



From texts to codes.  
Multivariate descriptive methods applied to  
open-ended survey questions.

Anne-Sophie Cousteaux  
ESRA Conference, June 29– July 3, 2009

### A challenge for survey research

“Processing free responses (i.e. responses to open questions) is a challenge for both statisticians and specialists in text analysis. In a corpus of free responses, lexical frequencies are artificial for the most part, because the same question is asked among hundreds or thousands of people. The juxtaposition of the responses results in a redundant text by construction, in which stereotypes are not uncommon.”

LEBART, SALEM, BERRY, *Exploring Textual Data*, 1998, p.14

### The example of lay perceptions of health

“What does being healthy mean to you?” (INSEE, EPCV, 2001)

### Objective

To present how multivariate descriptive methods (correspondence analysis and cluster analysis) can successfully be applied to textual data to produce a standardized classification of the individual answers.

### 1. Typological classifications of lay conceptions of health

- 1.1 From the analysis of in-depth interviews
- 1.2 From the manual post-coding of open questions

### 2. Data and methods : A lexicometric analysis of textual data

- 2.1 The open-ended question on health
- 2.2 Textual analyses carried out by SPAD statistical package
- 2.3 Counting words and accounting for contexts

### 3. Results : The final classification

- 3.1 Correspondence analysis of the lexical table
- 3.2 Hierarchical cluster analysis
- 3.3 Description of the ten classes

## Conclusions

## 1.1 From the analysis of in-depth interviews

HERZLICH (1969) : “health-in-a-vacuum”, “reserve of health”,  
“equilibrium”

PIERRET (1984) : “health-illness”, “health-tool”, “health-product”,  
“health-institutions”

**But** samples restricted to healthy men and women of working age  
⇒ Variations according to age? health status?

Questionnaire vs in-depth interviews  
= a large and representative sample

## 1.2 From the manual post-coding of open questions

d'HOUTAUD, FIELD (1984) a survey carried out among 4000 health centre users

“What is, according to you, the best definition of health?” => 10 themes

**But** coder bias

BLAXTER (1990) *Health and Lifestyle Survey* (n=9000 individuals)

2 questions : “Describe a healthy person you know” and “What is it like when you are healthy?” => 9 themes

**But** cost of applying manual technique to thousands of answers

Multivariate descriptive methods vs manual technique

= a more objective summary easily produced by statistical methods

## 2.1 The open-ended question on health

“What does being healthy mean to you?” (INSEE, EPCV, 2001)

A possible halo effect



Medical consumption	Diseases	Health system	Lifestyle behaviours	Working conditions	Emotional well-being
---------------------	----------	---------------	----------------------	--------------------	----------------------

4912 responses out of 5194 interviewees = 95%

### Two understandings of the question

- The definition of good health
- The value attached to health


A telegraphic writing style (interviewer bias)

## 2.2 Textual analyses carried out by SPAD<sup>®</sup> statistical package

### The lexicometric approach

- Based on frequencies of words and of repeated segments.

### But counting words is not so easy !

- Homonymy, synonymy, polysemy, negative expression  
=> The need for the context 
- Function words (articles, pronouns, conjunctions...)
- Hapax (words that appears only once)

### Methodology

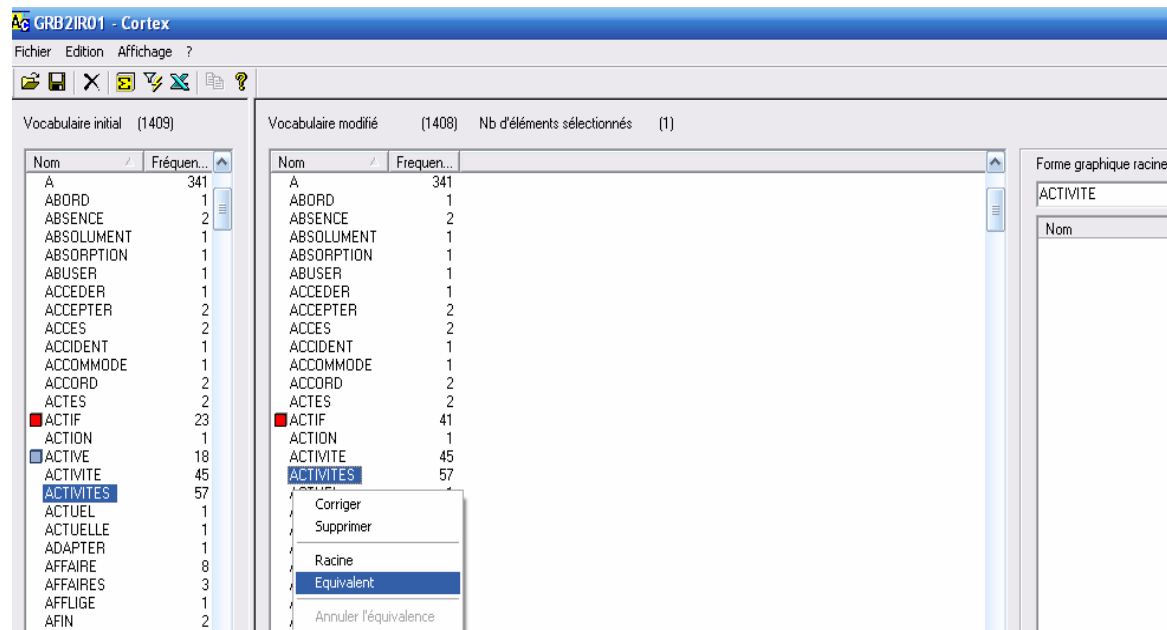
- Words turned into variables after lemmatization.
- Multivariate descriptive methods applied to texts :  
correspondence analysis of lexical tables then hierarchical cluster analysis

### 2.2 Textual analyses carried out by SPAD® statistical package

#### Lemmatization

Grouping words together

Some conventions in French : nouns put into the singular, adjectives put into the masculine singular, verbs in the infinitive (and possibly synonyms)



A semi-automatic process in SPAD : defining the root and its equivalents

1409 distinct words  
=> 122 lemmas

### 2.3 Counting words and accounting for contexts

#### A quantitative analysis of lemmatized vocabulary

Lemmas	Translation	Frequency
être	to be	1958
avoir	to have	1318
bien	well, good	1210
pouvoir	to be able to	899
faire	to do (sth)	631
forme	(on) form	574
malade	ill	482
sentir	to feel	448
tout	all	404
aller	to go	269
vivre	to live	265
travailler	to work	261
peau	skin (french expression : to feel good about oneself)	258
mal	(to feel) bad, (to have) pain	239
vie	life	237
bon	well, good	236
médecin	doctor	233

## 2.3 Counting words and accounting for contexts

### Repeated segments

	Positive wording	Negative wording
Avoir (to have) n = 1318	Avoir la forme (to be feeling great) n = 130 Avoir envie (to feel like sth, to want to do) n = 82 Avoir le moral (to be in good spirits) n = 52 Avoir une bonne hygiène (to have a healthy lifestyle) n = 12	Ne pas avoir mal (not to have pain) n = 197 Ne pas avoir de maladie (not to have illness) n = 126 Ne rien avoir (to have nothing) n = 35 Ne pas avoir recours au médecin (not to go to the doctor's) n = 14
Etre (to be) n = 1958	Etre bien (to feel good) n = 420 Etre bien dans sa peau (to feel good about oneself) n = 139 Etre en pleine forme (to be in fine form) n = 62 Etre heureux (to be happy) n = 50 Etre actif (to be active) n = 31	Ne pas être malade (not to be ill) n = 419 Ne pas être fatigué (not to be tired) n = 49
Pouvoir (to be able to) n = 899	Pouvoir faire (to be able to do sth) n = 245 Pouvoir travailler (to be able to work) n = 144 Pouvoir marcher (to be able to walk) n = 52 Pouvoir se lever (to be able to get up) n = 30 Pouvoir vivre normalement (to be able to lead a normal life) n = 27	

## 2.3 Counting words and accounting for contexts

### Specific vocabulary for sub-populations

Example : Lay conceptions of health over the lifecourse

18-29	not to be ill to be in fine forme (mentally and physically) sport, healthy lifestyle
30-44 45-59	to work, to be able to work, to be able to go to work not to go to the doctor's
60-74	wealth, the most important thing ( <i>value attached to health</i> ) to be happy not to suffer
75 and more	dependency, capability, to be self-sufficient, to manage on one's own to lead a normal life to be able to move, to walk, to do DIY (do-it-yourself), to do gardening, to do the housework not to suffer

### 3.1 Correspondence analysis of the lexical table

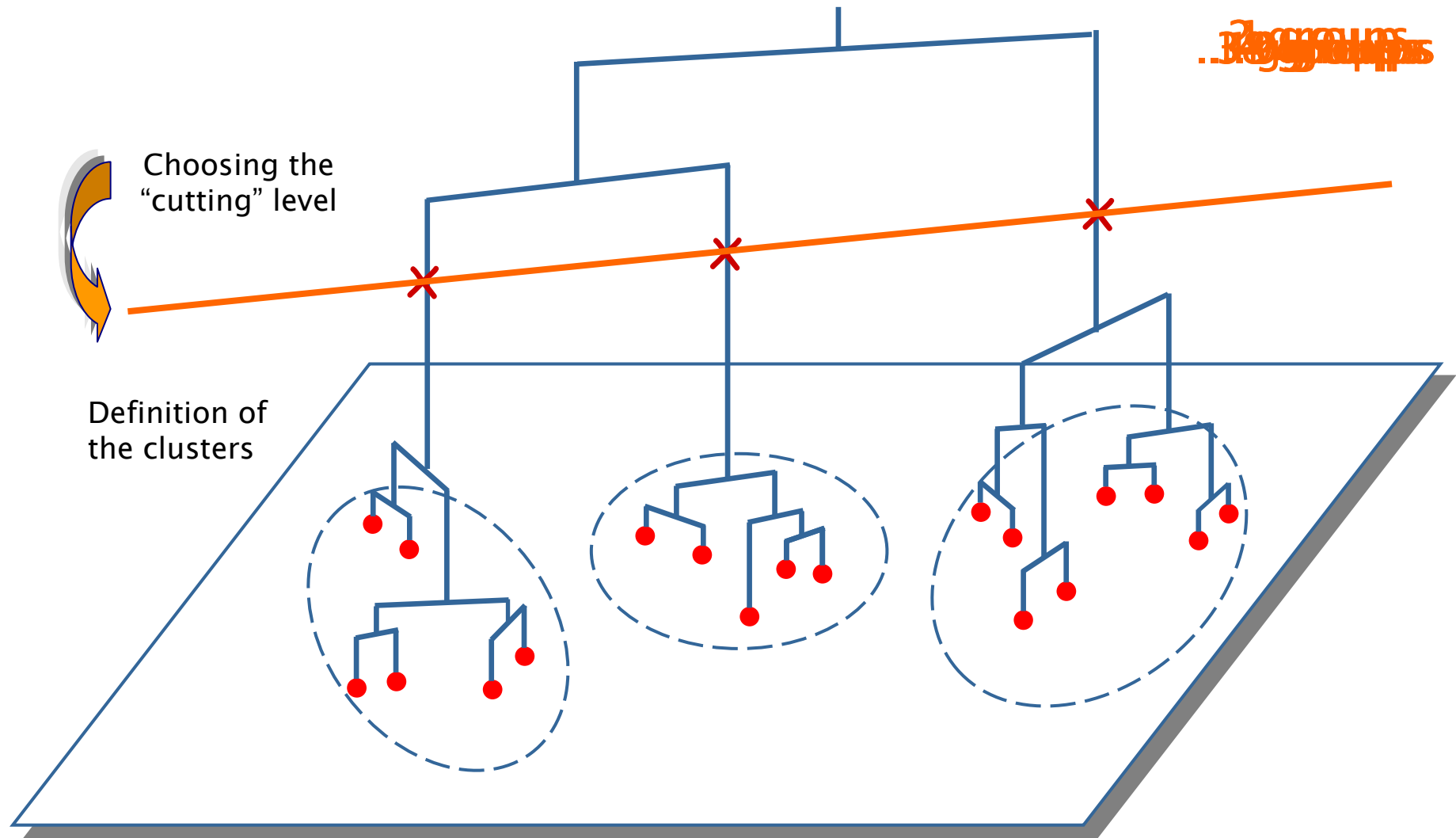
#### The sparse matrix T

	Active variables					Supplementary variables						
	L1	L2	L3	L4	...	L114	...	L122	Sex	Age	Education	Chronic disease
Ind 1	1	0	0	1		0		0	1	3	2	0
Ind 2	0	2	0	0		0		0	1	2	2	1
Ind 3	0	0	0	0		1		0	1	1	1	0
...												
Ind 4912	0	0	1	0		0		0	1	3	1	0

Correspondence analysis performed to obtain the individual coordinates on principal axes

3. Results : The final classification

3.2 Hierarchical cluster analysis



Source : Michel Tenenhaus

### 3.3 Description of the ten classes

#### Results of the cluster analysis

Frequency	Characteristic lemmas
14,8%	life, well/good, mental, physical, health problems, healthy lifestyle, hygiene, to have, to be in good spirits, healthy, trouble, nothing, healthy diet, smoking
19,0%	to be, ill, form, to suffer, autonomous, happy, tired, active, energy, vitality, young
19,7%	well/good, to feel, skin, head, body, to sleep, to eat, pain, to be
3,6%	illness, to have, serious
5,1%	doctor, to visit, to go, to need
3,1%	medicine, to take, to need, obligation, permanently
5,6%	pain, nowhere, to have, backache
3,7%	to live, delight, normal, to lead, pleasure, to be able to, life, happy
24,1%	to do, to be able to, to want, to work, to walk, sport, activities, to get up, to move, work, to take care, DIY, morning, housework, everyday, restraint, family, leisure activity, to keep, difficulty, help, to eat
1,3%	Alone, to be self-sufficient, to manage, the others, dependency, to be able to, autonomous

### 3.3 Description of the ten classes

#### The interpretation

Registers	Classes	Frequency	Characteristic lemmas
Complete well-being	<b>Healthy lifestyle</b>	14,8%	life, well/good, mental, physical, health problems, healthy lifestyle, hygiene, to have, to be in good spirits, healthy, trouble, nothing, healthy diet, smoking
	<b>Energy</b>	19,0%	to be, ill, form, to suffer, autonomous, happy, tired, active, energy, vitality, young
	<b>Well-being</b>	19,7%	well/good, to feel, skin, head, body, to sleep, to eat, pain, to be
Illness and its consequences	<b>Absence of serious illness</b>	3,6%	illness, to have, serious
	<b>Never go to the doctor's</b>	5,1%	doctor, to visit, to go, to need
	<b>Absence of medicine</b>	3,1%	medicine, to take, to need, obligation, permanently
	<b>Absence of pain</b>	5,6%	pain, nowhere, to have, backache
Instrument	<b>Being happy and leading a normal life</b>	3,7%	to live, delight, normal, to lead, pleasure, to be able to, life, happy
	<b>Being able to</b>	24,1%	to do, to be able to, to want, to work, to walk, sport, activities, to get up, to move, work, to take care, DIY, morning, housework, everyday, restraint, family, leisure activity, to keep, difficulty, help, to eat
	<b>Dependency</b>	1,3%	alone, to be self-sufficient, to manage, the others, dependency, to be able to, autonomous

### 3.3 Description of the ten classes

Registers	Classes	Frequency	Characteristic categories			
			Age	Sex	Education	Chronic disease
Complete well-being	Healthy lifestyle	14,8%	18-29		high	
	Energy	19,0%	18-29 30-44			
	Well-being	19,7%	30-44	men	high	no
Illness and its consequences	Absence of serious illness	3,6%	30-44	women	high	no
	Never go to the doctor's	5,1%	30-44	men		
	Absence of medicine	3,1%			high	yes
	Absence of pain	5,6%	45-59	women	low	yes
Instrument	Being happy and leading a normal life	3,7%	60-74			
	Being able to	24,1%	75+	men		
	Dependency	1,3%	75+	women	low	

#### From a methodological point of view

- Interesting results for open-ended questions
- Multivariate descriptive methods able to describe, quantify and characterize
- Computer-aided coding which produces less subjective categorizations and reduces treatment costs

**But** sensitivity to methodological choices (lemmatization, threshold, aggregated or whole lexical tables)



Thank you for your attention

[annesophie.cousteaux@sciences-po.fr](mailto:annesophie.cousteaux@sciences-po.fr)

### Contextualisation of words

#### Example : constraint

---

```
EDITION DES CONTEXTES DE MOTS
CONTEXTES DU MOT: CONTRAINTE
FREQUENCE DE REPETITION DU MOT :    25
      POUVOIR FAIRE TOUT CE QU ON VEUT SANS CONTRAINTE PHYSIQUE
              VIVRE SANS CONTRAINTE
              N AVOIR AUCUNE CONTRAINTE PHYSIQUE
      NORMALEMENT SANS ENTRAVE PRATIQUER LES ACTIVITES SANS CONTRAINTE
              FAIRE CE QUE J AI ENVIE DE FAIRE SANS CONTRAINTE
      FAIRE CE QUE L ON VEUT ET ASSURER LE QUOTIDIEN SANS CONTRAINTE PHYSIQUE ET MENTALE
D EXCERCER UNE ACTIVITE PROFESSIONNELLE ET SPORTIVE SANS CONTRAINTE NI RESTRICTION
      NE PAS ETRE MALADE NE PAS SOUFFRIR NE PAS AVOIR DE CONTRAINTE REGIME PRISE DE MEDICAMENTS
              AVOIR AUCUNE CONTRAINTE DE SUIVI MEDICAL
      POUVOIR FAIRE CE QUE L ON VEUT SANS CONTRAINTE
              S ALIMENTER NORMALEMENT SANS CONTRAINTE ABSORPTION D ALCOOL REDUITE AVOIR UN MINIMUM D ACTIVITE
      POUVOIR CONTINUER A VIVRE SANS CONTRAINTE
      POUVOIR SE DEPLACER ET MANGER SANS CONTRAINTE
      ETRE BIEN PHISIQUEMENT ET MORALEMENT PAS DE CONTRAINTE PHYSIQUE
              MENER UNE VIE NORMALE SANS CONTRAINTE
              FAIRE SANS CONTRAINTE CONTRAIRE DU HANDICAP
              VIVRE SANS MEDICAMENT SANS CONTRAINTE DE REGIME
      CHEZ LE MEDECIN ET POUVOIR FAIRE CE QUE L ON VEUT SANS CONTRAINTE
              POUVOIR FAIRE TOUT CE QUE JE VEUX SANS CONTRAINTE MEDICALE
      POUVOIR FAIRE CE QUE J AI ENVIE SANS AVOIR DE CONTRAINTE
      ETRE EN FORME POUVOIR SE DEPLACER SANS CONTRAINTE
      POUVOIR DISPOSER DE SES MOUVEMENTS SANS CONTRAINTE ET SANS DOULEURS
      POUVOIR FAIRE CE QUE L ON VEUT PAS DE REGIME PAS DE CONTRAINTE
      POUVOIR PRATIQUER LES LOISIRS SPORTS TRAVAIL SANS AUCUNE CONTRAINTE
      AVOIR LA POSSIBILITE DE FAIRE DU SPORT NE PAS AVOIR DE CONTRAINTE DANS LA VIE DE TOUS LES JOURS
```

