

Cours *Distant Reading* : Visualisation

4. Introduction à la stylométrie

Simon Gabay



Introduction

Définition

La stylométrie est une approche computationnelle et quantitative du texte, dont l'objectif est de mesurer des idiosyncrasies stylistiques, appelées stylomes.

“stylome”, a set of measurable traits of language products

H. Van Halteren, *et al.*. "New machine learning methods demonstrate the existence of a human stylome." 2005

Le postulat est donc que, un peu comme l'ADN, le style de chaque être humain lui est unique, et que ses caractéristiques sont mesurables.

Expériences et cas célèbres

- Van Halteren 2015: quelques étudiants écrivent 9 textes courts (entre 600 et 1300 mots), que l'on est capable d'attribuer
- Vosoughi, Zhou, Roy 2015: associer environ 5500 comptes facebook et twitter sur la base de la production écrite de ces comptes.
- Les cas J.K. Rowling (Juola 2015) et E. Ferrante (Rybicki 2018)

Le fond et la forme

Traditionnellement, on regarde "le fond". Le meilleur exemple sont les champs lexicaux, qui se limitent souvent aux substantifs/adjectifs/verbes. Les sondages stylistiques peuvent être plus précis (adverbe), mais le champ se trouve alors limité à des extraits pour des raisons évidentes.

En informatique, certaines techniques reprennent cette approche en proposant des analyses thématiques (*topic modeling*), géographiques (*information retrieval*)...

La stylométrie fait le pari inverse: elle s'intéresse à la masse, et donc aux fréquences. Elle utilise volontiers les mots outils comme les articles (*/e*, */a*), les prépositions (*à, de*), les conjonctions (*et, que*).

On appelle parfois ces mots outils (*function words* ou *stop words* en anglais, par opposition aux *content words*) des mots vides, ce qui est vrai sémantiquement, mais pas syntaxiquement.

MFW

Les *Federalist papers*

The Federalist Papers est un recueil de 85 articles publié en 1787. C'est encore aujourd'hui une source importante pour l'histoire du droit américain, car ces articles ont été l'une des principales sources d'interprétation de la constitution américaine.

Ce recueil est publié sous un pseudonyme unique (*Publius Valerius Publicola*) par trois auteurs: James Madison, Alexander Hamilton et John Jay. Si nous sommes certains de connaître tous les articles publiés par le dernier (N°2, 3, 4, 5 et 64), il n'en va pas de même pour les deux autres.

Nous avons donc plusieurs problèmes:

- Les articles 49-58, 62 et 63 ne sont attribués à aucun auteur.
- Les articles 18, 19 et 20 sont écrit à deux mains par Madison et Hamilton, sans qu'il ne soit possible de dire qui est responsable de quelle partie.

In the PRESS,
and speedily will be published,

THE
FEDERALIST,

A Collection of Essays written in fa-
vor of the New Constitution.

By a Citizen of New-York.

Corrected by the Author, with Additions
and Alterations.

This work will be printed on a fine Paper
and good Type, in one handsome Volume duo-
decimo, and delivered to subscribers at the
moderate price of one dollar. A few copies
will be printed on superfine royal writing pa-
per, price ten shillings.

No money required till delivery.

To render this work more complete, will be
added, without any additional expence,

PHILO-PUBLIUS,

AND THE

Articles of the Convention,
As agreed upon at Philadelphia, Septem-
ber 17th, 1787.

I
F
-
C
E
S
R
C
I
J
T

Problème

Comment savoir qui a écrit quoi? Un réflex naturel serait de regarder la longueur des phrases:

- 34,55 mots en moyenne (σ de 19,2) pour Hamilton
- 34,59 mots en moyenne (σ de 20,3) pour Madison

Il faut donc trouver autre chose... Le lexique est évidemment la meilleure option.

Le cas Adair

Douglass Adair (1912-1968)

- "The Authorship of the Disputed Federalist Papers", 1944
- "The Tenth Federalist Revisited", 1951

Intuition stylométrique: Adair regarde les fréquences des mots, et remarque notamment que Hamilton préfère *while*, là où Madison utilise plutôt *whilst*.

TABLE 1.4-1
INCIDENCE: NUMBERS OF PAPERS IN WHICH WORD OCCURRED
AT LEAST ONCE

	<i>enough</i>	<i>while</i>	<i>whilst</i>	<i>upon</i>	Number of papers examined
Hamilton	14	10	0	23	23
Madison	0	0	13	4	19
Disputed	0	0	5	1	12
Joint	1	0	2	2	3

TABLE 1.4-2
RATES PER 1000 WORDS

	<i>enough</i>	<i>while</i>	<i>whilst</i>	<i>upon</i>	Total words in 1000's
Hamilton	0.59	0.26	0	2.93	45.7
Madison	0	0	0.47	0.16	51.0
Disputed	0	0	0.34	0.08	23.9
Joint	0.18	0	0.36	0.36	5.5
					126.1

Mosteller & Wallace, *Applied Bayesian and Classical Inference: The Case of The Federalist Papers*, p. 11.

Une approche statistique

Frederick Mosteller et David Wallace vont mener les premières études statistiques permettant d'identifier un auteur à partir des fréquences des mots.

- « Inference in an Authorship Problem », *Journal of the American Statistical Association*, vol. 58, juin 1963, p. 275-309
- *Inference and Disputed Authorship: Federalist Papers*, 1964.

Un de leurs objectifs est de retrouver des marqueurs (*markers*)

TABLE 2.5-4
SURVIVORS OF THE LOW-FREQUENCY SCREENING STUDY

Hamilton markers (H, M)	Madison markers (H, M)
(14, 0) <i>enough</i>	(0, 13) <i>whilst</i>
(10, 0) <i>while</i>	(2, 13) <i>consequently</i>
(8, 0) <i>destruction</i>	(0, 8) <i>although</i>
(8, 0) <i>offensive</i>	(1, 9) <i>violate + s + d + ing</i>
(10, 1) <i>affect + ed</i>	(3, 12) <i>pass + es + ed + ing</i>
(9, 1) <i>commonly</i>	(1, 8) <i>voice</i>
(9, 1) <i>vigor + ous</i>	(1, 8) <i>throughout</i>
(6, 0) <i>city + cities</i>	(2, 10) <i>language</i>
(6, 0) <i>contribute</i>	(0, 5) <i>fortune + s</i>
(6, 0) <i>defensive</i>	(0, 5) <i>join + ed</i>
(8, 1) <i>direction</i>	(0, 5) <i>violence</i>
(5, 0) <i>disgracing</i>	(1, 7) <i>again</i>
(5, 0) <i>rapid</i>	(1, 7) <i>function + s</i>
(13, 4) <i>considerable + ly</i>	(1, 7) <i>innovation + s</i>

Mosteller & Wallace, *Applied Bayesian and Classical Inference: The Case of The Federalist Papers*, p. 40.

L'importance des *Function words*

Mike Kestemont 2014 propose plusieurs raisons à l'importance des *function words*.

- Ils sont (très) fréquents, ce qui est parfait pour des études statistiques
- Ils sont donc nécessairement utilisés de manière (plutôt) inconsciente
- Leur emploi est décorélé du thème, du genre, du registre du texte

Alternatives aux *MFW*

Les rimes

Etude sur Jacob van Maerlant (c.1230-†c.1288) par Kestemont, Daelemans et Sandra (2012)

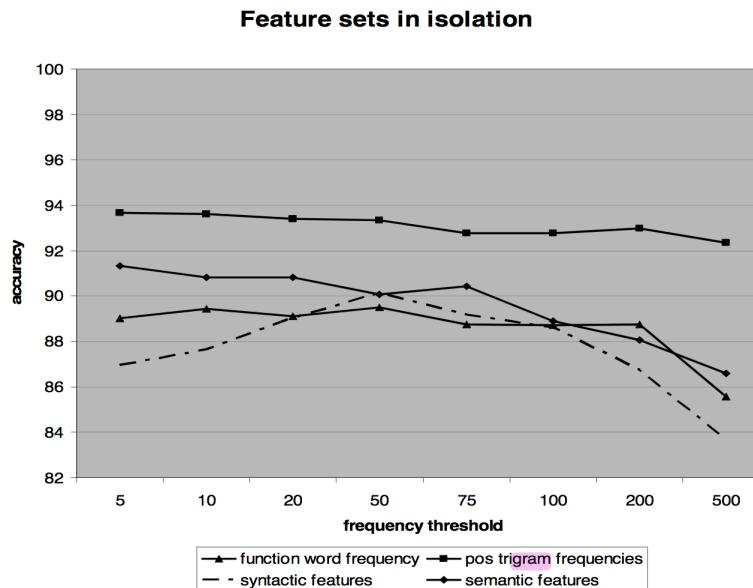
- Gros corpus de c. 200 000 vers
- Doute sur l'attribution de certains textes à Lodewijk van Velthem (c. 1260/1275-†c.1317)
- Problème: au Moyen Âge, les textes ne sont pas stables, et le contenu diffère d'un témoin à un autre.
- Hypothèse: les mots à la rime sont moins affectés par la variation textuelle, et doivent permettre une attribution fiable

Les POS

- Zhao & Zobel 2007 montre l'importante des POS pour l'attribution

	Function words			POS tags		
	the	of	a	cc	in	jj
Shakespeare	7.6	4.8	4.1	3.8	5.9	2.8
Marlowe	9.5	6.2	3.2	3.2	6.4	2.4

- Gamon 2004 travaille sur des *N-grams* (en l'occurrence des trigrammes) de POS



Les séquences de caractères

Sapkota, Bethard et Montes, 2015 proposent donc de travailler sur des *N-grams*, mais de caractères. Afin d'aller plus loin que les *N-grams*, il distinguent:

- Préfixe : un n-gramme de caractères couvrant les n premiers caractères d'un mots de longueur au moins n+1
- Suffixe :un n-gramme de caractères couvrant les n derniers caractères d'un mots de longueur au moins n+1
- Espace-préfixe : un préfixe commençant par un espace
- Espace-suffixe : un suffixe finissant par un espace

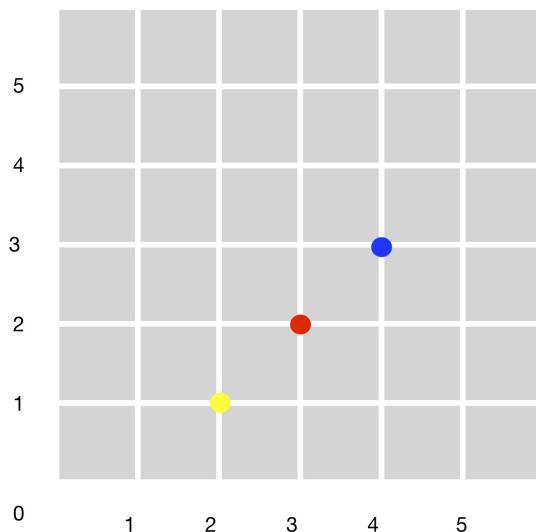
SC	Category	<i>N</i> -grams				
affix	<i>prefix</i>	tha	the	wit	con	hav
	<i>suffix</i>	ing	hat	ion	ent	ers
	<i>space-prefix</i>	_th	_of	_to	_an	_in
	<i>space-suffix</i>	he_	of_	to_	ed_	ng_
word	<i>whole-word</i>	the	and	for	was	not
	<i>mid-word</i>	tio	ati	iti	men	ent
	<i>multi-word</i>	e_t	s_a	t_t	s_t	n_t
punct	<i>beg-punct</i>	..T	's_	,_t	,_a	._I
	<i>mid-punct</i>	s,_	e,_	s..	e's	y's
	<i>end-punct</i>	es,	on.	on,	es.	er,

Sapkota, Bethard et Montes, 2015

Méthodes exploratoires

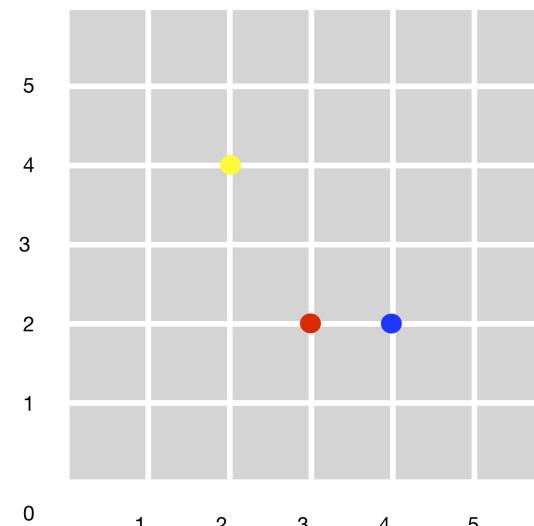
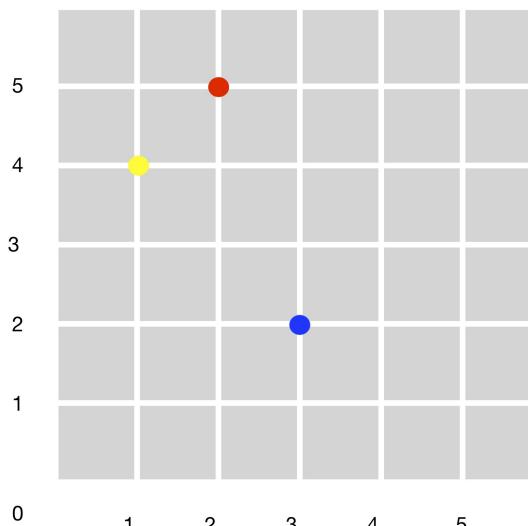
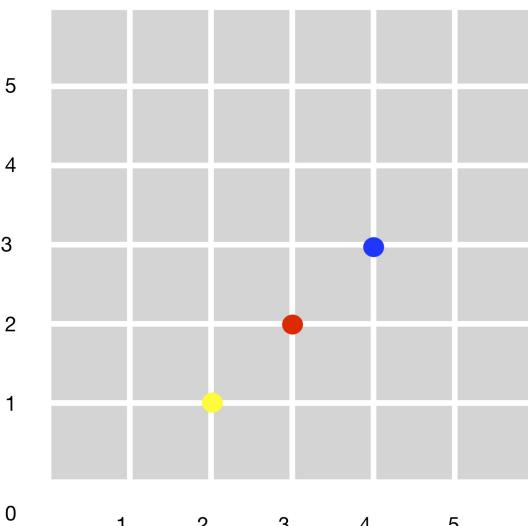
Deux dimensions

Elève	Avoir	Être
Molière	2	1
Corneille	3	2
Racine	4	3



Trois dimensions

Elève	Avoir	Être	Manger
Molière	2	1	4
Corneille	3	2	5
Racine	4	3	2



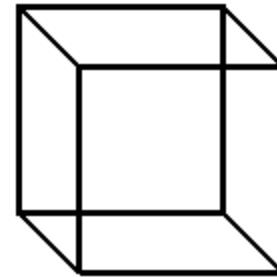
N-dimensions



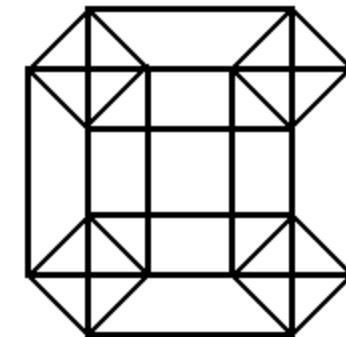
ligne



carré



cube



hypercube

Comment représenter de l'information à N-dimensions?

Principe

Les méthodes d'analyse factorielle s'utilisent pour décrire et hiérarchiser les relations statistiques qui peuvent exister entre des individus placés en ligne (par exemple des auteurs) et des variables placées en colonnes (leurs fréquences de mots) dans un tableau.

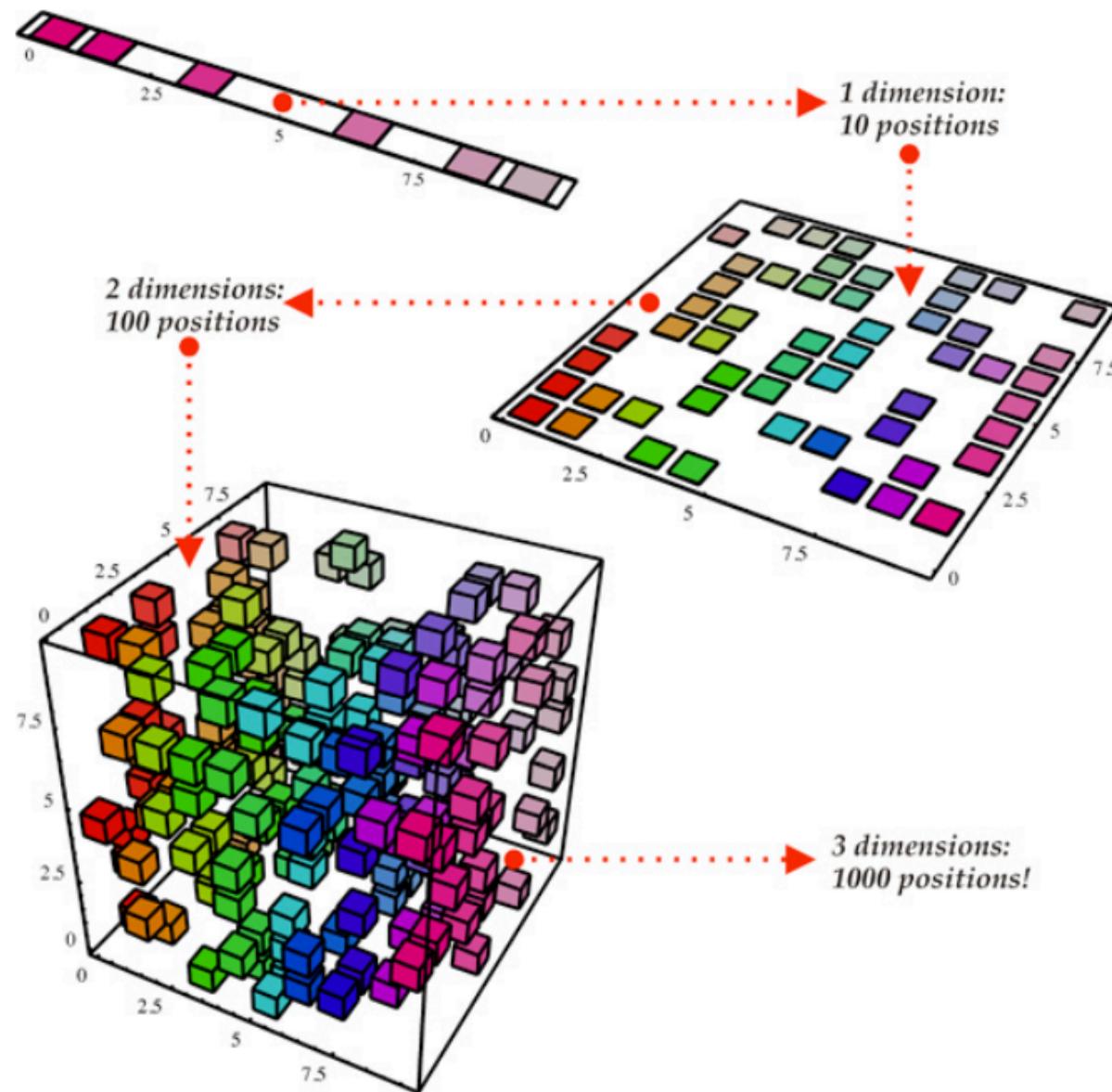
Elles considèrent le tableau de données comme un nuage de points dans un espace mathématique ayant autant de dimensions qu'il y a de colonnes dans le tableau de données ; elles cherchent à le projeter sur des axes ou des plans (appelés factoriels) de façon que l'on puisse en visualiser et étudier au mieux la forme et donc rechercher globalement des corrélations.

Le principe est de tenter de trouver une configuration optimale selon un critère de théorie de l'information pour respecter les proximités entre points : deux points qui sont proches (resp. éloignés) dans l'espace d'origine devront être proches (resp. éloignés) dans l'espace de faible dimension.

Réduction de dimension

Pour représenter un objet en 3D, je vais donner plusieurs images, prises sous différents angles. Ces angles ne sont pas pris au hasard: ils doivent donner un maximum d'information, car on ne peut pas multiplier les images.





Moshe Binieli, "An overview of Principal Component Analysis"

ACP

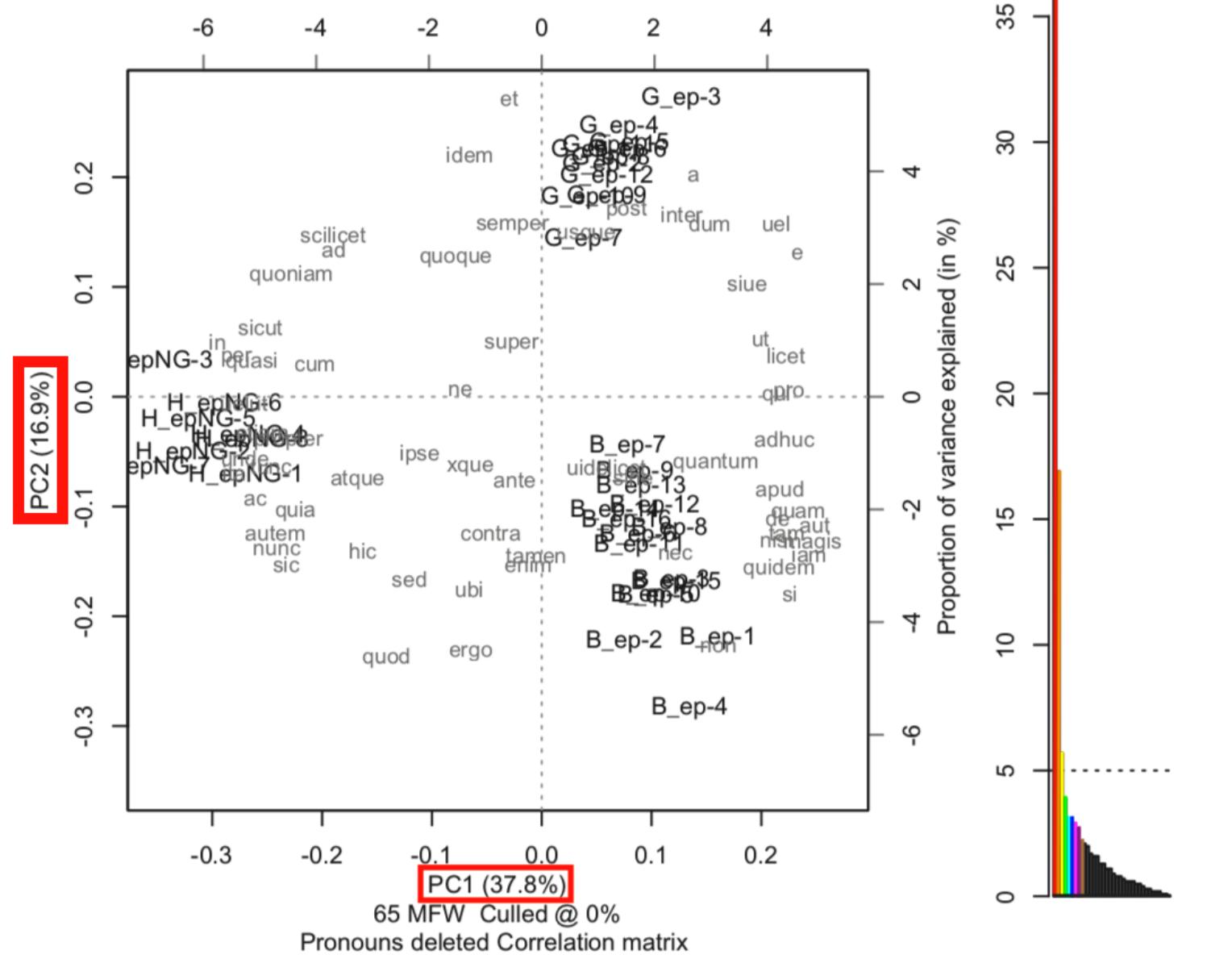
L'Analyse en composantes principales (ACP ou *PCA* en anglais pour *principal component analysis*) va déterminer les deux axes qui expliquent le mieux la dispersion de l'objet, interprété comme un nuage de points.

Prenons une classe notée sur 100 pour deux devoirs:

Elève	Mathématiques	Français
Cunégonde	80	30
Marcel	50	50
Josette	30	60

Si nous devons retenir deux axes pour décrire ce tableau, le premier sera la moyenne de chaque élève pour les deux matières, le second l'écart entre les deux notes. On pourrait multiplier les axes, mais ils n'apporteraient pas beaucoup plus d'information.

Principal Components Analysis



t-SNE

Au moment de la réduction de N à deux dimensions, nous allons perdre de l'information.

- Nous pouvons décider de préserver les longues distances entre les points, et ainsi maintenir les grands équilibres en "étirant" le graphique (PCA).
- Nous pouvons aussi décider de conserver les petites distances entre les points avec un algorithme *t-Distributed Stochastic Neighbor Embedding* (t-SNE).

Avec cette dernière approche, nous allons devoir définir le nombre de points voisins retenus lors du calcul: on parle de *perplexité*.

Un algorithme t-SNE ne donne jamais le même résultat: il doit être relancé plusieurs fois.

Quantité vs qualité

L'analyse en composante principale et l'algorithme t-SNE sont utilisés avec des grands ensembles de données. Quand les variables sont qualitatives, on se tourne plus volontiers

- vers une AFC (Analyse Factorielle des Correspondances) quand on a deux variables qualitatives
- vers une ACM (Analyse des Correspondances Multiples) lorsqu'on a plus de deux variables

AFC

L'AFC permet de croiser deux variables qualitatives. Elle s'applique donc sur un tableau de contingence, c'est-à-dire où les sommes des colonnes et des lignes ont un sens, et où X et Y s'expliquent l'un l'autre.

Elève	Mathématiques	Français	Moyenne
Cunégonde	80	30	55
Marcel	50	50	50
Josette	30	60	45

Doit devenir, par exemple:

Genre/résultat	Accepté.e	Recalé.e
Fille	1	1
Garçon	1	0

ACM

L'ACM est, pour aller vite, une ACP sur des catégories. Un ensemble d'individus (en lignes) est décrit par un ensemble de variables qualitatives (en colonnes), comme dans une enquête d'opinion.

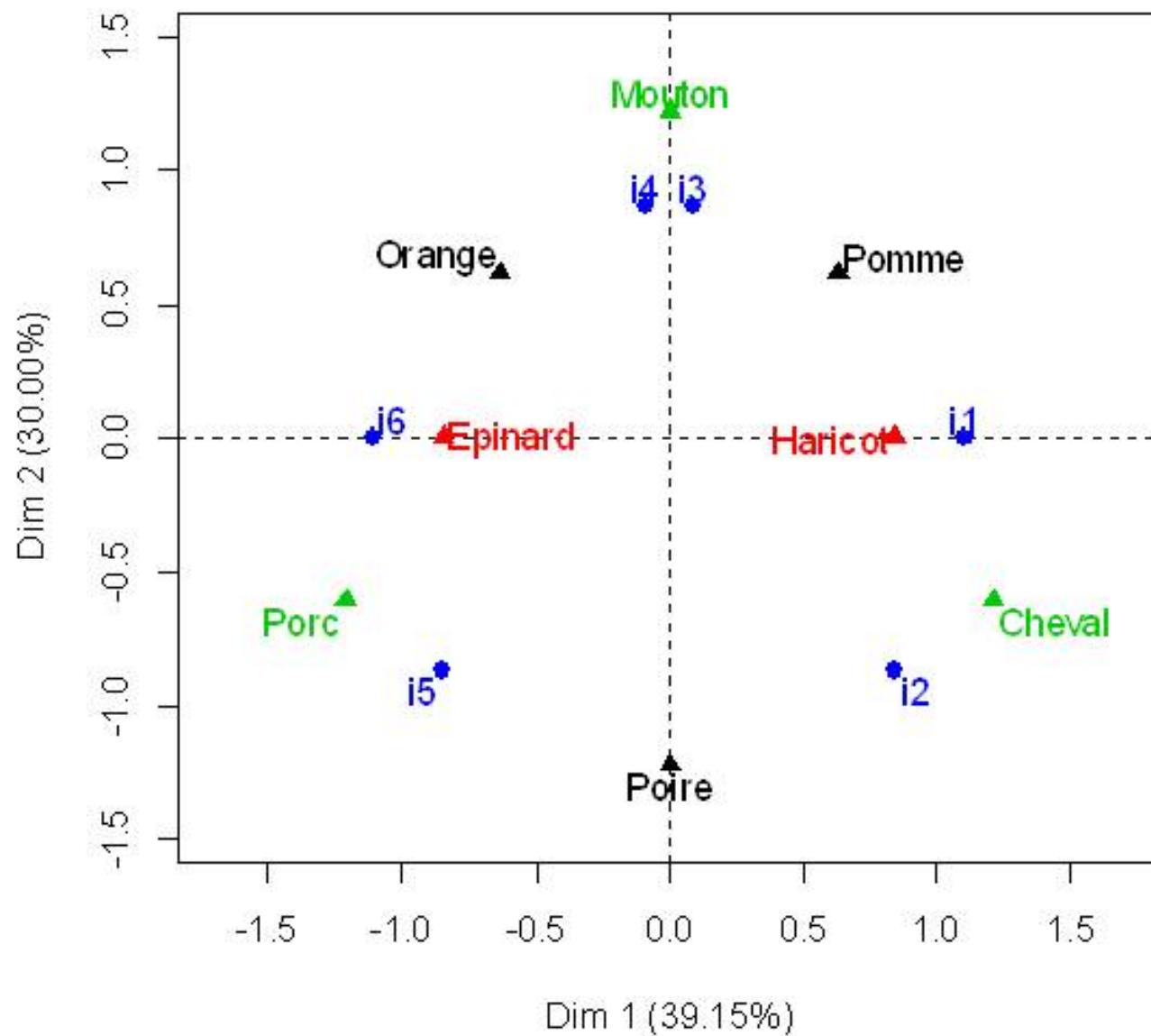
Ce type d'analyse, très utilisé en sociologie, a été popularisé par Pierre Bourdieu dans les années 70. Pour un individu i :

	Fruit	Légume	Viande
i_1	Pomme	Haricot	Cheval
i_2	Poire	Haricot	Cheval
i_3	Orange	Haricot	Mouton
i_4	Pomme	Epinard	Mouton
i_5	Poire	Epinard	Porc
i_6	Orange	Epinard	Porc

On transforme le premier tableau en tableau disjonctif complet:

	Fruit:Pomme	Fruit:Poire	Fruit:Orange	...
i_1	1	0	0	...
i_2	0	1	0	...
i_3	0	0	1	...
i_4	1	0	0	...
i_5	0	1	0	...
i_6	0	0	1	...

Individus et modalités



MDS

Le principe du MDS (*Multidimensional scaling*) est de positionner nos individus dans un espace à N dimensions (normalement 2, voire 3) en fonction de leur similarité/dissimilité. On va donc passer par une tableau représentant la distance entre les différents individus.

L'idée est donc de "comprimer" un ensemble d'informations en une matrice de similarité, sur laquelle on va réaliser une ACP.

Cette déformation étant un peu brutale, il va falloir évaluer la déformation produite. Pour cela on va évaluer le *stress*.

Le principe est le suivant. Notre tableau des notes:

Elève	Mathématiques	Français	Moyenne
Cunégonde	80	30	55
Marcel	50	50	50
Josette	30	60	45

Devient une matrice de similarité (ici basée sur le point de pourcentage).

Elève	Cunégonde	Marcel	Josette
Cunégonde	0	5	10
Marcel	5	0	5
Josette	10	5	0

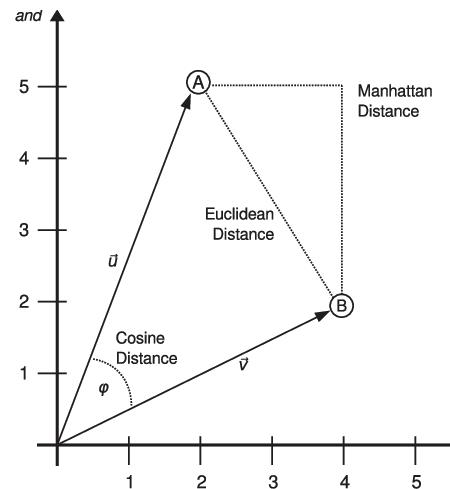


Distance

Quelle distance?

Pour retrouver des textes proches, il faut arriver à mesurer la distance qui les sépare: qu'est-ce que cela eut dire? Comment faire?

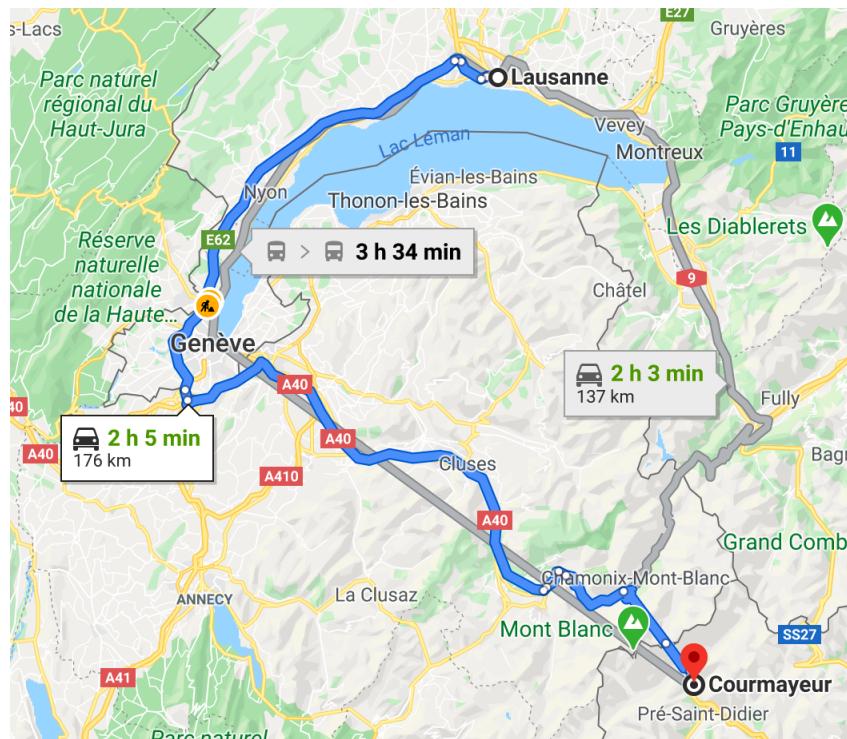
- Il existe une multitude de distances:



- Celles-ci peuvent être calculées sur différents types de données: le nombre d'occurrences, fréquence relative, cote Z...

Distances

La distance "à vol d'oiseau" est un bon indicateur de l'espace qui sépare deux points, mais pas nécessairement le meilleur: pour aller de Lausanne à Courmayeur en voiture, il vaut mieux éviter le Léman et passer par un col...



Source: Google maps

Distance de euclidienne vs distance de Manhattan

La distance euclidienne est la distance la plus connue car elle est la plus intuitive. Cependant, il n'est pas certain que la distance entre Proust et Zola soit de même nature que celle entre New York et Londres.

Il existe des mesures alternatives pour mesurer la distance entre deux points, comme par exemple la distance dite "de Manhattan".



Distance euclidienne:

$$\sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} = \sqrt{(5 - 1)^2 + (4 - 1)^2} = 5$$

Distance de Manhattan:

$$\sum_{i=1}^n |x_i - y_i| = |5 - 1| + |4 - 1| = 7$$



Source: Packtpub, "Calculating the distance"

Cote Z

Plutôt que de prendre les fréquences absolues ou les fréquences absolues, on peut prendre la "cote Z". Elle correspond au nombre d'écart types séparant un résultat de la moyenne:

$$CoteZ = \frac{X - \mu}{\sigma}$$

Où X est un mot, μ est la moyenne et σ est l'écart type.

On normalise les moyennes à 0 et l'écart type à 1: les chiffres tournent donc autour de ces valeurs.

Delta de Burrows

John Burrows 2002 propose de faire une distance de manhattan sur des cotes Z: c'est le "delta de Burrows".

On peut donc déduire le Delta (Δ), soit la différence entre la fréquence (f) d'un mot (i) entre deux documents (D et D') de la manière suivante:

$$Z(f_i(D)) = \frac{f_i(D) - \mu_i}{\sigma_i}$$

$$\Delta(D, D') = \frac{1}{n} \sum_{n=1}^n |Z(f_i(D)) - Z(f_i(D'))|$$

$$\Delta(D, D') = \frac{1}{n} \sum_{n=1}^n \left| \frac{f_i(D) - \mu_i}{\sigma_i} - \frac{f_i(D') - \mu_i}{\sigma_i} \right|$$

$$\Delta(D, D') = \frac{1}{n} \sum_{n=1}^n \left| \frac{f_i(D) - f_i(D')}{\sigma_i} \right|$$

A	B	C	D	F	G	I	J	K	L	N	O	P	Q	S	T	U	V	W	X	Y	Z		
1		Main set		Milton		<i>Paradise Lost</i>				<i>World's Infancy</i>				<i>Paradise Regained</i>				<i>Samson Agonistes</i>					
2		count				30				count				30				count					
3		sum				31.489				sum				36.164				sum					
4		mean (= "delta")				1.050				mean (= "delta")				1.205				mean (= "delta")					
5		stdev				0.770				stdev				1.163				stdev					
6		Mean	Stdev	Scores	z-scores	Scores	z-scores	Diff.	Abs. diff.	Scores	z-scores	Diff.	Abs. diff.	Scores	z-scores	Diff.	Abs. diff.	Scores	z-scores	Diff.	Abs. diff.		
7	1	the	4.242	0.630	4.719	0.757	4.091	-0.239	-0.996	0.996	7.866	5.753	4.996	4.996	3.619	-0.988	-1.746	1.746	2.809	-2.274	-3.031	3.031	
8	2	and	3.770	0.501	4.407	1.272	4.165	0.789	-0.483	0.483	3.474	-0.590	-1.862	1.862	4.441	1.340	0.068	0.068	3.298	-0.940	-2.212	2.212	
9	3	of	1.821	0.315	2.420	1.905	2.769	3.015	1.110	1.110	2.169	1.106	-0.799	0.799	2.765	3.002	1.097	1.097	2.561	2.353	0.448	0.448	
10	4	a	1.601	0.430	0.893	-1.645	0.696	-2.103	-0.458	0.458	1.296	-0.708	0.936	0.936	0.873	-1.691	-0.047	0.047	1.094	-1.177	0.468	0.468	
11	5	to(i)	1.419	0.272	1.247	-0.634	1.289	-0.480	0.154	0.154	0.918	-1.846	-1.212	1.212	1.389	-0.111	0.523	0.523	1.824	1.491	2.124	2.124	
12	6	in(p)	1.358	0.189	1.554	1.035	1.720	1.916	0.881	0.881	1.476	0.624	-0.411	0.411	1.536	0.940	-0.095	0.095	1.552	1.028	-0.007	0.007	
13	7	his	1.154	0.323	1.062	-0.284	1.532	1.171	1.454	1.454	1.359	0.635	0.919	0.919	1.287	0.413	0.696	0.696	1.009	-0.448	-0.165	0.165	
14	8	with	1.022	0.208	1.480	2.202	1.484	2.224	0.022	0.022	0.972	-0.239	-2.441	2.441	1.141	0.572	-1.630	1.630	1.436	1.991	-0.211	0.211	
15	9	to(p)	1.014	0.131	0.999	-0.119	1.245	1.761	1.880	1.880	0.819	-1.493	-1.373	1.373	1.663	4.957	5.077	5.077	1.428	3.161	3.281	3.281	
16	10	is	0.938	0.312	0.502	-1.397	0.239	-2.238	-0.841	0.841	1.233	0.944	2.341	2.341	0.465	-1.515	-0.118	0.118	0.442	-1.588	-0.191	0.191	
17	11	but	0.923	0.195	0.676	-1.268	0.696	-1.167	0.101	0.101	0.378	-2.801	-1.533	1.533	0.765	-0.814	0.453	0.453	0.916	-0.038	1.230	1.230	
18	12	he	0.803	0.241	0.465	-1.403	0.703	-0.413	0.990	0.990	0.603	-0.830	0.573	0.573	0.784	-0.079	1.324	1.324	0.435	-1.529	-0.126	0.126	
19	13	all	0.781	0.193	0.518	-1.366	0.836	0.283	1.649	1.649	0.720	-0.318	1.048	1.048	0.975	1.003	2.369	2.369	0.830	0.254	1.620	1.620	
20	14	I	0.766	0.391	0.882	0.297	0.700	-0.171	-0.467	0.467	0.711	-0.142	-0.438	0.438	1.198	1.103	0.806	0.806	1.676	2.326	2.030	2.030	
21	15	it	0.766	0.239	0.386	-1.591	0.151	-2.575	-0.984	0.984	0.558	-0.870	0.722	0.722	0.299	-1.953	-0.361	0.361	0.450	-1.322	0.270	0.270	
22	16	as	0.710	0.224	0.618	-0.410	0.737	0.119	0.529	0.529	0.540	-0.760	-0.350	0.350	0.701	-0.041	0.369	0.369	0.722	0.053	0.463	0.463	
23	17	their	0.641	0.237	0.513	-0.540	0.795	0.653	1.193	1.193	0.432	-0.880	-0.340	0.340	0.522	-0.498	0.042	0.042	0.761	0.506	1.046	1.046	
24	18	her	0.623	0.336	0.851	0.678	0.435	-0.560	-1.237	1.237	0.756	0.396	-0.282	0.282	0.312	-0.923	-1.601	1.601	0.287	-0.998	-1.675	1.675	
25	19	not	0.616	0.174	0.592	-0.138	0.847	1.324	1.462	1.462	0.432	-1.054	-0.916	0.916	0.841	1.290	1.428	1.428	1.180	3.231	3.369	3.369	
26	20	be	0.586	0.167	0.555	-0.187	0.401	-1.109	-0.921	0.921	0.459	-0.763	-0.576	0.576	0.503	-0.496	-0.309	0.309	0.520	-0.397	-0.209	0.209	
27	21	you	0.580	0.252	0.174	-1.608	0.037	-2.154	-0.546	0.546	0.261	-1.265	0.344	0.344	0.006	-2.275	-0.666	0.666	0.023	-2.208	-0.599	0.599	
28	22	they	0.564	0.234	0.270	-1.259	0.464	-0.428	0.830	0.830	0.396	-0.719	0.540	0.540	0.370	-0.831	0.427	0.427	0.310	-1.084	0.175	0.175	
29	23	for(p)	0.559	0.114	0.270	-2.539	0.000	-4.905	-2.366	2.366	0.342	-1.903	0.637	0.637	0.280	-2.444	0.095	0.095	0.466	-0.817	1.722	1.722	
30	24	by(p)	0.555	0.106	0.412	-1.349	0.689	1.260	2.608	2.608	0.432	-1.162	0.187	0.187	0.822	2.518	3.866	3.866	0.582	0.254	1.603	1.603	
31	25	my	0.512	0.370	0.587	0.201	0.258	-0.687	-0.888	0.888	0.351	-0.435	-0.636	0.636	0.472	-0.110	-0.311	0.311	1.226	1.928	1.727	1.727	
32	26	we	0.510	0.275	0.159	-1.279	0.265	-0.891	0.388	0.388	0.468	-0.153	1.126	1.126	0.127	-1.392	-0.113	0.113	0.124	-1.404	-0.125	0.125	
33	27	from	0.500	0.127	0.534	0.265	0.884	3.019	2.754	2.754	0.567	0.527	0.262	0.262	0.771	2.132	1.866	1.866	0.520	0.157	-0.108	0.108	
34	28	that(rp)	0.476	0.228	0.925	1.964	0.313	-0.715	-2.680	2.680	0.234	-1.061	-3.026	3.026	0.172	-1.333	-3.297	3.297	0.217	-1.135	-3.099	3.099	
35	29	or	0.471	0.165	0.856	2.333	0.906	2.636	0.302	0.302	0.153	-1.929	-4.263	4.263	1.064	3.595	1.261	1.261	0.908	2.648	0.315	0.315	
36	30	our	0.460	0.268	0.270	-0.711	0.354	-0.397	0.314	0.314	0.558	0.366	1.078	1.078	0.319	-0.528	0.183	0.183	0.225	-0.877	-0.166	0.166	

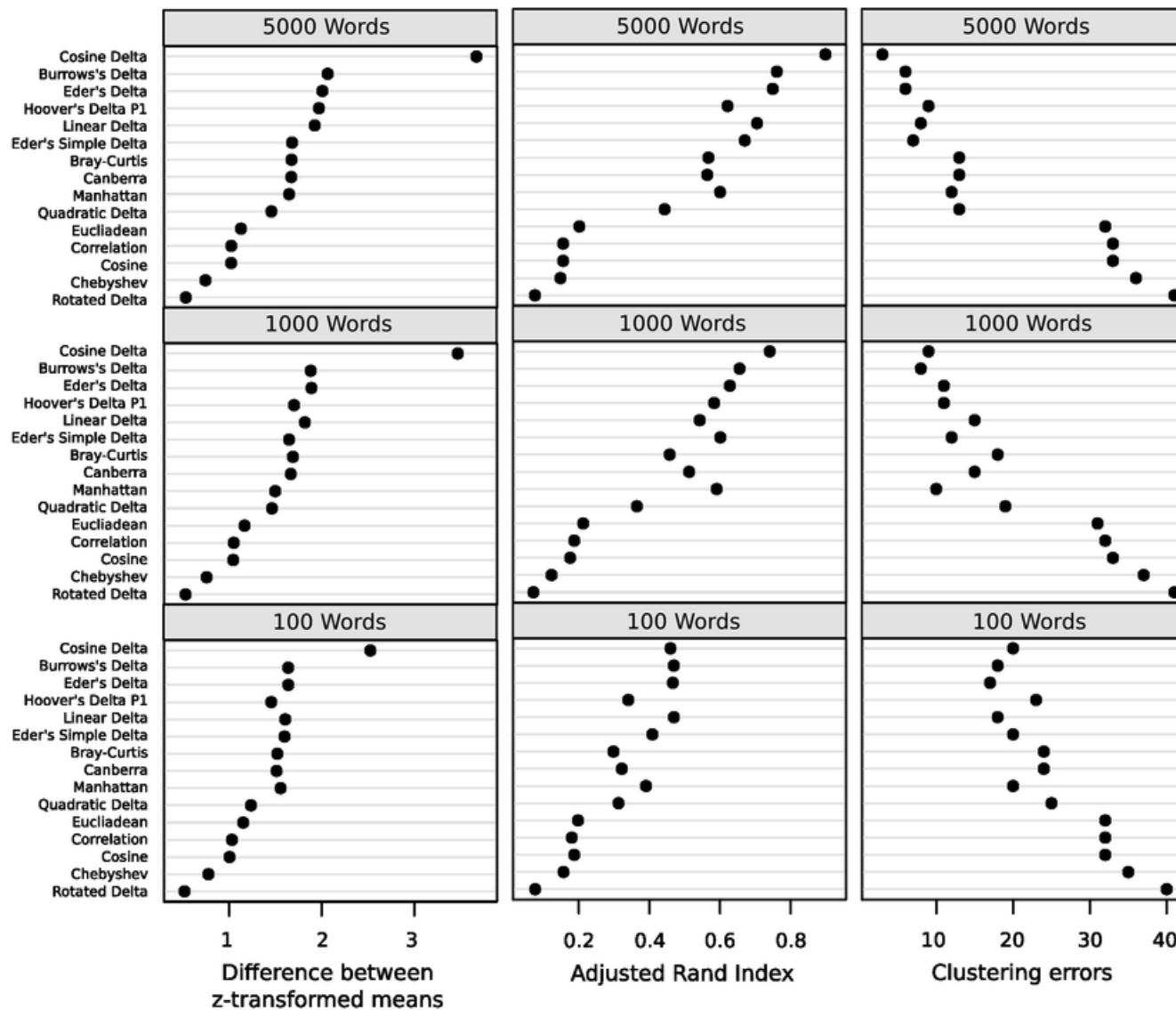
Source: Burrows, " 'Delta': a Measure of Stylistic Difference and a Guide to Likely Authorship"

Et tant d'autres...

Distance/Similarity measure	Formula
Manhattan Distance	$\sum_{i=1}^n x_i - y_i $
Euclidean Distance	$\sqrt{\sum_{i=1}^n (x_i - y_i)^2}$
Canberra Distance	$\sum_{i=1}^n \frac{ x_i - y_i }{ x_i + y_i }$
Cosine Distance	$\frac{\sum_{i=1}^n x_i y_i}{\sqrt{\sum_{i=1}^n x_i^2} \sqrt{\sum_{i=1}^n y_i^2}}$
Burrows' Delta	$\frac{1}{n} \sum_{i=1}^n \left \frac{x_i - \mu_i}{\sigma_i} - \frac{y_i - \mu_i}{\sigma_i} \right $
Argamon's Linear Delta	$\frac{1}{n} \sum_{i=1}^n \sqrt{\left \frac{(x_i - y_i)^2}{\sigma_i^2} \right }$
Eder's Delta	$\frac{1}{n} \sum_{i=1}^n \left(\left \frac{x_i - y_i}{\sigma_i} \right \cdot \frac{n - n_i + 1}{n} \right)$
Eder's Simple Delta [9]	$\sum_{i=1}^n \sqrt{x_i} - \sqrt{y_i} $
Argamon's Quadratic Delta	$\frac{1}{n} \sum_{i=1}^n \frac{1}{\sigma_i^2} (x_i - y_i)^2$
Bray-Curtis Dissimilarity	$\frac{\sum_{i=1}^n x_i - y_i }{\sum_{i=1}^n (x_i + y_i)}$
Kulczynski Distance	$\frac{\sum_{i=1}^n x_i - y_i }{\sum_{i=1}^n \min(x_i, y_i)}$
Jaccard Index	$\frac{2 \sum_{i=1}^n x_i - y_i }{\sum_{i=1}^n (x_i + y_i)}$ $\frac{\sum_{i=1}^n x_i - y_i }{1 + \sum_{i=1}^n (x_i + y_i)}$
Gower Similarity	$\frac{1}{n} \sum_{i=1}^n \frac{ x_i - y_i }{\max_i - \min_i}$
Alternative Gower Similarity	$\frac{1}{n_0} \cdot \sum_{i=1}^n x_i - y_i $
Horn's modification of Morisita's Overlap Index	$\frac{2 \sum_{i=1}^n x_i y_i}{\left(\frac{\sum_{i=1}^n x_i^2}{(\sum_{i=1}^n x_i)^2} + \frac{\sum_{i=1}^n y_i^2}{(\sum_{i=1}^n y_i)^2} \right) \sum_{i=1}^n x_i \sum_{i=1}^n y_i}$
Mountford Index	$\frac{1}{\alpha}$, where α is the parameter of Fisher's log-series
Binomial Index [5]	$\sum_{i=1}^n \frac{x_i \cdot \ln \frac{x_i}{2n} + y_i \cdot \ln \frac{y_i}{2n} - 2n \cdot \ln \frac{1}{2}}{2n}$

Source: Stanikūnas, Mandravickaitė & Krilavičius 2017

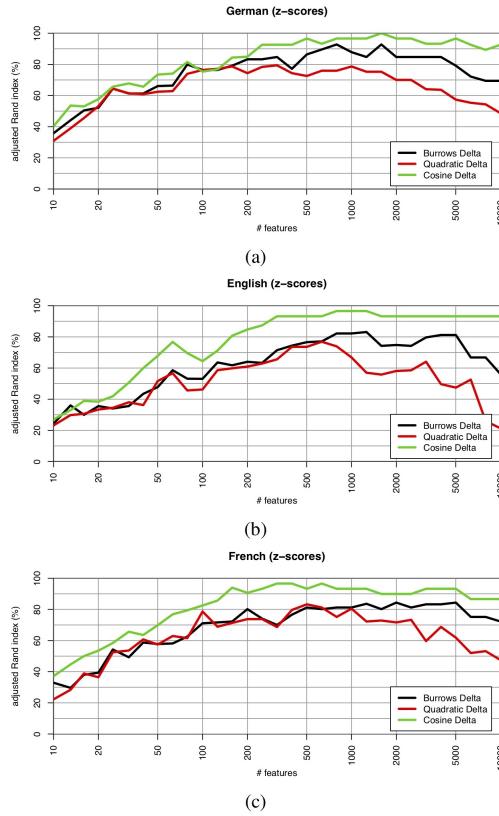
Performance of distance measures on 19th c. English novels:



Source: Evert, Proisl, Jannidis, Reger, Pielström, Schöch, Vitt, 2017.

Différences entre langues

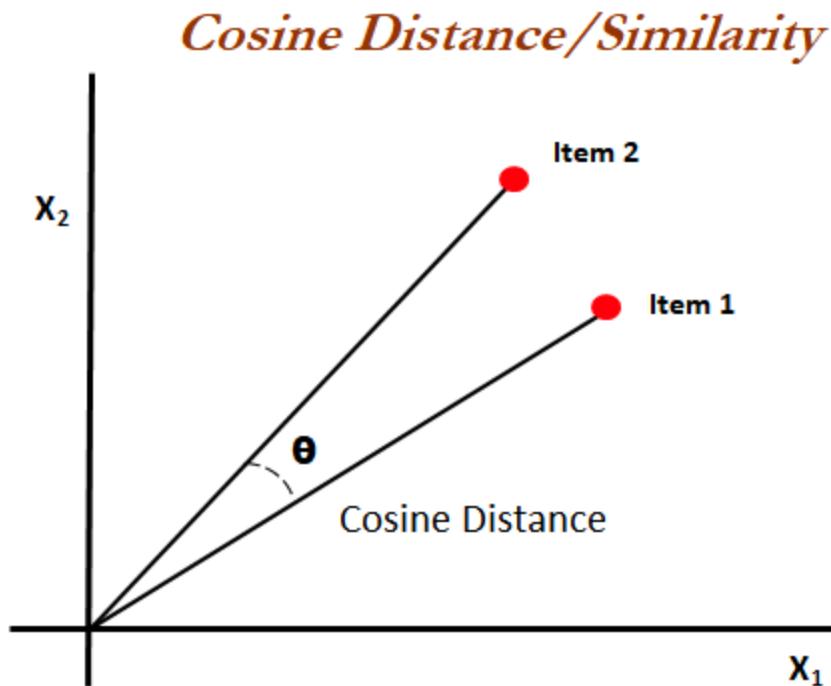
Attention, ces scores varient d'une langue à l'autre.



Source: Evert, Proisl, Jannidis, Pielström, Schöch, Vitt, "Towards a better understanding of Burrows's Delta in literary authorship attribution"

Similarité de cosinus

L'idée est de calculer l'angle θ formé par deux vecteurs A et B pour évaluer leur similarité:



L'angle θ s'obtient par le produit scalaire et la norme des vecteurs. Pour deux vecteurs x et y , et deux vecteurs A et B :

$$\cos(\theta) = sim(x, y) = \frac{x \cdot y}{\|x\| \|y\|}$$

$$sim(x, y) = \frac{\sum_{n=1}^n A_i \cdot B_i}{\sqrt{\sum_{n=1}^n A_i^2} \sqrt{\sum_{n=1}^n B_i^2}}$$

$$sim(x, y) = \frac{x_1 \times y_1 + x_2 \times y_2 \dots x_n \times y_n}{\sqrt{x_1^2 + x_2^2 \dots x_n^2} \quad \sqrt{y_1^2 + y_2^2 \dots y_n^2}}$$

Cosine delta

Prenons un exemple fictif avec deux textes de trois tokens, l'un écrit par Molière, l'autre par Corneille:

Auteur/token	Avoir	Être	Manger
Molière	2	1	4
Corneille	3	2	5

La similarité entre ces deux textes serait donc la suivante:

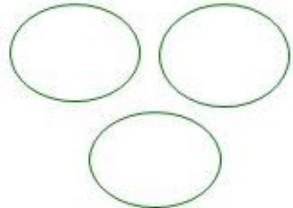
$$sim(Moliere, Racine) = \frac{2 \times 3 + 1 \times 2 + 5 \times 4}{\sqrt{2^2 + 1^2 + 5^2} \quad \sqrt{3^2 + 2^2 + 4^2}}$$

Comme nous parlons de *cosine delta*, il faudrait préalablement normaliser les vecteurs, et donc z-transformer les valeurs du tableau.

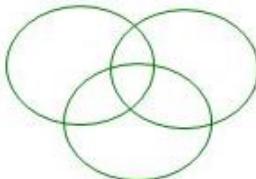
Classification

Stratégies d'agrégation

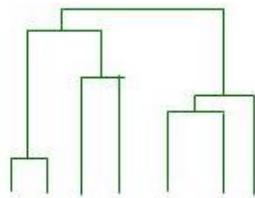
Non overlapping



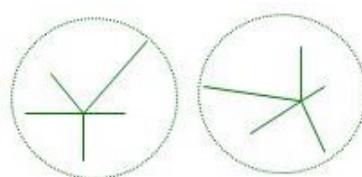
Overlapping



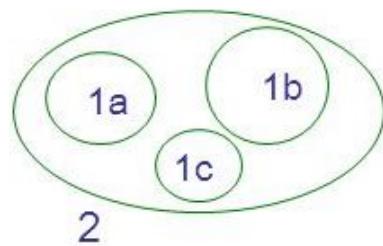
Hierarchical



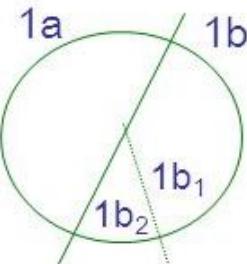
Non-hierarchical



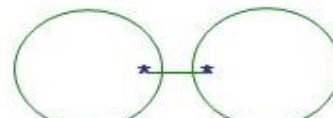
Agglomerative



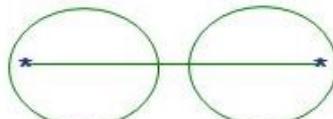
Divisive



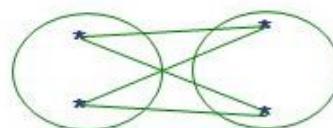
Single Linkage:
Minimum distance



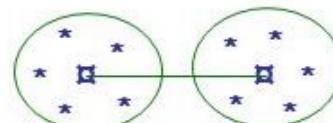
Complete Linkage:
Maximum distance



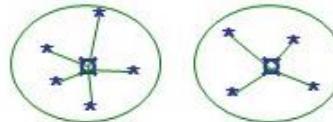
Average Linkage:
Average distance



Centroid method:
*Distance between
centres*

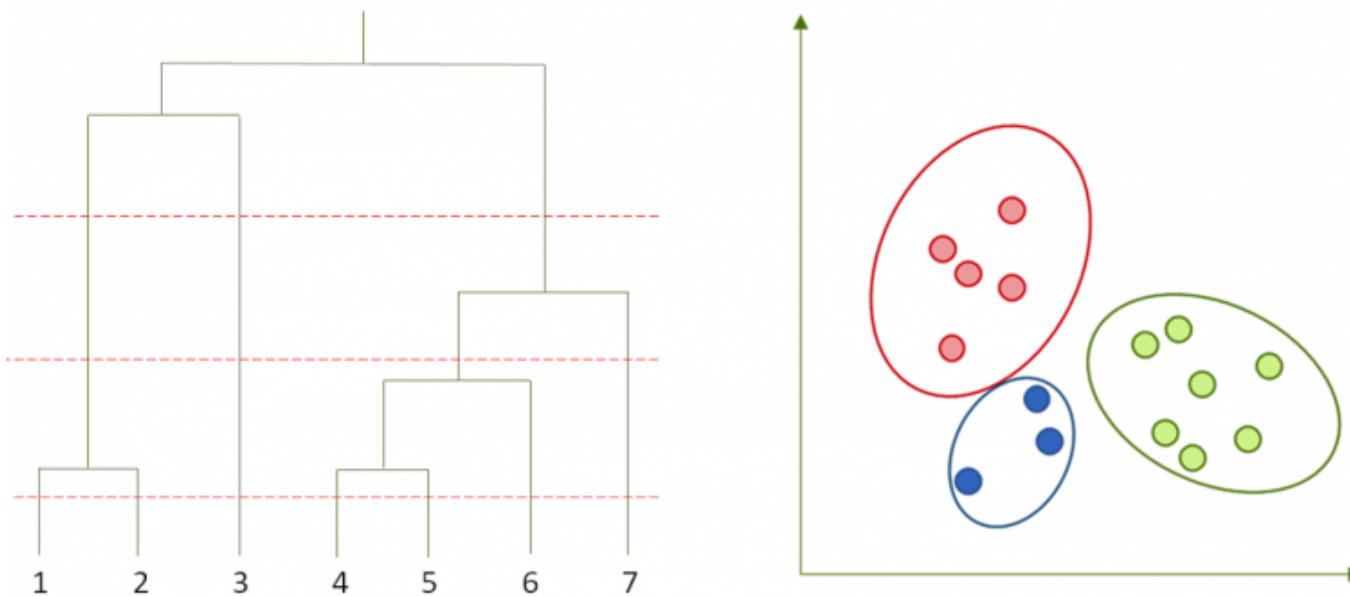


Wards method:
*Minimization of
within-cluster variance*



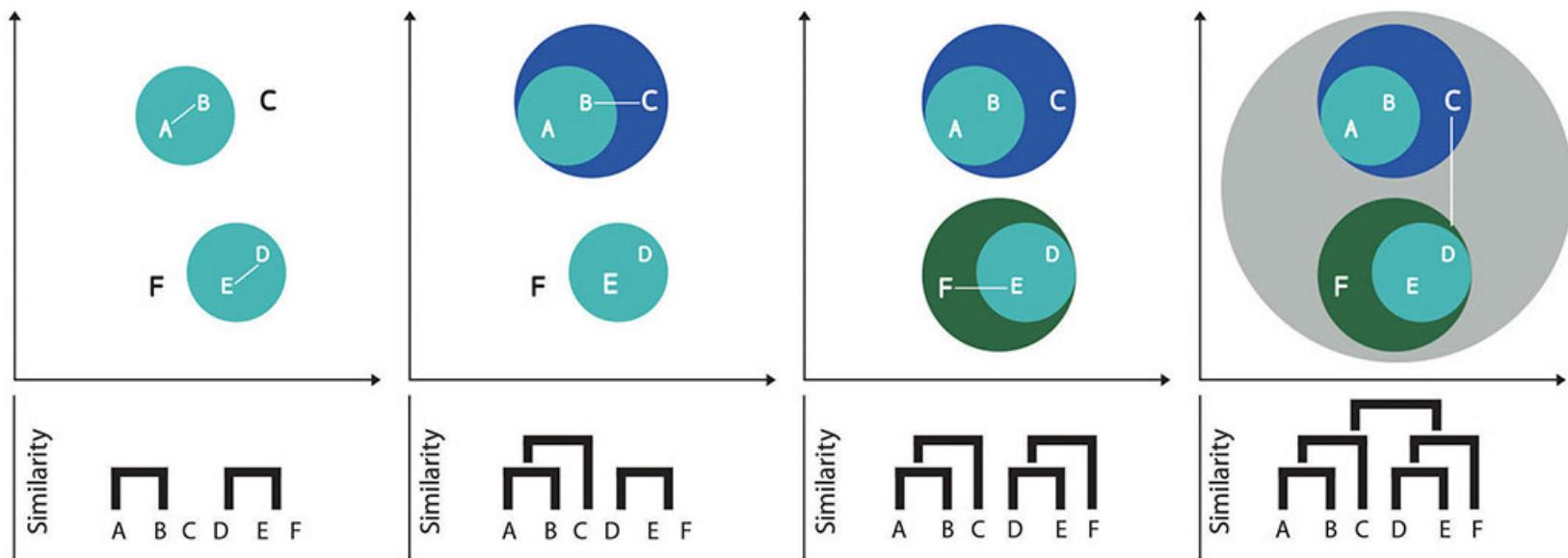
CAH

En stylométrie, on a tendance à adopter une approche hiérarchique (à gauche)



Cette représentation est un dendrogramme, et elle permet de représenter une classification ascendante hiérarchique (CAH).

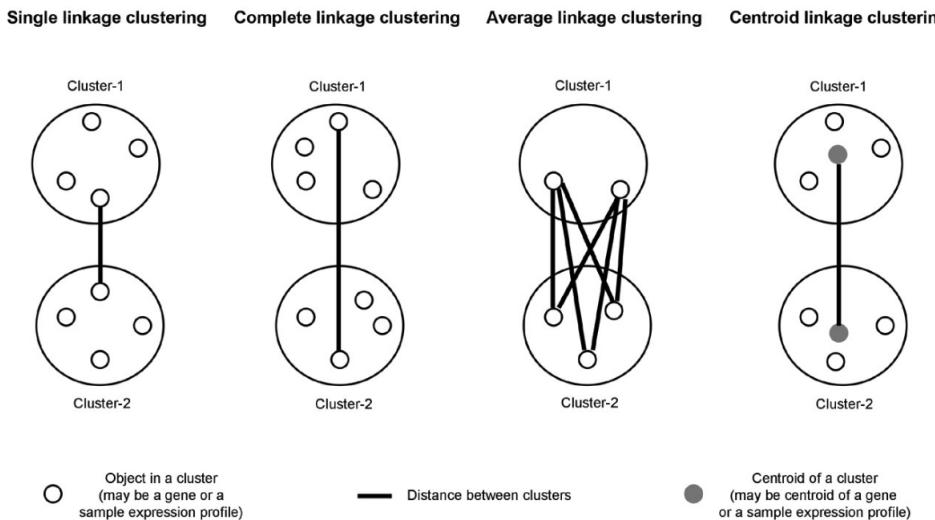
L'objectif d'une CAH est de constituer séquentiellement des paires à partir des mesures de distances/similarité jusqu'à n'obtenir qu'une seule classe:



En stylométrie, on va constituer des paires sur les calculs de distance.

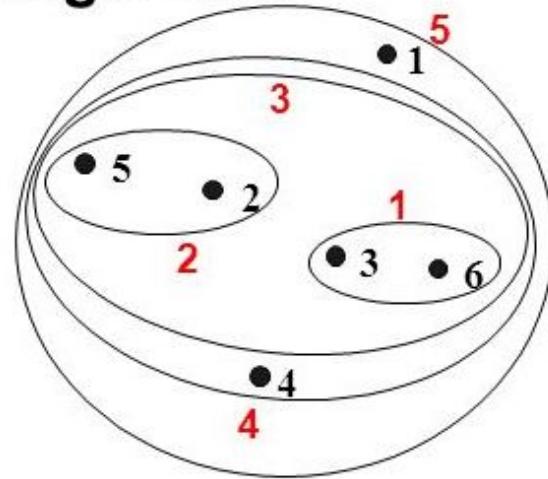
L'objectif est de regrouper deux classes d'une partition pour obtenir une partition plus agrégée: mais comment faire?

- Le *complete linkage*, qui calcule la distance maximale entre deux points est recommandée par Burrows 2002
- Le *single linkage* calcule la distance minimale entre deux points
- L'*average linkage* calcule toutes les distances entre les différents points et en fait la moyenne
- La distance entre les centroïdes calcule la distance entre les centres de gravité

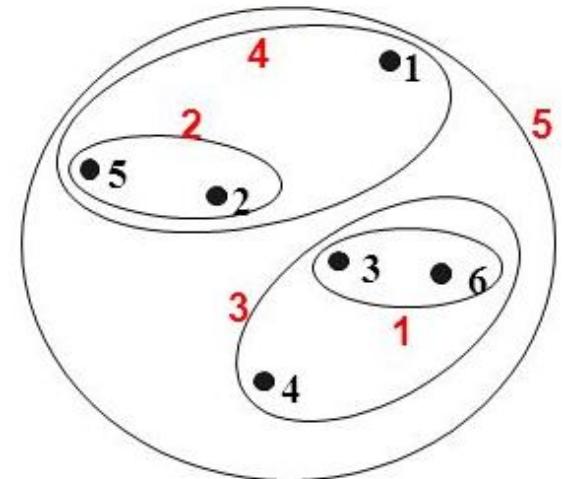


La méthode d'agrégation a un impact important sur le résultat

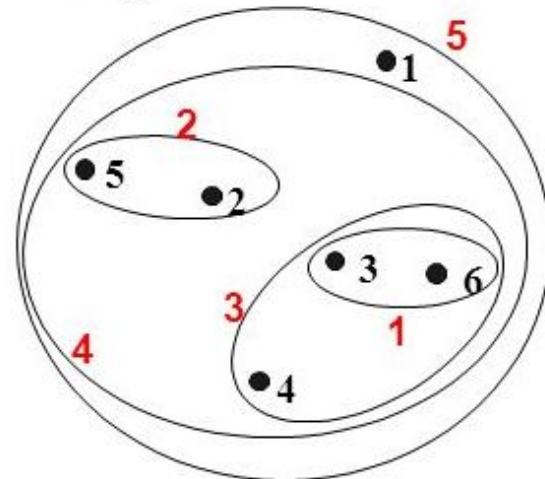
Single-link



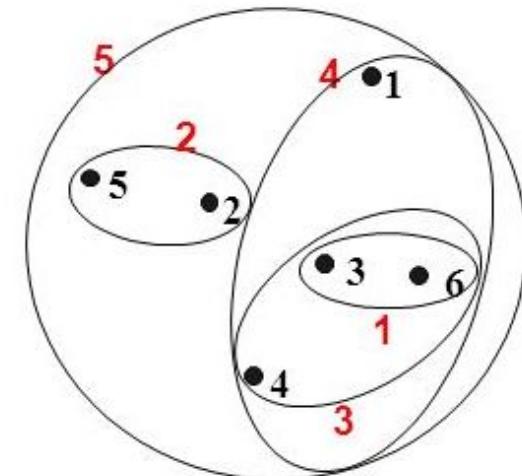
Complete-link



Average-link



Centroid distance



Source: [slideplayer](#)

La Méthode de Ward

La méthode de Ward, recommandé par Hoover 2003b et qui **est la plus utilisé aujourd'hui**, minimise la variance (ou inertie) intra-classe, et maximise (logiquement) la variance (ou inertie) interclasse.

Comme elle minimise la distance centroïde, elle constitue des groupes homogènes.

Remerciements/sources

Merci à JB Camps et Fl. Cafiero, qui retrouveront une partie de leur enseignement dans ces *slides*.