Informatique, linguistique & politique : un bien curieux mélange

Jacques Savoy
Institut d'informatique
Université de Neuchâtel

Avant-propos

- 2008, Année de l'informatique en Suisse
- Qu'est-ce que l'informatique ?
  - L'insomniaque (autiste) devant son écran ?
  - Le bricoleur amoureux du tournevis ?
  - Le beau parleur avec ses acronymes (HTTP, P2P, ADSL, flux RSS) ?
- Gérer l'information sous plusieurs formes (multilingues, mathématiques, ouverture)

Quels intérêts …

Qu'est-ce que l'informatique (statistiques) peut apporter des éléments de réflexion en sciences humaines ?

1. Analyse du discours politique
2. Affinités politiques entre cantons

Internet
Avec le risque de *surestimer* son impact à court terme et de *sous-estimer* son importance à long terme

Linguistique …

- Etude scientifique du language
- Parenté entre langues
- Phonologie, morphologie (mots et règles), syntaxe, sémantique
- Mais avec des liens avec la technologie

Correcteur d'orthographe

Traduction automatique

Moteur de recherche (question/réponse)

Analyse du discours …

- Statistique lexicale / textuelle
- Comment attribuer une œuvre littéraire à son auteur ? Ou à un homme de plume …
- Comment distinguer le discours de Ségolène et de Nicolas ?
- Comment distinguer les discours des divers présidents (mesurer leurs différences) ?

L'affaire Molière-Corneille

- Pierre Louys (octobre 1919) s'interroge sur la paternité des œuvres de Molière
- *Le Misanthrope*, *Don Juan*, *Amphitryon* et *Tartuffe* sont l'œuvre de Corneille totalement ou *en partie*
- Base : la versification, la prosodie, le style (dans *Amphitryon*)
- Qui sont les protagonistes ?

L'affaire Molière-Corneille

- Jean Baptiste Poquelin (1622-1673)
- 1645-1959 (14 ans) années difficiles production faible
- 1659-1673 (14 ans) production abondante comédien, directeur du théâtre du Roi
- 1658 Corneille & Molière à Rouen

L'affaire Molière-Corneille

- Pierre Corneille (1606-1684)
- *Le Cid* (1636)
- Se venger des critiques faites a *Polyeucte* (1643)
- 1647 élu à l'*Académie Française*
- Difficile de critiquer (La Bastille)
- Comédie, genre jugé indigne
- Besoin d'argent (?)

L'affaire Molière-Corneille

- Pour *Psyché* (1671), pas de doute, les deux auteurs ont écrits ensemble
- Possible dans d'autre cas (Lully)
- Pas de manuscrit retrouvé chez Molière après sa mort soudaine. Et pourtant Molière était un homme ordonné.
- Mais sujet à des vives critiques, on a jamais contesté la paternité de ses œuvres de son vivant.
- On n'a pas trouvé de manuscrit chez Racine ou peu chez Corneille

L'affaire Molière-Corneille

Les oeuvres discutables
L'Etourdi (1658), *Le Dépit amoureux* (1658), *Sganarelle ou le cocu imaginaire* (1660), *Dom Garcie de Navarre* (1661), *L'Ecole des Maris* (1661), *Les Fâcheux* (1661), *L'Ecole des Femmes* (1662), *La Princesse d'Elide* (1664), *Le Tartuffe* (1664), *Dom Juan* (1665), *Le Misanthrope* (1666), *Mélicerte* (1666), *Amphitryon* (1668), *L'Avare* (1668), *Psyché* (1671), *Les Femmes savantes* (1672)

L'affaire Molière-Corneille

- Des présomptions, un faisceau d'indices troublants (pas toujours concordants), des intérêts communs entre Corneille & Molière …
- Pierre Louys (1919) les confronte sur la base du style (versification, rythme des vers) et d'un étude minutieuse et comparative
- Et l'informatique dans tout cela …

L'affaire Molière-Corneille

- Notre but : mesurer une distance intertextuelle (D. Labbé)
  Si les deux textes sont similaires (proches) la distance doit être faible. Si les deux textes sont très différents, la distance doit être élevée.
  - Avec une valeur = 0
    (deux textes ayant le même vocabulaire)
  - Et valeur = 1
    (pas un mot en commun)
- Comment procéder ?

Distance intertextuelle

« Quoi ! tu ne me dis mot ! Crois-tu que ton silence
Puisse de tes discours réparer l'insolence ?
Des pleurs effacent-ils un mépris si cuisant,
Et ne t'en dédis-tu, traître, qu'en te taisant ?
Pour triompher de moi, veux-tu, pour toutes armes»
Corneille, *La Place royale*, III, 6.

> « Ah ! que vous savez bien ici, contre moi-même,
> Perfide, vous servir de ma faiblesse extrême,
> Et ménager pour vous l'excès prodigieux
> De ce fatal amour né de vos traîtres yeux !
> Défendez-vous au moins d'un crime qui m'accable,
> Et cessez d'affecter d'être avec moi coupable.»
> *Le Misanthrope*, IV, 3

Distanc

Sur la base des lettres et de leur distribution ?

- Voyez-vous, ce qui me dérange, moi, chez Molière, c'est cette surabondance de "a". C'est pourquoi je lis plus volontiers du Racine …
- Même s'il y a plus de "i" et de "o" ?
- Ah oui, c'est vraiment le "a" qui m'insupporte !

Distance intertextuelle

- Sur la base du vocabulaire
- Quelles sont les formes les plus fréquentes ?
- Sont-elles vraiment très fréquentes par rapport aux autres ? Ou est-ce que les mots possèdent tous des fréquences plus ou moins similaires ?
- L'apport de l'informatique devient évidente. Donnons à l'ordinateur un bon corpus de textes à "digérer"

Le vocabulaire

Quelles sont les formes les plus fréquentes ?

Le journal *Le Monde* et l'*Agence Télégraphique Suisse*

nombre de mots 60 520 416
nombre formes 389 613

Français	
fréquence	forme
3 498 779	de
1 766 953	la
1 341 260	l
1 222 098	le
1 062 055	et
1 061 040	des
1 030 069	les
996 578	d
819 757	en
788 890	du

Le vocabulaire

Les formes correspondent à des mots-outils, peu liées à un contenu sémantique précis.

Français	
fréquence	forme
3 498 779	de
1 766 953	la
1 341 260	l
1 222 098	le
1 062 055	et
1 061 040	des
1 030 069	les
996 578	d
819 757	en
788 890	du

Dans d'autres langues

1.	der	de	di	the
2.	die	la	e	of
3.	und	l	il	to
4.	in	le	la	a
5.	den	et	che	and
6.	von	des	a	in
7.	das	les	un	s
8.	mit	d	per	that
9.	im	en	l	for
10.	zu	du	del	is

Les dix mots les plus fréquents

16 % de l'allemand ou l'italien

23,5% du français, 21,6% de l'anglais

Distance intertextuelle

Distance entre deux auteurs A1 et A2

$$D(A1, A2) = 1 - \frac{|A1 \cap A2|}{|A1 \cup A2|}$$

$$D(A1, A2) = 1 - \frac{9}{10} = 0.1$$

Auteur 1		Auteur 2	
fréquence	forme	fréquence	forme
4 250	de	5 637	de
3 254	la	4 466	la
2 996	et	2 869	l
2 317	les	2 516	que
2 126	l	2 393	les
1 855	le	2 360	le
1 578	que	2 333	et
1 521	**des**	2 248	est
1 474	est	1 980	qui
1 469	qui	1 862	**je**

Les discours politiques

- Autre exemple : les discours présidentiels
- Plusieurs auteurs différents

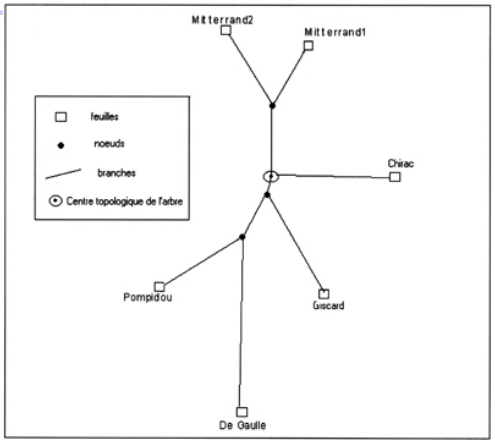
de Gaulle (1958-1969)
Pompidou (1969-1974)
Giscard (1974-1981)
Mitterand1 (1981-1988)
Mitterand2 (1988-1995)
Chirac (1995-2002)

Les discours politiques

- Lesquels sont les plus similaires / distants ?
- Quel président se rapproche le plus de de Gaulle ?

Analyse des discours

Le discours présidentiel français sous la Ve République (1958-2002)



Les discours politiques

- Deux grands discours sous la Ve république
  - le gaulliste et le mitterandien (les deux extrêmes)
  - le centre par Giscard et Chirac
- Et les distances selon le vocabulaire
  - Distance (De Gaulle - Mitterand2) = 0,229
  - Distance (Mitterand1 - Mitterand2) = 0,106
  - Distance (De Gaulle - Pompidou) = 0,158
  - Distance (De Gaulle - Chirac) = 0,218
- La chronologie n'est pas respectée
- Différence de terminologie
  "Immigration" pour Chirac, "Immigrants" pour Mitterand

Les discours politiques

- Mais le style change Discours nominal ou verbal
  - de Gaulle & Pompidou : noms, adjectifs
  - Mitterand : pronoms, verbes, adverbes
- La mort du politique (D. Mayaffre & X. Luong)
  - la surabondance du "je"
  - le verbe devient de plus en plus fréquent
  - Apparition d'une "novlangue" (G. Orwell, *1984*), d'un politiquement correct avec ses formes simples, rassurantes et sans ambiguïté ?

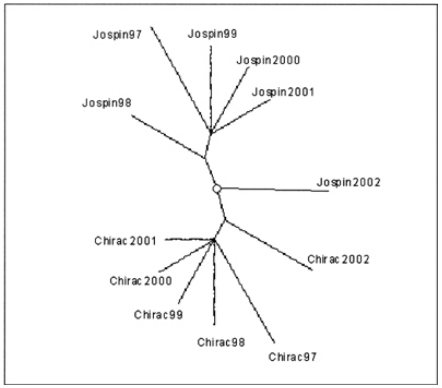
Les discours politiques

- Et les années 1997 - 2002 ?
- Même période mais deux hommes politiques ayant une fonction dans le pouvoir exécutif

- Cohabitation du président (Chirac) et du 1er ministre (Jospin)
- Distance entre les discours de l'un et l'autre

Analyse des discours

- Le discours de la cohabitation (1997-2002)
- Chaque auteur est bien distinct
- La chronologie est (plus ou moins) respectée



Les discours politiques

- Sauf le discours Jospin 2002 et Chirac 2002
  On va de la plus grande différence (1997) vers, d'année en année, un rapprochement
- L'année électoral 2002
  - Les deux vont vers le centre
  - Les deux discours se rapprochent
  - mais Jospin02 est un discours assez éloigné des autres années (le plus distant est Jospin97). Les mots ont changé et les électeurs l'ont compris.

Les discours politiques

Ségolène Royale (11 discours)
Nicolas Sarkozy (17 discours)

Richesse lexicale après 90 000 formes
7 970 mots chez Nicolas
8 031 mots chez Ségolène
10 671 mots dans les articles de presse

Comparer des textes

Nicolas		Ségolène		*Le Monde-ATS*	
fréquence	**forme**	**fréquence**	**forme**	**fréquence**	**forme**
4 250	de	5 637	de	3 498 779	de
3 254	la	4 466	la	1 766 953	la
2 996	et	2 869	l	1 341 260	l
2 317	les	2 516	que	1 222 098	le
2 126	l	2 393	les	1 062 055	et
1 855	le	2 360	le	1 061 040	des
1 578	que	2 333	et	1 030 069	les
1 521	**des**	2 248	est	996 578	**d**
1 474	est	1 980	qui	819 757	**en**
1 469	qui	1 862	**je**	788 890	**du**

Le discours politique se distingue des autres

Les discours politiques

- Nicolas, Ségolène et la presse
- Le pronom "je" (11e et 10e vs. 97e dans la presse)
  le discours politique (électoral) a sa propre saveur
- idem avec "m", "me" ou "moi" (88e, 94e, 945e)
- Abondance de pronoms (nous, vous)
- Noms
  France (25e, 27e, 74e) et "français", "politique", "république" vs. "ans", "francs", "président"

Les discours politiques

- Différence Nicolas & Ségolène
- "femmes" (337e) vs. "femmes" (80e)
- "hommes" (111e) vs. "homme" (282e)
- "parler" (101e) vs. "parler" (378e)

Nicolas: "état", "culture", "enfants", "peut", "faut", "veut"

Ségolène: "jeunes", "pacte", "Europe", "entreprises", "salariés", "ensemble"

Comparer des textes

forme	rang Nicolas	rang Ségolène
sécurité	225	123
identité	126	1 063
emploi	161	103
droit	140	104
histoire	116	192

trigrammes	Fréquent chez
une France qui	Ségolène
la lutte contre	Ségolène
je vous propose	Ségolène
Si je suis [élu]	Nicolas
parce que je	Nicolas
je veux être	Nicolas

Les discours politiques en CH

- La plate-forme électorale proposée par les quatre grands partis de Suisse (leur site Internet)
- La distance entre les programmes restent assez faible mais tous se diffèrent des dépêches d'agence de l'ATS (de l'UDC: 0,345 au PRD: 0,415).

PS - PRD : 0,275 & PS - UDC : 0,275
PDC - UDC : 0,28
PRD - PDC : 0,285

Les discours politiques en CH

Et les mots (pleins) les plus fréquents ?

PS	PDC	PRD	UDC
nous 19	suisse 19	suisse 22	suisse 13
politique 21	nous 29	doit 23	politique 26
doit 24		politique 25	
suisse 28		nous 26	
		sécurité 30	

Les discours politiques en CH

et les sept mots les plus fréquents

PS	PDC	PRD	UDC
nous	*suisse*	*suisse*	*suisse*
politique	*nous*	doit	*politique*
doit	*politique*	*politique*	AI
suisse	doit	*nous*	droit
culturelle	formation	sécurité	état
sociale	jeune	doivent	étranger
droit	enfant	armée	fédéral

En résumé

- Le discours politique change nettement avec le temps (en France pour le moins)
- La fréquence des formes (mais on peut également traiter les catégories grammaticales comme nom, verbe, pronom). Le "je" en France, le "nous" en Suisse.
- Attribuer une œuvre a son auteur possède d'autres applications (homme de plume, poème de Shakespeare)
- Outil complémentaire à l'analyse sémantique

Affinités entre cantons

- Deuxième question
  Vers un nouveau découpage politique en Suisse
- Applications de méthodes informatiques utilisées dans la gestion des cartes de fidélité

Analyse politique …

- Comment analyser / comprendre les oppositions en Suisse ?
- Le « Röstigraben » ?
- Durant son histoire
  - Opposition Ville - Campagne
  - Opposition Catholique - Protestant
- Nous désirons expliquer les votations fédérales (pas les élections)
- Pas une seule votation mais un groupe

Analyse politique …

- Comment comprendre la Suisse du XXIe siècle ?
- Simple : Le « Röstigraben »
- Oui mais cela implique
  - Une seule entité « Suisse Romande »
  - Une entité monolithique « Suisse Alémanique »
  - et le Tessin ?
- Clé pour expliquer une votation donnée

La Suisse du XXI^{e} siècle

- Comment définir des cantons "proches"
- Comment calculer une distance politique entre cantons
- Pourcentage de "oui"
  - Pas le taux de participation
  - Pas le nombre de "oui"
  - Pas les objets acceptés ou refusés
  - Donc une différence entre 49 % et 51 % sera plus faible qu'entre 35 % et 40 %

La Suisse du XXIe siècle

- Les pourcentages d'acceptation depuis 1950 à 2007
- Subdivisé en tranches d'environ dix ans
  1950-59, 1960-69, 1970-78, 1979-1989, 1990-1999, 2000-2007
  Stabilité des mentalités durant les dix ans
- La dernière période représente 73 dernières votations fédérales

La Suisse du XXe siècle

Quelques cas extrêmes (plus forte différence)

- 1950-1959
- Arrêté fédéral concernant la construction d'abris antiaériens dans les bâtiments existants (5 octobre 1952)
- 21,8 % (GR) – 7,6 % (UR) = 14.2 %
- Ce n'est pas entre Romands et Alémaniques ni entre ville et campagne …

La Suisse du XXe siècle

Quelques cas extrêmes

- 1960-1969
- Arrêté fédéral modifiant l'article 72 de la constitution (élection du Conseil national) (4 novembre 1962)
- 93,0 % (GE) – 14,0 % (GL) = **79 %**

La Suisse du XXe siècle

Plus forte variabilité entre cantons

- Arrêté fédéral concernant l'initiative populaire « demandant l'harmonisation du début de l'année scolaire dans tous les cantons » (22 septembre 1985)
- Arrêté fédéral abrogeant les articles de la constitution fédérale sur les jésuites et les couvents (20 mai 1973)

La Suisse du XXIe siècle

Quelques cas extrêmes

- 2000-2007
- Arrêté fédéral concernant la réforme de la péréquation financière (28 novembre 2004)
- 81,9 % (UR) – 16,3 % (ZG) = **65,6** %

- « Pour une caisse maladie unique et sociale » (11 mars 2007)
- « Acquisition de la nationalité suisse par la troisième génération » (26 septembre 2004)

La Suisse du XXIe siècle

Quelques cas extrêmes (2000-2007)

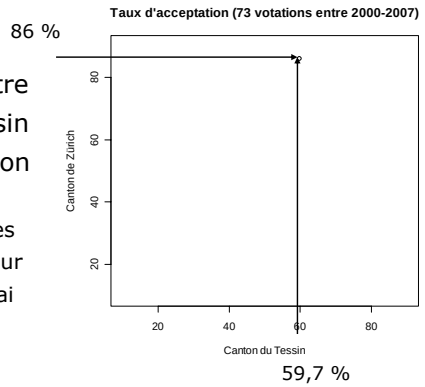
- Les cantons votent d'une même voix
- « Pour des coûts hospitaliers moins élevés » (26 novembre 2000)
- écart type = **3,1** %

- Variation forte sur un votation donnée mais sur un ensemble ?
- Peut-on visualiser les données que nous avons

La Suisse du XXIe siècle

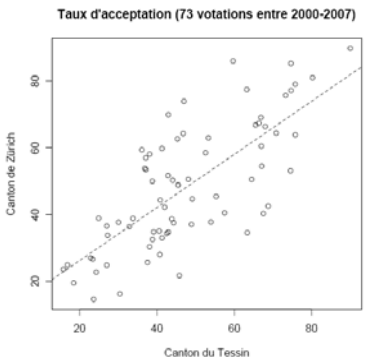
Par exemple entre Zürich et le Tessin une seule votation

Arrêté fédéral modifiant les articles de la Constitution sur la formation (21 mai 2006)



La Suisse du XXIe siècle

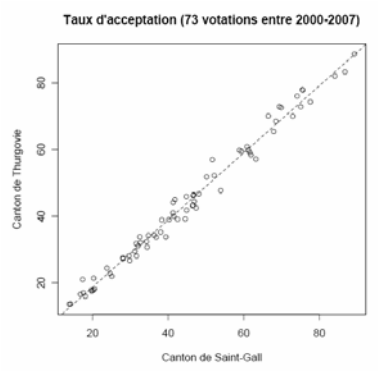
Par exemple entre le canton de Zürich et celui du Tessin



La Suisse du XXI^{e} siècle

La paire la plus similaire ?

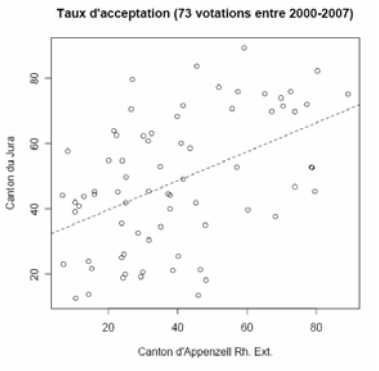
Les cantons de Thurgovie et de St-Gall



La Suisse du XXI^{e} siècle

La paire la plus différente ?

Les cantons du Jura et d'Appenzell Rhodes Extérieures

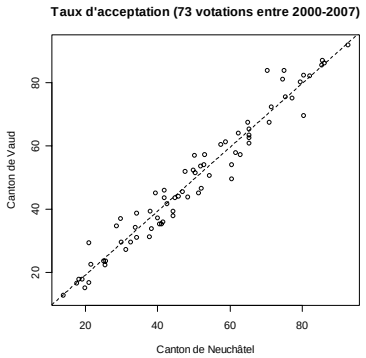


La Suisse du XXI^{e} siècle

Autre exemple et entre Romands cette fois

Pour Neuchâtel, le canton le plus proche c'est …

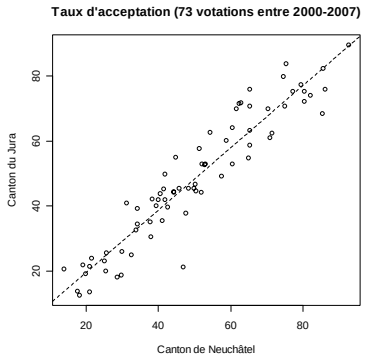
Vaud



La Suisse du XXI^{e} siècle

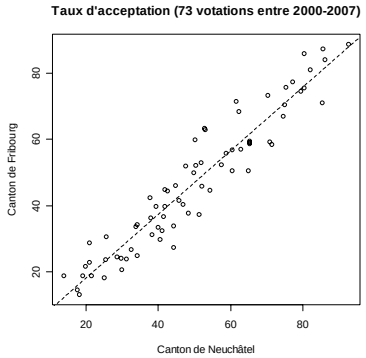
Autre exemple plus près de nous

Neuchâtel - Jura



La Suisse du XXIe siècle

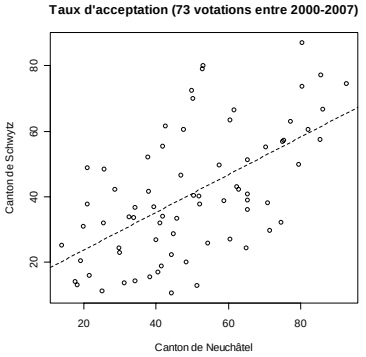
Avec notre voisin du Sud du lac …
Neuchâtel - Fribourg



La Suisse du XXIe siècle

La différence la plus forte entre le canton de Neuchâtel et ...

Schwytz



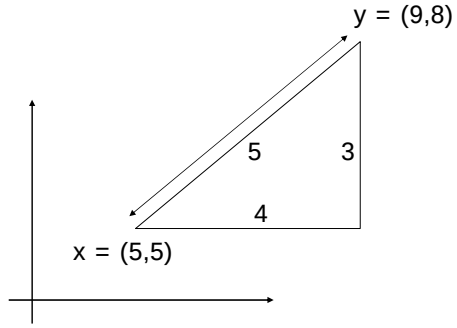
La Suisse du XXIe siècle

- Et comment mesurer la similarité / différence entre deux cantons sur un ensemble de votations ?

Vote \ Canton	A	B	C	D
Vote 1	40 %	36 %	60 %	50 %
Vote 2	49 %	52 %	38 %	38 %

La distance entre A et B sera de
$(40 - 36)^2 + (49 - 52)^2 = 4^2 + 3^2 = 16 + 9 = 25$
Et on prend la racine carrée de 25 donc 5
La distance entre C et D = $\sqrt{[(60-50)^2 + (38-38)^2]} = \sqrt{100}$

Calcul de distances



$$Distance(x,y) = \sqrt{(5-9)^2 + (5-8)^2} = \sqrt{4^2 + 3^2} = \sqrt{16+9} = 5$$

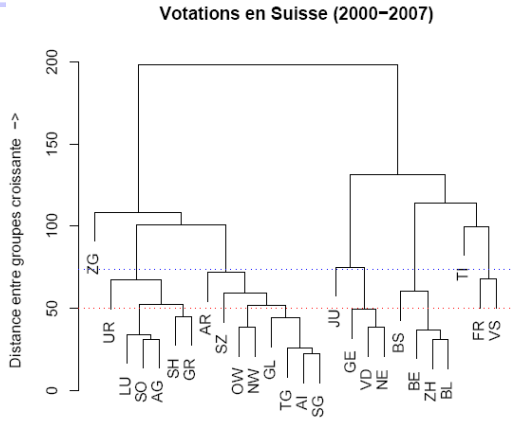
La Suisse du XXIe siècle

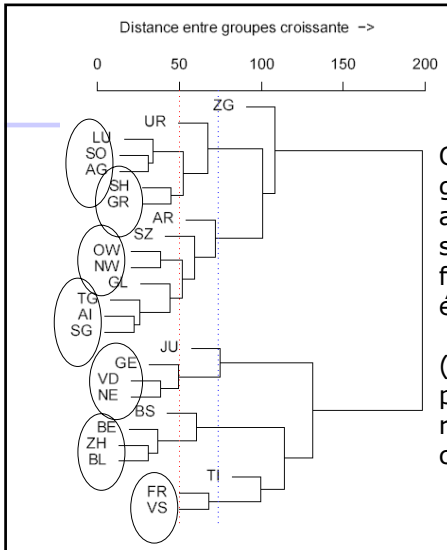
- On calcule toutes les distances entre tous les cantons. Cela fait beaucoup de nombres (26x25/2)

- Faire un graphique pour regrouper les paires de cantons les plus similaires (des couples)
- Puis inclure d'autres cantons ou classes (mais la distance / dissimilarité va croître)
- Continuer jusqu'à former une seule classe

La Suisse du XXIe siècle

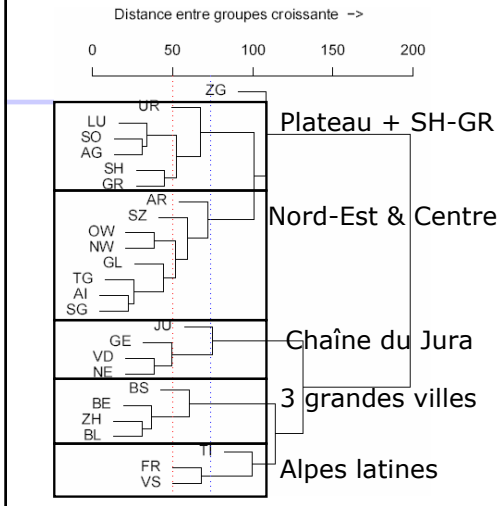
Comment lire ce dessin ?





On retrouve sept groupes de cantons assez homogènes selon les votations fédérales (et non les élections)

(les candidats pouvant former des regroupements de cantons)

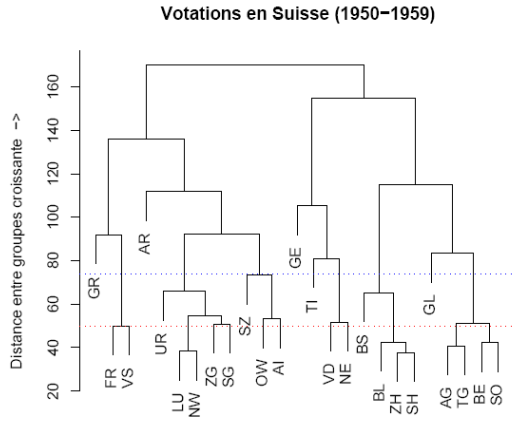


La Suisse du XX^{e} siècle

- Pour 2000-2007
- Pas une Suisse romande homogène
  «VD-NE-GE-JU», «FR-VS»
- Une Suisse du Nord-Est «AI, SG, TG»
- LU regarde vers le plateau «SO, AR, LU»
- Les centres «BL, ZH» (ou «BL, ZH, BE, BS») se rapprochent de la Suisse romande
- BS très proche de la Suisse Romande
- ZG un électron libre

La Suisse du XX^{e} siècle

Et en 1950 ?



La Suisse du XX^{e} siècle

- La dynamique : ce qui est stable et ce qui change
- Des regroupements stables «FR-VS», «VD-NE» ou «ZH-BL»
- Nouveau en 2000-2007
  - Une Suisse du Nord-Est («AI, SG, TG, AR»)
  - Le Mittelland évolue
    «BE, SO, AG, TG» à «SO, AR, LU»
  - Les centres «BL, BS, ZH, BE» se rapprochent de la Suisse romande
  - Deux électrons libres : TI et ZG