ANALYSE DE DONNÉES TEXTUELLES EN ÉDUCATION: PRINCIPES ET APPLICATIONS POTENTIELLES

16 juin 2021

12h à 13h

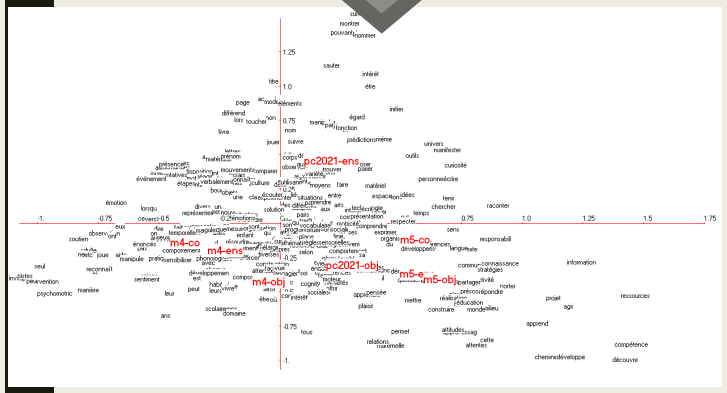
ADT: qu'est-ce que c'est?

- **Analyse de contenu à l'aide d'opérations statistiques (associations).**
- Plusieurs disciplines:
  - *Linguistique, analyse du discours, statistique, informatique, enquêtes socio-économiques, psychosociologie, marketing…*
- Applications potentielles:
  - *Pour explorer des corpus volumineux: curriculum, discours politiques, mémoires déposés, recherches…*
  - *Pour inférer et induire des structures sous-jacentes (explicatives ou non, p.ex. des attitudes, idéologies, représentations sociales).*
  - *Pour modéliser (p.ex. comportements ou pratiques).*

Des dénominations changeantes, mais trois grandes approches

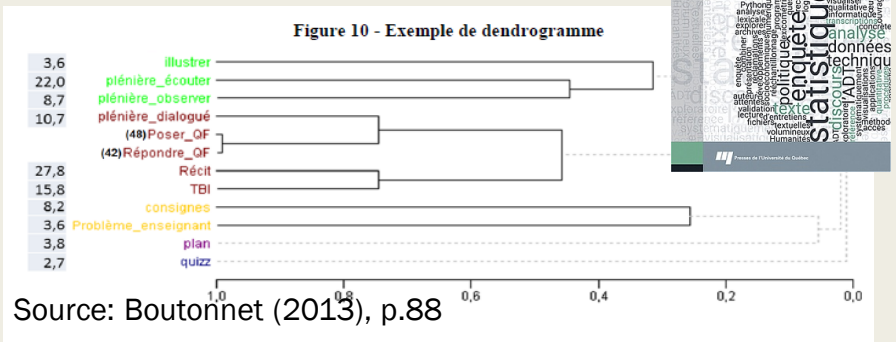
Lexicométrie

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



Textométrie

Analyse du *tissu textuel* pour décrire des attirances contextuelles, l'organisation internet de textes, les contrastes intertextuels, les caractéristiques de textes, des indicateurs d'évolution lexicale.



Source: Boutonnet (2013), p.88

Logométrie

Analyse de textes à travers l'hypertextualité (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.

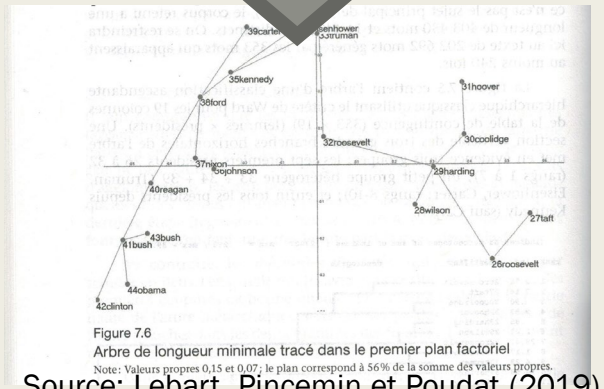
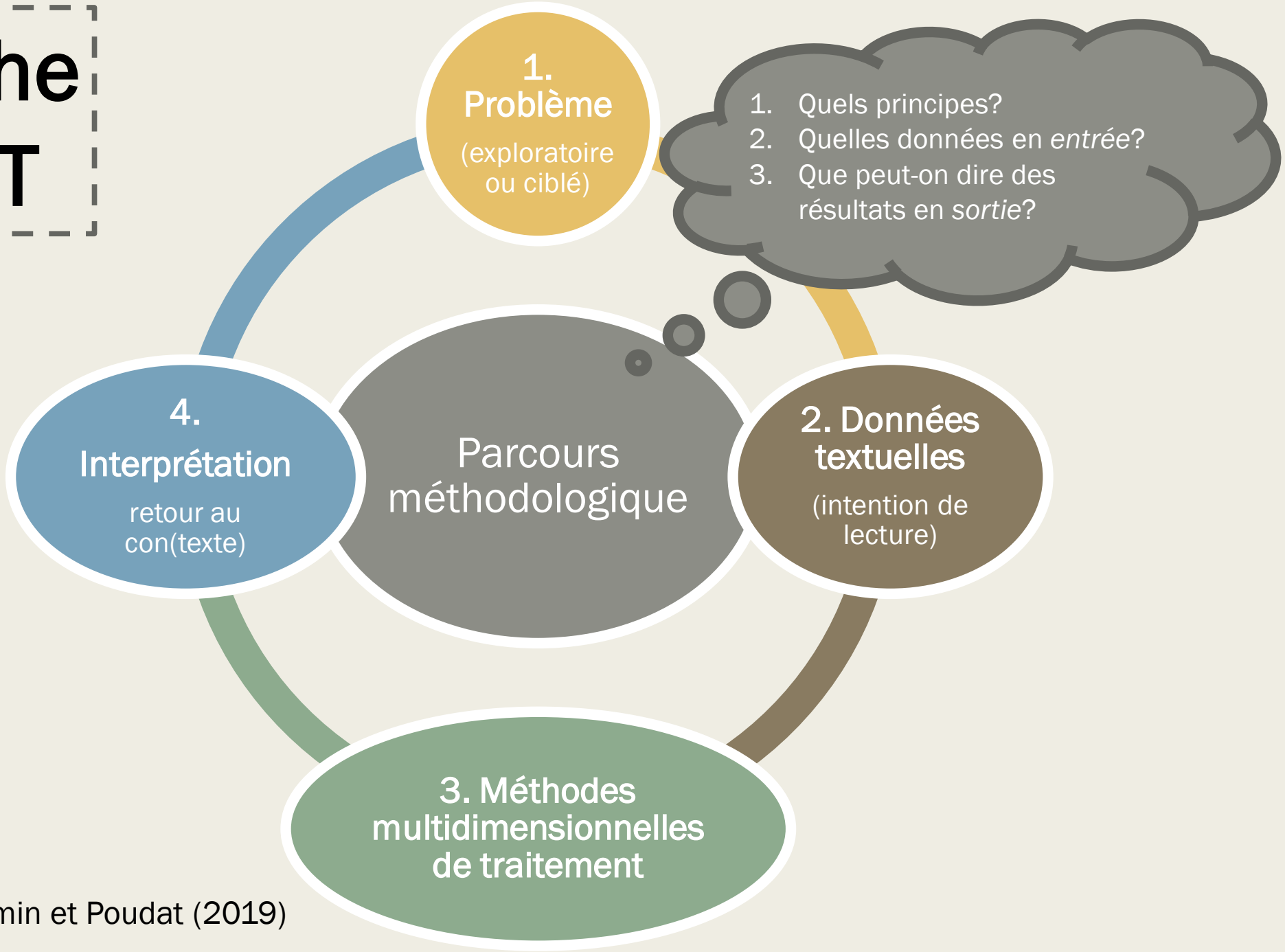


Figure 7.6
Arbre de longueur minimale tracé dans le premier plan factoriel
Note: Valeurs propres 0,15 et 0,07; le plan correspond à 56% de la somme des valeurs propres.

Source: Lebart, Pincemin et Poudat (2019), p.274

Démarche de l'ADT



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…*
- 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)

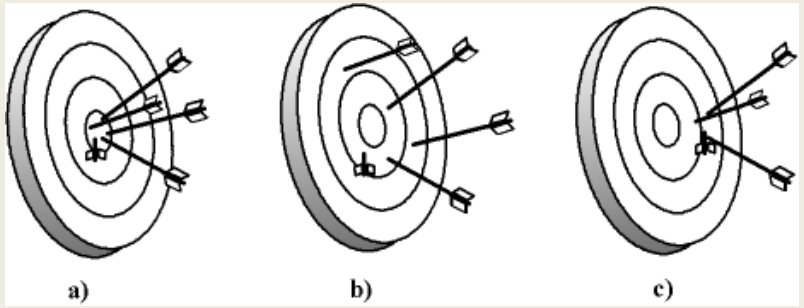
Exemple *Attributions enseignantes*
Quelles théories de l'apprentissage fondent les attributions enseignantes au regard de l'enseignement d'un mode de pensée historique (Moreau et Smith, 2017)

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

2. Données textuelles

(intention de lecture)

- 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*.
- Attention au « Gros, c'est beau » (Habert, 2000) -> *base documentaire*?
  - *Risques d'incertitude (random error) et de déformation (biais error).*



Source: https://commons.wikimedia.org/w/index.php?curid=20589720

2. Constitution du corpus Enquêtes socio-économiques

Principe d'intention de lecture

- L'intention de lecture dicte en amont la formulation de questions (discours organisé):
  - *Avoir une liste de questions préétablies dans leur libellé (nécessité!)*
  - *Avoir une liste de réponses préétablies (possibilité)*
- Variabilité selon la nature du questionnement (ouvert ou fermé) conduisant à des informations différentes et non comparables (Lazarsfeld, 1944).

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?

3. Méthodes multidimensionnelles de traitement

Composition du *Tableau lexical entier*

- 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)

3. Méthodes multidimensionnelles de traitement

Unités d'analyse

- 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)!*

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.

Statistique descriptive et exploratoire

- Établir des relations entre des variables ou des proximités entre des individus par une représentation graphique multidimensionnelle (nuage de points):
  1. *Méthodes factorielles (axes principaux):*
     - Analyse en composantes principales (ACP)
     - Analyse factorielle des correspondances (AFC)
     - Méthode de rééchantillonnage (*bootstrap*) (non abordée)
  2. *Méthodes de classification*

3. Méthodes multidimensionnelles de traitement

Statistique descriptive et exploratoire
Méthodes factorielles (axes principaux)

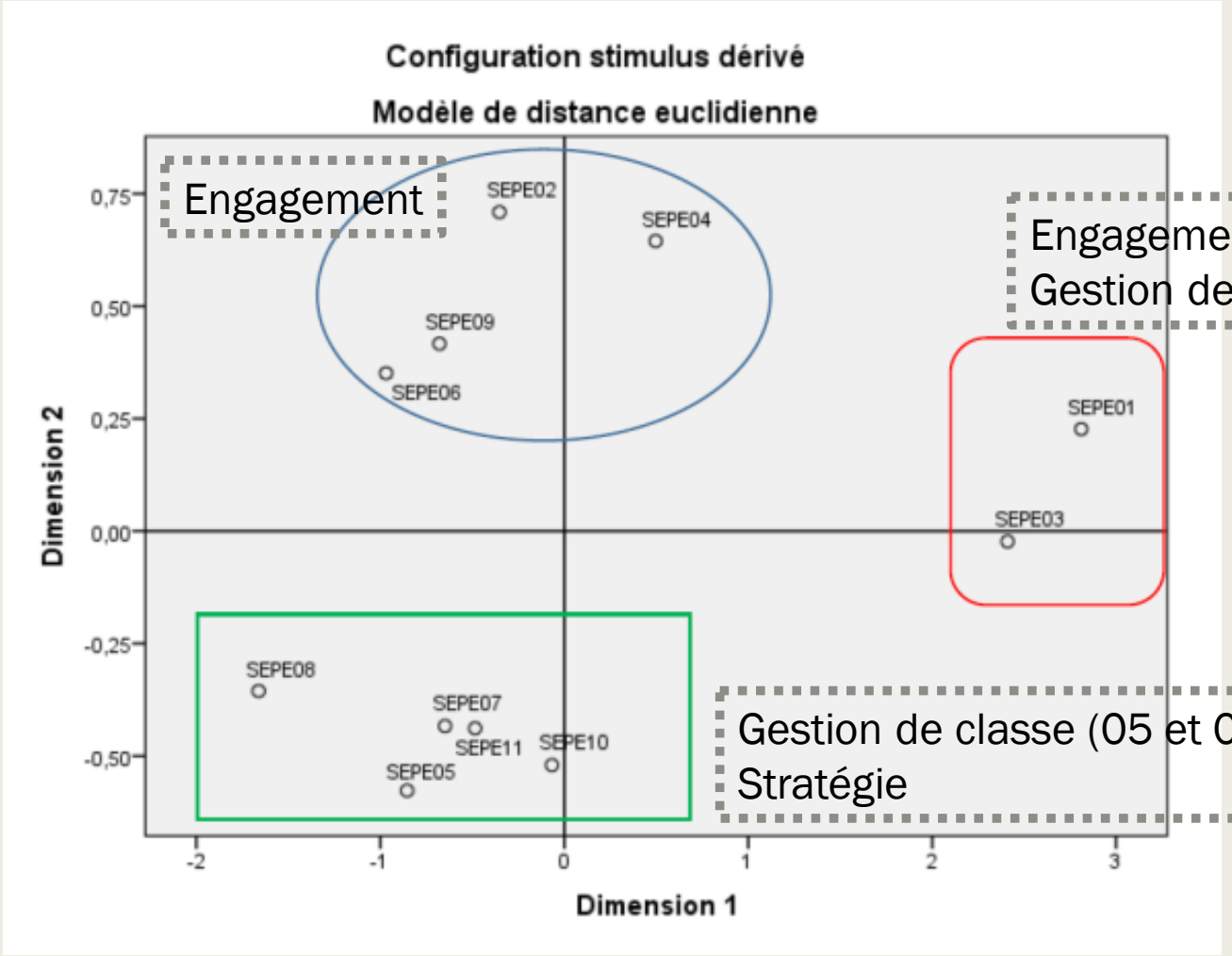
Analyse en composantes principales (ACP)

- 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)

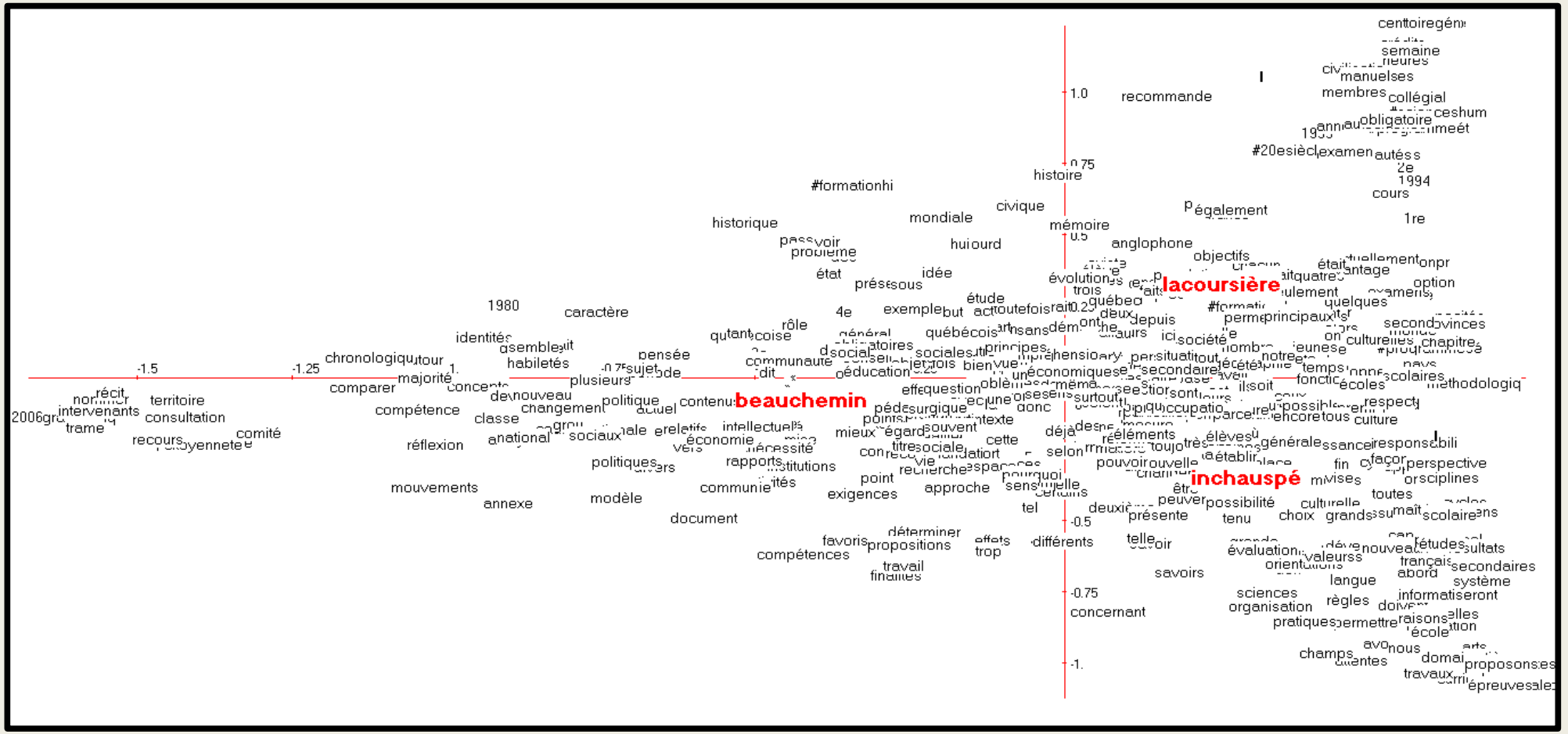
Analyse factorielle des correspondances (AFC)

- 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

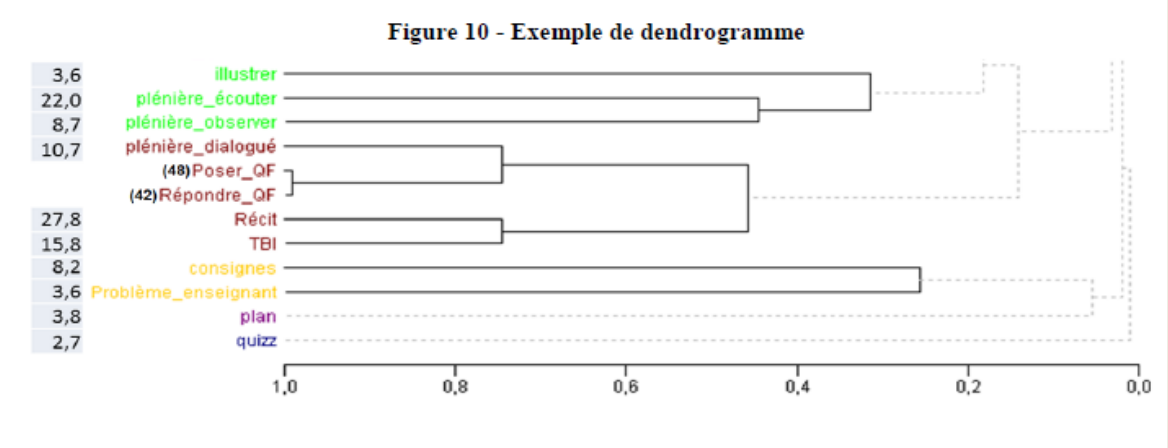


Exemple d'AFC
Débat sur l'histoire nationale



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 ascendante ou descendante*
  - *Partitionnement: cartes auto-organisées de Kohonen (non abordée).*



Source: Boutonnet (2013), p.88

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

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

Ce qui nous intéresse…

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

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
- 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: lexico5*

« *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)

JE VOUS REMERCIE DE VOTRE ATTENTION

Période de questions