

*CFIES- 2010, ULB, Septembre 2010*

---

# **Statistique Textuelle et Histoire**

Ludovic Lebart,  
*Telecom-ParisTech, Paris, France.*  
*ludovic@lebart.org*

---

# Statistique textuelle et Histoire

---

- 1. Rencontres de la statistique et du texte
  - 2. Quelques outils statistiques
  - 3. Stylométrie, Attribution d'auteurs
  - 4. Exemples de corpus historiques
  - 5. Exemple de corpus médiéval
  - 6. Conclusion
-

# 1. Rencontres de la statistique et du texte

Les distributions lexicales initialement découvertes comme des lois empiriques devant permettre des améliorations dans le domaine de la transcription Sténographique .

Estoup J. B. (1916) - *Gammes sténographiques*, 4<sup>ème</sup> Edition, Paris.

Elles sont par la suite étudiées sous le signe de la "psycho-biologie du langage" par *G.K. Zipf (1935)*.

Zipf G. K. (1935) - *The Psychobiology of Language, an Introduction to Dynamic Philology*, Boston, Houghton-Mifflin.

Travaux de pionnier de Mandelbrot:

Mandelbrot B. (1968) - Les constantes chiffrées du discours, *Le Langage*, Encyclopédie de la Pléiade, vol XXV, Gallimard, Paris.

# 1. Rencontres de la statistique et du texte (suite)

Dans un second temps, la **statistique lexicale** (G. U. Yule, P. Guiraud puis Ch. Muller) entreprend de résoudre une série de problèmes et d'études comparatives sur le vocabulaire des "grands auteurs" .

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

Guiraud P. (1954) - *Les caractères statistiques du vocabulaire*, P.U.F., Paris.

Guiraud P. (1960) - *Problèmes et méthodes de la statistique linguistique*, P.U.F., Paris.

et en particulier des auteurs du théâtre classique français du 17<sup>e</sup> siècle.

Muller C. (1964) - *Essai de statistique lexicale : L'illusion comique de P. Corneille*, Klincksieck, Paris.

Muller C. (1968) - *Initiation à la statistique linguistique*, Larousse, Paris.

# 1. Rencontres de la statistique et du texte (suite)

Parallèlement, les méthodes développées dans ce même cadre seront présentées par G. Herdan, sous le nom de **linguistique statistique**, comme "la quantification de la théorie saussurienne du langage".

Herdan G. (1964) - *Quantitative Linguistics*, Londres, Butterworths.

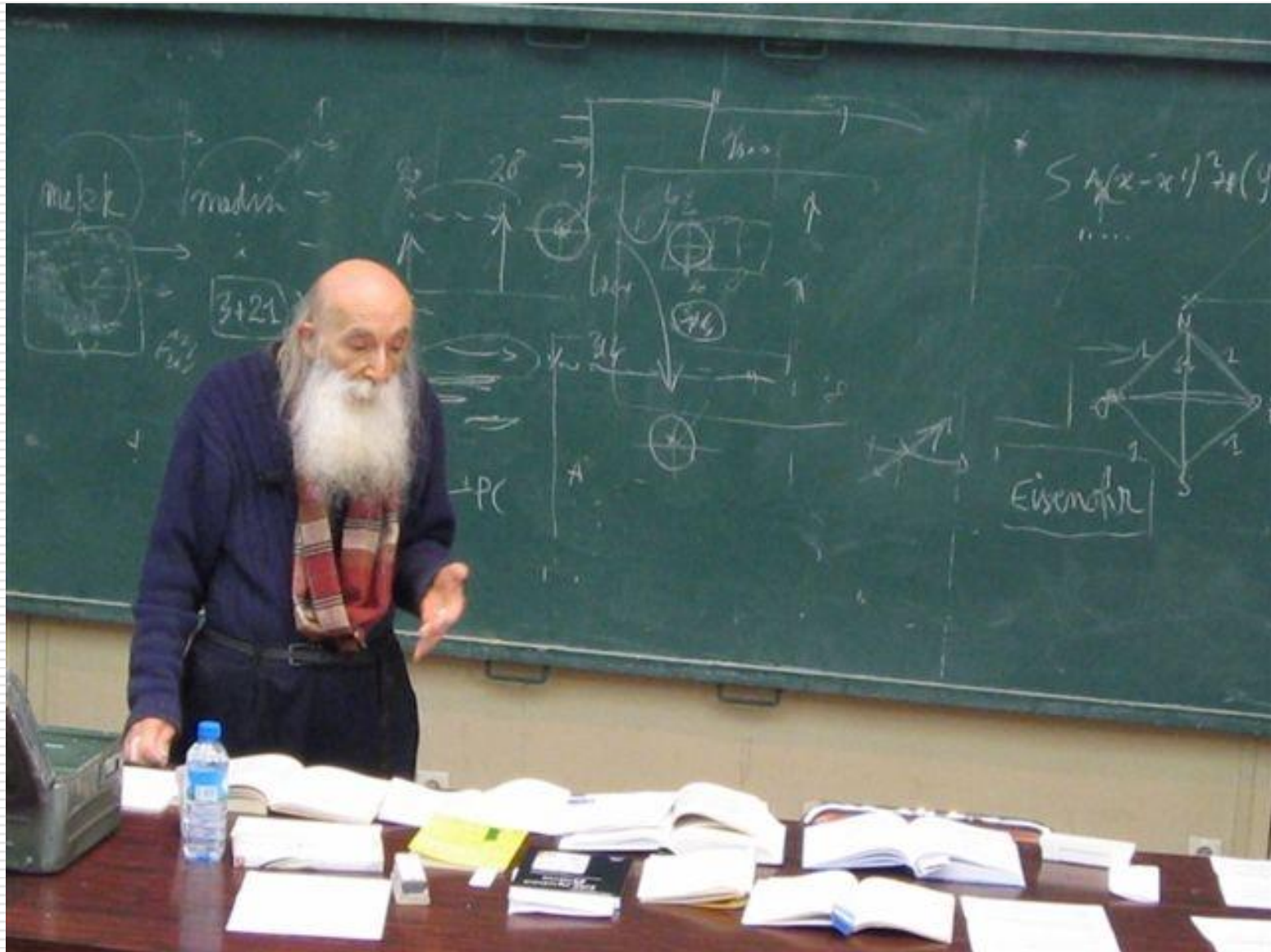
Selon Herdan (1964) cette discipline se présente comme une branche de la linguistique structurale, avec pour principale fonction la description statistique du fonctionnement (dans des corpus de textes) des unités définies par le linguiste.

Enfin dans les années 60 apparaissait l'analyse multidimensionnelle lexicale.

Benzécri J.-P.(1964) - *Cours de linguistique mathématique*, Faculté des Sciences de Rennes.

Benzécri J.-P. & coll. (1973) - *La taxinomie*, Vol. I ; *L'analyse des correspondances*, Vol. II, Dunod, Paris.

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



# 1. Rencontres de la statistique et du texte (suite et fin)

## Des banques de données textuelles... à la linguistique de Corpus

A une époque plus récente, la statistique textuelle, délaissant les dépouillements expérimentaux réalisés manuellement, s'oriente vers des comparaisons portant sur de plus vastes ensembles de textes.

E. Brunet (1981) réalise à partir des données du Trésor de la langue française de telles études comparatives.

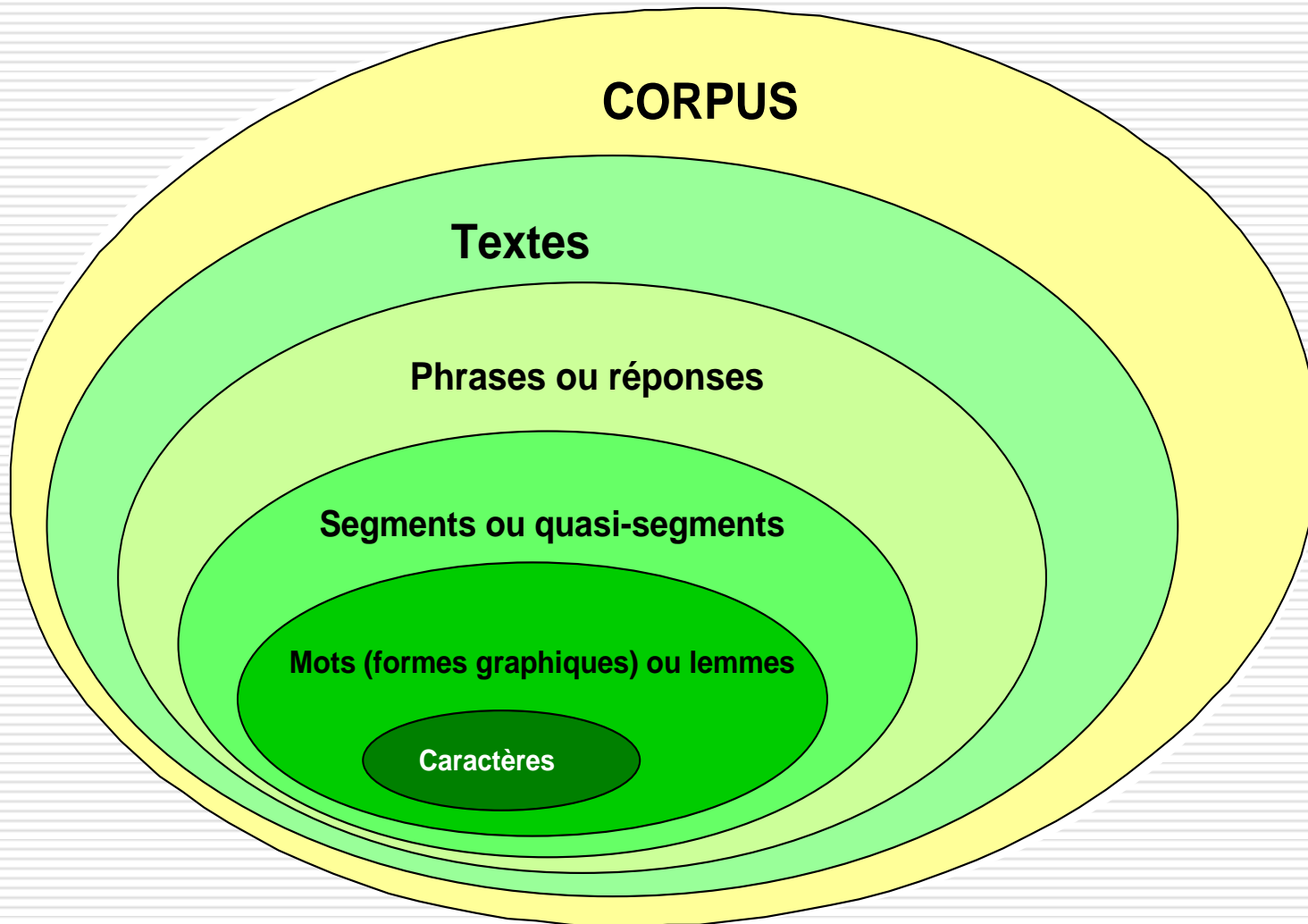
L'équipe du *Trésor général des langues et parlers français* gère ainsi, pour les besoins d'une communauté scientifique qui s'élargit chaque jour, un ensemble de textes qui compte désormais plus de 160 millions d'occurrences pour ce qui concerne les 19e et 20e siècles (FRANTEX).

Le BNC (British National Corpus) contient ainsi plus de 100 millions de mots étiquetés.

Les années récentes ont vu s'étendre de façon spectaculaire ces **banques** devenus **corpus**, avec apparition du paradigme de la **linguistique de Corpus**

## 2. Quelques outils statistiques

Unités Statistiques extraites des textes



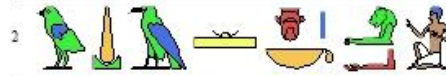
Exemple de texte  
Hiéroglyphique

(André Salem,  
Romuald Schummer)

*Le conte du naufragé*



I10:D46-M17-N35-T18-G43-A1-M17-N29:D21:Y1



G43-U28-G1-Y1-F34\*Z1:V31-F4:D36-A1



G17-D36:V31-F22:D54-N35:N35:Z2-F26:N35-W24-G43-O1



O42:Q3-D40-J1:D21-Q3\*Z7:M3-V28-A25-A24-Y5:N35-M17-X1-P11-M3



F4:X1\*X1-Z7-D21:D36:X1-D2:Z1-N16:Z1\*N23



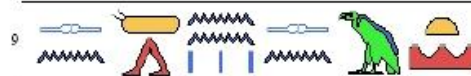
D21:D36-V28-V31:N35-W24:Z7-A2-R8-N14-A30-A2-O34:A1\*Z1-V30-D2:Z1-V28-Q3:X1-D32:D36-T22-N35:W24-G43-A1-Z4:I9



M40-G43:X1-A1:Z2-N35:Z2-M18-M17-X1:D54-V26:D46-X1:Y1-D35:N35-N35:O4-G43-G37:N35-A12-A1:Z2-N35:Z2



F22:D54-N35:N35:Z2-F22-G43-Z4-V4-V4-X1:N25



O34:N35-X5:D54-N35:N35:Z2-O34:N35-G14-X1:N25



G17-D36:V31-D21:I9-N35:Z2-M18-M17-D54-N35:Z2-G17-R4:X1-Q3:Y1

21



..... car·c'est·fatigant·de·te·parler··Laisse-moi·donc·te·raconter¶¶

22



..... quelque·chose·de·semblable·qui·m'est·arrivé, #

125



..... Laisse-moi·donc·te·raconter·quelque·chose·de·semblable·qui·est¶¶

..... arrivé·sur·cette·île¶¶

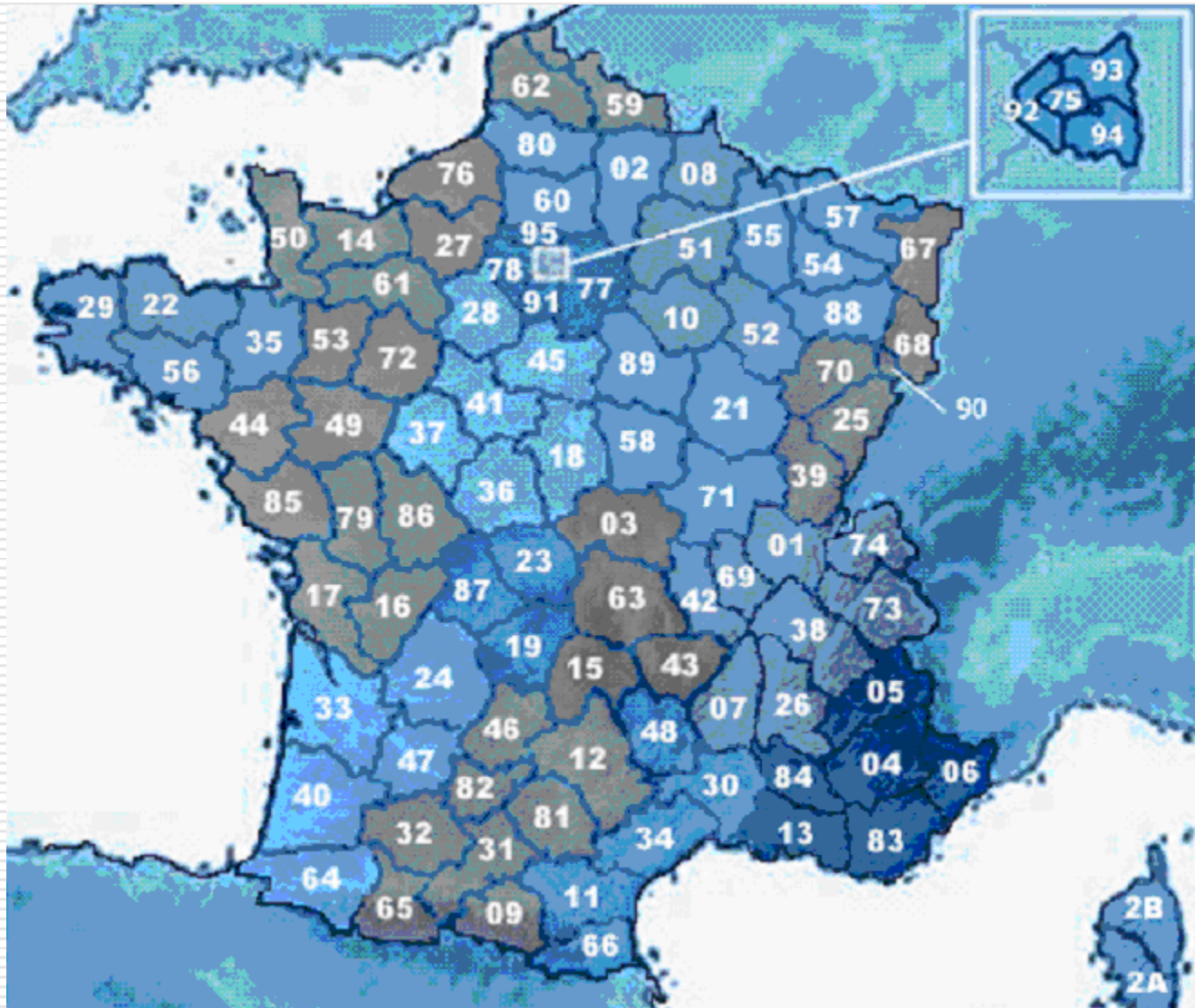
#

⌘

⌘

## 2. Quelques outils statistiques (suite)

Description de données structurées sous forme de chaînes de caractères



\*\*\*\* *Ain*  
Ain Isere Jura Rhone Hte\_Saone Savoie Hte\_Savoie

\*\*\*\* *Aisne*  
Aisne Ardennes Marne Nord Oise Seine\_Marne Somme

\*\*\*\* *Allier*  
Allier Cher Creuse Loire Nièvre Puy\_de\_Dome Hte\_Saone

\*\*\*\* *Alpes\_Prov*  
Alpes\_Prov Alpes\_Hautes Alpes\_Marit Drome Var Vaucluse

\*\*\*\* *Alpes\_Hautes*  
Alpes\_Hautes Alpes\_Prov Drome Isere Savoie

\*\*\*\* *Alpes\_Marit*  
Alpes\_Marit Alpes\_Prov Var

\*\*\*\* *Ardeche*  
Ardeche Drome Gard Loire Hte\_Loire Lozere

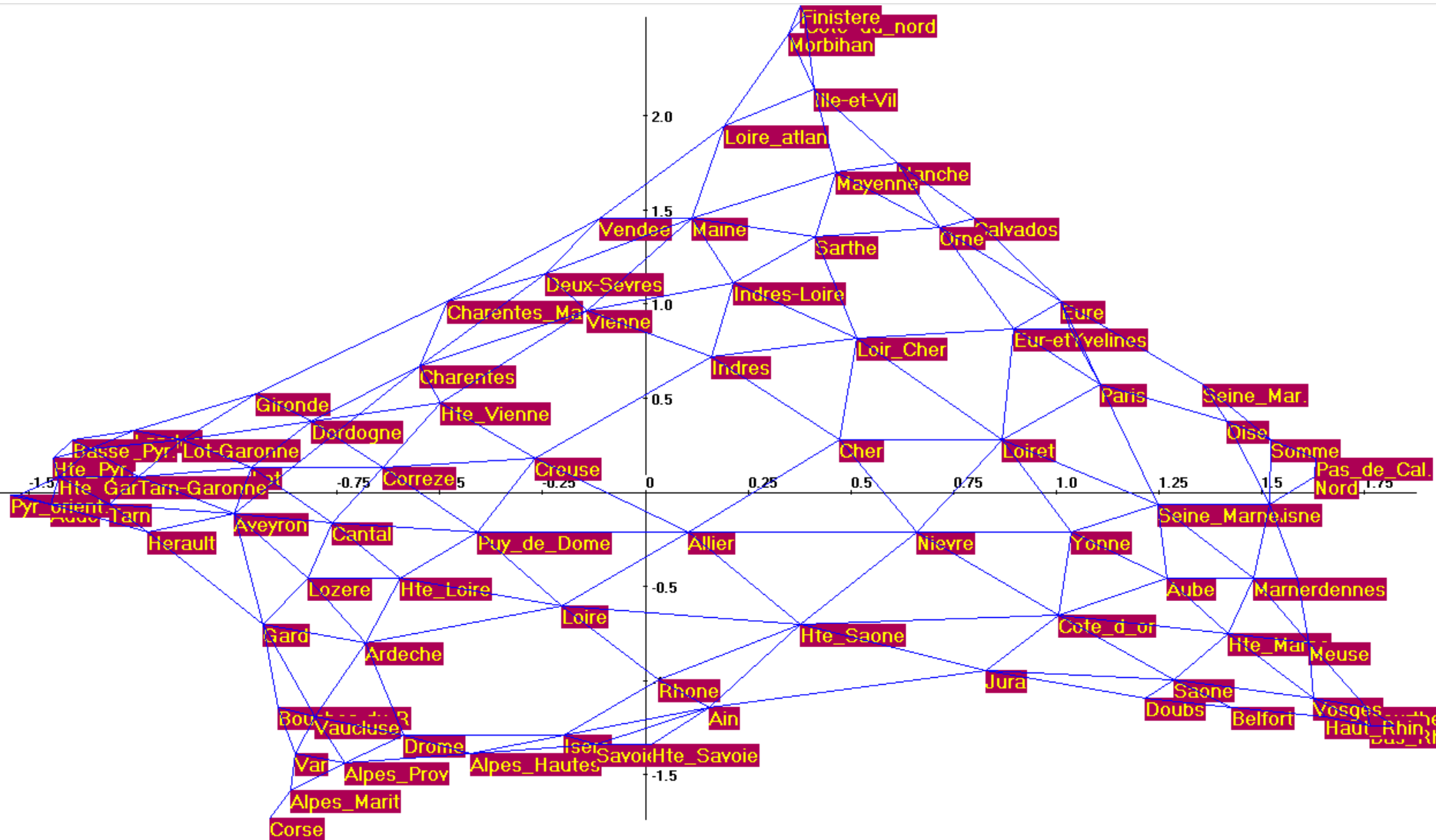
\*\*\*\* *Ardennes*  
Ardennes Aisne Marne Meuse

\*\*\*\* *Ariege*  
Ariege Aude Hte\_Garonne Pyr\_orientales.

.....



## 2. Quelques outils statistiques (suite)



## 2. Quelques outils statistiques (*suite*)

### Exemples d'indices mesurant la richesse du vocabulaire

$R = V / T$  (*type -token ratio*) ou encore  $R' = \text{Log}V / \text{Log}T$ .

McKinnon et Webster (1971) (pour une langue donnée,  $V$  est borné supérieurement, alors que  $T$  n'a pas de limite a priori).

L'indice  $D$  de Simpson (1949) est le quotient :

$$D = \sum_r r(r-1)V_r / T(T-1)$$

où  $V_r$  est le nombre de formes distinctes apparaissant exactement  $r$  fois dans le texte.

C'est donc le quotient du nombre de paires d'occurrences d'une même forme par le nombre total de paires d'occurrences.

*[Probabilité que 2 occurrences prises au hasard proviennent d'un même mot]*

$$K = 10^4 D (T-1)/T \quad (\text{Yule})$$

### 3. La stylométrie : Attribution d'auteurs

---

#### Des pionniers: Mosteller et Wallace

Travail de Mosteller et Wallace (1964) sur l'attribution des "Federalist papers". Parmi ces 77 textes politiques, publiés anonymement à New York à la fin du dix-huitième siècle, 12 textes n'ont pu être attribués à un auteur parmi deux possibles (Alexander Hamilton et James Madison).

Le traitement statistique permet de désigner l'auteur le plus probable de ces 12 textes litigieux.

On pourra aussi consulter sur un thème voisin un travail plus récent de Holmes (1992) sur l'homogénéité des "Mormon Scriptures", et une revue assez complète de ce même auteur sur les méthodes de la *stylométrie* (1995).

---



## Modèle statistique de « Capture » pour les fréquences

### Valeurs théoriques et observées pour le poème *Taylor*

<i>Fréquences</i>	<i>Taylor observé</i>	<i>Taylor théorique</i>
0	9	6.97
1	7	4.21
2	5	3.33
3-4	8	5.36
5-9	11	10.24
10-19	10	13.96
20-29	21	10.77
30-39	16	8.87
40-59	18	13.77
60-79	8	9.99
80-99	5	7.48

**Distribution des formes dans huit poèmes, selon leur fréquence d'apparition dans l'oeuvre de Shakespeare.**

<i>Freq.</i>	<b>BJon</b>	<b>Marl</b>	<b>Donn</b>	<b>Cymb</b>	<b>Mids</b>	<b>Phoe</b>	<b>Sonn</b>	<b>Tayl</b>	<i>Total</i>
0	8	10	17	7	1	14	7	9	73
1	2	8	5	4	4	5	8	7	43
2	1	8	6	3	0	5	1	5	29
3-4	6	16	5	5	3	9	5	8	57
5-9	9	22	12	13	9	8	16	11	100
10-19	9	20	17	17	6	18	14	10	111
20-29	12	13	14	9	9	13	12	21	103
30-39	12	9	6	12	4	7	13	16	79
40-59	13	14	12	17	5	13	12	18	104
60-79	10	9	3	4	9	8	13	8	64
80-99	13	5	10	4	3	5	8	5	53
+100	148	138	145	120	103	111	155	140	1060
<b>Total</b>	<b>243</b>	<b>272</b>	<b>252</b>	<b>215</b>	<b>156</b>	<b>216</b>	<b>264</b>	<b>258</b>	<b>1876</b>

**Les huit poèmes élisabéthains :**

**Ben Jonson**

An Elegy

**C. Marlowe** Four poems

**J. Donne**

The Ecstasy -----

**Shakespeare** Cymbeline (extraits)

**Shakespeare**

A Midsummer Night's Dream **Shakespeare** The Phoenix and Turtle

**Shakespeare**

Sonnets (extraits)

**Shakespeare** (?) Taylor's Poem

## Autres approches du même problème

Le  $\chi^2$  (test d'indépendance classique sur les tables de contingence), calculé (malgré la faiblesse de certains effectifs), vaut, pour 77 degrés de liberté ( $77 = [8 - 1] \times [12 - 1]$ ), et avec les notations usuelles :

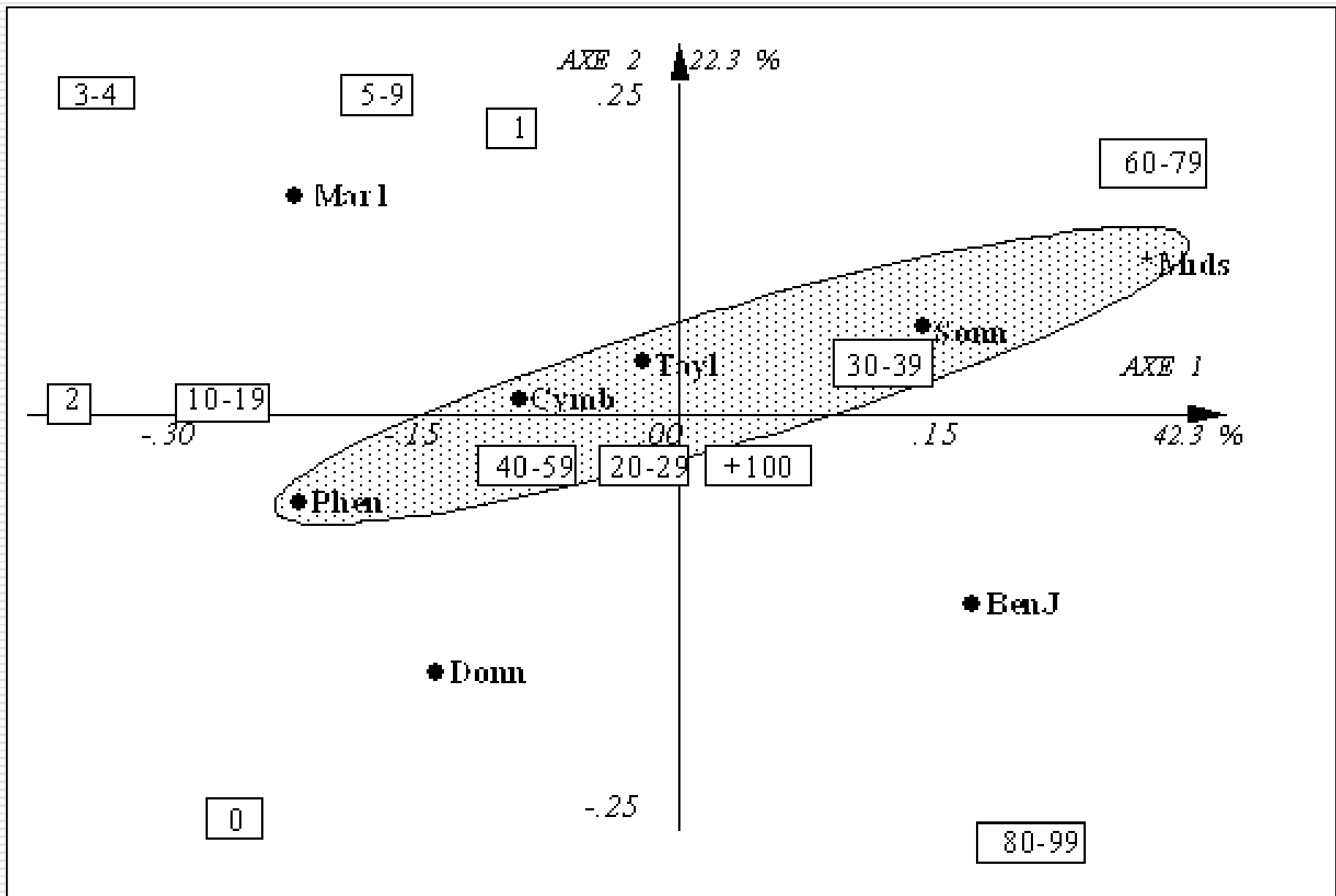
$$\chi^2 = 1876 \times \sum_{ij} (f_{ij} - f_{i.} f_{.j})^2 / f_{i.} f_{.j} = 104.7$$

Cette valeur a sensiblement une chance sur 100 d'être dépassée dans l'hypothèse d'homogénéité, qui correspond ici à l'indépendance des lignes et des colonnes de la table.

Cette hypothèse est donc douteuse.

L'un des mérites de l'analyse des correspondances est de nous décrire, dans le cas d'un rejet de l'hypothèse d'indépendance, comment cette hypothèse est rejetée.

# Visualisations de la table de contingence Auteurs x Fréquences



# 4. Exemples de corpus historiques

---

## **Quelques corpus emblématiques :**

Les discours de l'Etat de l'Union.  
(Présidents des Etats-Unis de Washington à G.W. Bush)

Les séries des discours de Fidel Castro, de Mao Tsé Toung,  
de Charles De Gaulle, de François Mitterrand, de Francisco Franco.

Un corpus littéraire : « Le trésor de la langue Française (TLF).

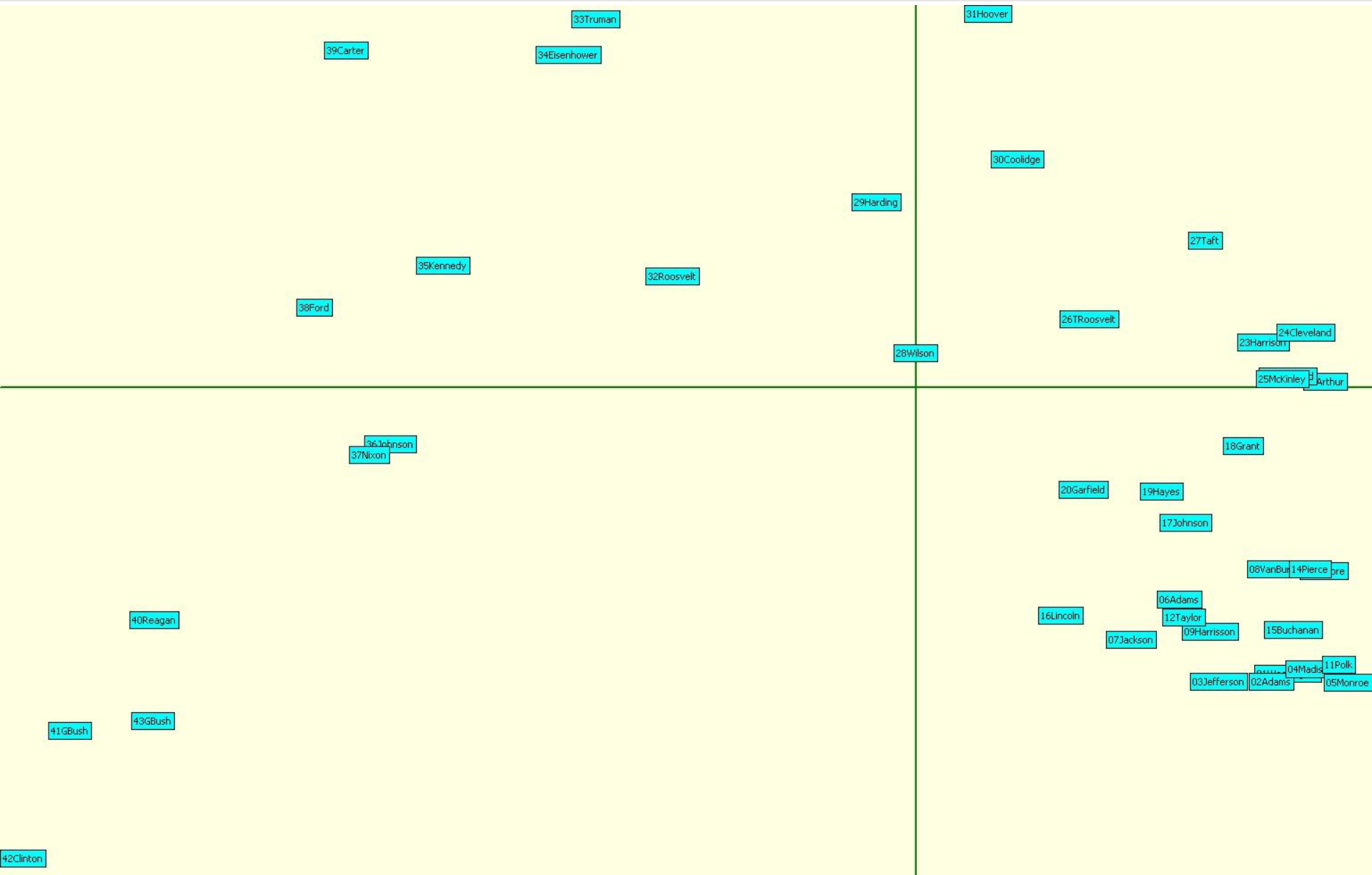
Le British National Corpus.

---

# Principales caractéristiques de la partition: « 43 Présidents des USA »

Partie	occurrences	formes	hapax	Fréq. Max	Forme
01Washington	18121	3179	1603	1560	the
02Adams	9486	2085	1181	788	the
03Jefferson	24613	3619	1776	1806	the
04Madison	24118	3649	1795	2240	the
05Monroe	50388	4480	1819	4422	the
06Adams	2920	999	700	287	the
07Jackson	2311	844	596	183	the
08VanBuren	15366	2974	1578	1264	the
09Harrisson	8472	1881	1096	795	the
11Polk	55954	5081	2071	4666	the
12Taylor	1094	495	373	96	the
13Fillmore	31811	4480	2139	2817	the
14Pierce	45493	5374	2396	4174	the
15Buchanan	27294	3811	1796	2419	the
16Lincoln	4352	1256	778	295	the
17Johnson	9298	2172	1294	931	the
18Grant	12639	2766	1597	1072	the
19Hayes	2496	830	534	226	the
20Garfield	2996	1016	678	289	the
21Arthur	33353	5057	2421	2888	the
22Cleveland	55870	6435	2927	4869	the
23Harrison	59886	6304	2773	5105	the
24Cleveland	61875	7123	3086	4884	the
25McKinley	81527	8076	3471	7569	the
26TRoosevelt	131755	9343	3662	10441	the
27Taft	97183	8246	3525	8886	the
28Wilson	38173	4627	2156	2717	the
29Harding	14798	3085	1675	1020	the
30Coolidge	56217	5900	2596	3670	the
31Hoover	29867	4248	1980	2125	the
32Roosevelt	54954	6313	3139	3640	the
33Truman	72084	6204	2599	4734	the
34Eisenhower	59798	6535	2986	3589	the
35Kennedy	21252	3953	2077	1208	the
36Johnson	35146	4417	2104	1971	the
37Nixon	29255	3622	1715	1843	the
38Ford	14149	2803	1483	788	the
39Carter	47139	5889	2762	2805	the
40Reagan	38194	5197	2624	1873	the
41GBush	18442	3283	1753	985	the
42Clinton	65808	5849	2583	3030	the
43GBush	30208	4322	2087	1372	the

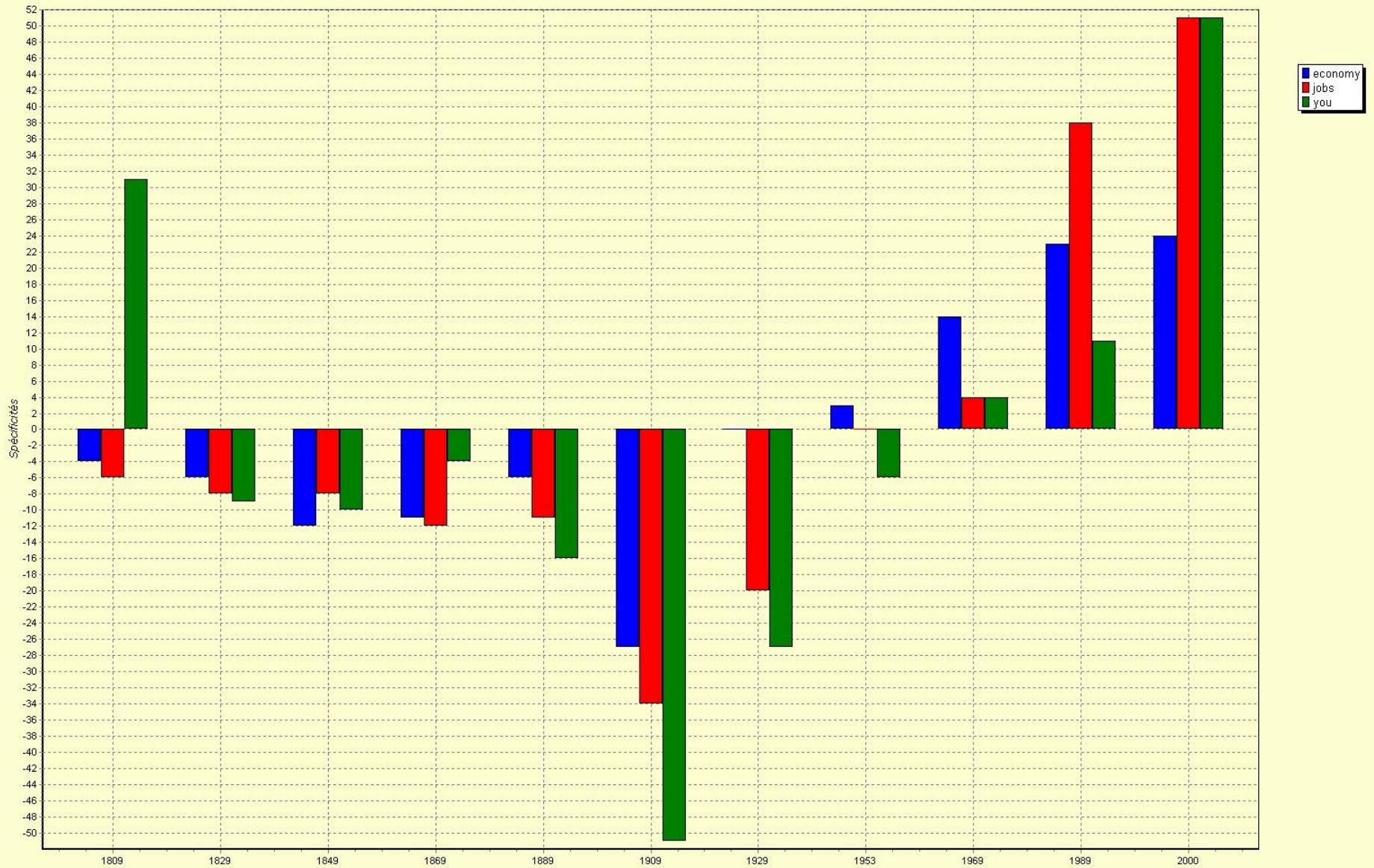
# Typologie (Analyse des correspondances) des 43 discours



## Mots caractéristiques de l'époque récente

Forme	Frq. Tot.	Fréquence	Coeff.
you	1789	1082	***
jobs	301	270	***
economy	715	396	***
percent	332	278	***
technology	113	110	***
do	1803	795	***
parents	115	108	***
U	131	125	***
And	1396	1134	***
commitment	134	130	***
But	1457	717	***
Social	123	115	***
families	300	228	***
America	1478	1147	***
world	2315	1156	***
new	2232	1202	***
Il	99	99	***
us	2377	1147	***
1980	92	92	***
s	2394	1621	***
tax	665	417	***
years	1947	907	***
Let	336	235	

# Mots spécifiques: Economy, You, jobs



# Les formes *travailleurs* et *salariés* à la CGT

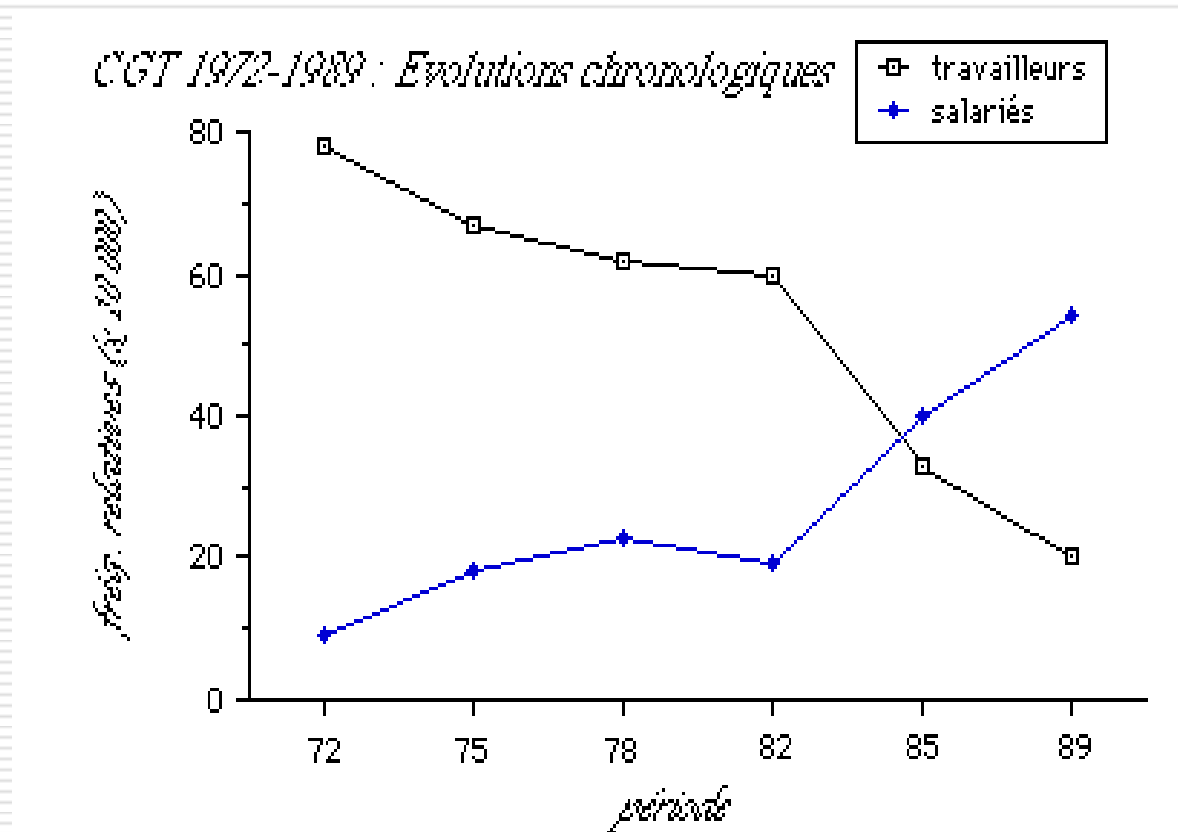


TABLEAU 2 : ÉVOLUTION DE QUELQUES SUBSTANTIFS ET ADJECTIFS DE 1789 A 1964 (Etienne BRUNET, Analyse du TLF)  
 La « tendance » est mesurée par l'indice de Spearman ou coefficient des rangs significatif au seuil de 5 % quand  $p > /0.52$ ).

P E R I O D E		1789	1816	1833	1842	1850	1860	1870	1880	1893	1908	1919	1927	1933	1938	1946
		1815	1832	1841	1849	1859	1869	1879	1892	1907	1918	1926	1932	1937	1945	1964
MOT	TENDANCE TOTAL	E C A R T S R E D U I T S														
ADJECTIFS EN PROGRESSION																
PIRE	+0,97 3126	-14,4	-07,7	-07,9	-06,5	-05,4	-05,3	-03,2	-04,5	+05,0	+04,4	+04,4	+05,4	+19,6	+08,9	+06,9
TECHNIQUE	+0,93 1663	-10,2	-10,6	-10,7	-08,7	-05,5	-09,4	-08,7	-06,4	-06,7	-04,7	-03,9	+07,5	+16,4	+12,2	+47,9
ESTHETIQUE	+0,91 1379	-11,0	-08,7	-09,7	-07,2	-32,8	-03,6	-05,5	+03,1	+00,4	-00,1	+11,4	+03,9	+02,7	+18,6	+08,6
INCONSCIENT	+0,90 1863	-12,9	-12,0	-11,9	-10,6	-10,8	-07,5	-05,5	+06,4	+04,9	-00,2	+00,4	+02,8	+00,8	+20,1	+35,8
MULTIPLE	+0,89 1009	-06,9	-06,8	-07,0	-05,3	-04,7	-02,5	-04,1	-01,8	+06,1	-02,7	+04,7	+03,6	-01,2	+05,0	+22,5
IDIOT &c.	+0,88 1656	-11,6	-09,2	-09,8	-07,1	-02,2	+01,3	+02,4	+03,8	+00,7	+02,3	+00,9	+05,5	+11,5	+08,2	+05,1
ADJECTIFS EN REGRESSION																
IMMORTEL	-0,91 1094	+10,5	+03,0	+08,9	+03,0	+00,6	-00,2	+02,5	-02,6	+01,2	-01,3	-03,4	-06,3	-06,9	-05,0	-05,4
SAGE	-0,90 5902	+12,0	+13,5	+06,3	-01,4	+01,6	-02,1	-01,8	+02,5	-02,6	-02,4	-05,2	-06,5	-02,4	-06,2	-07,2
INFORTUNE	-0,90 1315	+40,8	+10,9	+02,6	+02,6	-01,4	-01,4	-04,3	-07,9	-04,3	-05,0	-07,4	-06,5	-08,5	-08,5	-07,1
ROMAIN	-0,87 3185	+17,3	+33,7	-00,7	+13,2	-02,4	+09,2	-07,1	-09,7	-06,3	-03,0	-07,8	-08,8	-06,2	-09,0	-12,2
BARBARE	-0,80 2939	+14,2	+19,0	-02,1	+01,7	-04,1	+13,9	-04,2	+03,2	-00,0	+00,5	-07,4	-09,5	-07,7	-07,5	-09,0
VENERABLE	-0,78 1093	+05,4	+03,0	+00,4	+07,0	-01,4	-00,2	+01,9	-01,4	+01,6	+00,0	-00,1	-05,2	-01,5	-03,4	-06,2
SOLITAIRE	-0,75 3871	+05,6	-00,6	+10,0	+13,2	-00,1	+01,9	-05,8	-01,0	-00,9	-03,8	-01,5	-08,4	-05,1	-02,5	-01,9
AUGUSTE	-0,74 1170	+06,2	+03,4	+00,1	+01,2	+02,2	+01,6	+08,0	-01,0	+01,0	+05,2	-03,3	-04,5	-06,7	-05,3	-07,8
INGRAT	-0,74 1530	+07,8	+00,4	+05,8	+04,3	+01,1	+00,0	-01,4	-04,7	+01,6	+02,7	-02,2	-04,1	-03,8	-05,7	-02,7
CHASTE &c.	-0,72 1286	-00,8	+01,0	+09,8	+03,1	+08,4	+01,4	+01,0	+03,4	-00,4	-00,2	-03,4	-04,1	-06,4	-05,5	-06,7
SUBSTANTIFS EN PROGRESSION																
SOUCI	+0,98 3742	-15,0	-10,4	-05,6	-07,4	-07,2	-03,7	-02,5	+00,0	+03,7	+03,3	+05,5	+05,5	+06,7	+17,8	+10,0
TENTATION	+0,98 1965	-09,6	-08,0	-03,4	-03,4	-02,9	-02,1	-00,8	-00,5	+00,7	-00,6	+02,3	+04,0	+03,5	+13,0	+03,5
RYTHME	+0,97 2070	-12,0	-10,3	-11,5	-08,1	-07,9	-06,8	-05,4	-05,9	+05,3	+04,4	+09,1	+07,4	+08,1	+15,7	+10,5
APPEL	+0,97 4007	-14,8	-12,7	-11,7	-10,0	-08,0	-04,2	-00,9	+00,0	+05,1	+03,2	+08,0	+13,5	+05,6	+12,5	+15,4
CIGARETTE	+0,97 1577	-11,9	-11,1	-10,6	-07,7	-07,1	-07,1	-00,9	+02,3	-02,0	+04,6	+05,0	+06,5	+16,3	+12,9	+12,2
PROBLEME	+0,97 5936	-19,9	-14,7	-13,4	-13,5	-10,9	-08,6	-13,1	-08,3	-00,3	-05,5	+04,6	+25,3	+17,5	+18,7	+42,1
CONTACT	+0,97 3043	-10,6	-10,3	-09,5	-08,4	-06,4	-02,0	-06,3	-04,1	+03,3	-01,3	+09,5	+09,3	+03,4	+10,3	+22,7
USINE	+0,96 1356	-09,9	-08,7	-09,4	-06,3	-07,0	-05,4	-02,4	-00,8	-00,4	-01,2	+01,5	+05,1	+28,4	+05,0	+12,9
VACANCES &c.	+0,96 1596	-11,3	-09,3	-06,3	-02,9	-04,9	-02,8	-00,6	-02,5	-00,1	+08,5	+03,9	+04,8	+08,6	+05,7	+10,9
SUBSTANTIFS EN REGRESSION																
PLEURS	-0,95 2699	+14,7	+10,1	+25,8	+00,6	+08,7	-03,1	+00,6	-04,2	-02,4	-06,4	-06,5	-09,8	-09,1	-10,8	-10,8
FLAMBEAU	-0,95 1511	+11,0	+08,8	+09,3	+05,1	+08,8	+02,0	+01,5	-03,1	-01,9	-04,2	-07,7	-07,8	-08,6	-06,7	-07,7
FORTUNE	-0,95 9414	+08,1	+16,6	+19,1	+31,1	+05,8	+05,1	-02,9	+03,7	-08,7	-09,2	-13,5	-10,7	-12,3	-15,6	-16,5
VOILE	-0,95 5290	+08,4	+03,1	+14,6	+06,0	+07,2	+04,8	+00,2	-01,2	-02,7	-02,9	-04,7	-08,2	-07,5	-08,3	-09,3
EPOUX	-0,94 3398	+44,7	+05,8	+05,6	-04,5	+04,5	+02,8	-02,6	-04,2	-06,3	-06,3	-08,3	-07,4	-08,9	-08,0	-11,8
JOUG	-0,94 1004	+17,8	+08,8	+04,7	+01,8	+00,8	-02,9	-01,7	-03,3	-00,1	-03,2	-03,8	-03,0	-05,2	-06,1	-06,8
ANTIQUITE	-0,93 1646	+17,5	+08,1	+10,2	+06,0	+02,4	-00,5	-05,2	-01,6	-04,5	-05,9	-04,4	-06,3	-06,0	-05,9	-06,3
TOMBEAU &c.	-0,93 4209	+24,1	+04,3	+16,6	+20,4	+04,5	+02,0	-01,3	-07,3	-05,9	-02,9	-06,9	-13,3	-11,6	-10,8	-11,7

## 5. Exemple de corpus médiéval

---

### ■ Description du corpus

- Visualisation par AC
  - Sériation
  - Inférence statistique
  - Conclusion
-

## Facteurs étroitement interdépendants connus

- variation des graphies
- médiation fondamentale des copistes
- effets régionaux et temporels marqués
- absence de corpus de référence et de normes

Conséquence : difficulté d'utilisation systématique d'outils de traitement automatique de la langue

---

## 5. Exemple de corpus médiéval

---

12ème\_Vers

15 textes de la Base de Français Médiéval du 12e et du 13e siècle  
riches en variantes graphiques

383 193 occurrences de 27 459 formes graphiques

---

## 12eme\_en\_vers

---

- Nom dans le corpus BFM: **stbrend**
- Auteur : Benedeit
- Titre : **Voyage de saint Brendan**
- Date: **début XII**
- Ed. sc.: I. Short, B. Merrilees
- Maison d'éd. : Manchester Un. Press
- Date d'édition : 1979
- Domaine : religieux
- Genre : hagiographie
- Dialecte : anglo-normand
- Nombre d'occurrences: 10829

- Nom dans le corpus: **gormont**
  - Auteur : anonym
  - Titre : **Gormont et Isembart**
  - Date : autour : **1130**
  - Ed. sc.: A. Bayot
  - Maison d'éd. : Champion; Coll. CFMA
  - Date d'édition : 1931
  - Domaine : littéraire
  - Genre : épique
  - Dialecte : n d (+ centre ou s-o de Paris)
  - Nombre d'occurrences: 3815
-

## 12eme\_en\_vers

---

- Nom dans le corpus: **roland**
- Auteur : anonyme
- Titre : **Chanson de Roland**
- Date : autour : **1100**
- Ed. sc.: G. Moignet
- Maison d'édition : Bordas
- Date d'édition: 1969
- Domaine : littéraire
- Genre : épique
- Dialecte : anglo-normand
- Nombre d'occurrences: 29338

- Nom dans le corpus: **thebes**
  - Auteur : anonyme
  - Titre : **Novel de Thèbes**
  - Date : **1150**
  - Éd. sc.: G. Raynaud de Lage
  - Maison d'éd. : Champion; - Coll : CFMA
  - Date d'édition: 1968
  - Domaine : littéraire
  - Genre : nouvelle
  - Dialecte : n d
  - Nombre d'occurrences: 62698
-

## 12eme\_en\_vers

---

- Nom dans le corpus: **eracle**
- Auteur : Gautier d'Arras
- Titre : **Eracle**
- Date : **1176 - 1184**
- Éd. sc.: G. Raynaud de Lage
- Maison d'éd. : Champion; Coll : CFMA
- Date d'édition: 1976
- Domaine : littéraire
- Genre : nouvelle
- Dialecte : n d
- Nombre d'occurrences: 40839

- Nom dans le corpus: **beroul**
  - Auteur : Béroul
  - Titre : **Tristan**
  - Date: entre **1165 et 1200**
  - Éd. Sc. : L. M. Defourques, E. Muret
  - Maison d'édition : Champion; Coll : CFMA
  - Date d'édition: 1947
  - Domaine : littéraire
  - Genre : nouvelle
  - Dialecte : franco-picard
  - Nombre d'occurrences: 27257
-

## 12eme\_en\_vers

---

- Nom dans le corpus: **renart10**
- Auteur : anonyme
- Titre : **Roman de Renart** (branche X)
- Date : **early XIII**
- Ed. sc.: M. Roques
- Maison d'édition : Champion; Coll : CFMA
- Date d'édition: 1948-1963
- Domaine : littéraire
- Genre : short stories
- Dialecte : n d
- Nombre d'occurrences: 13472

- Nom dans le corpus: **renart11**
  - Auteur : anonyme
  - Titre : **Roman de Renart** (branche XI)
  - Date : **early XIII**
  - Ed. sc.: M. Roques
  - Maison d'édition : Champion; Coll : CFMA
  - Date d'édition: 1948-1963
  - Domaine : littéraire
  - Genre : short stories
  - Dialect : n d
  - Nombre d'occurrences: 8563
-

## 12eme\_en\_vers

---

- Nom dans le corpus: **amiamil**
- Auteur : anonyme
- Titre : **Ami et Amile**
- Date : autour : **1200**
- Ed. sc.: P.F. Dembowski
- Maison d'éd. : Champion; Coll : CFMA
- Date d'édition: 1969
- Domaine : littéraire
- Genre : épique
- Dialecte : n d
- Nombre d'occurrences: 25283

- Nom dans le corpus: **belinc**
  - Auteur : Renaut de Beaujeu
  - Titre : **Bel Inconnu**
  - Date : **avant 1214**
  - Ed. sc.: P. Williams
  - Maison d'édition : Champion; Coll : CFMA
  - Date d'édition: 1929
  - Domaine : littéraire
  - Genre : nouvelle
  - Dialecte : n d
  - Nombre d'occurrences: 36692
-

## 12eme\_en\_vers

---

- Nom dans le corpus: **thomas**
- Auteur: Guernes de Pont-Sainte Maxence
- Titre : **Vie de saint Thomas**
- Date : **entre 1172 et 1174**
- Éd. sc.: E. Walberg
- Maison d'éd.: Champion; Coll : CFMA
- Date d'édition: 1936
- Domaine : religieux
- Genre : hagiographie
- Dialecte : n d
- Nombre d'occurrences: 53947

- Nom dans le corpus: **louis**
  - Auteur : anonyme
  - Titre : **Couronnement de Louis**
  - Date : autour : **1130**
  - Ed. sc.: E. Langlois
  - Maison d'édition : Champion; Coll. : CFMA
  - Date d'édition: 1925
  - Domaine : littéraire
  - Genre : épique
  - Dialect : n d
  - Nombre d'occurrences: 19786
-

## 12eme\_en\_vers

---

- Nom dans le corpus: **escoufle**
- Auteur : Jean Renart
- Titre : **Escoufle**
- Date: **1200 -1202**
- Ed. sc.: F. P. Sweester
- Maison d'édition : Droz
- Collection : TLF
- Date d'édition: 1974
- Domaine : littéraire
- Genre : nouvelle
- Dialecte : picard
- Nombre d'occurrences: 57967

- Nom dans le corpus: **dole**
  - Auteur : Jean Renart
  - Titre : **Roman de la Rose  
ou de Guillaume de Dole**
  - Date:: **1210 ou 1228**
  - Ed. sc.: F. Lecoy
  - Maison d'édition : Champion; Coll : CFMA
  - Date d'édition: 1962
  - Domain : littéraire
  - Genre : nouvelle
  - Dialecte : n d
  - Nombre d'occurrences: 34555
-

## 12eme\_en\_vers

---

- Nom dans le corpus: **vergy**
  - Auteur : anonyme
  - Titre : **Châtelaine de Vergy**
  - Date: **milieu XIII, avant 1288**
  - Ed. sc.: G. Raynaud, L. Foulet
  - Maison d'édition : Champion; Coll : CFMA
  - Date d'édition: 1921
  - Domain : littéraire
  - Genre : nouvelle
  - Dialect : n d
  - Nombre d'occurrences: 6117
-

## 5. Exemple de corpus médiéval (suite)

---

- Description du corpus

- **Visualisation par AC**

- Sériation

- Inférence statistique

- Conclusion

---

## Table Mots - Textes

	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.
affaire	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.
amis	133.	32.	30.	24.	58.	74.	0.	13.	10.	5.	5.	2.	22.	33.	7.
amor	5.	28.	26.	35.	35.	37.	1.	5.	15.	5.	5.	0.	11.	26.	30.
amors	5.	49.	4.	25.	22.	37.	0.	1.	0.	0.	0.	0.	1.	1.	8.
amur	0.	0.	0.	0.	0.	0.	0.	0.	0.	0.	12.	2.	0.	119.	0.
an	36.	1.	24.	2.	7.	12.	1.	2.	37.	97.	3.	10.	20.	2.	2.
andui	10.	8.	3.	0.	2.	24.	0.	0.	3.	2.	1.	0.	5.	3.	0.
anel	2.	0.	14.	7.	0.	20.	0.	1.	4.	0.	0.	0.	0.	15.	0.
ans	12.	8.	0.	1.	16.	22.	0.	0.	1.	0.	0.	0.	0.	0.	0.





## 5. Exemple de corpus médiéval (suite)

---

- Description du corpus
- Visualisation par AC

### ▪ Sériation

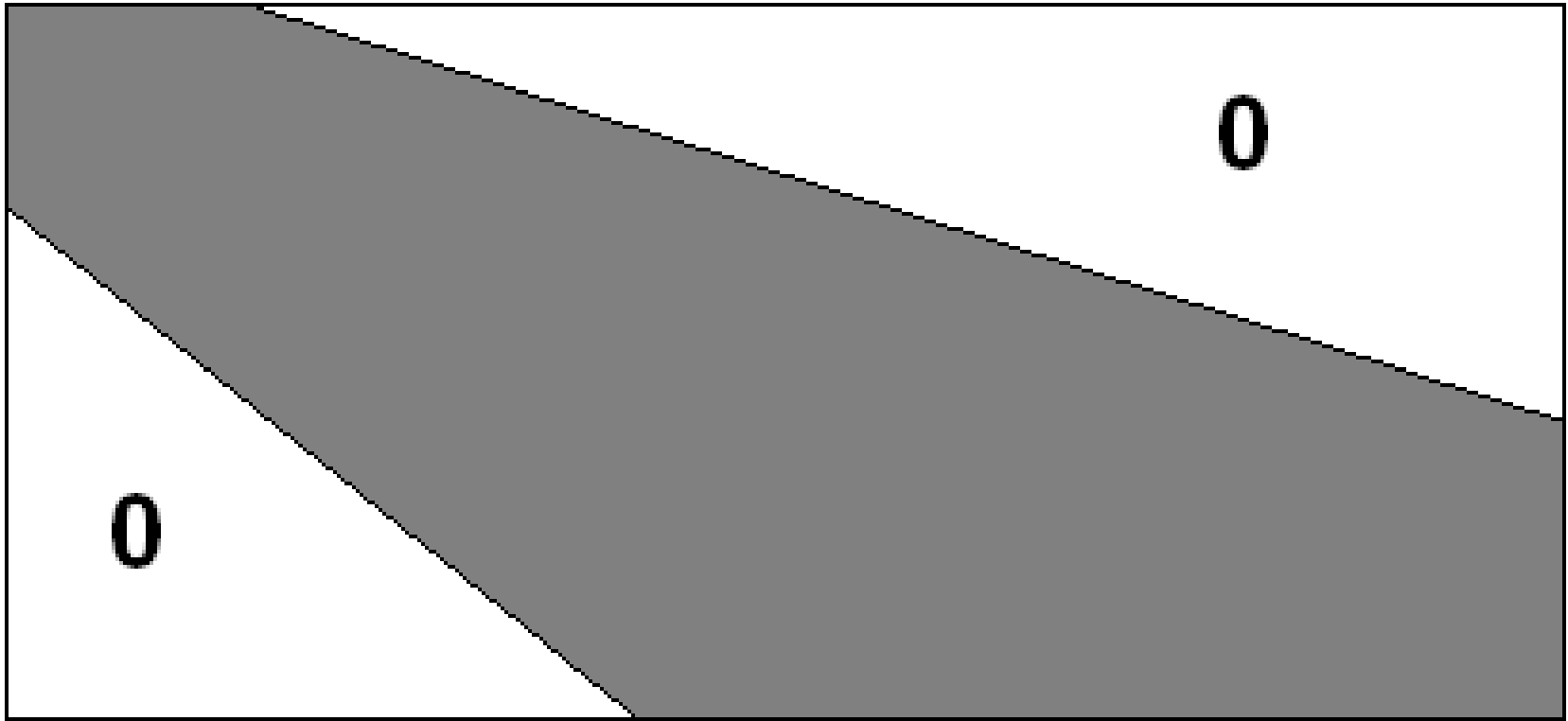
- Inférence statistique
  - Conclusion
-

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.

### **Une propriété fondamentale de l'analyse des correspondances**

Si le tableau initial de données, après ré-ordonnement des lignes et des colonnes, peut prendre la forme du tableau esquissé sur la figure suivante ...

alors ce ré-ordonnement est fourni par l'ordre des coordonnées des lignes et des colonnes sur le premier axe de l'analyse des correspondances du tableau initial de données.



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

Les lignes et les colonnes de la table ci-dessous ont été ré-ordonnées selon le premier axe de l'analyse des correspondances de la table lexicale initiale.

	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	0	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
amur	12	2	119	0	0	0	0	0	0	0	0	0	0	0	0
abes	0	48	0	0	10	0	0	0	0	0	0	0	0	0	0
jo	123	21	126	2	0	0	0	20	0	0	2	0	0	2	0

Même table  
(suite)

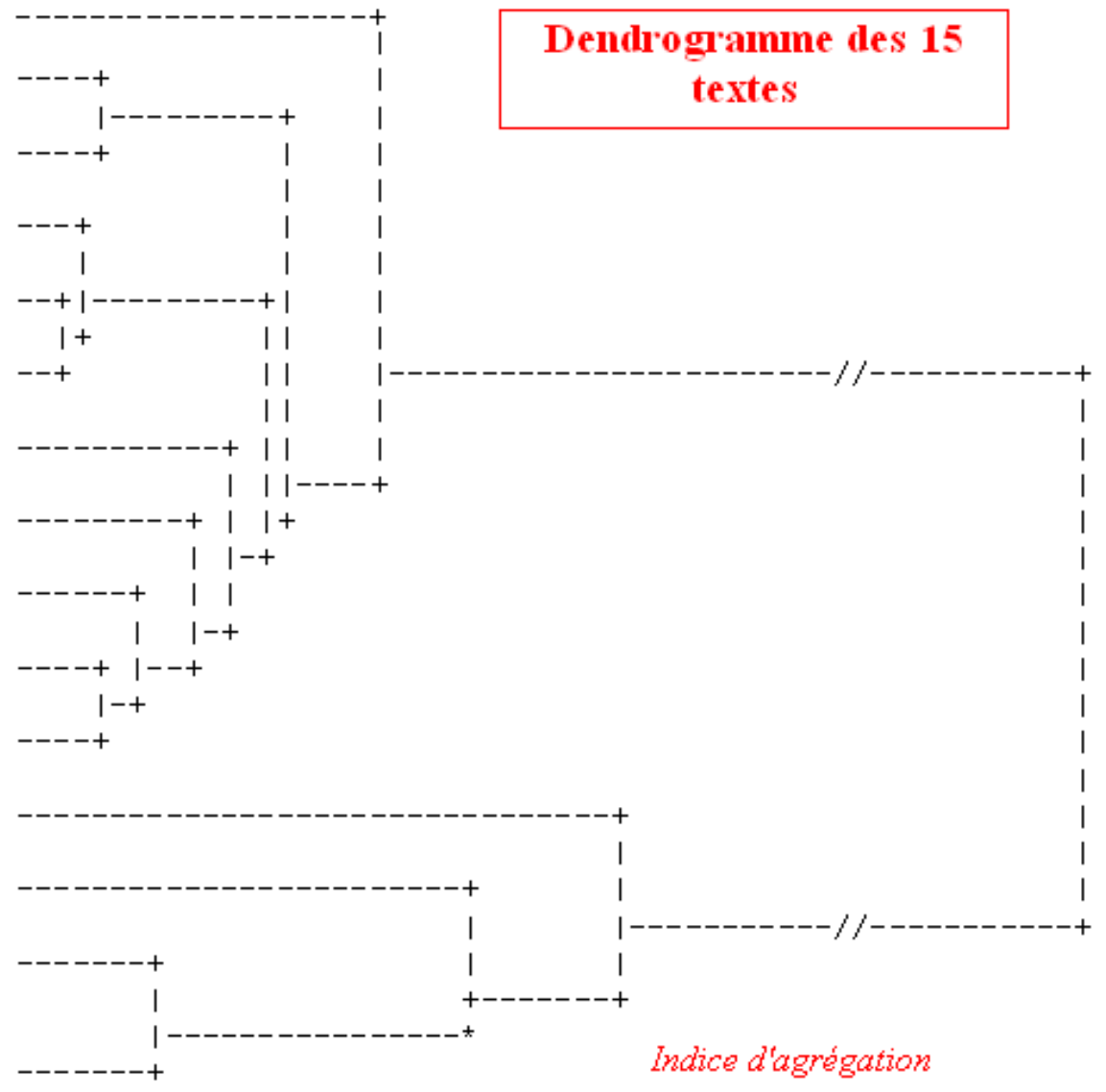
rola stbr thom gorm loui bero theb beli amia rena erac rena dole esco verg

nef	0	42	17	1	0	0	0	0	2	0	1	0	0	0
pur	94	83	238	20	0	0	0	0	0	0	0	1	5	3
vus	35	35	192	21	0	0	0	0	0	0	0	0	0	0
amur	12	2	119	0	0	0	0	0	0	0	0	0	0	0
abes	0	48	0	0	0	0	0	0	0	0	0	0	0	0
jo	123	21	126	2	0	0	0	20	0	0	2	0	0	2
dolur	1	1	62	0	0	0	0	0	0	0	0	0	0	0
sunt	154	59	33	12	0	0	43	0	0	0	0	0	0	0
fut	59	2	2	24	0	3	2	1	0	0	0	0	4	0
altre	62	21	50	0	22	0	0	0	0	0	0	0	0	0
tere	68	0	9	0	0	0	3	12	0	0	0	0	0	1
aveir	29	6	37	2	12	0	0	0	0	0	0	0	0	0
paien	83	0	0	3	13	0	0	0	0	1	4	0	0	6
ço	124	61	125	1	0	0	0	60	0	0	0	0	0	0
reis	134	5	23	8	56	0	0	0	0	0	0	0	0	0
mei	43	17	59	7	31	0	0	0	0	0	0	0	1	0
dei	8	0	34	2	7	1	0	0	0	0	0	0	0	0
seit	42	16	29	0	24	0	0	0	6	0	0	0	0	0
deit	17	11	30	1	19	0	0	0	0	0	0	0	0	0
sis	34	7	7	1	1	0	3	1	0	0	6	0	2	4
rei	59	2	26	30	38	0	0	0	0	0	0	0	0	0
asez	42	8	8	0	0	19	0	0	0	2	0	2	0	0
veit	40	7	21	1	38	0	0	0	0	0	0	0	0	0
paiens	25	0	0	10	5	0	0	0	0	0	10	0	0	3
nen	39	8	20	4	2	6	2	3	3	0	22	0	1	0
dreit	15	11	13	2	29	0	0	0	0	0	0	0	0	0
vent	2	21	27	1	2	4	4	1	2	0	2	0	4	7
volt	20	8	72	0	17	0	8	1	20	0	8	0	0	8
cels	31	0	2	0	9	0	0	9	0	0	1	0	0	10
out	44	39	6	1	0	73	0	1	0	1	0	0	0	1
ki	213	6	47	2	0	0	0	19	0	0	2	0	0	184
abat	25	0	0	4	6	1	5	4	2	1	1	1	0	4

Même table : 25 dernières lignes

	rola	stbr	thom	gorm	loui	bero	theb	beli	amia	rena	erac	rena	dole	esco	verg
puceles	0	0	0	0	0	0	10	11	0	1	7	0	9	37	0
sachiez	0	0	0	0	2	3	13	0	0	14	0	4	66	0	10
ançois	0	0	0	0	0	4	1	13	0	0	17	1	0	42	0
biauté	0	0	0	0	0	0	4	11	0	0	10	0	5	31	0
devise	0	0	0	0	0	2	2	6	1	5	8	2	13	17	0
vielle	0	0	0	0	0	1	6	0	1	3	30	0	2	20	0
coi	0	0	0	0	0	0	9	14	0	9	20	21	3	45	0
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
amui	0	0	0	0	0	0	2	9	0	7	17	9	11	27	3
affaire	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
ensemble	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

1 7.49 amiamile  
 2 1.32 louis  
 3 5.57 belinc  
 4 .79 beroul  
 5 .11 renard  
 6 4.89 renardx  
 7 4.09 thebes  
 8 3.32 eracle  
 9 1.97 escoufle  
 10 1.03 vergy  
 11 44.01 dole  
 12 13.30 stbrendan  
 13 9.68 thomas  
 14 2.42 gormont  
 15 ----- roland



**Dendrogramme des 15  
 textes**

*Indice d'agrégation*

## 5. Exemple de corpus médiéval (suite)

---

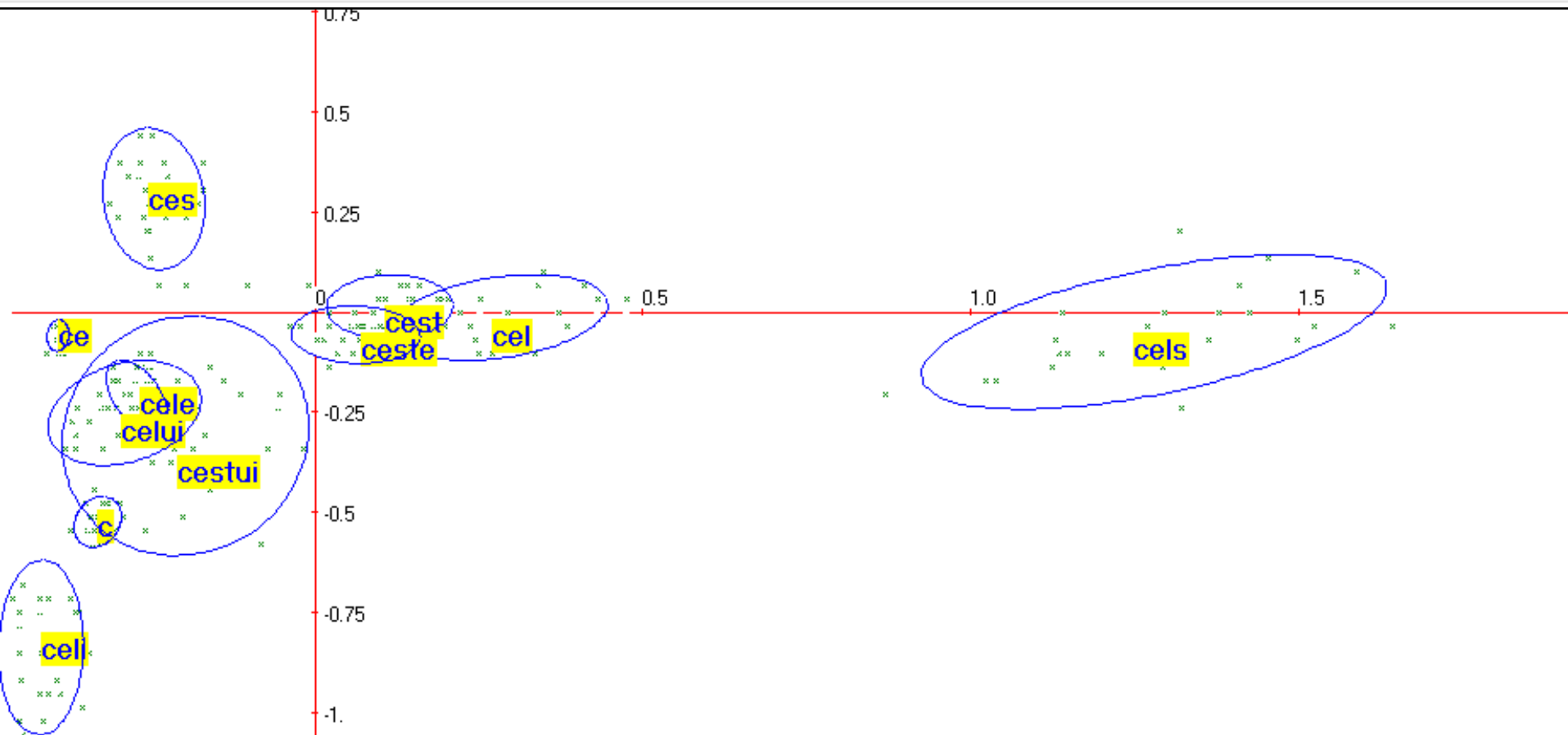
- Description du corpus médiéval
- Visualisation par AC
- Sériation

### ▪ Inférence statistique

- Conclusion
-



# Zones de confiance des différents démonstratifs dans le plan factoriel (sans les autres mots)



## 6. Conclusion

- Nouveaux matériaux et concepts pour une analyse linguistique / historique
  - Outils variés, stratégie complexe
  - Une implémentation interactive est indispensable
  - Vers un statut scientifique des visualisations ?
-

Le logiciel (Dtm-Vic) peut être librement téléchargé sur les sites des auteurs.

[www.dtmvic.com](http://www.dtmvic.com)

**Data and Text Mining , Visualization, Inference, Classification**

---

*Danke*

*Thank You*

---

*Obrigado*

*Merci*

*Grazie*

*Gracias*

*Ευχαριστω πολι*

*Cámõn*

*Choukrane*

*Domo Arigato*

---

---

**Le logiciel (DTM)**  
**est disponible gratuitement à l'adresse**  
**[www.lebart.org](http://www.lebart.org)**

---