A partir de documents écrits, étude de l'efficience et de la parcimonie dans la sélection d'extraits textuels

Comparaison d'un mode de sélection par le chercheur et d'un mode d'extraction automatisée.

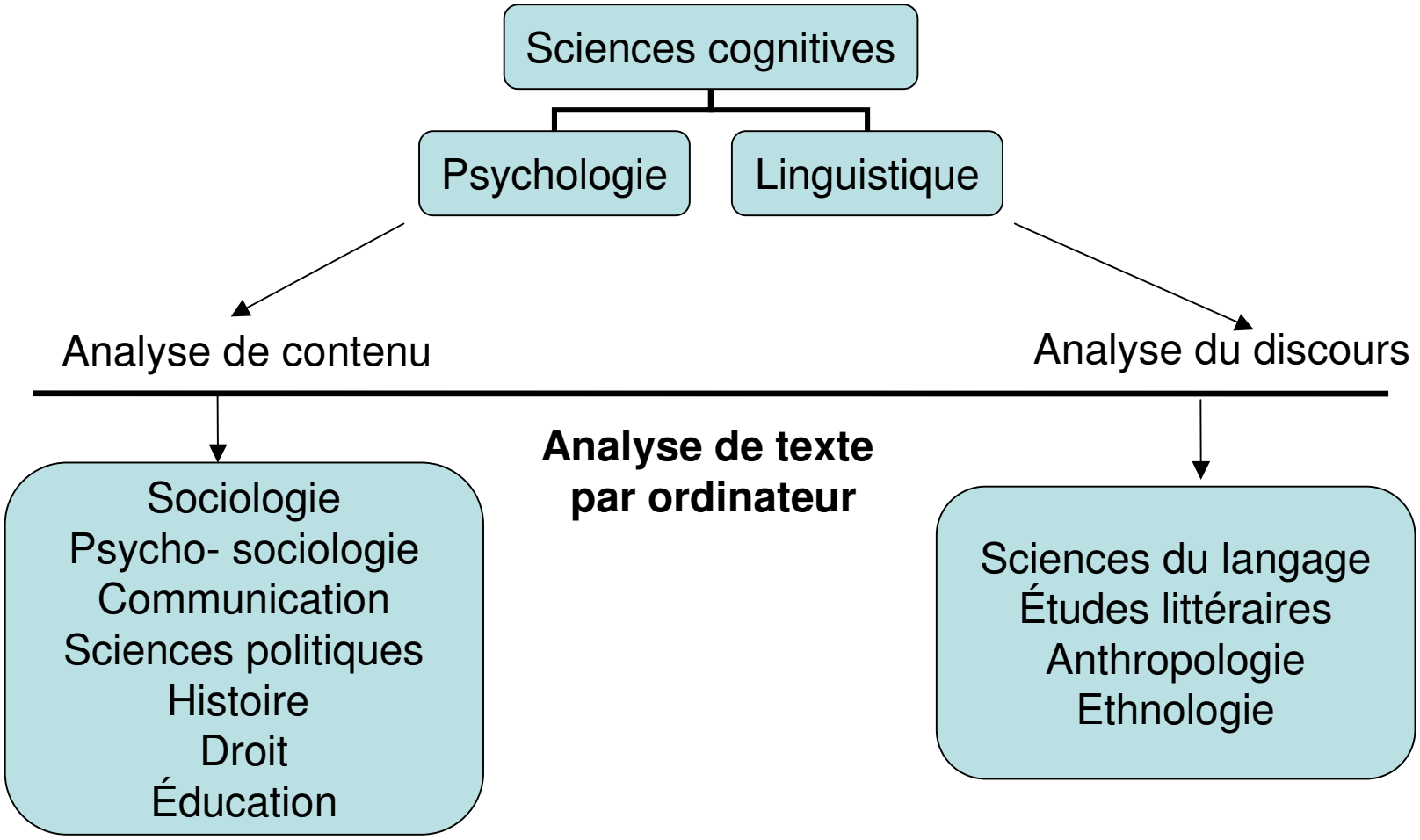
© Landry N., Pitman S. et Auger R.

Plan de la communication

- Description du contexte
- Instrumentation: base de données GRÉ, logiciel ALCESTE
- Critères de comparaison
- Résultats (efficience, parcimonie)
- Conclusion et discussion

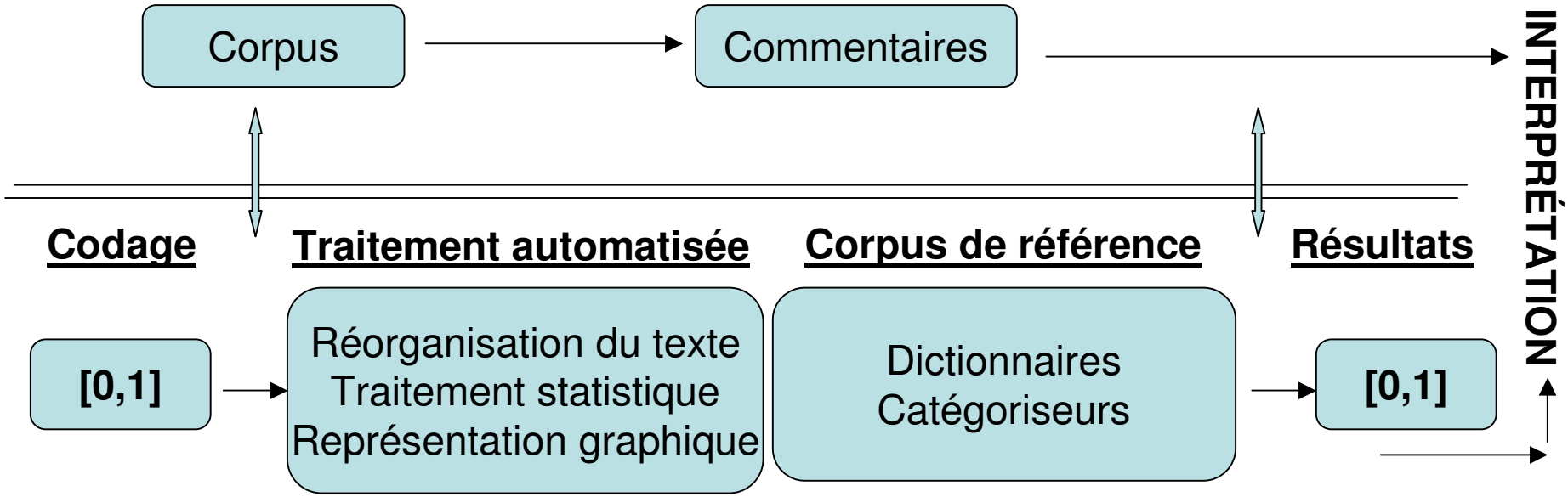
© Landry N., Pitman S., Auger R.

ARQ Colloque d'automne 2004 UQTR



Duchastel J. (2004). Formation en analyse de texte assistée par ordinateur. École d'été, UQAM.

Analyse de textes par le chercheur



Traitement automatisé de données textuelles

Corpus analysé: Patton 2002, chap. 9 et Silverman (1997)

Salem A. (2004). Formation en analyse de texte assistée par ordinateur. École d'été, UQAM.

Analyse textuelle fait par le chercheur à l'aide d'un outil de gestion des données

Base de données « Gestion de la Recension des Écrits » (GRÉ)

Catégorisation:

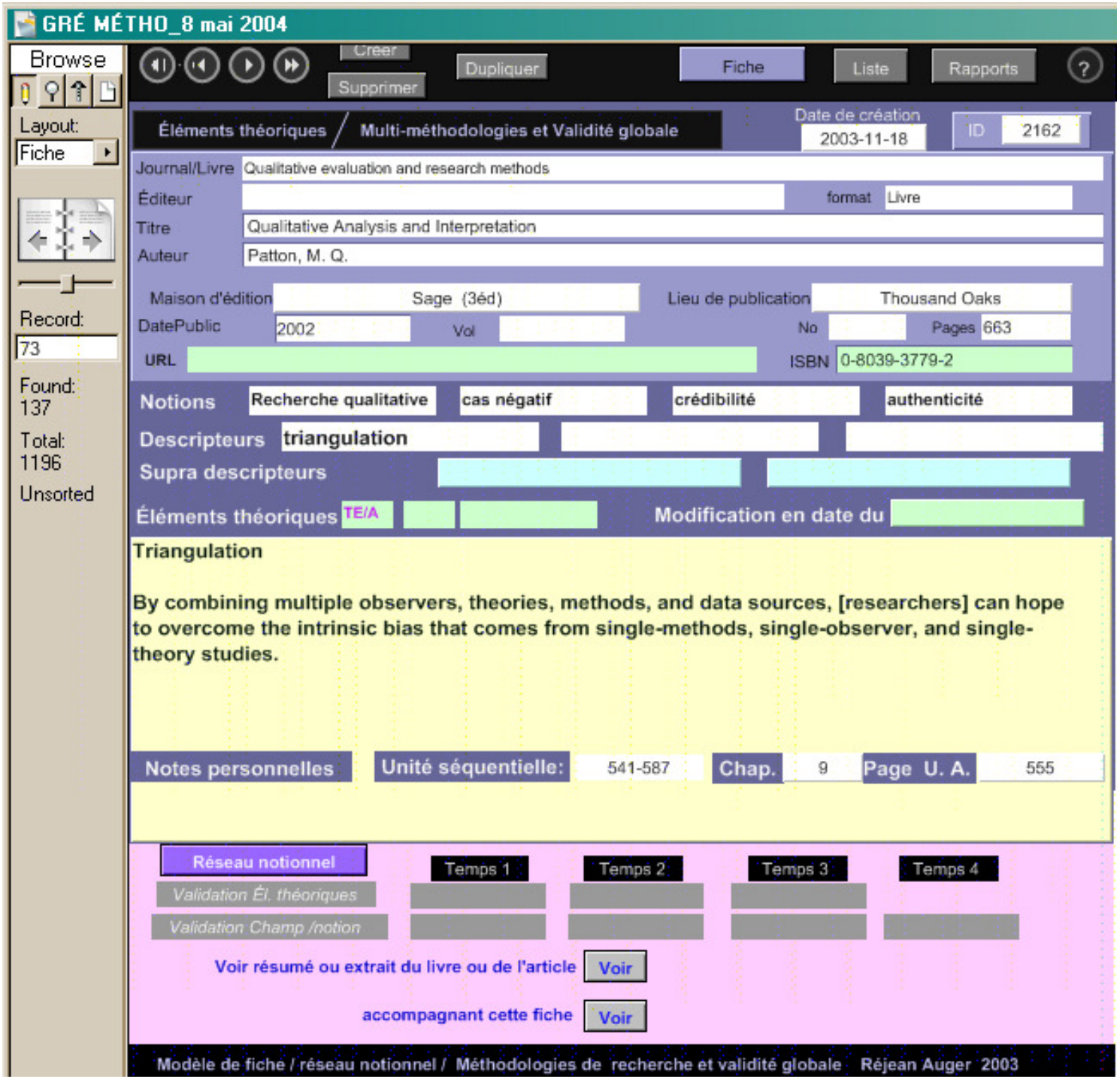
- **en fonction d'éléments théoriques:** TA: axiologiques, TF: formelles, TP: praxiologiques, TE: explicatifs.
- **de notions** tenant compte de l'expression explicite dans un texte
- **de descripteurs** de premier niveau d'inférence
- **de supra descripteurs** de deuxième niveau d'inférence

Validation:

- **par éléments théoriques (TA, TF, TP, TE)**
- **par notions, descripteurs, supra descripteurs**

Visualisation ou rapport:

- **d'extraits de l'article ou du livre**
- **de figures, d'images**
- **de notions par auteurs, dates**
- **d'éléments de synchronie, de diachronie**


GRÉ MÉTHO_8 mai 2004
Browse
Layout:
Fiche
Record:
73
Found:
137
Total:
1196
Unsorted
Créer
Supprimer
Dupliquer
Fiche
Liste
Rapports
Éléments théoriques / Multi-méthodologies et Validité globale
Date de création
2003-11-18
ID
2162
Journal/Livre
Qualitative evaluation and research methods
Éditeur
format
Livre
Titre
Qualitative Analysis and Interpretation
Auteur
Patton, M. Q.
Maison d'édition
Sage (3éd)
Lieu de publication
Thousand Oaks
DatePublic
2002
Vol
No
Pages
663
URL
ISBN
0-8039-3779-2
Notions
Recherche qualitative
cas négatif
crédibilité
authenticité
Descripteurs
triangulation
Supra descripteurs
Éléments théoriques
TE/A
Modification en date du
Triangulation
By combining multiple observers, theories, methods, and data sources, [researchers] can hope to overcome the intrinsic bias that comes from single-methods, single-observer, and single-theory studies.
Notes personnelles
Unité séquentielle:
541-587
Chap.
9
Page U. A.
555
Réseau notionnel
Temps 1
Temps 2
Temps 3
Temps 4
Validation Él. théoriques
Validation Champ /notion
Voir résumé ou extrait du livre ou de l'article
Voir
accompagnant cette fiche
Voir
Modèle de fiche / réseau notionnel / Méthodologies de recherche et validité globale Réjean Auger 2003

Traitement automatisé de données textuelles à l'aide du logiciel ALCESTE

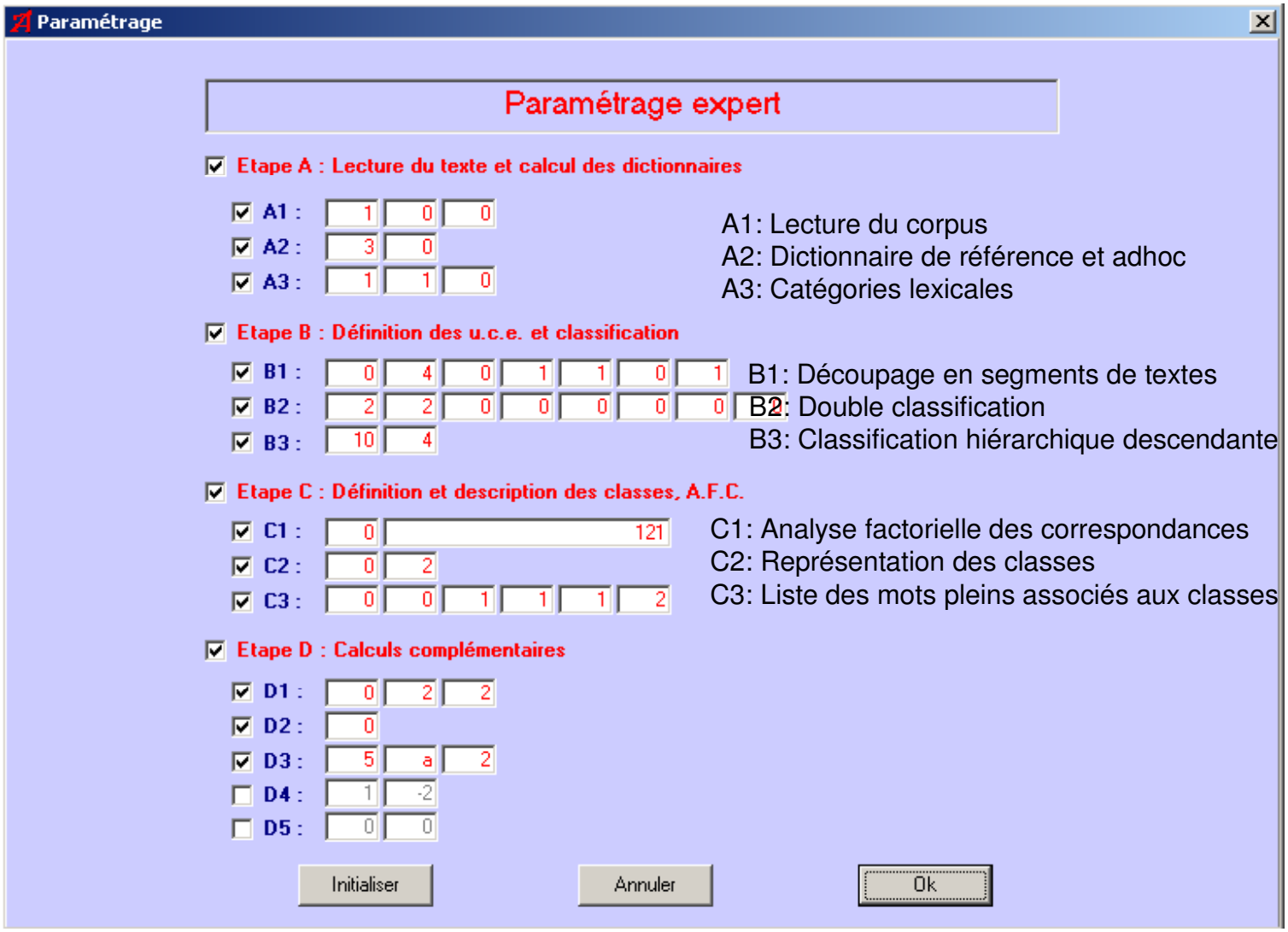
Analyse des Lexèmes Cooccurrents dans un Ensemble de Segments de Texte

- **Objectif:**
  - **quantifier un texte pour en extraire les structures signifiantes les plus fortes.**

- **Deux conditions:**
  - **1) Le corpus se présente comme un tout ayant une certaine cohérence**
  - **2) Document suffisamment volumineux pour que l'élément statistique entre en ligne de compte.**

- **Préparation du texte:**
  - **Identification d'une ligne étoilée:**

**** *aut_Patton *dat_2002 ou **** *aut_Silverman * dat_1997


Paramétrage
Paramétrage expert
Etape A : Lecture du texte et calcul des dictionnaires
A1 : 1 0 0
A2 : 3 0
A3 : 1 1 0
A1: Lecture du corpus
A2: Dictionnaire de référence et adhoc
A3: Catégories lexicales
Etape B : Définition des u.c.e. et classification
B1 : 0 4 0 1 1 0 1
B2 : 2 2 0 0 0 0 0
B3 : 10 4
B1: Découpage en segments de textes
B2: Double classification
B3: Classification hiérarchique descendante
Etape C : Définition et description des classes, A.F.C.
C1 : 0 121
C2 : 0 2
C3 : 0 0 1 1 1 2
C1: Analyse factorielle des correspondances
C2: Représentation des classes
C3: Liste des mots pleins associés aux classes
Etape D : Calculs complémentaires
D1 : 0 2 2
D2 : 0
D3 : 5 a 2
D4 : 1 -2
D5 : 0 0
Initialiser
Annuler
Ok

Comparaison du mode de sélection par le chercheur et du mode d'extraction automatisée des segments de textes

Sur la base de deux critères:

Efficience — **Relation entre deux facteurs s'exprimant sous la forme de ratio entre deux mesures.**

Parcimonie — **Diminution des informations tout en conservant toute l'information disponible par son caractère englobant et essentiel.**

Visant à réduire la complexité du réel.

La conduite rationnelle implique la substitution à la réalité complexe d'un schéma de la réalité assez simple pour pouvoir être pris en charge par une activité résolutoire.

Statistiques descriptives de l'analyse textuelle par ALCESTE

Éléments de comparaison	Classe 1	Classe 2	Classe 3	Classe 4	Total	Total absolu
UCI	Patton (1- 563)		Silverman (564-729)		2	2
UCE retenues	72	119	63	163	417 (57%)	729
% relatif d'uce	17,28%	28,54%	15,10%	39,09%		
% absolu d'uce	9,88%	16,32%	8,64%	22,36%		
Mots pleins retenus	97	181	133	155	566 (66%)	860
Mots pleins significatifs	38	45	32	71	186 (33%)	
Mots pleins significatifs / nombre de formes distinctes dans le corpus					186 (4%)	4353

Temps d'exécution: 2 minutes 13 secondes pour 4 353 formes distinctes

© Landry N., Pitman S., Auger R.

Extrait textuel par Alceste versus par le chercheur [UCE]

lone #analysts, and single perspective #interpretations. however, a #common misconception about #triangulation #involves thinking that the purpose is to demonstrate that #different #data #sources or inquiry approaches #yield essentially the #same #result.

However, a common misconception about triangulation involves thinking that the purpose is to demonstrate that different data sources or inquiry approaches yield essentially the same result. The point is to test for such consistency. Different kinds of data may yield somewhat different results because different types of inquiry are sensitive to different real-world nuances. Thus, understanding inconsistencies in findings across different kinds of data can be illuminative and important.

Classe: 1 Descripteur: diversité des sources

Statistiques descriptives de l'analyse textuelle par le chercheur

Éléments de comparaison	Classe 1	Classe 2	Classe 3	Classe 4	Classe 5	Total	Total absolu
UCI	Patton (1- 255)		Silverman (256-304)			2	2
UCE retenues	70	89	2	104	40	304 (42%)	729
% relatif d'uce	23,03%	29,28%	0,007%	34,21%	13,16%		
% absolu d'uce	9,60%	12,21%	0,0002%	14,27%	5,49%		
MOTS pleins retenus	246	307	8	389	141	1091 (25%)	4353
Mots pleins retenus / nombre de formes distinctes dans le corpus							

Temps d'exécution: **135 heures pour 4 353 formes distinctes**

Mesures d'efficience

Mesure d'efficience absolue :
Le rapport entre le nombre de formes distinctes du corpus intégral sur le nombre de classes et la durée du traitement en secondes.

Mesure relative d'efficience :
Le ratio entre la mesure de l'efficience du logiciel Alceste et la mesure de l'efficience du chercheur.

Efficience (MOTS)	Alceste	Chercheur
absolue	(4 353 formes / (4 classes * 133 sec.) = 8,18	(4 353 formes / (5 classes * 8100 sec.) = 0,11
relative	8,18 / 0,11 = 74,36	

Indices de parcimonie

Mesure absolue

(Total des UCE retenues / Total des UCE potentielles du corpus) *100

(Total de mots pleins significatifs / Grand total des formes distinctives) *100

Interprétation

Plus la valeur est petite, plus grande est la parcimonie.

Parcimonie / mesure absolue	Alceste	Chercheur
UCE	(417 / 729) *100 = 57%	(304 / 729) *100 = 42%
MOTS	(186/ 4 353) *100 = 4%	(1091/ 4 353) *100 = 25%

En ce qui concerne la classification et la catégorisation des UCE

Cas de la notion de « crédibilité » selon chacune des classes identifiées par le chercheur		Classes et descripteurs				
		Diversité des sources (Classe 1)	Rigueur méthodologique (Classe 2)	L'impact du chercheur (Classe 3)	Types d'approche (Classe 4)	Crédibilité du chercheur (Classe 5)
		Fréquence	Fréquence	Fréquence	Fréquence	Fréquence
Crédibilité	N1	0	2	1	2	2
	N2	0	0	0	1	0
	N3	2	1	0	2	2
	N4	4	1	0	1	0
Total:		6	4	1	6	4

UCE retenues	Classe 1	Classe 2	Classe 3	Classe 4	Classe 5
Alceste (417 uce)	72	119	63	163	.....
Chercheur (304 uce)	70	89	2	104	41

En somme…

Le logiciel Alceste est un instrument susceptible d'augmenter **l'efficience** dans le traitement de données textuelles (collecte et classification).

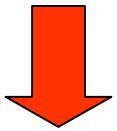
Le logiciel Alceste est un instrument susceptible de favoriser une collecte **parcimonieuse** de données textuelles, très précisément en ce qui concerne le repérage et la classification des **mots** pleins du corpus.

Tandis que le chercheur pourrait s'avérer plus **parcimonieux** que le logiciel Alceste lorsqu'il s'agit d'extraire des unités contextuelles élémentaires (UCE) ou **unités de texte**.

L'intégration des deux modes d'extraction de données textuelles (par le chercheur et automatique) est-elle souhaitable à l'intérieur d'une démarche générale d'analyse de textes ?

Et dans quelles limites ?

Représentativité des classes et des catégories
Adéquation de la catégorisation



Inférence de connaissances valides et réplicables

Fin de la présentation

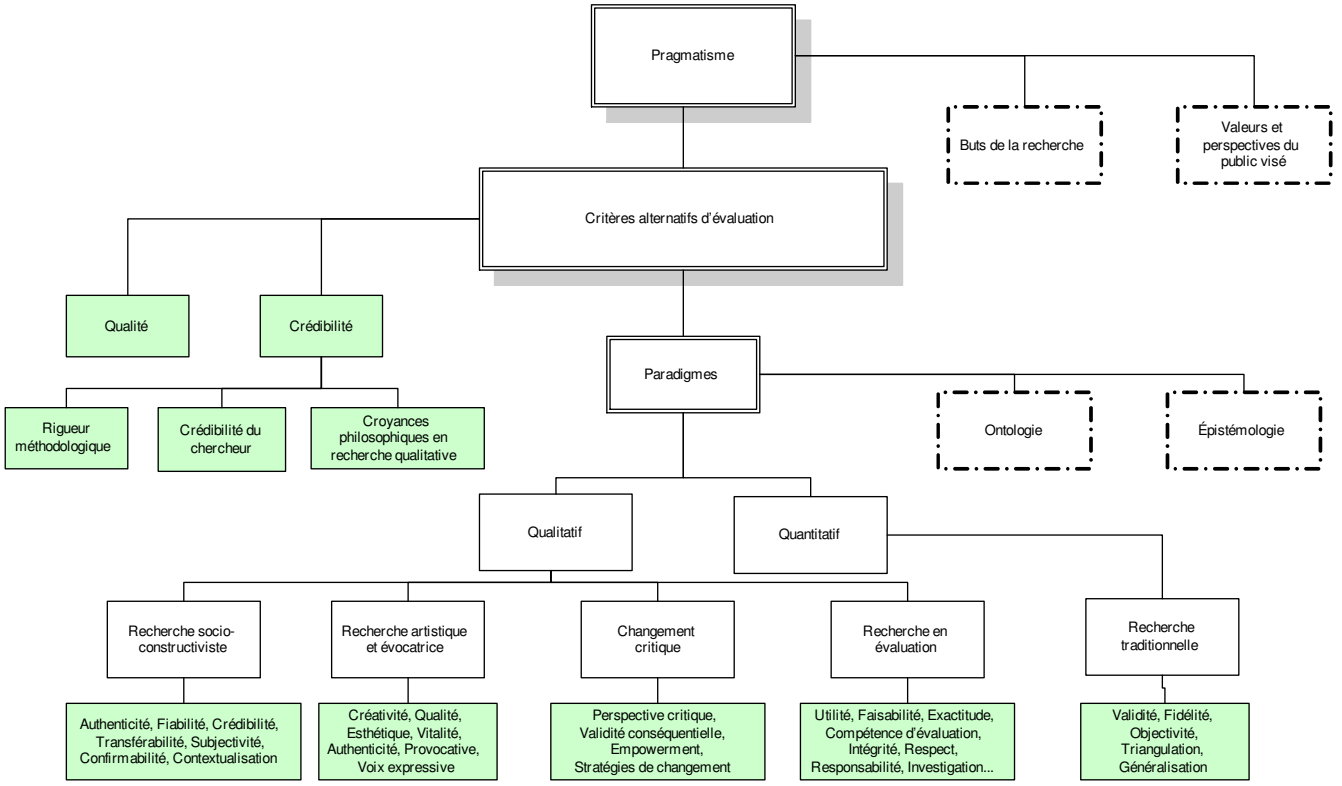
© Landry N., Pitman S., Auger R.

Diapositives supplémentaires

AU CAS OÙ…

Le réseau notionnel du chercheur

Qualité et crédibilité d'une recherche



Réseau notionnel / chercheur — Mode combiné — 2003-09-23

Efficience relative entre Alceste et le chercheur

Ratios calculés (Alceste / chercheur) en fonction de différentes bases de comparaison

	Classe 1	Classe 2	Classe 3	Classe 4	Classe 5	Total
UCE retenues	**1,03**	**1,34**	**31,5**	**1,57**	**------**	**0,73**
ALCESTE/Chercheur	**(72 / 70)**	**(119 / 89)**	**(63 / 2)**	**(163 / 104)**		**[1,75 / 2.40]**
MOTS	**Formes distinctes dans le corpus intégral 4 353 / nb. classes / durée**					**74,36**
Alceste	**(4 353 formes / (4 classes * 133 sec.) = 8,18**					**[8,18 / 0,11]**
Chercheur	**(4 353 formes / (5 classes * 8100 sec.) = 0,11**					
Temps d'exécution						**0,016**
Alceste	2 minutes 13 secondes ou **133 secondes**					**[133/ 8100]**
Chercheur	135 heures ou **8100 secondes**					

Efficience relative entre Alceste et le chercheur

Selon différents éléments de comparaison

	Classe 1	Classe 2	Classe 3	Classe 4	Classe 5	Total
UCE retenues	1,03	1,34	31,5	1,57	------	0,73
ALCESTE/Chercheur	(72 / 70)	(119 / 89)	(63 / 2)	(163 / 104)		
MOTS	Formes distinctes dans le corpus intégral 4 353 / nb. classes / durée					74,36
Alceste	(4 353 formes / (4 classes * 133 sec.) = 8,18					[8,18 / 0,11]
Chercheur	(4 353 formes / (5 classes * 8100 sec.) = 0,11					