ATELIER D'INTRODUCTION À L'ANALYSE DE DONNÉES TEXTUELLES

19 décembre 2019
11h30 à 13h
A1-228 (Sherbrooke)
L1-3650 (Longueuil)

Une précision avant de débuter…

Atelier sur les données textuelles
La statistique textuelle, une approche originale aux débats socioéducatifs.

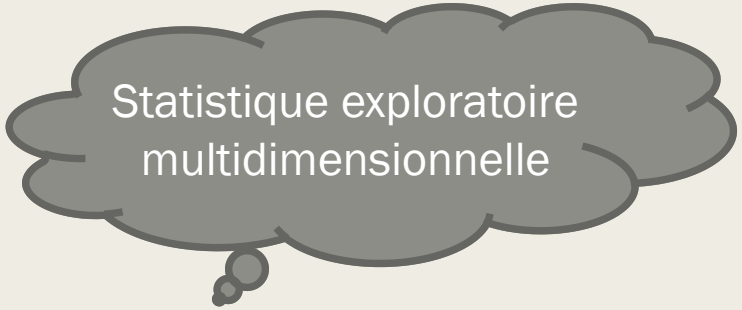
Sommaire

- L’analyse de données textuelles (ADT): pourquoi?
- Les origines de l’ADT
- La démarche de recherche en ADT (et quelques logiciels appropriés)
- Quelques principes à respecter en ADT

L’ANALYSE DE DONNÉES TEXTUELLES : POURQUOI?

Intention: définir des attentes réalistes à l’égard de l’ADT

Pour différentes raisons… à différents moments…



- Pour explorer des données textuelles (p.ex. opinions) de volume plus ou moins important (voire massif) à l'aide de méthodes statistiques.
- Pour inférer et induire des structures sous-jacentes (explicatives ou non, p.ex. des attitudes).
- Pour modéliser (p.ex. comportements ou pratiques).
- Pour expérimenter (p.ex. dispositifs d'activités transactionnelles).

À l'origine des approches quantitatives Années 1960 et 1970

Statistique lexicale

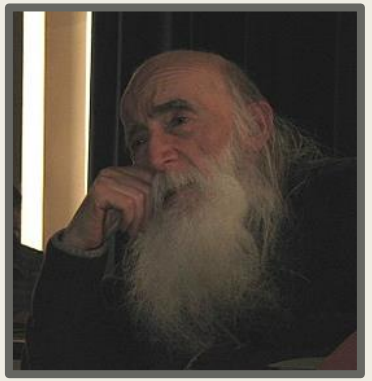
- Décrire les normes et les usages associés à un corpus de textes
- Textes: *sacs de mots*
- Question de représentativité : Échantillonnage/écarts à la moyenne

Des pionniers
Pierre Guiraud (1954, 1960)
Charles Muller (1968, 1977)

Analyse multidimensionnelle lexicale

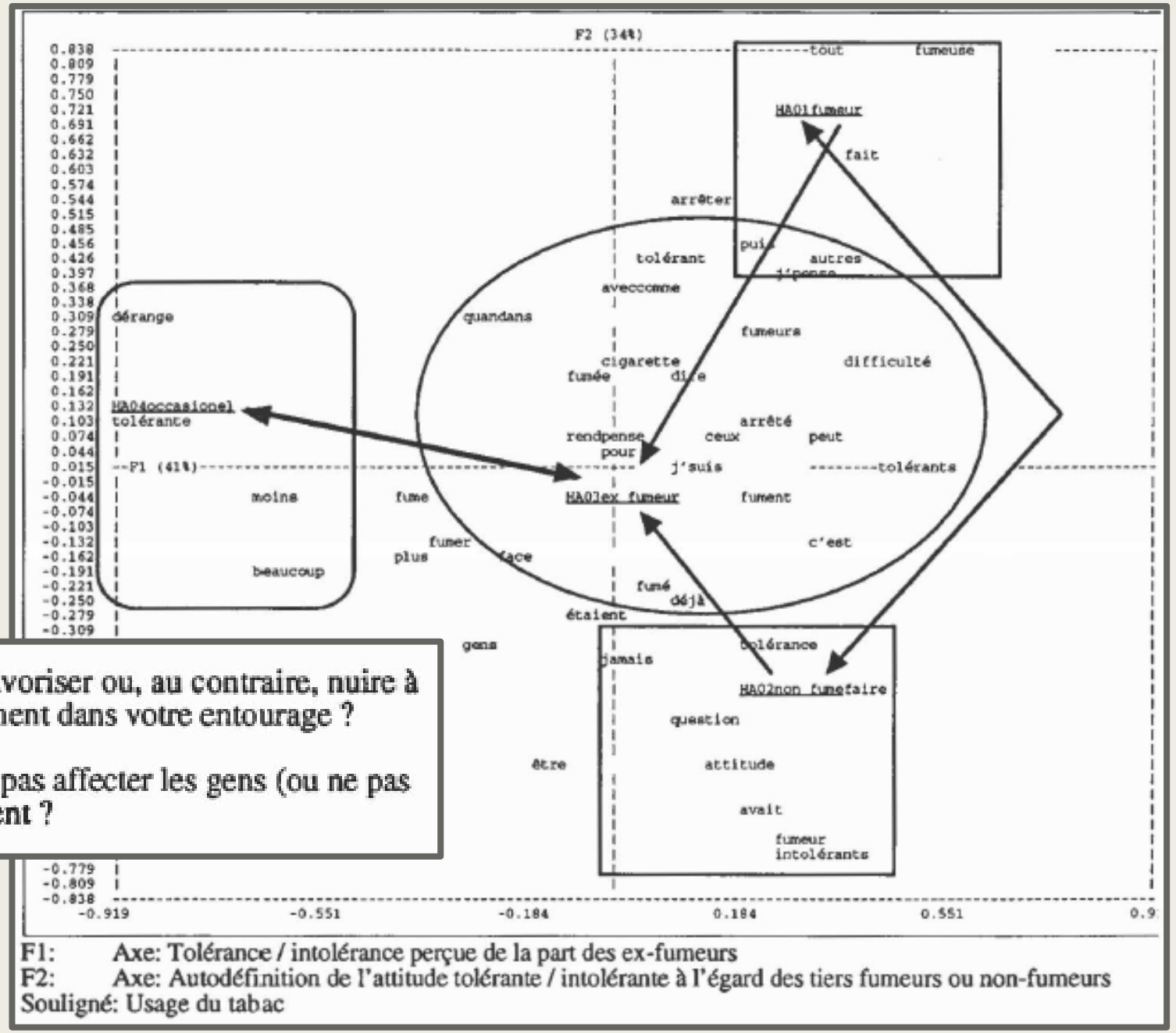
- Analyse en axes principaux
- Production de cartographies synthétiques et visuelles de textes (nuage de points)

Jean-Paul Benzécri (1973)



Exemple d'analyse factorielle

Enquête sur les attitudes des employées et des employés ainsi que des usagères et des usagers de l'Université de Sherbrooke en regard du tabagisme (Larose et Audette, 1995)



Croyez-vous que le fait d'avoir été fumeur peut favoriser ou, au contraire, nuire à une attitude de tolérance à l'égard des gens qui fument dans votre entourage ?

Dans quelles conditions le fait de fumer peut-il ne pas affecter les gens (ou ne pas nuire de façon sérieuse aux gens) qui nous entourent ?

F1: Axe: Tolérance / intolérance perçue de la part des ex-fumeurs
F2: Axe: Autodéfinition de l'attitude tolérante / intolérante à l'égard des tiers fumeurs ou non-fumeurs
Souligné: Usage du tabac

L’ADT aujourd’hui…

- À l’intersection de plusieurs disciplines:
  - *Linguistique, analyse du discours, statistique, informatique, enquêtes socio-économiques, psychosociologie, marketing…*
- Application de la méthode statistique dans l’étude de textes:
  - *Statistique multidimentionnelle*
  - *Analyse de données*
  - *Théorie de l’apprentissage*
  - *Fouille de données (*data mining*)* et *fouille de textes (*text mining*)*

À l'intersection de plusieurs disciplines

Linguistique	Traitement automatique des langues (TAL)	Fouille de données et de textes	Analyse qualitative
Visée descriptive: décrire la diversité des usages langagiers: 1. *Attirances* (collocation) entre des mots 2. Constructions inédites	**Visée applicative**: formaliser les descriptions linguistiques pour développer des applications (dialogue homme-machine)	**Visée applicative**: extraire de l'information *cachée* à partir de grands volumes de données pour les exploiter (p.ex. modéliser comportements d'achat)	**Visée descriptive**: réduire des données en thèmes par codage (ouvert, axial et sélectif)
Analyse des contextes, des registres discursifs, des spécificités, des mots-clés: Condordances (contexte) Cooccurrences (voisinage) Séquences (segments répétés, n-grammes, etc.) Représentativité et contrôle (rééchantillonnage)	Analyse des régularités structurant les corpus oraux (phonétique, phonologie) et écrits (morphologie, syntaxe, sémantique, texte et corpus) Annotation linguistique automatique (étiqueteurs morphosyntaxiques et analyseurs syntaxiques)	Transformation de données brutes en données *analysable*s et *synthétisables*: Analyses régressives Analyses factorielles Méthodes de classification (proches voisins, k-moyennes, arbres de décision, bayes, etc.) Pas de *retour au texte*	Attention du chercheur moins sur le texte et plus sur un objet d'analyse. Analyse par construction progressive d'une grille d'annotation. Subjectivité assumée. Pas ou peu d'intérêt pour les fréquences.

Des dénominations changeantes

(statistique lexicale, statistique linguistique, linguistique quantitative, lexicométrie…)

Trois grandes approches

Lexicométrie

Analyse de textes comme ensemble de mots pour décrire des caractéristiques, des communalités et des spécificités

Textométrie

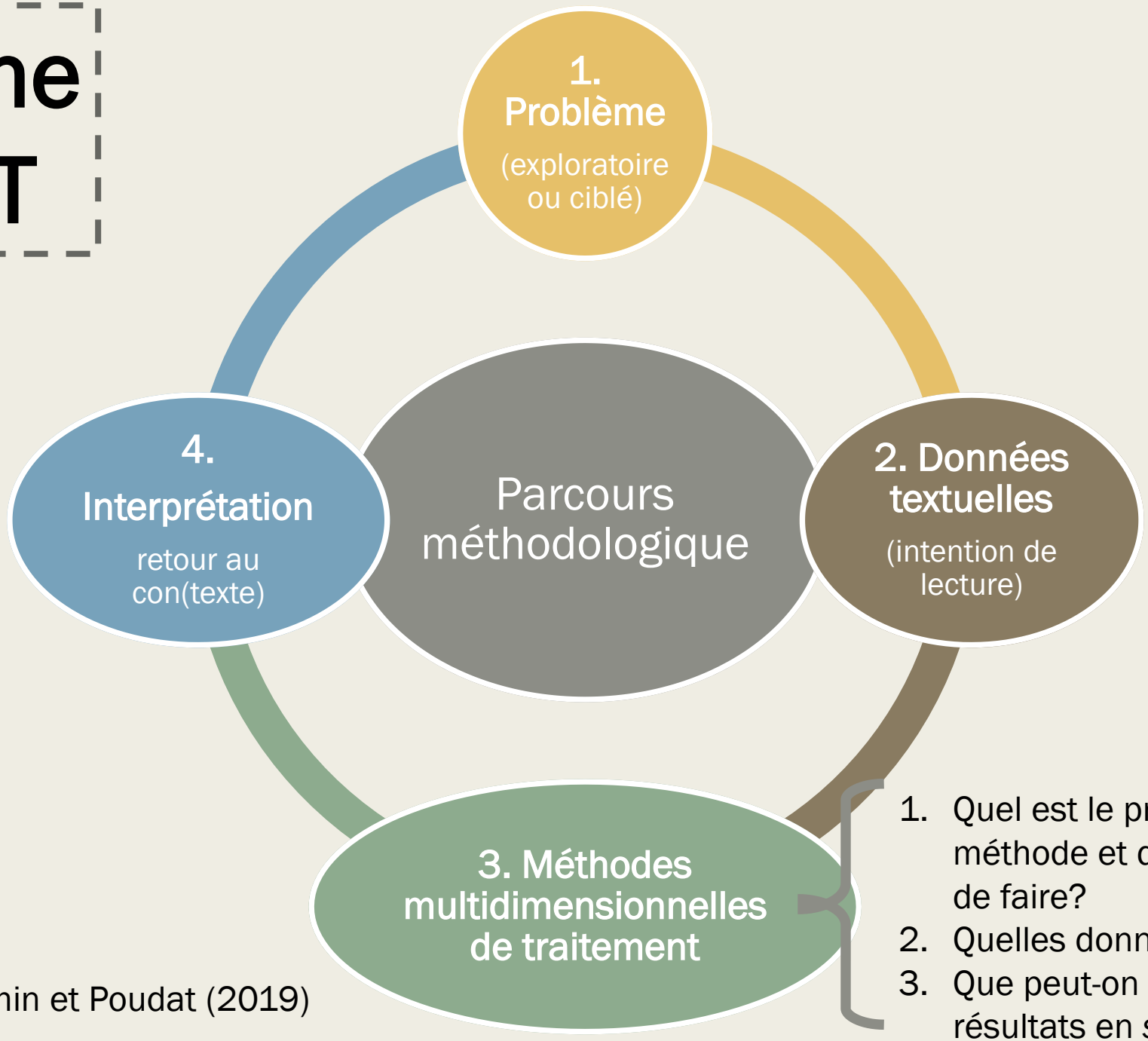
Analyse du *tissu textuel* pour identifier et décrire des caractéristiques:

- Attirances contextuelles: phraséologie, champs thématiques;
- Organisation interne du texte: unités et séquences linéaires: mots répartis dans le texte ou présentés en *rafale*;
- Diversité des informations linguistiques: contrastes intertextuels (sur-emploi/sous-emploi de mots dans un texte);
- Repérage de mots/phrases caractéristiques d'un texte;
- Indicateurs d'évolution lexicale: période caractéristique d'un terme, détection des ruptures significatives.

Logométrie

Analyse de textes à travers l'hypertextualité des corpus (à l'aide de navigateurs hypertextuels, d'index et de concordanciers) pour décrire le contexte, la régularité et la saillance d'*unités de discours* par une procédure semi-automatisée (qualitative et quantitative dans une tension lecture globale/locale de corpus souvent numériques).

Démarche de l'ADT



1. Quel est le principe de la méthode et que permet-elle de faire?
2. Quelles données en *entrée*?
3. Que peut-on dire des résultats en *sortie*?

Source: Lebart, Pincemin et Poudat (2019)

1. Problème: exploratoire ou ciblé

1. Problème (exploratoire ou ciblé)

- Approche exploratoire:
  - *Textes et corpus (écrits ou oraux) pré-existants*
  - *p.ex. analyse de curriculum dans le contexte du débat sur l'enseignement de l'histoire « nationale »*
- Approche ciblée:
  - *Questions (plus ou moins) ouvertes et données d'enquête*
  - *p.ex. analyse de situations professionnelles*

Exemple *Débat sur l'histoire nationale*
Les divergences d'interprétation et les ambiguïtés caractérisant ces rapports [recommandations curriculaires sur l'enseignement de l'histoire *nationale*] nous amènent à nous interroger sur leurs orientations. À quel point convergent/divergent-elles? (Moreau et Smith, accepté)

Exemple *Attributions enseignantes*
Quelles théories de l'apprentissage fondent les attributions enseignantes au regard de l'enseignement d'un mode de pensée historique (Moreau et Smith, 2017)

2. Données textuelles ou la *constitution d'un corpus*

2. Données textuelles

(intention de lecture)

- *Un corpus est une collection de données langagières qui sont sélectionnées et organisées selon des critères linguistiques explicites pour servir d'échantillon du langage*. (Habert, Nazarenko et Salem (1997, p.11)
- Principe d'*intention de lecture*: l'objet et ses fondements
- Principe de *représentativité*: échantillon de l'**usage** langagier relatif à l'objet, autrement c'est une *base documentaire*.

2. Données textuelles ou la *constitution d'un corpus*

2. Données textuelles

(intention de lecture)

- Taille du corpus: critères
  - *Nature du corpus (variation observable)*
  - *Moyens pour constituer le corpus*
- Attention au « Gros, c'est beau » (Habert, 2000). Risque de trop en prendre et de ne constituer qu'une *base documentaire*:
  - *Incertitude (random error)*
  - *Déformation (biais error)*

Exemple *Débat sur l'histoire nationale*

Inclure ou non les recommandations curriculaires du rapport Corbo (1994) introduisant l'idée de domaine d'apprentissage?

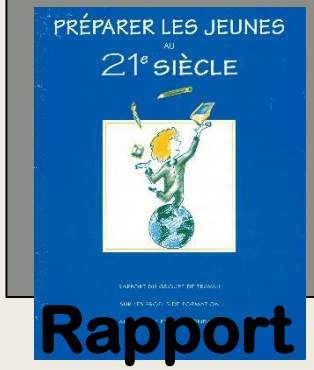
Exemple *Attributions enseignantes*

Formulation des questions…

Exemple *Débat sur l'histoire nationale*

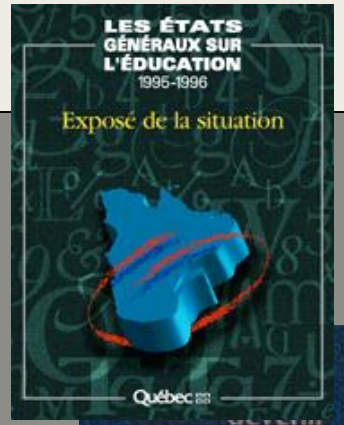
Les pièges d'une documentation abondante

1994



Rapport Corbo

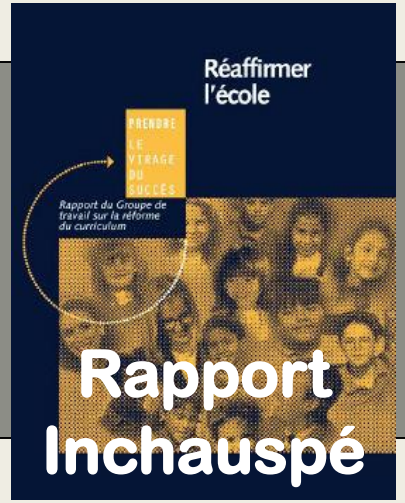
1996



Rapport Lacoursière

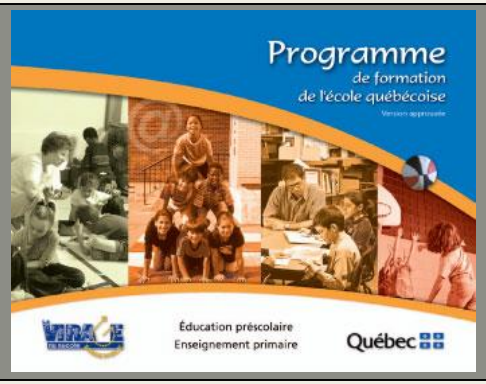
1996

1998



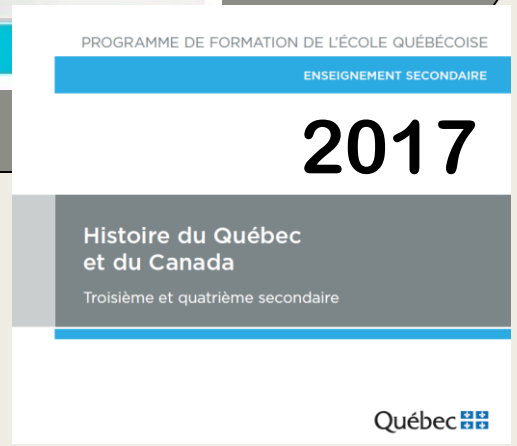
Rapport Inchauspé

2001



Rapport Beauchemin-Fahmi-Eid

2014



2017

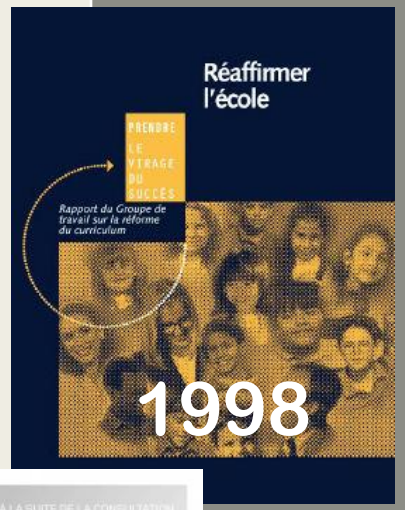
Exemple *Débat sur l'histoire nationale*

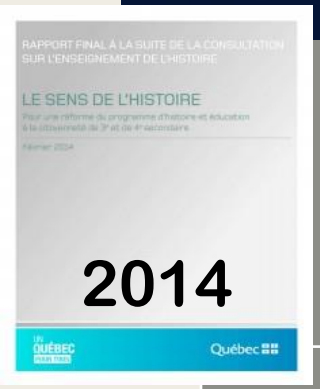
Trois documents retenus:

Principe de l'objet: *recommandation curriculaire* (pas de curriculum, de programme d'études, de mémoires ministériels, l'avis de Paul Inchauspé, etc.) sur l'*enseignement de l'histoire*

Principe de l'usage (non pas tel qu'il est dans le débat, mais qu'il doit être en fonction de l'objet): les recommandations curriculaires sont *immédiatement en amont* des programmes d'études.

1996





Rapports	Pages retranscrites	Pages écartées de l'analyse
Rapport Lacoursière	p.i, ii, ix, x et xiii p.1 à 80	p.iii à viii (table des matières et lettre à la Ministre)
Rapport Inchauspé	p.13 à 38 p.39 à 43 p.45 à 48 p.52 à 53 p.55 p.61 p.64 à 65 p.72-75 p.91 à 95 p.97 à 102 p.104 à 107 p.109 à 116 p.121-124 p.125 à 132 p.136 à 138 p.139 (en haut) p.141 à 142	p.1 à 12 p.38 à 39 p.43 p.48 à 51 p.53 à 55 p.56 à 61 p.62 à 63 p.65 à 71 p.75 à 90 p.95 à 97 p.102 à 103 p.108 à 109 p.117 à 120 p.124 à 125 p.133 à 136 p.139 à 143
Rapport Beauchemin	p.i à ii p.1 à 64	p.iii à iv

2. Constitution du corpus Enquêtes socio-économiques

Principe d'intention de lecture

- Production de *textes* qui n'en sont pas au sens de la linguistique:
  - *L'intention de lecture dicte en amont la formulation de questions (discours organisé)…*
  - *…intervenant comme stimuli dans la production de données de type représentationnel (discours descriptif/explicatif) par des répondants.*
- Variabilité selon la nature du questionnement (ouvert ou fermé) conduisant à des informations différentes et non comparables…

Principe de représentativité (usage langagier?)

Le gouvernement des États-Unis…

TABLE I

Should allow speeches against democracy	["Yes" to 2(a)]	21%
Should not forbid speeches against democracy	["No" to 2(b)]	39
Should not allow speeches against democracy	["No" to 2(a)]	62
Should forbid speeches against democracy	["Yes" to 2(b)]	46
Don't know [2(a)]		17
Don't know [2(b)]		15

Recherche de Rugg (1941): constats!

- Absence de symétrie entre les questions ouvertes ou fermées (p.ex. Pensez-vous que les États-Unis doivent autoriser/interdire les discours publics contre la démocratie?
- Écart accentué par le niveau d'instruction des répondants et l'origine socio-géographique.

Source: Rugg, D. (1941). Experiments in wording questions II. *The public opinion quarterly, 5*(1), 91-92,

2. Constitution du corpus Enquêtes socio-économiques

2. Données textuelles

(intention de lecture)

- Quelques balises, pour contrôler le degré d'ouverture du questionnement:
  - *Avoir une liste de questions préétablies dans leur libellé (nécessité!)*
  - *Avoir une liste de réponses préétablies (possibilité)*
- Quelques considérations (Lazarsfeld, 1944)…
  - *Questionnement ouvert pertinent pour recueillir des informations spontanées, réduire le temps d'entrevue ou du questionnaire.*
  - *Prolonger par une question fermée (pourquoi?).*

Exemple *Attributions enseignantes*

Comment apprend-t-on à penser historiquement?

Exemple *Attributions enseignantes*

Considérez-vous que cette situation d'enseignement-apprentissage a favorisé l'apprentissage de la pensée historique?

Si oui, expliquez pourquoi?

Si non, expliquez pourquoi?

Norme lexicologique
ou de la nécessité de *lire* les corpus…

- Ensemble de règles pour articuler au mieux les unités textuelles/linguistiques et le niveau lexical (mots) et pour indexer le texte en vue des traitements ultérieurs.
- Regroupement de lexèmes (p.ex. aujourd et hui).
- Désambiguïsation (séparation des formes homographes appartenant à des vocables distincts, p.ex. formation, cours, mets, etc.).
- Dictionnaire d’abréviations et siglaisons

Lemmatiser… ou pas? Un vieux débat!

2. Données textuelles

(intention de lecture)

- Lemmatiser neutralise les variations attribuables aux flexions, donc:
  - *Diminue la taille du vocabulaire (sans réduire la taille du corpus)…*
  - *Diminue la densité des points dans le plan factoriel et accroît leur contribution respective (donc leur valeur explicative)…*
- Mais…
  - *Peut induire une projection idéologique…*
  - *Et conduire à une perte d'informations (l'usage du singulier correspond souvent à un sens plus abstrait, le pluriel, un sens concret).*
- Actuellement, la lemmatisation est « une option recommandée comme *complément* des analyses des textes bruts » (Lebart, Pincemin et Poudat, 2019, p.56)

Unités d'analyse

2. Données textuelles

(intention de lecture)

- Mots-formes graphiques (mot tel qu'il apparaît dans le texte)
- Lemmes: neutralisation des variations contextuelles de flexion (accord, conjugaison) ou de typographie.
- Parties du discours: les mots en fonction de leur catégorie grammaticale (annotation syntaxique automatique)

Décomptage réalisé par n'importe lequel logiciel d'analyse textuelle (Lexico, DTM-Vic, Sphinx…)

Outils d'annotation morphosyntaxique automatique (ou étiqueteurs morphosyntaxiques) utilisés en TAL: *étiquettes* associées à un lemme et une catégorie morphosyntaxique.

TreeTagger et *Cordial Analyseur* (paramètres)
Brill Tagger (« entraînement » de l'outil par inférence de règles locales)
Le Trameur, *PrimeStat*, *SYNTEX* (annot. syntaxique)
Par contre, *plus les étiquettes sont précises (richesse des informations), plus le risque d'erreur est grand (robustesse)!*

Élaboration des corpus
Modalité de travail

- Copier-coller les parties de discours pertinentes dans un fichier *word*.
- Balisage (représentant chaque *locuteur* dans un corpus):
  - **** *01*
  - [discours]
  - **** *02*
  - [discours]
  - **** *03*
  - [discours]
  - *Etc.*

Constitution du *Tableau lexical entier* par décomptage (segmentation)

- Segmentation: opération de découpage des données en unités textuelles minimales, ne pouvant plus être décomposées davantage
  - *Mots reconnus en fonction des espaces et des ponctuations*
- Formes-lignes (mots)
  - *L'ensemble des formes graphiques définit le vocabulaire du corpus (V)*
  - *L'ensemble des occurrences de chacune des formes constitue la taille du corpus (T)*
  - *Seules les formes apparaissant au-delà d'un certain seuil de fréquence seront soumises à l'analyse. Celles moins fréquentes et les hapax (f=1) ne sont pas retenus…*
- Sujets-colonnes (auteurs des programmes d'études)

Tableau lexical entier: exemple

		obje	cont	appr	éval
#aujourdhui	i	5.	25.	0.	0.
#faits	i	12.	2.	1.	0.
#histoireeteduca	i	22.	6.	10.	0.
#nouvellefrance	i	0.	22.	0.	0.
#ressources	i	0.	13.	0.	0.
#saintlaurent	i	0.	12.	1.	0.
#vivreensemble	i	7.	3.	0.	1.
a	i	4.	23.	4.	2.
abord	i	2.	10.	0.	1.
acquis	i	8.	3.	0.	0.
acteurs	i	11.	3.	3.	0.
action	i	11.	10.	0.	0.
actuelle	i	0.	10.	0.	0.
afin	i	4.	9.	1.	0.
aide	i	16.	19.	6.	0.
ailleurs	i	6.	16.	1.	0.
ainsi	i	17.	17.	6.	0.
alors	i	4.	12.	0.	0.
amener	i	4.	15.	0.	0.
amenés	i	3.	13.	0.	0.
amène	i	9.	0.	3.	0.
amérique	i	0.	14.	0.	0.
analyse	i	7.	3.	1.	3.
angle	i	4.	34.	0.	1.
année	i	4.	40.	0.	4.
années	i	0.	19.	0.	1.
appelés	i	2.	10.	0.	0.
apprendre	i	5.	1.	8.	0.
apprentissage	i	7.	14.	16.	5.
apprentissages	i	7.	3.	7.	2.
après	i	0.	13.	0.	0.
articulation	i	0.	13.	1.	0.
aspects	i	7.	9.	5.	0.
assemblée	i	0.	14.	0.	0.
assurer	i	0.	6.	4.	0.
au	i	46.	150.	28.	1.
aussi	i	12.	25.	5.	1.
autochtones	i	0.	17.	0.	1.
autre	i	6.	6.	1.	1.
autres	i	13.	32.	9.	0.
aux	i	28.	55.	14.	1.
avec	i	12.	25.	10.	2.
bien	i	2.	10.	0.	1.
britannique	i	1.	15.	0.	0.
britanniques	i	0.	13.	0.	0.
c	i	16.	43.	2.	0.
cadre	i	9.	4.	0.	2.
canada	i	0.	15.	0.	0.
canadienne	i	1.	10.	0.	0.
canadiens	i	0.	16.	0.	0.
cas	i	3.	6.	3.	0.
ce	i	12.	39.	6.	2.
cela	i	8.	7.	2.	0.
celles	i	4.	10.	2.	0.
cependant	i	0.	14.	0.	1.
ces	i	18.	41.	4.	0.
cet	i	2.	9.	1.	1.
cette	i	16.	34.	4.	1.
chambre	i	0.	16.	0.	0.
changement	i	10.	6.	5.	0.
chaque	i	3.	7.	0.	0.
chercher	i	7.	8.	0.	0.
choix	i	4.	11.	3.	0.
ci	i	2.	8.	2.	0.
citoyen	i	7.	16.	0.	0.
citoyenneté	i	31.	12.	6.	0.
citoyens	i	11.	9.	0.	0.
colonie	i	1.	26.	0.	0.
colonies	i	0.	12.	0.	0.
comme	i	15.	39.	8.	1.
comment	i	4.	23.	0.	0.
commerce	i	1.	9.	1.	0.
complexité	i	5.	2.	3.	0.
composantes	i	9.	2.	0.	4.
comprendre	i	14.	4.	0.	0.
compréhension	i	2.	8.	1.	0.
compte	i	9.	11.	3.	3.
compétence	i	25.	4.	9.	4.
compétences	i	18.	21.	7.	7.
concept	i	1.	10.	0.	0.
conception	i	1.	15.	0.	0.
concepts	i	13.	31.	3.	1.
connaissances	i	8.	21.	2.	1.
conquête	i	2.	8.	0.	0.
conscience	i	12.	6.	1.	0.
consolider	i	10.	3.	2.	1.
constitue	i	3.	9.	0.	0.
constituent	i	4.	10.	2.	1.
construction	i	2.	8.	2.	0.
contenu	i	5.	23.	4.	0.
contexte	i	7.	5.	9.	0.
continuité	i	3.	0.	6.	1.
cours	i	8.	37.	0.	3.
crise	i	0.	10.	0.	0.
critique	i	17.	2.	7.	4.
culture	i	2.	19.	0.	0.
culturel	i	1.	13.	0.	0.
culturelles	i	0.	11.	0.	0.
culturels	i	4.	6.	1.	0.
cycle	i	25.	32.	14.	5.
d	i	152.	298.	94.	23.
dans	i	67.	141.	34.	7.
davantage	i	3.	6.	0.	1.
de	i	408.	825.	206.	28.
depuis	i	0.	18.	1.	0.
des	i	226.	393.	125.	22.
deux	i	3.	14.	6.	0.
deuxième	i	13.	19.	0.	0.
devraient	i	1.	12.	0.	0.
diagramme	i	0.	1.	9.	0.
difficultés	i	12.	3.	0.	0.
différences	i	5.	6.	3.	0.
différentes	i	2.	9.	3.	0.
différents	i	6.	11.	6.	0.
disciplinaires	i	2.	19.	4.	1.
discipline	i	9.	1.	5.	0.
divers	i	4.	4.	2.	0.
diverses	i	3.	10.	2.	0.
diversité	i	9.	2.	1.	0.

Principe de base
Loi de Zipf (1935)

- Relation constante entre le rang et la fréquence d'un mot. (rang X fréquence = constance approximative).
- Exception: les plus hautes fréquences.
- Exemple: *Ulysses* de James Joyce
  - *Rang 10: le mot 's (f=2502)*
  - *Rang 100: le mot street (f=282)*
  - *Rang 1 000: le mot Marion (f=27)*
  - *Rang 10 000: le mot clogs (f=2)*

Source: Lebart, Pincemin et Poudat (2019)

3. Méthodes multidimensionnelles de traitement

3. Analyse des données

Deux démarches successives et complémentaires

1. Statistique descriptive et exploratoire:

- *Établir des relations entre des variables ou des proximités entre des individus*
- *Aucune inférence sur une population plus large.*
- *Indicateurs de moyenne et de dispersion*

2. Statistique inférentielle et confirmatoire (non abordée):

- *Valider/inférer des hypothèses (modèle.s)*
- *Stylométrie (aspects formels)*
- *Discrimination globale*

Quel est le principe de la méthode et que permet-elle? Quelles données en entrée? Que peut-on dire des résultats en sortie?

Source: Lebart, Pincemin et Poudat (2019)

Statistique descriptive et exploratoire

- Représentations graphiques et techniques descriptives multidimensionnelles
  1. *Méthodes factorielles (axes principaux):*
     - Analyse en composantes principales (ACP)
     - Analyse factorielle des correspondances (AFC)
     - Méthode de rééchantillonnage (*bootstrap*)
  2. *Méthodes de classification*

Statistique descriptive et exploratoire
Méthodes factorielles (axes principaux)

- Représentation géométrique des lignes et des colonnes des tableaux:
  - *Construction de deux nuages de point, chaque dimension du tableau permettant de définir les distances (euclidiennes) entre les éléments de l'autre dimension (calcul du chi-2).*
- Proximités entre les points (vecteurs-lignes et vecteurs-colonnes) représentent des *associations statistiques*.

Statistique descriptive et exploratoire
Méthodes factorielles (axes principaux)

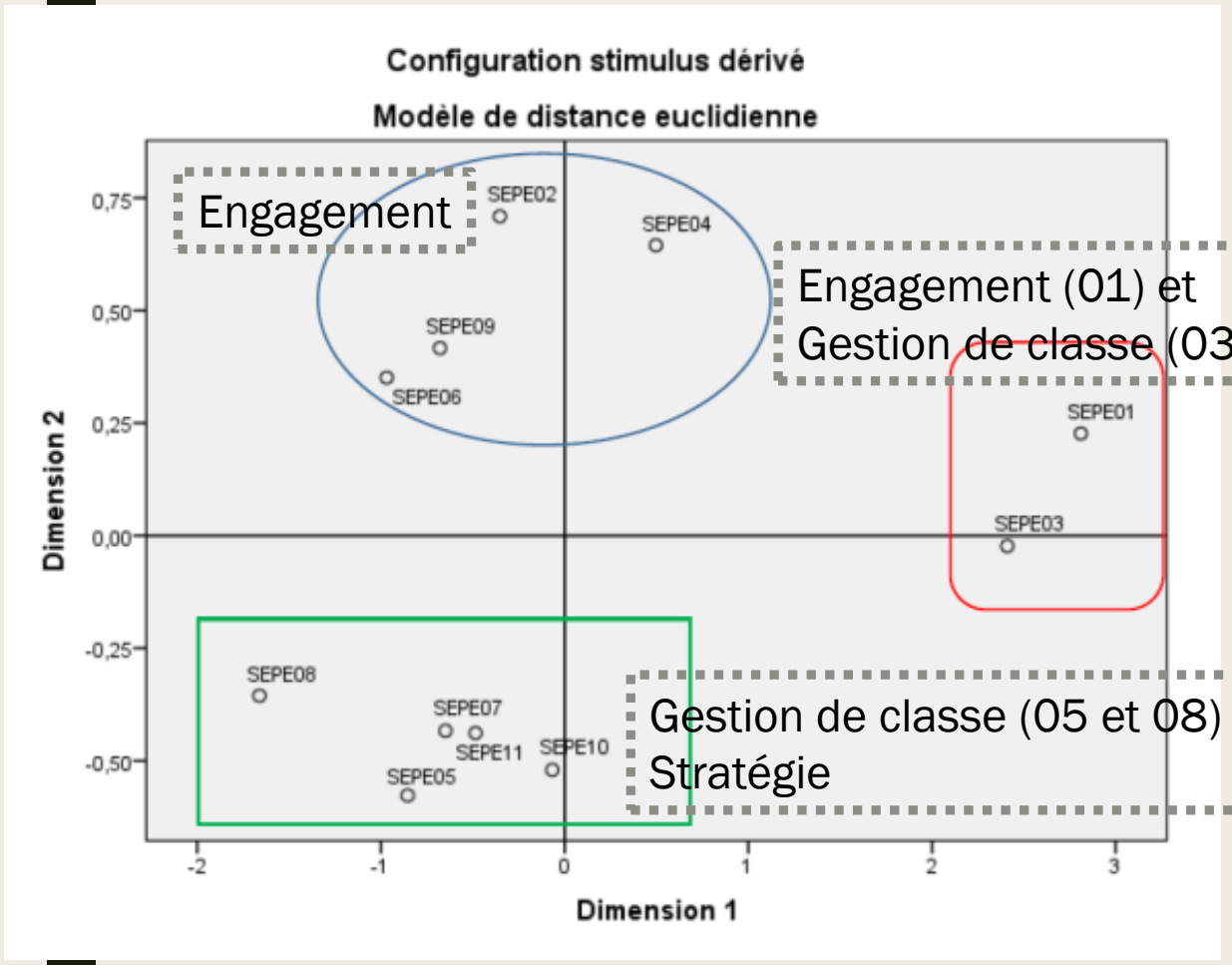
Analyse en composantes principales (ACP)

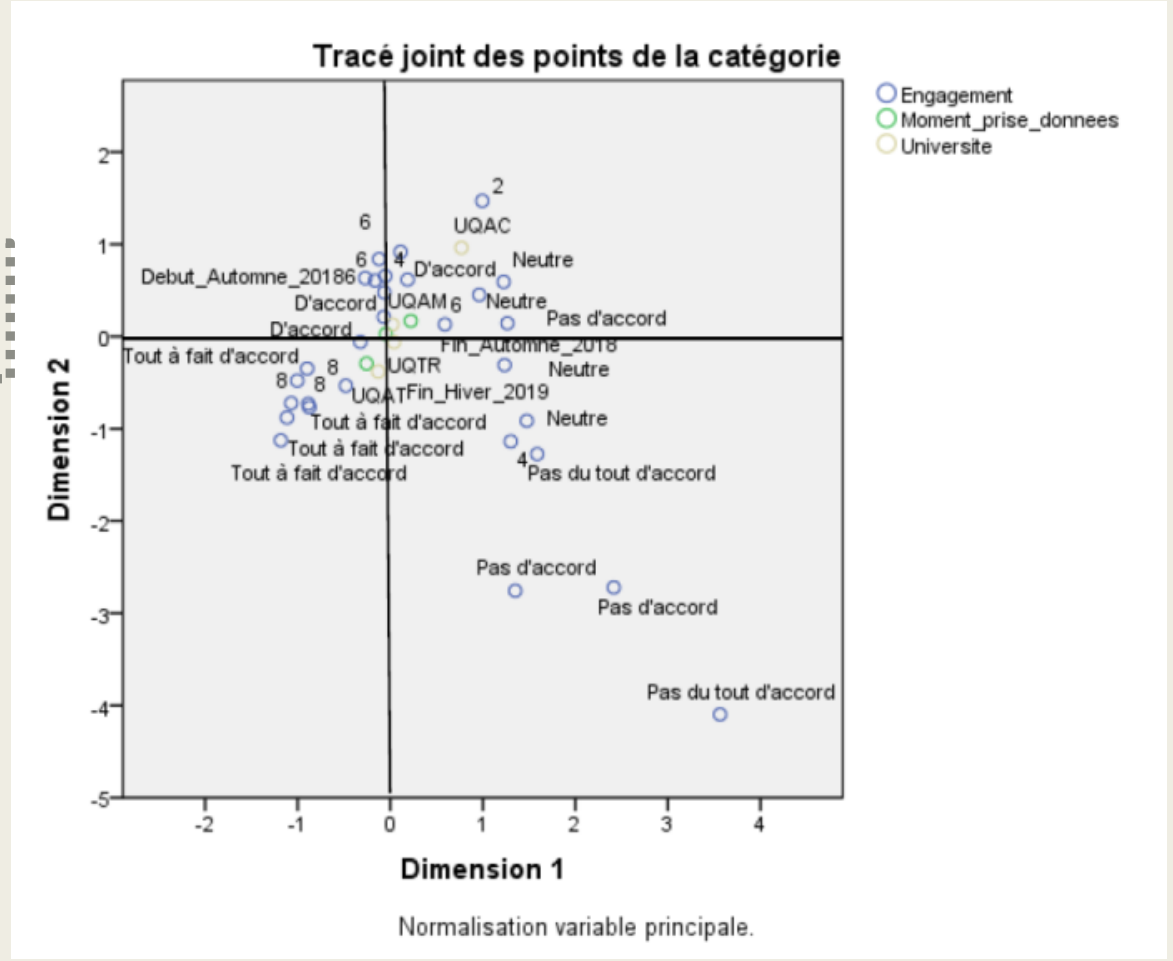
- Enquêtes sémiométriques (attribution de notes à des mots)
- Tableau de variables à valeurs numériques
- Colonnes: variables mesurées (p.ex. items d'un questionnaire)
- Lignes: individus/observations.
- Analyse du positionnement multidimensionnel (MDS)

Analyse factorielle des correspondances (AFC)

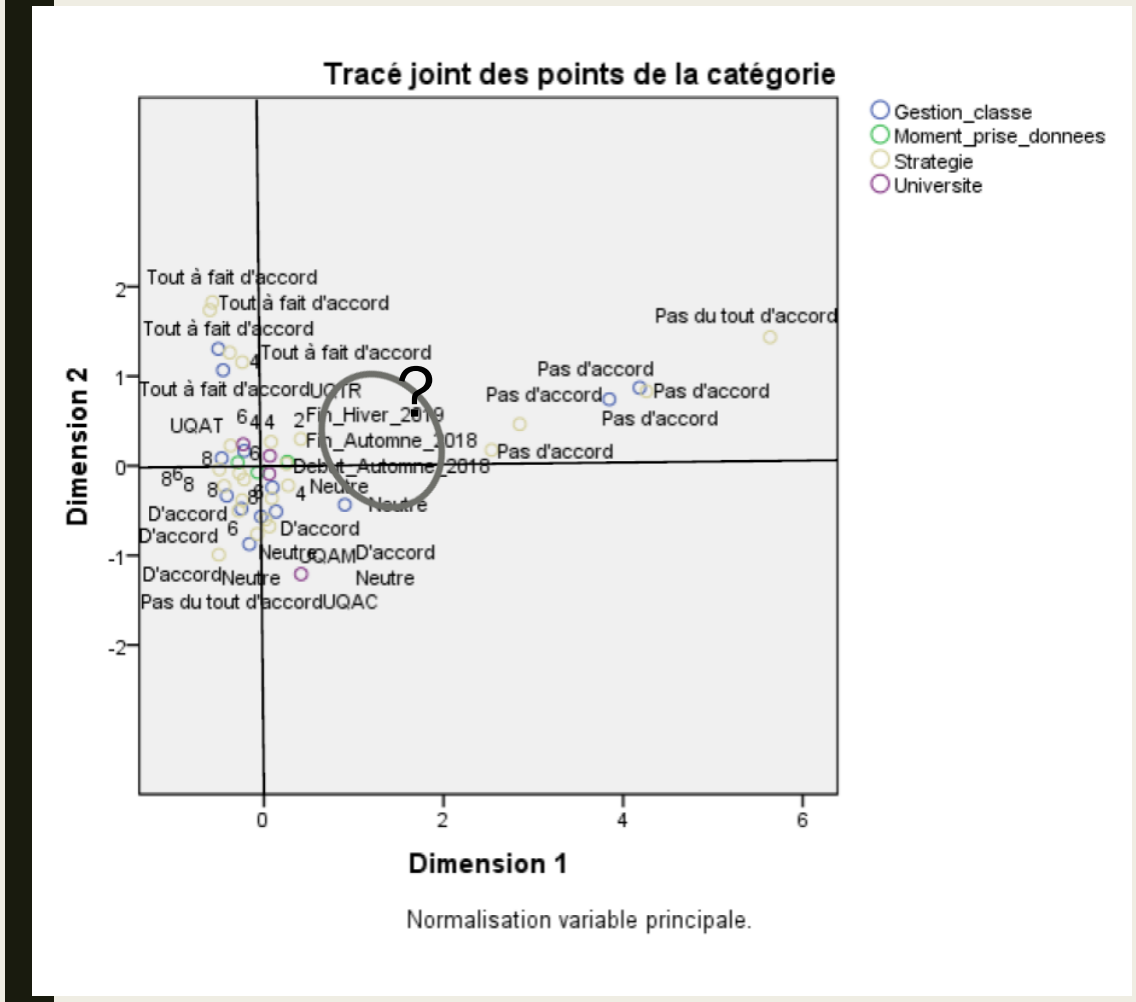
- Enquêtes lexicométriques
- Tables de contingence croisant des variables nominales
- Colonnes: variable nominale (p.ex. locuteur)
- Lignes: variable nominale (p.ex. mots)
- Identification de liens sémantiques en fonction des axes factoriels

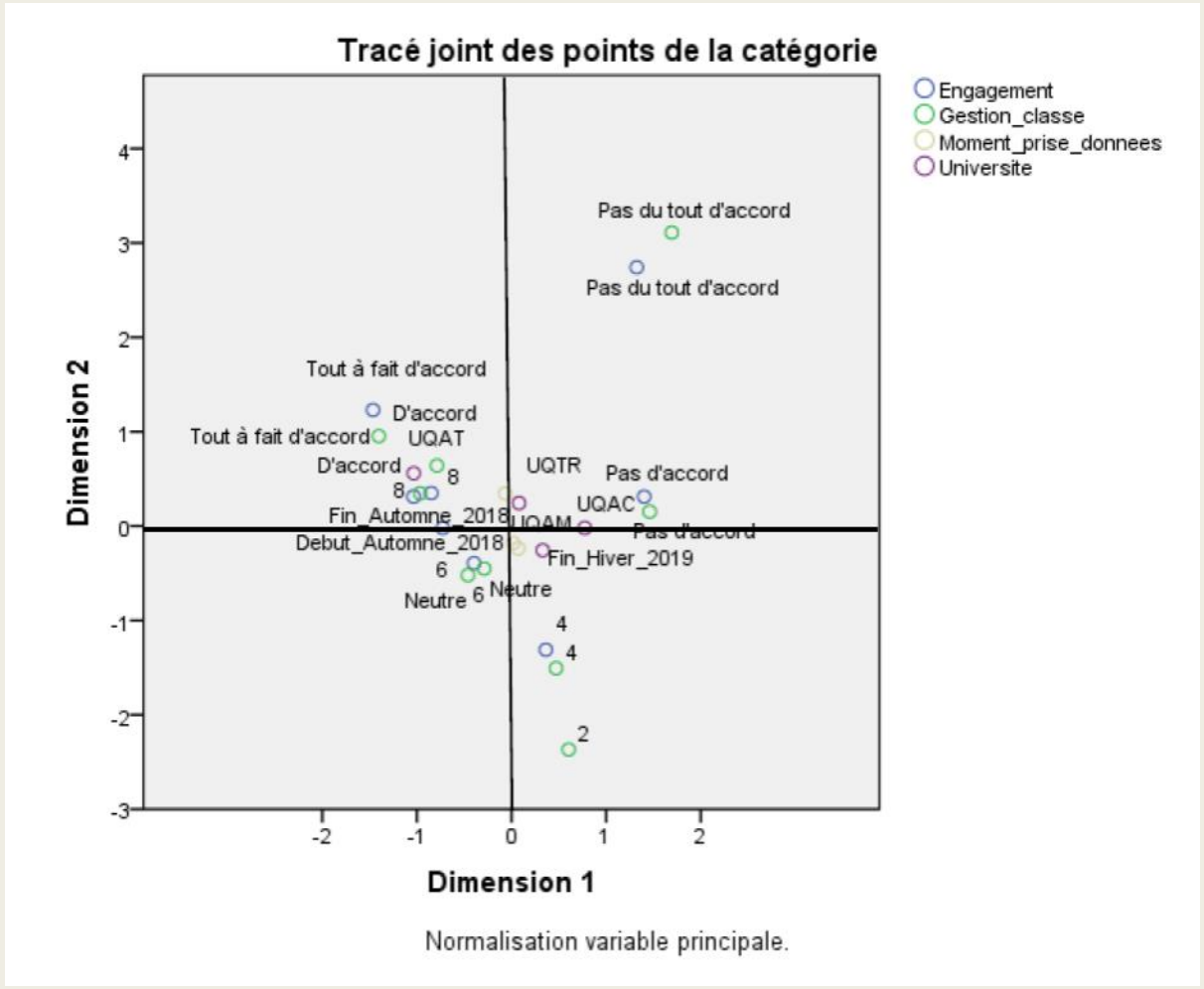
Exemple d'ACP.
Sentiment d'efficacité personnelle enseignant



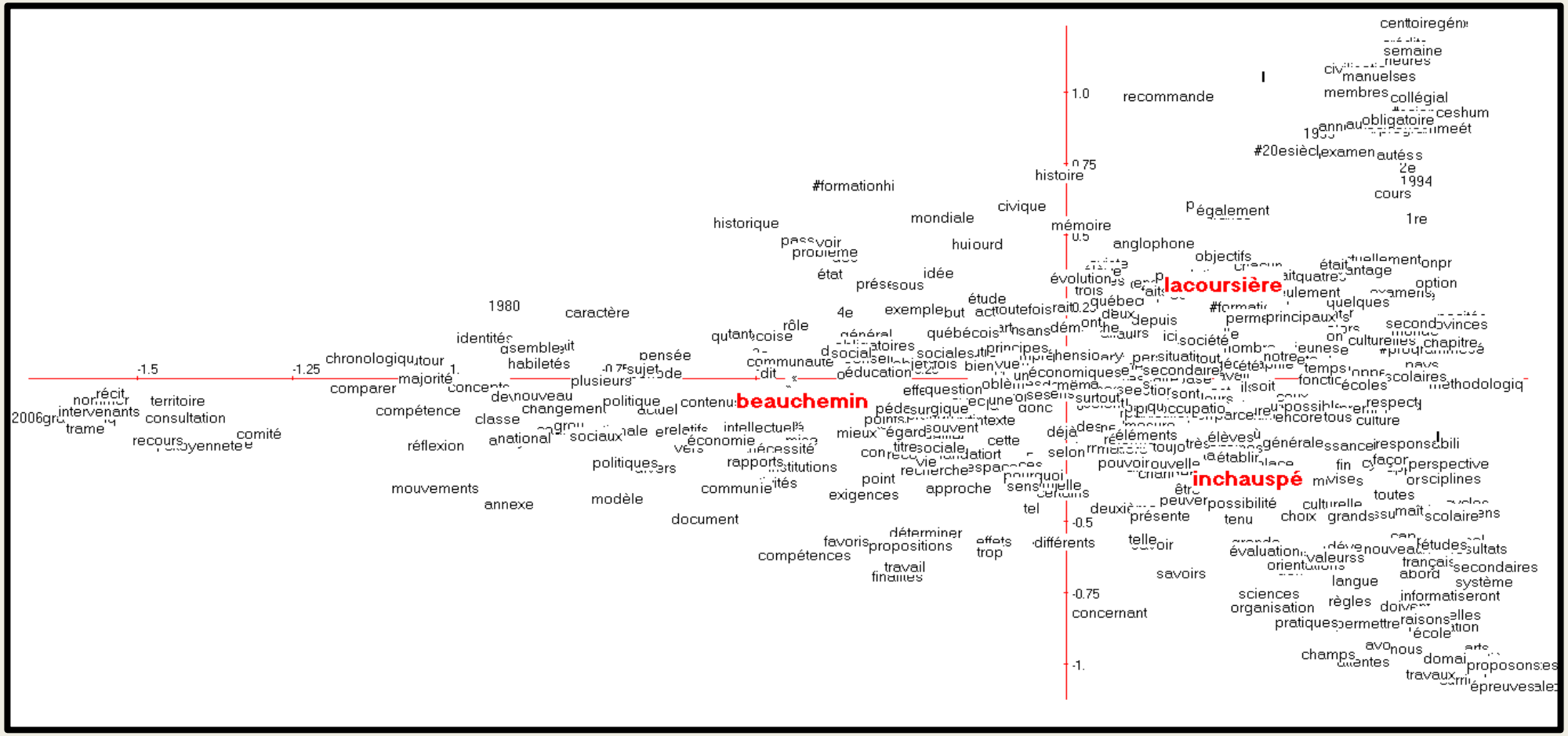


Exemple d'ACP. Recherche sur le sentiment d'efficacité personnelle enseignant





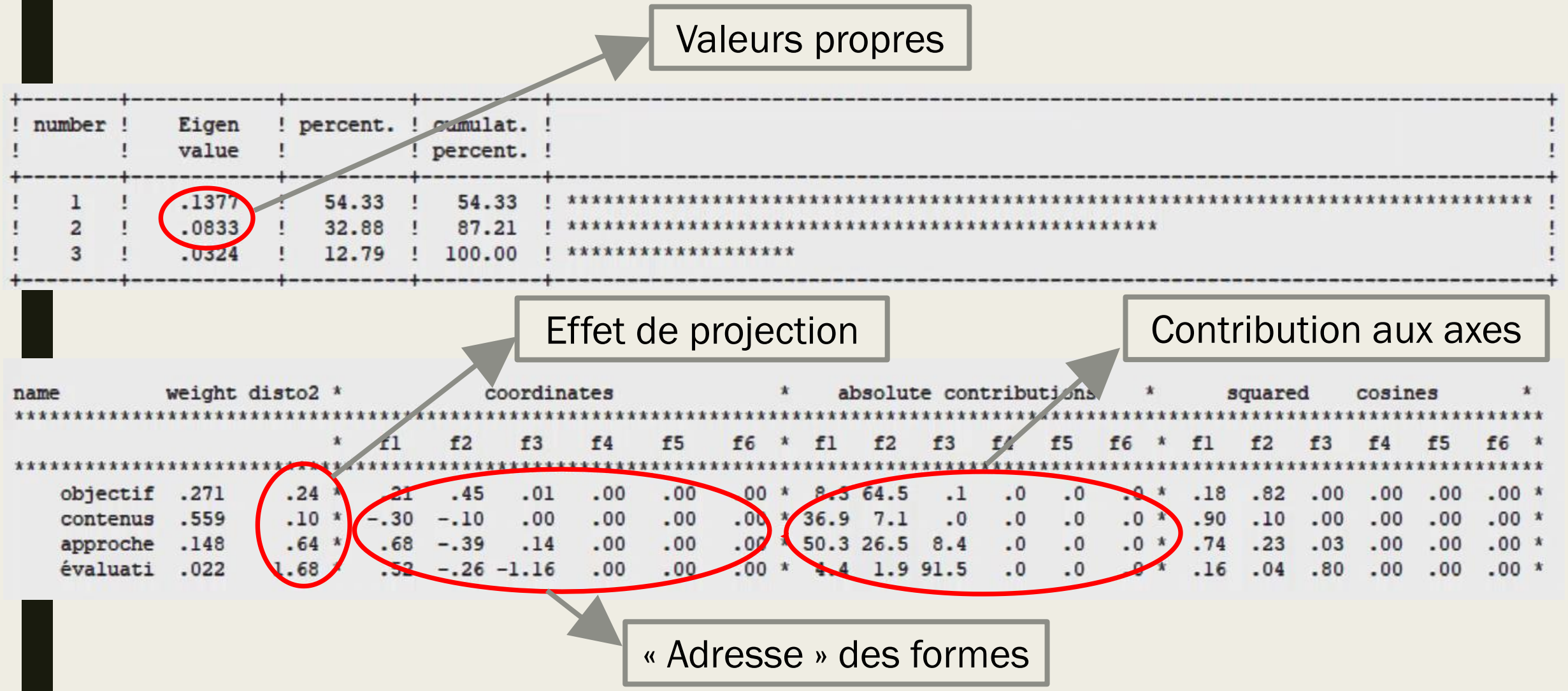
Exemple d'AFC
Débat sur l'histoire nationale



Analyse factorielle des correspondances
Quelques clés de lecture

- Valeur propre (valeur explicative à l'égard de l'inertie totale)
- Disto: effet de projection (attention!)
- Coordonnées: « adresse » des points dans le nuage
- Contributions (absolues): importance des éléments dans la construction de chaque axe (somme en colonne: 100)
- Cosinus carrés (contributions relatives): importance des différents axes dans l'explication de chaque élément (somme en ligne: 1)

Valeurs propres



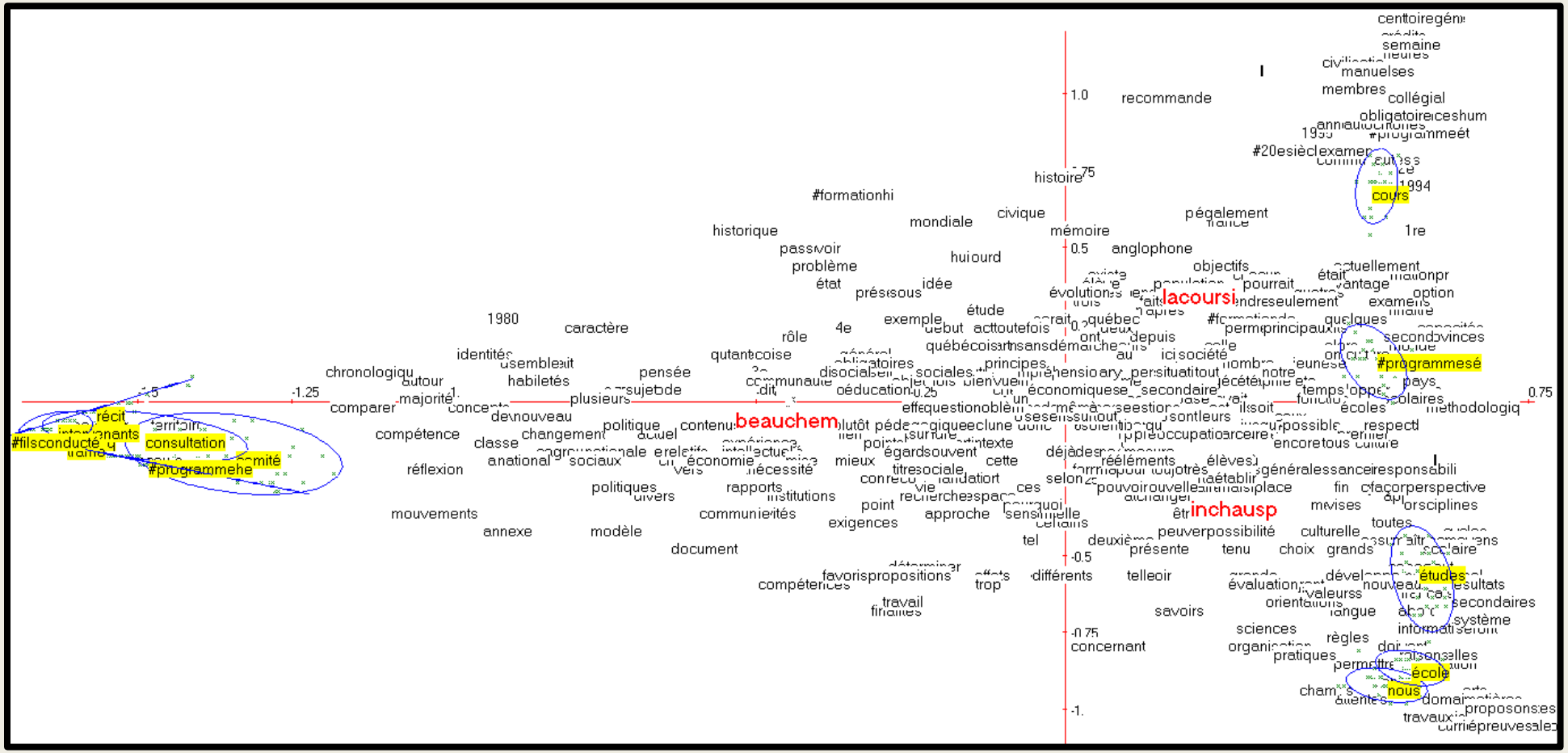
number	Eigen value	percent.	cumulat. percent.	
1	.1377	54.33	54.33	**********
2	.0833	32.88	87.21	******
3	.0324	12.79	100.00	**

name	weight	disto2	coordinates						absolute contributions						squared cosines					
			f1	f2	f3	f4	f5	f6	f1	f2	f3	f4	f5	f6	f1	f2	f3	f4	f5	f6
objectif	.271	.24	.21	.45	.01	.00	.00	.00	8.3	64.5	.1	.0	.0	.0	.18	.82	.00	.00	.00	.00
contenus	.559	.10	-.30	-.10	.00	.00	.00	.00	36.9	7.1	.0	.0	.0	.0	.90	.10	.00	.00	.00	.00
approche	.148	.64	.68	-.39	.14	.00	.00	.00	50.3	26.5	8.4	.0	.0	.0	.74	.23	.03	.00	.00	.00
évaluati	.022	1.68	.52	-.26	-1.16	.00	.00	.00	4.4	1.9	91.5	.0	.0	.0	.16	.04	.80	.00	.00	.00

Statistique descriptive et exploratoire
Méthode de rééchantillonnage

- Validation des analyses factorielles en testant la stabilité de structures observées (logiciel: DTM-Vic)
- Simulations de perturbations du tableau de données par tirage au hasard avec remise (entre 10 et 30).
  - *Comparaison des configurations entre les réplications et le tableau initial.*
  - *Loi hypergéométrique: même probabilité des éléments d'être choisis (1/*n*).*

Exemple de validation par rééchantillonnage *Débat sur l'histoire nationale*



Statistique descriptive et exploratoire
Méthodes de classification

- Représentation (dendogramme) des proximités entre les éléments d'un tableau lexical par regroupements ou classes plus homogènes jusqu'à *épuisement* des éléments.
- Deux familles:
  - *Classification hiérarchique: classes emboîtées les unes dans les autres en fonction d'un algorithme ascendant (agglomération d'objets) ou descendant (dichotomisation des objets).*
  - *Partitionnement: découpage de la population étudiée par cartes auto-organisées de Kohonen (non abordée).*

4. Interprétation des données *Pour revenir au texte…*

4. Interprétation

retour au con(texte)

- Du point de vue de la linguistique, aucun mot ne « contient son sens » et le contexte est fondamental dans l'interprétation:
  - *Localement (phrase et paragraphe)*
  - *Globalement (locuteur, genre textuel, contexte d'énonciation, etc.)*
- Un retour balisé par un rapport aux statistiques

Ce qui nous intéresse…

1. Proximités et oppositions linguistiques
2. Caractéristiques du corpus
3. Spécificités (sous-ensembles du corpus)

Mais ces critères d'analyse n'ont de sens qu'en fonction du corpus de référence.

Source: Lebart, Pincemin et Poudat (2019)

Différentes façons de *revenir au texte*

Synthétique

- Relevé de termes
- Concordance: tous les contextes d'occurrence d'un mot/expression (pivot)
- Coocurrence (*quels mots s'attirent entre eux?*) et segments répétés
- Calcul des spécificités (répartition des mots entre les parties du corpus)
- Relevé d'extraits
- Retour au texte intégral (utile si éléments paratextuels, iconographiques, multimédias); le moins synthétique

Synthétique

- Quelques logiciels: Le Trameur; TXM; IRaMuTeQ

Les unités séquentielles simples Pour comprendre la *structure* du texte

- Méthode des segments répétés:
  - *Unités adjacentes récurrentes endogènes au corpus*
  - *Triés par longueur décroissante, fréquence décroissante et ordre alphabétique*
  - *Logiciel: lexico*

« *Le texte n'est pas qu'un sac de mots*, mais un système muni d'une structure linéaire sur laquelle s'ordonnent et se combinent les unités. » (Lebart, Pincemin et Poudat, 2019, p.72)

Les unités séquentielles complexes D’autres possibilités envisagées…

- Collocations (linguistique de corpus)
- *Prefabs* (approches cognitives)
- Quasi-segments
- Motifs textuels

JE VOUS REMERCIE DE VOTRE ATTENTION

Période de questions