11.	2.	2.
altre	0.	0.	0.	0.	0.	0.	0.	22.	0.	0.	62.	21.	0.	50.	0.
ame	1.	1.	1.	7.	25.	27.	0.	0.	8.	3.	0.	0.	6.	0.	5.
amer	14.	16.	2.	13.	17.	14.	0.	2.	1.	1.	6.	0.	5.	27.	6.
ami	72.	10.	8.	6.	7.	27.	1.	3.	1.	5.	10.	0.	4.	16.	3.
amie	4.	46.	20.	17.	20.	72.	0.	0.	1.	3.	1.	0.	9.	24.	14.

Tableau 2. Extrait de la même table dont les lignes et les colonnes ont été réordonnées selon le premier axe principal de l'AC (premières lignes de la table réordonnée)

	rola	stbr	thom	gorm	loui	bero	theb	beli	amia	rena	erac	rena	dole	esco	verg
respunt	49	8	5	0	0	0	0	0	0	0	0	0	0	0	0
mult	186	88	58	2	0	0	0	0	0	0	0	0	0	0	0
unt	104	52	21	8	0	0	0	0	0	0	0	0	0	0	0
lur	93	144	38	10	0	0	0	0	0	0	0	0	0	0	0
ad	442	65	100	26	0	0	0	0	0	0	0	0	0	0	0
tuz	42	38	22	2	0	0	0	0	0	0	0	0	0	0	0
tute	34	8	14	1	0	0	0	0	0	0	0	0	0	0	0
vunt	19	27	18	0	0	0	0	0	0	0	0	0	0	0	0
ben	97	1	40	3	0	0	0	0	0	0	0	0	0	0	0
fud	0	51	23	4	0	0	0	0	0	0	0	0	0	0	0
sun	231	38	130	40	0	0	0	0	0	0	0	0	0	0	0
tut	58	50	28	20	0	0	0	0	0	0	0	0	0	1	1
cum	46	83	63	10	0	0	0	2	0	0	1	0	0	0	0
od	48	44	30	5	0	0	0	0	0	0	4	0	0	0	0
mun	36	12	38	5	0	0	0	0	0	0	0	0	0	0	0
dunc	17	38	41	7	0	0	0	0	0	0	0	0	0	0	0
dunt	17	9	27	1	0	0	0	0	0	0	0	0	0	0	0
sur	77	31	29	24	0	0	0	1	0	0	2	0	0	0	0
e	1040	344	635	107	0	2	0	5	0	0	4	0	8	41	0
sei	21	12	20	1	1	1	0	0	0	0	0	0	0	0	0
nef	0	42	17	1	0	0	0	0	2	0	1	0	0	0	0
pur	94	83	238	20	0	0	0	0	0	0	0	1	5	3	0
vus	35	35	192	21	0	0	0	0	0	0	0	0	0	0	0

On observe de façon claire le vocabulaire exclusif des quatre textes classés en tête (*roland, stbrendan, thomas, gormont*) mais on note aussi des exceptions intéressantes dans la partie droite du tableau (colonne *escoufle* notamment). Selon Lejeune-Dehousse (1935) plusieurs copistes sont intervenus dans le manuscrit de l'Escoufle.

Tableau 3. 25 dernières lignes de la même table dont les lignes et les colonnes ont été réordonnées selon le premier axe principal de l'AC

	rola	stbr	thom	gorm	loui	bero	theb	beli	amia	rena	erac	rena	dole	esco	verg
cis	0	0	0	0	0	1	0	7	1	0	30	0	1	30	0
comment	0	0	0	0	5	0	0	0	0	6	8	8	0	75	8
tex	0	0	0	0	0	6	0	0	7	3	0	2	17	30	0
çou	0	0	0	0	0	0	0	2	0	0	71	0	0	22	0
velt	0	0	0	0	0	0	1	10	0	0	44	0	0	43	0
tans	0	0	0	0	0	0	0	14	6	3	11	5	0	67	0
maison	0	0	0	0	0	2	0	0	7	5	9	12	1	29	0
anui	0	0	0	0	0	0	2	9	0	7	17	9	11	27	3
afaire	0	0	0	0	0	2	5	6	0	8	11	4	1	47	0
ainc	0	0	0	0	0	0	0	18	0	0	35	0	4	62	3
ame	0	0	0	0	0	1	6	1	1	8	25	3	7	27	5
biax	0	0	0	0	0	0	5	0	18	5	0	2	8	71	0
ains	0	0	0	0	0	0	0	24	1	0	79	0	2	106	0
maniere	0	0	0	0	0	2	3	5	1	4	14	7	17	24	7
moult	0	0	0	0	0	0	0	0	160	0	0	0	0	512	0
damoisele	0	0	0	0	0	1	6	7	0	0	0	1	17	40	1
vallet	0	0	0	0	0	0	1	7	0	0	3	0	18	24	0
tous	0	0	0	0	0	0	0	0	0	0	81	0	0	62	0
ensamble	0	0	0	0	0	0	0	0	9	0	0	0	10	37	1
assés	0	0	0	0	0	0	0	0	0	0	32	0	0	35	0
ausi	0	0	0	0	0	0	2	5	0	3	1	2	12	31	5
lués	0	0	0	0	0	0	0	2	0	0	10	0	57	48	0
samblant	0	0	0	0	0	0	0	0	6	0	4	0	5	25	15
jou	0	0	0	0	0	0	0	0	0	0	32	1	0	134	0
comme	0	0	0	0	1	0	0	1	0	1	1	0	0	71	18

On note en bas du tableau ré-ordonné les mots absents chez les premiers auteurs, mais on voit qu'il existe aussi des situations intermédiaires formant un continuum le long du premier axe. La présence des chiffres bruts permet une interprétation beaucoup plus circonstanciée que la visualisation graphique des plans factoriels.

4.3. Application au tableau lexical amputé des quatre premiers textes/auteurs

Une fois détecté et analysé le principal facteur d'hétérogénéité (présence surtout chez quatre auteurs ou copistes de graphies particulières), il convient évidemment de ne pas s'arrêter là. La méthode la plus simple pour approfondir l'investigation consiste à éliminer les quatre auteurs fortement contributeurs au premier axe principal, et à recommencer une analyse sur la table (941 x 11) restante.

Le nouveau premier axe trouvé sur cette table réduite est, comme on pouvait s'y attendre, très voisin du second facteur de l'analyse globale. Mais la situation n'est pas toujours aussi caricaturale, et la suppression de textes ne permet pas, en général, de retrouver un axe connu.

On note sur les tableaux 4 et 5 que les formes concernées par les rangs extrêmes ne sont pas du tout les mêmes que celles des tableaux 2 et 3.

D'où un nouvel « épluchage » (*peeling* est le terme technique...) du tableau lexical, avec cette fois de nouvelles oppositions entre dialectes ou régions (d'auteurs ou de copistes).

Tableau 4. Premières lignes (sur 941) de la table lexicale (11 textes en colonne) dont les lignes et les colonnes ont été réordonnées selon le premier axe principal de l'AC opérée sur 11 textes.

	esco	erac	beli	verg	dole	loui	amia	rena	rena	bero	theb
k	266	6	8	0	0	0	0	0	1	0	0
jou	134	32	0	0	0	0	0	0	1	0	0
ki	184	2	19	0	0	0	0	0	0	0	0
contesse	47	1	0	0	2	0	0	0	0	0	0
assés	35	32	0	0	0	0	0	0	0	0	0
tous	62	81	0	0	0	0	0	0	0	0	0
ains	106	79	24	0	2	0	1	0	0	0	0
velt	43	44	10	0	0	0	0	0	0	0	1
cascuns	32	24	12	0	0	0	0	0	0	0	0
cis	30	30	7	0	1	0	1	0	0	1	0
ainc	62	35	18	3	4	0	0	0	0	0	0
cose	25	76	23	0	0	0	0	0	0	0	0
jamais	53	11	4	0	0	0	0	0	0	9	0
gens	99	25	17	0	1	0	15	0	0	0	0
quens	167	3	4	0	16	0	0	0	8	0	7
voel	24	37	0	0	8	0	0	0	0	0	0

Tableau 5. Dernières lignes (sur 941) de la table lexicale (11 textes en colonne) dont les lignes et les colonnes ont été réordonnées selon le premier axe principal de l'AC opérée sur 11 textes.

	esco	erac	beli	verg	dole	loui	amia	rena	rena	bero	theb
val	3	5	0	0	2	3	1	2	5	0	33
piez	0	0	0	3	7	28	0	22	15	16	26
foiz	0	0	0	3	9	0	0	20	17	16	14
pou	0	0	0	0	1	16	0	8	17	0	15
mout	0	3	0	26	297	0	0	129	104	0	299
granz	0	0	1	1	52	26	0	12	7	8	66
chascun	1	0	0	0	12	0	10	9	6	23	33
ainçois	0	0	1	1	18	0	0	8	5	0	25
conme	0	8	0	0	1	0	10	9	23	15	62
unne	11	0	0	0	0	0	11	0	0	0	70
dedenz	0	0	0	2	8	10	1	0	0	16	33
vet	0	0	0	0	36	0	0	0	1	31	61
filz	0	0	0	0	12	21	0	5	0	6	50
onc	0	0	0	0	10	21	0	1	0	2	42
touz	0	0	0	0	3	0	37	8	6	0	83
leur	4	0	3	0	0	0	0	0	1	0	146
y	0	0	0	0	0	0	0	0	0	0	62

Ici encore, les exceptions peuvent être intéressantes, qu'il s'agisse d'interpréter ou simplement d'apurer les documents de base.

5. Inférence statistique locale

Alors même que les cartes factorielles sont reconnues comme irremplaçables pour décrire à grands traits les principales structures d'associations dans les tables lexicales, leur rôle pour procéder à des inférences statistiques plus fines est moins connu. Le caractère suggestif de ces cartes leur est souvent reproché comme générateur de complaisance ou de laxisme dans les interprétations. Enfin et surtout, la précision sur la position des points est rarement prise en compte. Les zones de confiance mentionnées puis utilisées dans la présente section vont répondre à ces objections et donner un statut plus scientifique à ces visualisations.

5.1. Principe des zones de confiance "bootstrap"

On sait que la technique de *bootstrap* (cf. Efron et Tibshirani, 1993) permet de tracer des zones de confiance (en général : ellipses) autour des points représentés sur les plans principaux, que ces points représentent des mots ou des textes. La méthode consiste à construire *n* "réplications" de l'échantillon par tirage *avec remise* des unités statistiques, qui sont ici les occurrences de formes. Dans une réplication, certaines unités apparaîtront ainsi deux fois ou plus, d'autres n'apparaîtront pas.

On crée par ces tirages une variabilité autour du tableau de données de départ. Sous des hypothèses faibles, on montre que la variabilité observée sur les *n* réplications est de l'ordre de grandeur de celle que l'on aurait observée dans la population. Autrement dit, on va pouvoir disposer de *n* réplications de paramètres complexes (comme les vecteurs propres, et donc les coordonnées factorielles) et obtenir à partir de ces réplications des intervalles de confiance pour ces paramètres.

5.2. Cas des composantes principales et de la SVD

5.2.1. Bootstrap total

On relève plusieurs variantes de la méthode : le *bootstrap total* consiste à refaire une analyse complète pour chaque réplication. Mais les axes répliqués ne sont pas forcément homologues d'une réplication à une autre, il peut y avoir des interversions d'axes, voir des rotations. Il faut alors faire coïncider par des techniques d'*analyses procustéennes* les axes homologues.

5.2.2. Bootstrap partiel

Le *bootstrap partiel* permet de lever cette difficulté. Il part de la constatation que le tableau initial est plus proche de la réalité observée que tous les tableaux répliqués, qui en sont des perturbations. L'analyse et les plans principaux de ce tableau initial servent de référence pour la projection de tous les tableaux répliqués (lignes et colonnes) en tant qu'éléments supplémentaires. Des expérience intensives (cf. Lebart *et al.,* 2006) ont montré l'efficacité de cette méthode.

5.3. Cas de l'analyse des correspondances et des tables lexicales

Rappelons que dans le cas d'une table lexicale, la technique consiste à tirer avec remise les occurrences de formes retenues. Ce tirage se fait selon un schéma multinomial comportant autant de catégories que la table a de cellules, et dont les fréquences théoriques sont celles des

cellules. Les lignes et colonnes des tables répliquées sont alors projetées comme éléments supplémentaires sur les plans factoriels de l'analyse de la vraie table lexicale (*bootstrap partiel*). Des ACP des « nuages de réplications » correspondant à chaque point-élément (ligne ou colonne) fournissent les ellipses de confiance cherchées.

5.4. Deux exemples de zones de confiance

On veut montrer par les deux exemples qui vont suivre que les analyses exploratoires de tables lexicales ne servent pas seulement à dégager de grands traits structuraux, mais qu'elles permettent aussi de procéder à des *focus* précis, à tester des hypothèses spécifiques. L'analyse factorielle de base sera celle opérée sur les 941 mots et les 15 textes avec le seuil de 40.

La figure 3 nous montre par exemple les positions des différents démonstratifs et de leurs graphies dans le premier plan. Les zones de confiances permettent maintenant de codifier l'interprétation de ces positions. Si la graphie « *cels* » est bien caractéristique du groupe « anglo-normand », on voit que certaines positions sont indiscernables (groupe : *cest, cel, ceste* ; groupe : *cele, celui, cestui*).

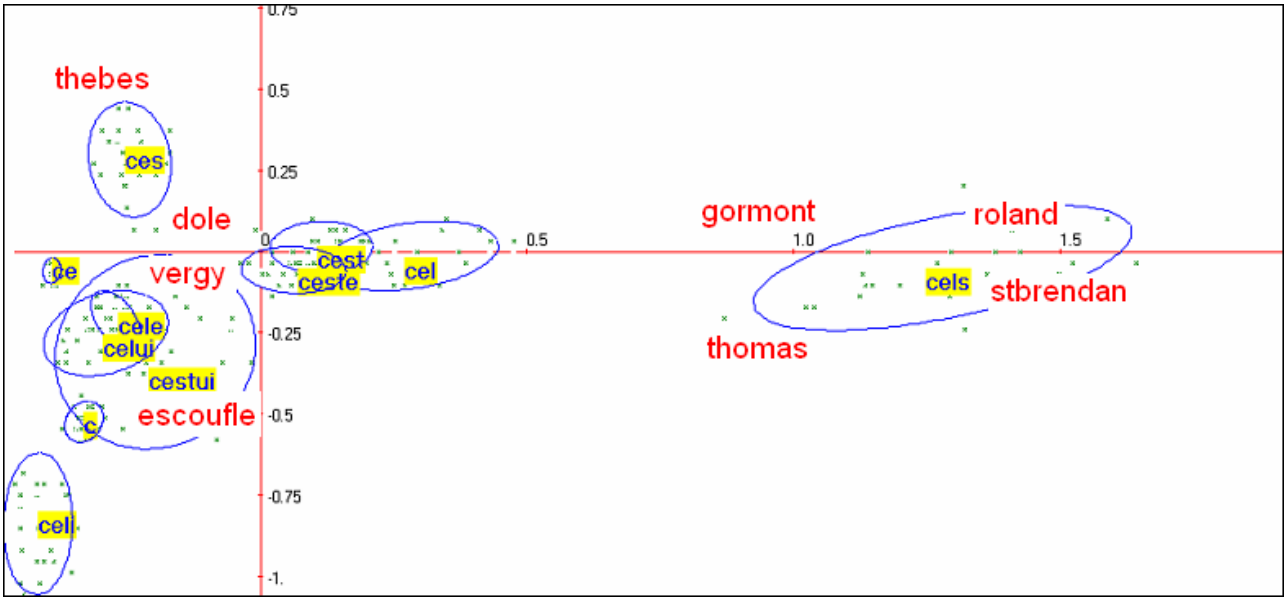

Figure 3. Zones de confiance de différents démonstratifs dans le plan factoriel (dont l'orientation est inversée par rapport à celui de la figure 1)

En revanche, l'opposition entre « *cel* » et « *ces* » le long de l'axe vertical peut être interprétée en termes de répartitions privilégiées chez certains auteurs.

On note à propos de l'opposition « cels / cel , ces » que « cels » est toujours pronom régime pluriel en anglo- normand (où il s'oppose à « celi », pronom régime singulier).

« cel » est très majoritairement déterminant régime singulier et s'oppose à « ces », déterminant régime pluriel.

La figure 4 montre (sur un nombre plus restreint de textes, après regroupements des graphies, des élisions...) que les répartitions des « et » et des « que » en début de vers sont significativement distinctes des répartitions ailleurs dans les vers (ellipses clairement disjointes). Compte tenu des effectifs important mis en jeu, il est ainsi possible d'interpréter des distances qui auraient pu être jugées *a priori* peu importantes.

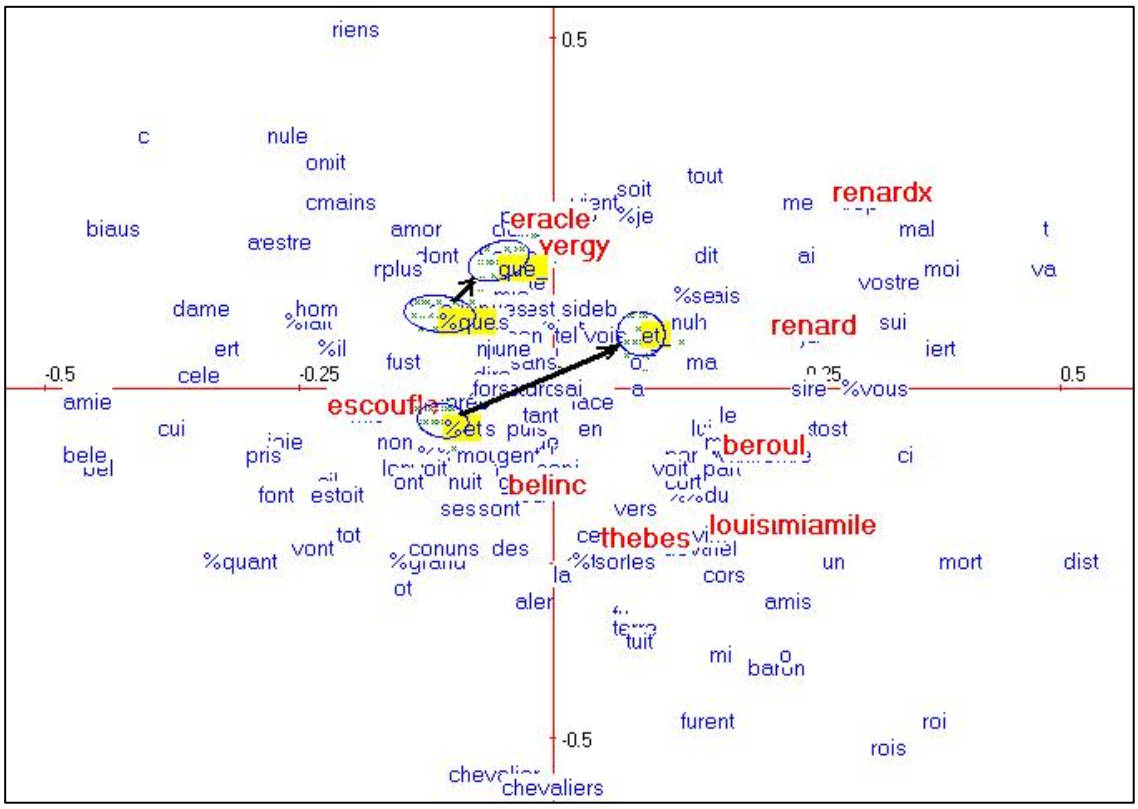

Figure 4. Comparaison des mots « que » et « et » occupant une position quelconque (base des flèches) et des mêmes mots positionnés en début de vers (extrémité des flèches).

Cette répartition significative pourrait nous permettre de mesurer l'effet structurant de ces éléments lexicaux et servir à étoffer l'hypothèse selon laquelle on passe en cours d'évolution de la subordination parataxique (par simple juxtaposition) à hypotaxique (avec conjonction, « que » par ex.).[7]

Une étude exhaustive des formes de « et » dans le corpus montre en effet que la position de cet élément induit des comportements syntaxiques distincts. On sait que cette conjonction cumule au moins deux rôles dans la grammaire du français médiéval. « et » coordonne par addition des syntagmes (nominal, verbal, etc.) ou des phrases. En outre, « et » peut, (Moignet 1973, pp 330-331) figurer en tête d'énoncé sans constituer une coordination syntaxique avec l'énoncé précédent. Certains grammairiens parlent alors d'éléments à valeur purement discursive. L'analyse du corpus révèle une différence syntaxique entre « et » de début de vers « et » ailleurs dans le vers. Le corpus contient plus de 13000 « et » dont 5300 en début de vers. La quasi totalité des « et » qui figurent à l'intérieur du vers sont du premier type :

> Vos li durrez urs **e** leons e chens,
> Set cenz camelz **e** mil hosturs muers,
> D'or **e** d'argent .IIII.C. muls cargez *La chanson de Roland*

En revanche, dans les cas où « et » début de vers joue le rôle de marqueur discursif, environs 400 constructions s'apparentent aux exemples suivants :

> Tristran l' entent, fist un sospir
> **Et dist : "Roïne de parage,**

[7] Voir Moignet (1973) p. 367.

Tornon ariere a l' ermitage;/ *Tristan*

Et dist li rois : "De gréz et volentiers,/ *Ami et Amile*

Avrum nos la victorie del champ ?"
E cil respunt : "Morz estes, Baligant ! *La chanson de Roland*

Cette syntaxe particulière où « et » marqueur discursif précède des verbes déclaratifs comme *dire* ou *répondre* pour introduire le style direct constitue une des caractéristiques de « et » début de vers.[8] Ces exemples illustrent la subordination parataxique où l'argument implicite du verbe du premier énoncé, l'énoncé citant, est situé dans l'énoncé suivant, l'énoncé cité. Les deux énoncés sont considérés comme syntaxiquement indépendants.

En cours d'évolution du français, durant la période médiévale, on passe graduellement de la subordination parataxique du style direct dans *La chanson de Roland* (début du XIIe siècle) à la subordination explicite en *que* du style indirect.

Et qui li dist : "Fole, demeure.
Vels tu hounir tot ton lignage? *Escoufle*

Et li cuens dist **qu**' a tous donroit/
Reubes, chevax, cels qui n' en orent. *Escoufle*

Et dist li quens **qu'** il se departent *Escoufle*

Se li a le castel mostré.
Por l'esgarder sont aresté
Et dient **que** bials est et gens,
Millor n'en ot ne rois ne quens *Bel Inconnu*

Se li demande qu'el fera.
Et dist **que** ele s'en ira Bel Inconnu *Bel Inconnu*

On le voit dans les exemples ci-dessus, le style direct alterne avec la subordination explicite dans l'Escoufle et le Bel Inconnu, romans du XIIIe siècle.

6. Conclusion

Cette recherche menée sur un corpus homogène (vers octosyllabique ou décasyllabique) montre, à partir des exemples ponctuels présentés, la possibilité de caractériser des traits syntaxiques sans catégorisation préalable, et donc d'explorer avec profit des textes faiblement enrichis. Cette exploration n'est ni intuitive, ni impressionniste. Elle se fonde sur deux adjuvants précieux de l'analyse des correspondances de tables lexicales : La sériation qui remet sous les yeux du chercheur les chiffres bruts originaux dans un contexte où ils prennent plus de signification ; les zones de confiance qui permettent d'extraire des *patterns* valides et de rejeter des proximités illusoires.

On a pu noter en passant que l'on observait peu de variation intratextuelle dans les hautes fréquences. Ce type d'analyse montre la nécessité de documenter les caractéristiques externes des textes (manuscrits, copistes, localisation ...).

[8] À 8 exceptions près.

On peut enfin espérer que notre approche permettra d'allier des méthodes statistiques de bon niveau aux analyses variationnistes utilisées en théorie du changement linguistique des dernières décennies.[9]

Elle devrait permettre, dans une étape ultérieure, de mettre en lumière des différences typologique, par exemple des phénomènes dont l'évolution diffère selon le genre.

Toutes les procédures de calcul et de tracé graphique sont implémentées dans le logiciel académique DTM qui peut être librement téléchargé à partir du site www.lebart.org.

Les vérifications des constructions dans les textes ont été effectuées avec le logiciel SATO (site www.ling.uqam.ca/ato).

Bibliographie

Benzécri J.-P. & collaborateur (1981). *Pratique de l'analyse des données,* tome 3, Linguistique & Lexicologie, Dunod, Paris.

Bertin J. (1973). *La graphique et le traitement graphique de l'information*. Flammarion, Paris.

Dees, A. (1987). *Atlas des formes linguistiques des textes littéraires de l'ancien français.* Max Niemeyer Verlag, Tübingen.

Dupuis F. et Lemieux M. (2006). Vérification d'hypothèse(s) et choix de corpus. *À la quête du sens.* ENS Éditions, Paris.

Dupuis F., Lemieux M. and Gosselin D. (1993). Conséquences de la sous-spécification des traits de Agr dans l'identification de pro. In *Language Change and Variation*. Vol. 3 no. 3, pp. 275-299.

Efron B., Tibshirani R. J. (1993). *An Introduction to the Bootstrap*. Chapman and Hall, New York.

Hartigan J. A. (1972). Direct clustering of a data matrix. *Journal of the American Statistical Association*. 6, 123-129.

Hill M. O. (1974). Correspondence analysis : a neglected multivariate method. *Applied Statistics*, 23, 340-354.

Lebart L., Salem A., Berry E. (1998). *Exploring Textual Data*, Kluwer Ac. Publisher, Dordrecht.

Lebart L., Piron M., Morineau A. (2006). *Statistique Exploratoire Multidimensionnelle*, *Visualisation et Inférence en Fouille de Données.* Dunod, Paris.

Lejeune-Dehousse R. (1935). *L'oeuvre de Jean Renard : Contribution à l'étude du genre romanesque au Moyen Age.* Genève, Slatkine Reprint.

Lerman I. C. (1972). Analyse de phénomène de la sériation. *Mathématique et Sciences Humaines*, 38, 39-57.

Lerman I. C. (1981). *Classification et Analyse Ordinale des Données*. Dunod, Paris.

Marchello-Nizia C. (2006). From personal to spatial deixis : the semantic evolution of demonstratives from Latin to French**.** In M. Hickman et S. Robert eds., *Space in languages, linguistic systems and cognitive categories*, Amsterdam, Benjamins Publishing Company : chapitre 5.

Moignet G. (1973). *Grammaire de l'ancien français.* Klincksieck, Paris.

Morin Y.-C. (à paraître). Histoire du corpus d'Amsterdam : le traitement des données dialectales. *Le Nouveau Corpus d'Amsterdam, Actes de l'atelier de Lauterbad*, Stuttgart. Steiner.

Petrie W. M. F. (1899). Sequence in prehistoric remains. *Journal of the Anthropological Instituted of Great Britain and Ireland*. 29, 295-301.

[9] Pour une illustration de cette approche, voir Dupuis *et al* (1993).