SENSOMETRICS - 2012

AgroCampus Ouest, Rennes, July 10-13

Processing Texts and Open-ended Questions in Sample Surveys

Ludovic Lebart

Centre National de la Recherche Scientifique
Telecom-ParisTech, Paris, France

www.lebart.org

Processing Texts and Open-ended Questions in Sample Surveys

Summary / Outline

1) Principles of Data Mining and Text mining: A reminder

2) Open-ended Questions: Why? How?

3) From texts to numerical data

4) Basic statistical tools: Visualization, Characteristic words, Bootstrap..

5) Applications: Open questions, sample surveys, texts

6) About textual data in general

7) Conclusions

Text Mining and Open-ended Questions in Sample Surveys

Summary / Outline

1) Principles of Data Mining and Text mining: A reminder

2) Open-ended Questions: Why? How?

3) From texts to numerical data

4) Basic statistical tools: Visualization, Characteristic words, Bootstrap.

5) Applications: Open questions, sample surveys, texts

6) About textual data in general

7) Conclusions

“Text Mining” and Multivariate exploratory statistical analysis of texts

Initial paradigm:

- Extracting statistical units from texts

- Complementing lexicometry with a multivariate approach

- Applying visualization tools to lexical tables

- Statistical validation and inference.

The fields of Text Mining

WEB

Press

Scientific papers, abstracts

Information Retrieval

Open-ended questions, free responses

Qualitative interviews, Discourses, Reports

Complaints

Text Mining and Open-ended Questions in Sample Surveys

Summary / Outline

1) Principles of Data Mining and Text mining: A reminder

2) Open-ended Questions: Why? How?

3) From texts to numerical data

4) Basic statistical tools: Visualization, Characteristic words, Bootstrap.

5) Applications: Open questions, sample surveys, texts

6) About textual data in general

7) Conclusions

Open questions : Why?

- ***To shorten interview time:***

Open ended questions are less costly in terms of interview time, and generate less fatigue and tension (voluminous lists of items)

- ***To gather spontaneous information:***

Marketing survey questions contain many questions of this type.
" What do you recall (or: what do you like) about this ad?

Open questions : Why? *(continuation)*

◆ ***To probe the response to a closed-end question:***

This is the follow up additional question *"Why?"*.
Explanations concerning a response already given have to be provided in a spontaneous fashion.

◆ ***To get information relating to non-comparable variables:***

Example : Environmental activism, dietary habits….

Open questions : Drawbacks and Advantages

DRAWBACKS

Cost

Complexity

Specificity

ADVANTAGES

Speed

Freedom
Specificity

Comparison between open and closed questions

A classical experiment, quoted by Schuman and Presser (1981), stresses the difficulty of comparing the two types of questionning.

When asked:

"What is the most important problem facing this country [USA] at present?",

16% of Americans mention *crime* and *violence* (open question), whereas the same item asked in a closed question produces **35%** of the same response.

The explanation given by authors is the following:
lack of security is often considered as a local, not a national problem,
so that the item *crime and violence* is not often mentioned spontaneously .

Closing the question indicates that this response is a relevant or *possible* response, resulting in a higher response percentage.

Heuristic value of open-ended questions

In some particular contexts, the absence of a response item list can play a positive role.

It can establish a climate of confidence and communication, and lead to better results when certain subjects are brought up.

This is what is indicated by the works of Sudman and Bradburn (1974) concerning questions having to do with "threats", and of Bradburn *et al.* (1979) concerning questions about alcohol and sexuality.

In international studies, it is important to know whether people interviewed in different countries understand the closed questions in the same way. (case of the follow up :**”Why”**).

As a matter of fact, it is also legitimate to raise this same issue of understanding with respect to regional and generational differences.

Heuristic value of open-ended questions *(continuation)*

The cultural gap between those who have conceived the questionnaire and the interviewees is often hidden by the purely numerical coding of the closed questions.

In a national survey about the attitudes of economically impaired people towards the minimum wage system in France, a classical open question was asked at the end of the interview:

"Would you like to add something about some topics that could be missing in this questionnaire, about the minimum wage system ?"

One answer (among many others of the same vein) was
" We eat potatoes and eggs, despite my diabetes and my cholesterol, because there are cheap."

Another: "Thank you for coming. It proves that you are thinking of me".

[Some respondents are far from the problematic "Attitude towards an institution"]

Empirical Post-Coding of free responses

(Drawbacks of this type of processing)

- ***Coder bias:*** Coder bias is added to interviewer bias, since the coder makes decisions and formulates interpretations, introducing a «personal touch ».

- ***Alteration of form:*** Information is destroyed in its form and often weakened in its content: quality of expression, level of vocabulary, and general interview tonality are lost.

- ***Weakening of content:*** (case of responses that are composed, complex, vague and diversified).

- ***Infrequent responses*** are eliminated a priori.

Example 1: Comments about Spanish wines: Examples of “responses”

The following comments about 443 bottles of wine can be considered as responses to the open-ended question:

"What do you think about this wine?"

Various closed questions (colour, type of grape, region, price, characteristics of the vineyard, vintage, etc.) complement the open question.

Example 1: Comments about Spanish wines: Examples of “responses”

---- I001

Manzana reineta, pomelo maduro, flores blancas. en boca suave y frutoso, con un agradable toque de acidez al final.

---- I003

Expresivo en sus notas florales y frutales, lirio, manzana verde, pera de agua, pétalos blancos. en boca suave, taninos muy sedosos de la fruta, bayas blancas y una acidez perfecta

---- I007

Nariz extremadamente perfumada: flores azules y blancas y cáscara de nuez . limón y frutos secos en boca.

---- I009

Boca muy equilibrada, con destellos de madera sobre un fondo de fruta amarilla madura. Buena persistencia. en nariz, sin embargo, algo insípido y dominado por notas de hierbas y un toque dulce de levaduras.

---- I010

……………………………………

Example 1: Comments about Spanish wines: Examples of “responses”

(English translation)

---- I001

Pippin apple, ripe grapefruit, white flowers. soft and fruity on the palate with a pleasant touch of acidity in the end.

---- I003

Expressive in its floral and fruity notes, lily, green apple, pear, water, white petals. in mouth soft, silky tannins of fruit, white berries, and perfect acidity.

---- I007

Extremely perfumed nose: blue and white flowers nutshell. lemon and nuts in mouth.

---- I009

Mouth very balanced, with flashes of wood on a background of ripe yellow fruit. good persistence. Nose, however, rather bland, dominated by notes of herbs and a hint of yeast sweetness.

---- I010……………………………………

Example 2: Open Questions / Copy-Test

Following a viewing of a television commercial on breakfast cereals (copy-test), several open questions were asked.

One of them is : ***What was the main idea of this commercial?***

In addition a number of closed questions were also asked (socio-demographic characteristics of respondents, purchase intent toward product seen).

Purchase intent , being an important issue will play a major role in the discussions that follow.

Two examples of responses to that open question.

1 - That it has complex carbohydrates in it, it has energy releaser and it tastes good... It showed people eating grape nuts.

2 - It gives you energy in the morning, nothing else.

Example 3: International survey (Tokyo Gas Company)

A survey in three cities (Tokyo, New York, Paris) about dietary habits.

The common open-ended questions were:

"*What dishes do you like and eat often?*
(With a probe: **"*Any other dishes you like and eat often?*").**
" What would be an ideal meal?"

Akuto H.(Ed.) (1992). *International Comparison of Dietary Cultures*,
Nihon Keizai Shimbun, Tokyo.

Akuto H., Lebart L. (1992). Le Repas Idéal. Analyse de Réponses
Libres en Anglais, Français, Japonais. *Les Cahiers de*
l'Analyse des Données, vol XVII, n° 3, Dunod, Paris

Example 3: International survey (*continuation*)

"What dishes do you like and eat often?
“What would be an ideal meal?”
[Four responses (New York)]

---- 1
SPAGHETTI,CHINESE
++++
CAESAR SALAD,LOBSTER TAILS,BAKED POTATO, CHOCOLATE MOUSSE

---- 2
SEAFOOD,GREEN SALAD,CHINESE FOOD
++++
CHAMPAGNE,CAVIAR,GREEN SALAD,GRILLED SEAFOOD

---- 3
CHINESE FOOD
++++
CHINESE FOOD,FRENCH FOOD,VEAL,BREAD
---- 4
PASTA
++++
BEARNAISE BEEF,CHINESE FOOD,ITALIAN FOOD,PASTA

Text Mining and Open-ended Questions in Sample Surveys

Summary / Outline

1) Principles of Data Mining and Text mining: A reminder

2) Open-ended Questions: Why? How?

3) From texts to numerical data

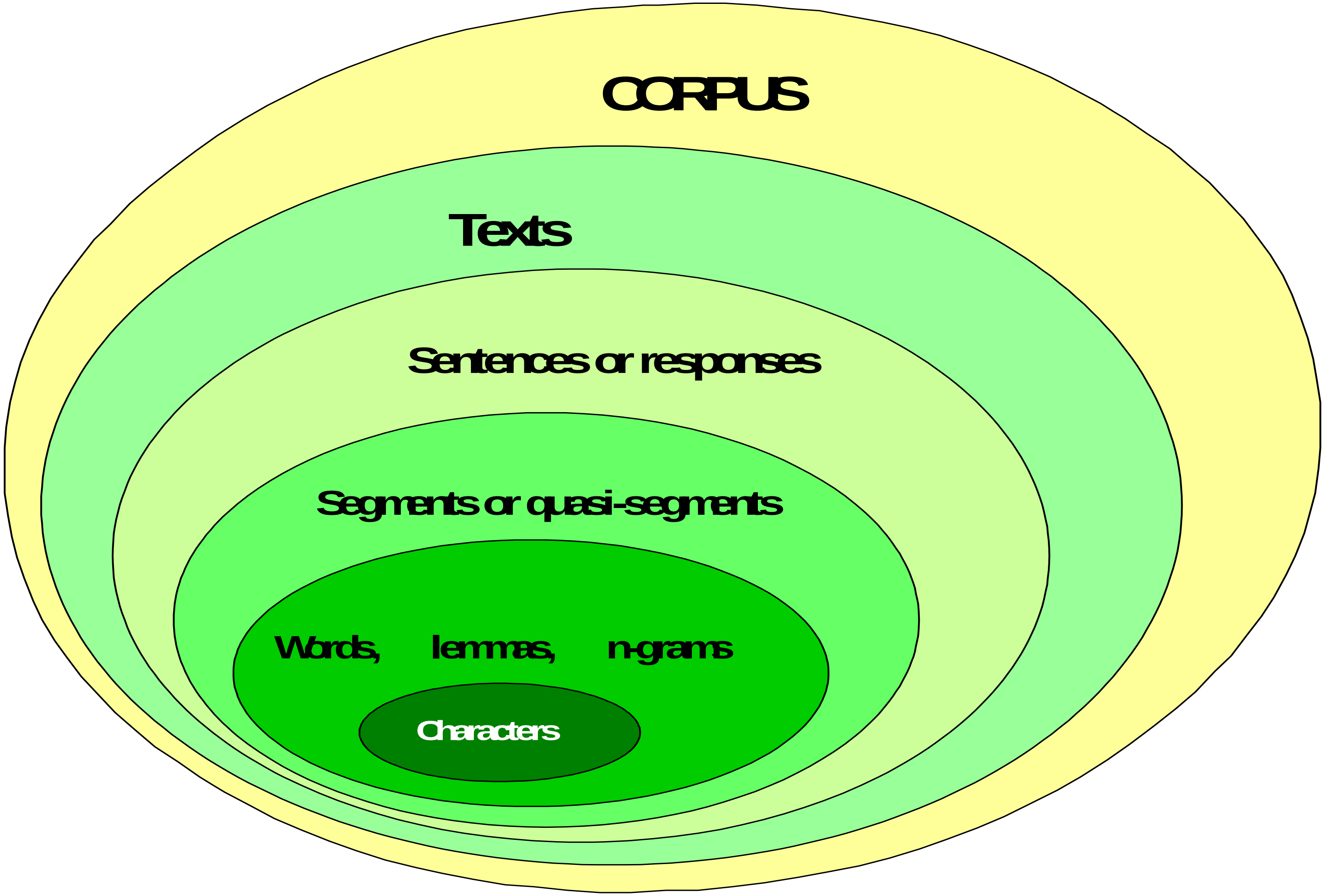
4) Basic statistical tools: Visualization, Characteristic words, Bootstrap.

5) Applications: Open questions, sample surveys, texts

6) About textual data in general

7) Conclusions

Statistical units derived from texts



Example 1: Comments about Spanish wines

Counts for the first phase of numeric coding:

Summary of results
----

total number of responses = **443**
total number of words = **14,061**
number of distinct words = **1394**

Selection of words

When the words appearing at least ***4*** times are selected, **12,404** occurrences (tokens) of these words remain, with **395** distinct words (types).

- **Distribution of words:**

« Zipf law » (a.k.a.: « Pareto law », « Power law »).

3- From texts to numerical data

Example 1: Comments about Spanish wines

Selected statistical units

Words (frequency order)

```
!--------!----------------!---------!
!  num.  ! used words     !  freq.  !
!--------!----------------!---------!
!    101 ! de             !    891  !
!    393 ! y              !    806  !
!    129 ! en             !    694  !
!     46 ! boca           !    433  !
!     87 ! con            !    356  !
!    174 ! fruta          !    334  !
!    378 ! un             !    308  !
!    261 ! nariz          !    246  !
!    259 ! muy            !    237  !
!    215 ! la             !    211  !
!    271 ! notas          !    211  !
!    309 ! que            !    168  !
!    355 ! taninos        !    167  !
!    123 ! el             !    158  !
!    379 ! una            !    152  !
!    232 ! madera         !    140  !
!--------!----------------!---------!
```

3- From texts to numerical data

Example 1: Comments about Spanish wines

Selected statistical units

Words (Alphabetical order)

```
+-------+-------------+--------+
!     1 ! a           !    66  !
!     2 ! abierto     !     9  !
!     3 ! acarameladas !    9  !
!     4 ! accesible   !    14  !
!     5 ! acidez      !    79  !
!     6 ! agradable   !    68  !
!     7 ! agradables  !    17  !
!     8 ! agua        !     6  !
!     9 ! ahora       !     5  !
!    10 ! al          !    27  !
!    11 ! albaricoque !     5  !
!    12 ! algo        !    72  !
!    13 ! alguna      !    20  !
!    14 ! algunas     !     5  !
!    15 ! algún       !    35  !
!    16 ! alta        !     8  !
!    17 ! amable      !     7  !
+-------+-------------+--------+
```

Example 2: "What is the main idea in this commercial"

Words appearing more than 9 times (100 responses)

Number	Word	Frequency	Number	Word	Frequency
1	I	14	25	in	27
2	a	59	26	is	37
3	About	15	27	it	133
4	all	21	28	it's	28
5	and	42	29	long	14
6	are	25	30	morning	9
7	been	12	37	nothing	25
8	carbohydrate	14	32	nutritional	9
9	carbohydrates	33	33	nutritious	12
10	cereal	34	34	nuts	25
11	complex	25	35	of	25
12	crunchy	9	36	people	28
13	eaten	10	37	showed	11
14	eating	19	38	taste	11
15	energy	33	39	that	80
16	for	57	40	that's	13

Example 2: "What is the main idea in this commercial"

```
 SEGM FREQ  LENGTH "TEXT of SEGMENT"
-------------------------------------
-----------------------------------------a
   1    8    3 a long time
-----------------------------------------are
   2    6    4 are good for you
-----------------------------------------carbohydrates
   3    5    3 carbohydrates in it
-----------------------------------------complex
   4   15    2 complex carbohydrates
-----------------------------------------for
   5   37    2 for you
-----------------------------------------give
   6    7    3 give you energy
-----------------------------------------gives
   7   11    2 gives you
   8    9    3 gives you energy
-----------------------------------------good
   9   24    2 good for
  10   22    3 good for you
-----------------------------------------grape
  11   25    2 grape nuts
-----------------------------------------have
  12    6    3 have been eating
-----------------------------------------healthy
  13    6    3 healthy for you
-----------------------------------------is
  14    9    4 is good for you
-----------------------------------------it
  15   26    2 it has
  16   19    2 it is
  17   14    2 it was
  18    8    3 it gives you
  19    8    3 it has a
  20    6    3 it has complex
  21    5    3 it is good
  22    6    4 it gives you energy
-----------------------------------------people
```

Examples of “segments”

Example 3: An international survey (Tokyo Gas Company)

```
!-------------------------------------!
!      words (frequency order)        !
!--------!---------------------!------!
!  num.  !      used words     ! freq.!
!--------!---------------------!------!
!     12 ! CHICKEN             !  254 !
!     73 ! STEAK               !  101 !
!     49 ! PASTA               !   95 !
!     22 ! FISH                !   87 !
!     60 ! SALAD               !   85 !
!      1 ! AND                 !   85 !
!     23 ! FOOD                !   82 !
!     52 ! PIZZA               !   62 !
!     79 ! VEGETABLES          !   57 !
!      4 ! BEEF                !   56 !
!     71 ! SPAGHETTI           !   55 !
!     13 ! CHINESE             !   54 !
!     80 ! WITH                !   48 !
!     59 ! ROAST               !   47 !
!     58 ! RICE                !   45 !
!     67 ! SHRIMP              !   45 !
!     43 ! MACARONI            !   42 !
!     56 ! POTATOES            !   39 !
!     35 ! HAMBURGERS          !   36 !
!     75 ! TUNA                !   35 !
!     26 ! FRIED               !   33 !
!     77 ! VEAL                !   33 !
!     38 ! ITALIAN             !   31 !
!      2 ! BAKED               !   29 !
!     48 ! PARMESAN            !   29 !
!     55 ! POTATO              !   27 !
!     46 ! MEATBALLS           !   25 !
!      3 ! BEANS               !   24 !
!     45 ! MEAT                !   24 !
!     76 ! TURKEY              !   24 !
!     14 ! CHOPS               !   23 !
!     34 ! HAMBURGER           !   22 !
!-------------------------------------!
```

City of New York

The common open-ended question : ***"What dishes do you like and eat often?”***

(With a probe: ***"Any other dishes you like and eat often?"***).

634 individuals.
(**6511** occurrences of **638** distinct words).

The processing takes into account the **83** words appearing at least **12** times.

Text Mining and Open-ended Questions
in Sample Surveys

Summary / Outline

1) Principles of Data Mining and Text mining: A reminder

2) Open-ended Questions: Why? How?

3) From texts to numerical data

4) Basic statistical tools: Visualization, Characteristic words, Bootstrap.

5) Applications: Open questions, sample surveys, texts

6) About textual data in general

7) Conclusions

Main techniques for performing data reductions:

- ***Principal axes methods*****,** largely based upon linear algebra, produce graphical representations on which the geometric proximities among row-points and among column-points translate statistical associations among rows and among columns. Correspondence analysis belongs to this family of methods. Assessment via Bootstrap techniques.

- ***Clustering or classification methods*** that create groupings of rows or of columns into clusters (or into families of hierarchical clusters) including the SOM (Self Organizing Maps, or Kohonen maps).

These two families of methods can be used on the same data matrix and they complement one another very effectively.

- ***Selection of characteristic units and responses (or: sentences)***

Characteristic units (words, segments, lemmas)Selecting « Modal responses »

Visualization through principal coordinates

Techniques such as **Principal Component Analysis** or **Correspondence Analysis** could be considered as variant of **Singular Value Decomposition**.

These techniques will be used as mere instruments of observation of the multidimensional reality.

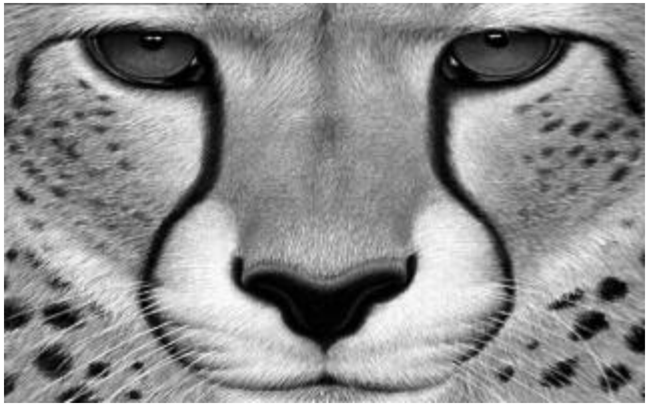
(such as a microscope or a telescope).

Two examples will illustrate these techniques.

- An example of **image compression**.

- An example of **graph description**.

Image “Cheetah” (*Data Compression*, Mark Nelson) and table (200 x 320) containing levels of grey.



95	88	88	87	95	88	95	95	95	106	95	78	65	71	78	77	77	etc.
143	144	151	151	153	170	183	181	162	140	116	128	133	144	159	166	170	
153	151	162	166	162	151	126	117	128	143	147	175	181	170	166	132	116	
143	144	133	130	143	153	159	175	192	201	188	162	135	116	101	106	118	
123	112	116	130	143	147	162	183	166	135	123	120	116	116	129	140	159	
133	151	162	166	170	188	166	128	116	132	140	126	143	151	144	155	176	
160	168	166	159	135	101	93	98	120	128	126	147	154	158	176	181	181	
154	155	153	144	126	106	118	133	136	153	159	153	162	162	154	143	128	
159	153	147	159	150	154	155	153	158	170	159	147	130	136	140	150	150	
151	144	147	176	188	170	166	183	170	166	153	130	132	154	162	120	135	
155	181	183	162	144	147	147	144	126	120	123	129	130	112	101	135	150	
166	147	129	123	133	144	133	117	109	118	132	112	109	120	136	120	136	
136	130	136	147	147	140	136	144	140	132	129	151	153	140	128	153	147	
130	133	140	124	136	152	166	147	144	151	159	140	123	130	123	109	112	
126	120	143	145	162	153	155	175	154	144	136	130	120	112	123	123	144	
144	159	155	155	162	166	158	147	140	147	126	123	132	135	136	144	147	
136	143	162	175	136	110	112	135	120	118	126	151	150	130	129	133	147	
133	151	143	106	85	93	128	136	140	140	144	143	126	117	116	129	124	

……………………………. . .etc.

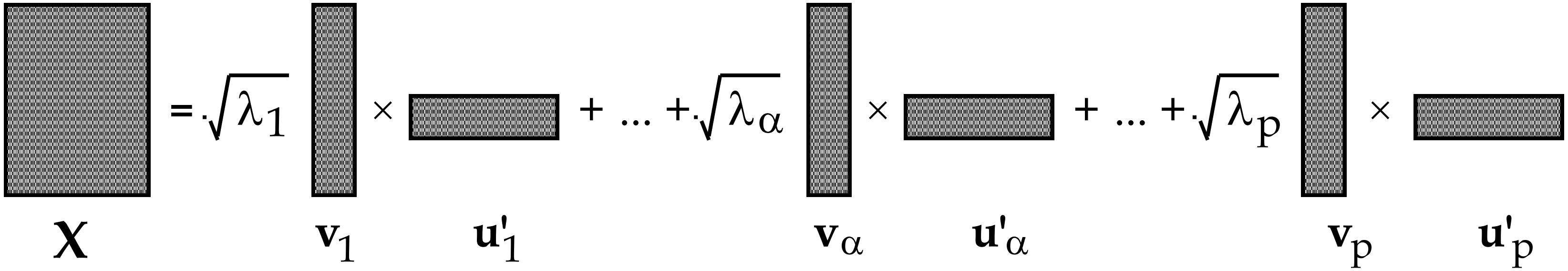
4) Basic statistical tools: Visualization, Characteristic words, Bootstrap.

Harold Hotelling, 1895-1973

Develops PCA as a technique of mathematical statistics. Recommends the use of the iterated power algorithm for computing eigenvalues. Proposes Canonical Analysis (1936).

- **Hotelling H.** (1933) - Analysis of a complex of statistical variables into principal components. *J. Educ. Psy.* 24, p 417-441, p 498-520.

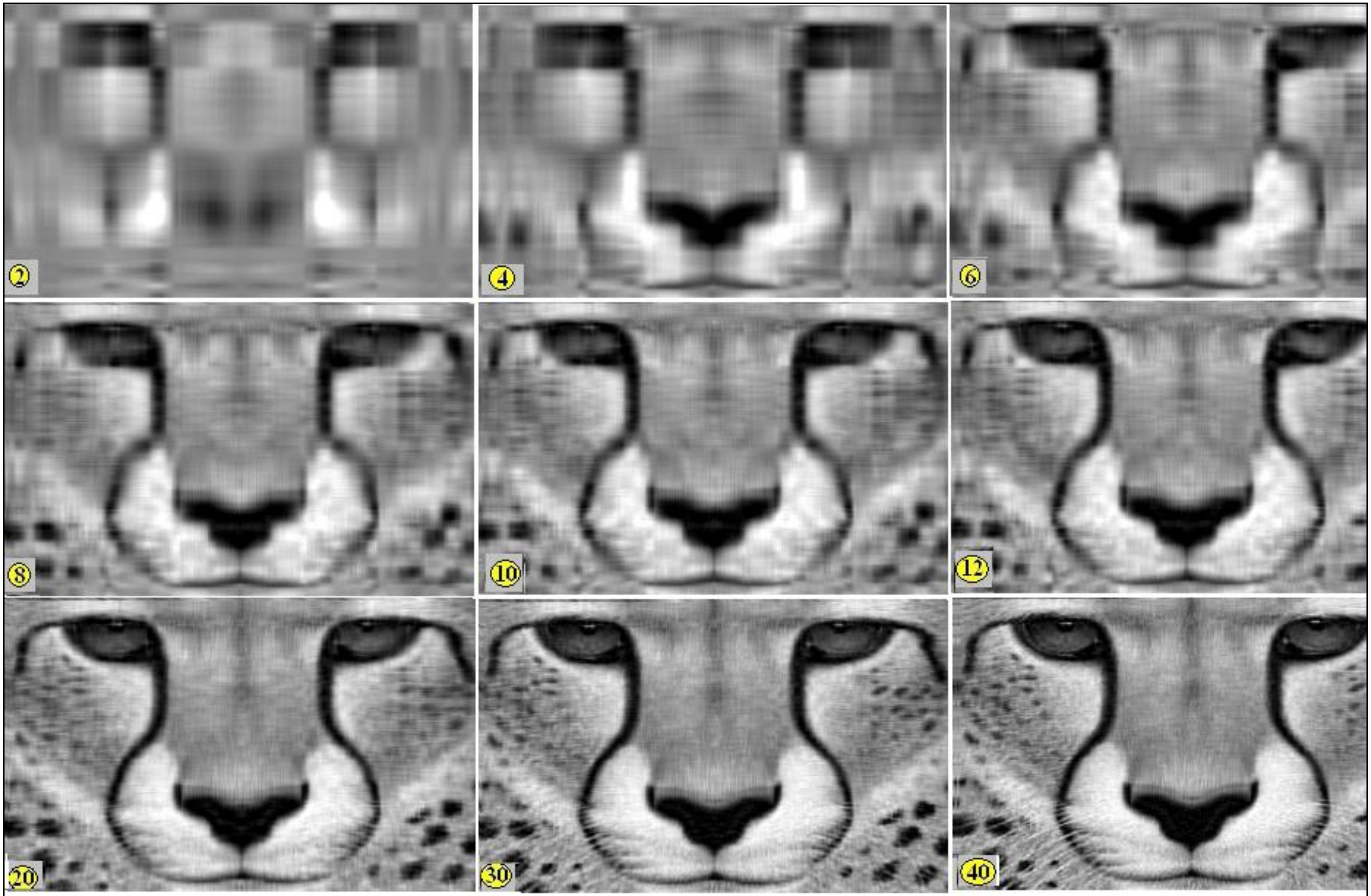
With Hotelling and Eckart & Young, principal axes techniques are connected to both *multivariate analysis* and *modern linear algebra*.



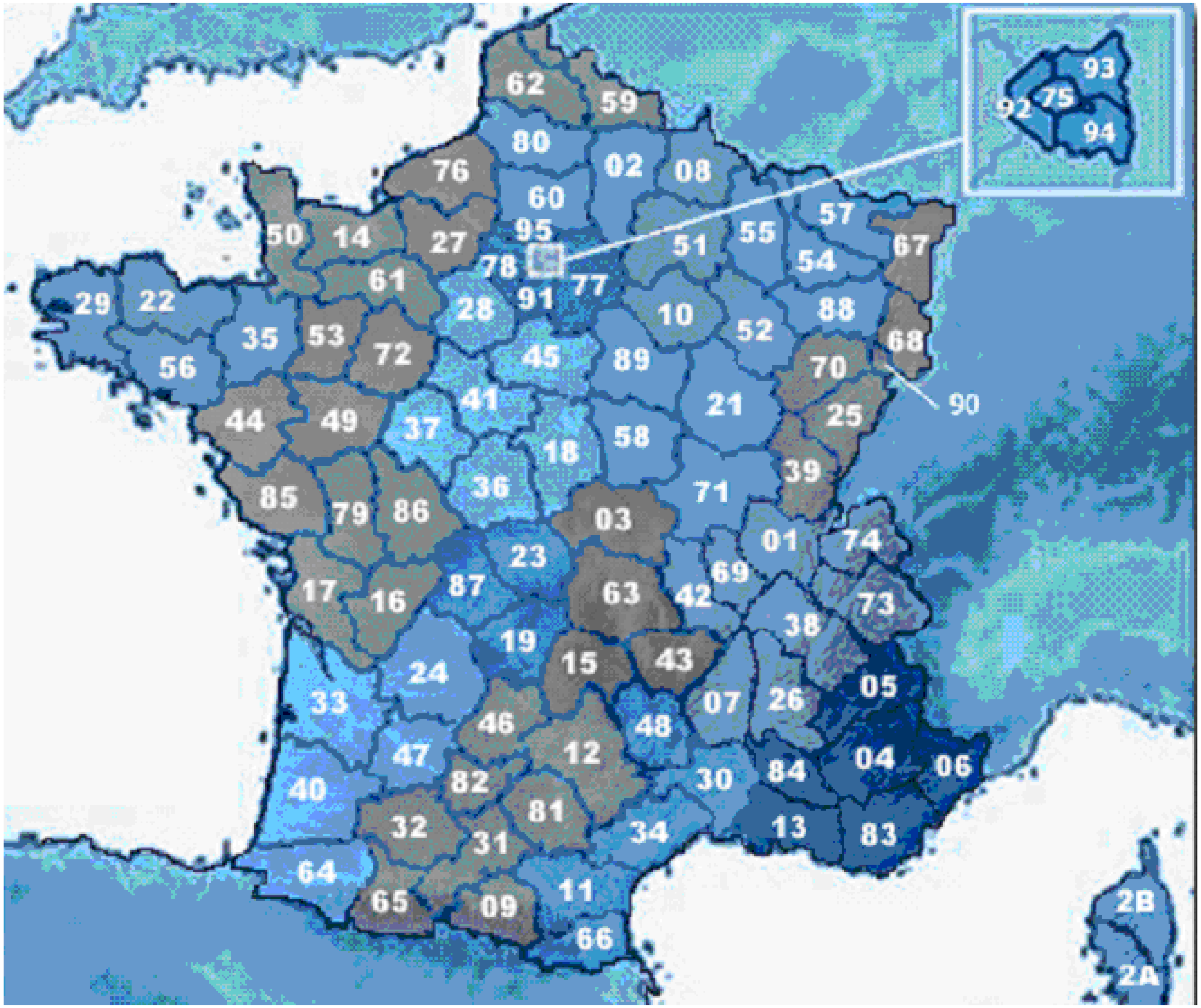
$$\mathbf{X} = \sqrt{\lambda_1}\,\mathbf{v}_1 \times \mathbf{u}'_1 + \dots + \sqrt{\lambda_\alpha}\,\mathbf{v}_\alpha \times \mathbf{u}'_\alpha + \dots + \sqrt{\lambda_p}\,\mathbf{v}_p \times \mathbf{u}'_p$$

- **Eckart C., Young G.** (1936) - The approximation of one matrix by another of lower rank. *Psychometrika*, l, p 211-218.

Reconstitution of the Cheetah with 2, 4, 6, 8, 10, 12, 20, 30, 40 principal axes



A pedagogical example: Description of « Textual Graphs »



Each area “answers” to the fictitious “open-question” :
Which are your neighbouring areas?

**** *Ain*
Ain Isere Jura Rhone Hte_Saone Savoie Hte_Savoie

**** *Aisne*
Aisne Ardennes Marne Nord Oise Seine_Marne Somme

**** *Allier*
Allier Cher Creuse Loire Nievre Puy_de_Dome Hte_Saone

**** *Alpes_Prov*
Alpes_Prov Alpes_Hautes Alpes_Marit Drome Var Vaucluse

**** *Alpes_Hautes*
Alpes_Hautes Alpes_Prov Drome Isere Savoie

**** *Alpes_Marit*
Alpes_Marit Alpes_Prov Var

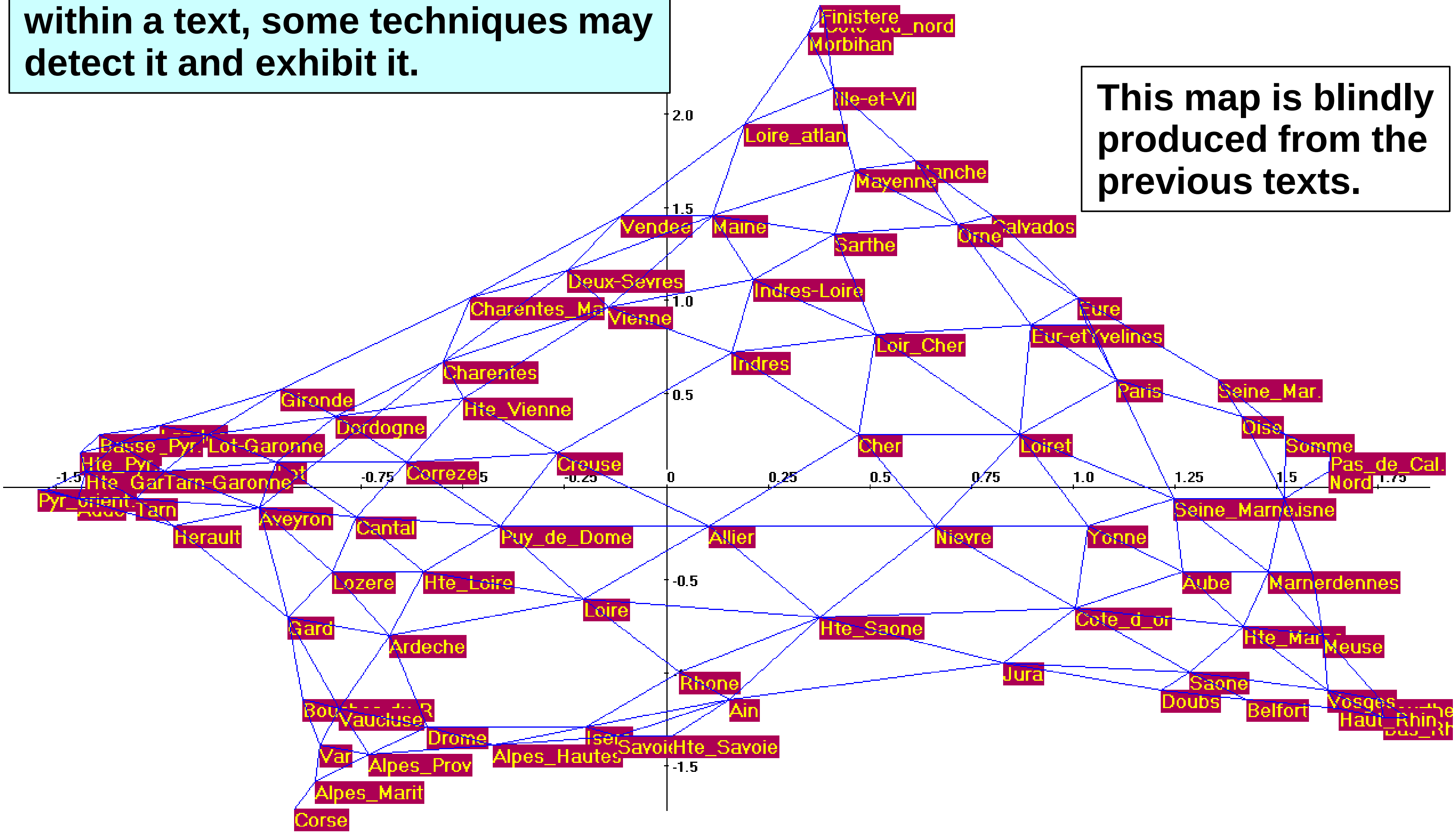
**** *Ardeche*
Ardeche Drome Gard Loire Hte_Loire Lozere

**** *Ardennes*
Ardennes Aisne Marne Meuse

…………………………

4) Basic statistical tools: Visualization, Characteristic words, Bootstrap.

The idea: When a pattern exists within a text, some techniques may detect it and exhibit it.



This map is blindly produced from the previous texts.

Characteristic elements (words, lemmas, segments)

A corpus contains several parts (categories of respondents).

Notations:

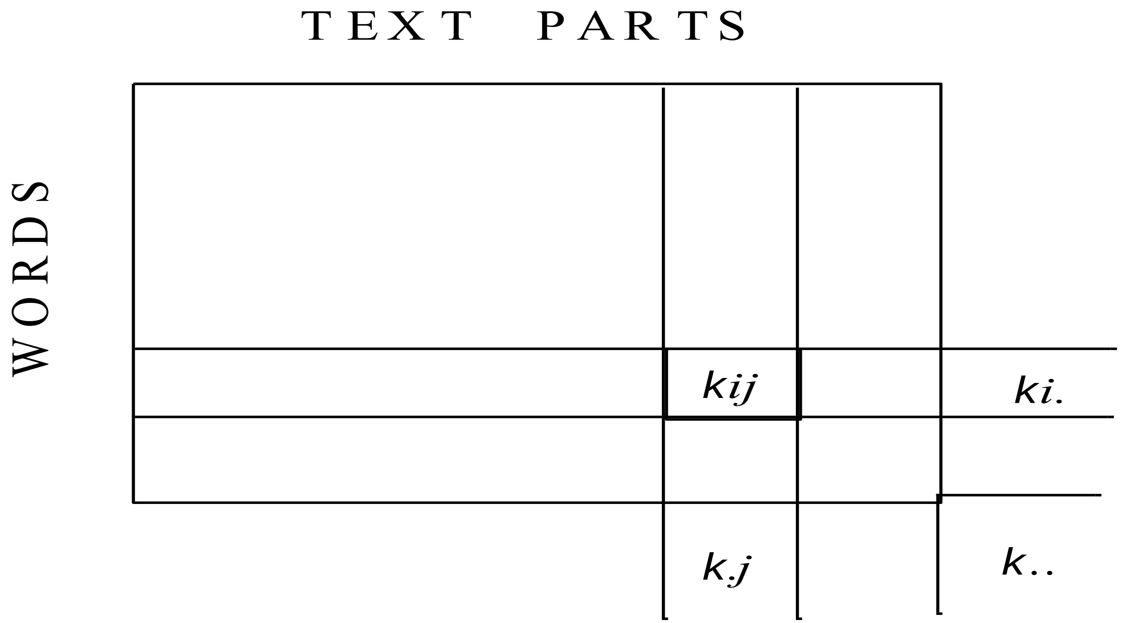
k_{ij}	-sub-frequency of word i in the part j of the corpus;
$k_{i.}$	-frequency of word i in the whole corpus;
$k_{.j}$	-frequency (size) of part j;
$k_{..}$	-size of the corpus (or, simply, k).

We are interested in the statistical significance of sub-frequency k_{ij} .

Is the word i abnormally frequent in part j ? Is it abnormally rare?

The comparison between the relative frequency of word i in part j and the relative frequency of word i in the entire corpus leads to a classical statistical test using either the hypergeometric distribution or its normal approximation.

The 4 parameters for computing characteristic elements



$k_{..}$	size of corpus
$k_{i.}$	frequency of word in corpus
k_{ij}	frequency of word in text part
$k_{.j}$	size of text part

Resampling techniques:
Bootstrap, opportunity of the method

- In order to compute estimates precision, many reasons lead to the Bootstrap method :
  - highly complex computation in the analytical approach
  - to get free from beforehand assumptions, no assumption about the underlying distributions
  - possibility to master every statistical computation for each sample replication

Reminder about the bootstrap

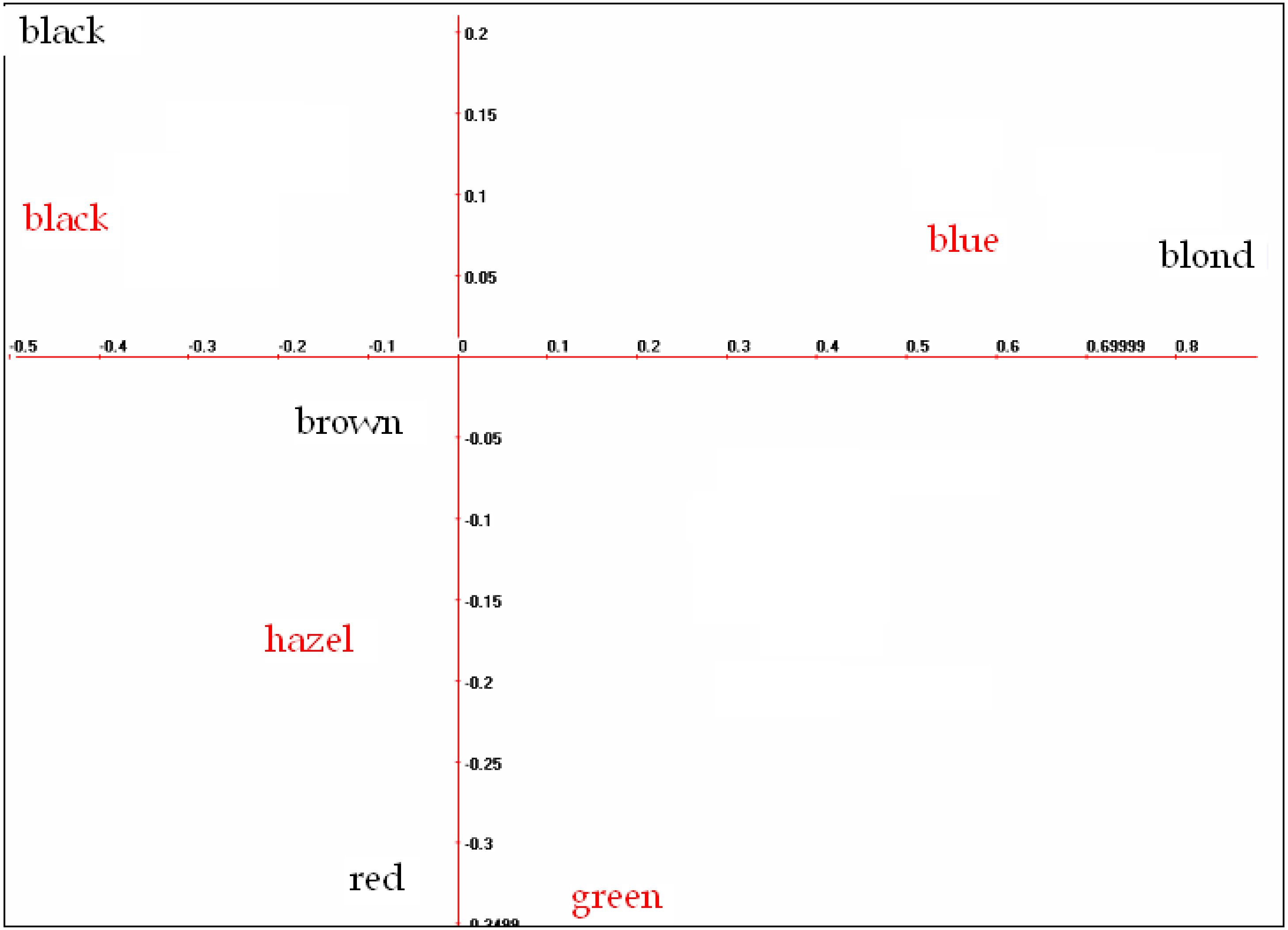
Contingency table,

592 women: Hair and eyes colour.

Eye colour	**Hair colour**				
	black	brown	red	blond	Total
black	*68*	*119*	*26*	*7*	*220*
hazel	*15*	*54*	*14*	*10*	*93*
green	*5*	*29*	*14*	*16*	*64*
blue	*20*	*84*	*17*	*94*	*215*
Total	*108*	*286*	*71*	*127*	*592*

Source : Snee (1974), Cohen(1980)

Principal plane (1, 2) ***Snee data. Hair - Eye***



Reminder about the bootstrap

Associations between eye and hair colour Example of replicated tables

Original

			Hair colour		
		Black	*Brown*	*red*	*blonde*
eye colour	black	*68*	*119*	*26*	*7*
	hazel	*15*	*54*	*14*	*10*
	green	*5*	*29*	*14*	*16*
	blue	*20*	*84*	*17*	*94*

Replicate 1

eye colour	black	*79*	*120*	*23*	*9*
	hazel	*14*	*60*	*15*	*12*
	green	*3*	*29*	*16*	*9*
	blue	*21*	*82*	*20*	*110*

Replicate 2

eye colour	black	*72*	*111*	*32*	*7*
	hazel	*14*	*47*	*13*	*14*
	green	*5*	*30*	*15*	*19*
	blue	*20*	*89*	*16*	*98*

Principle of partial bootstrap

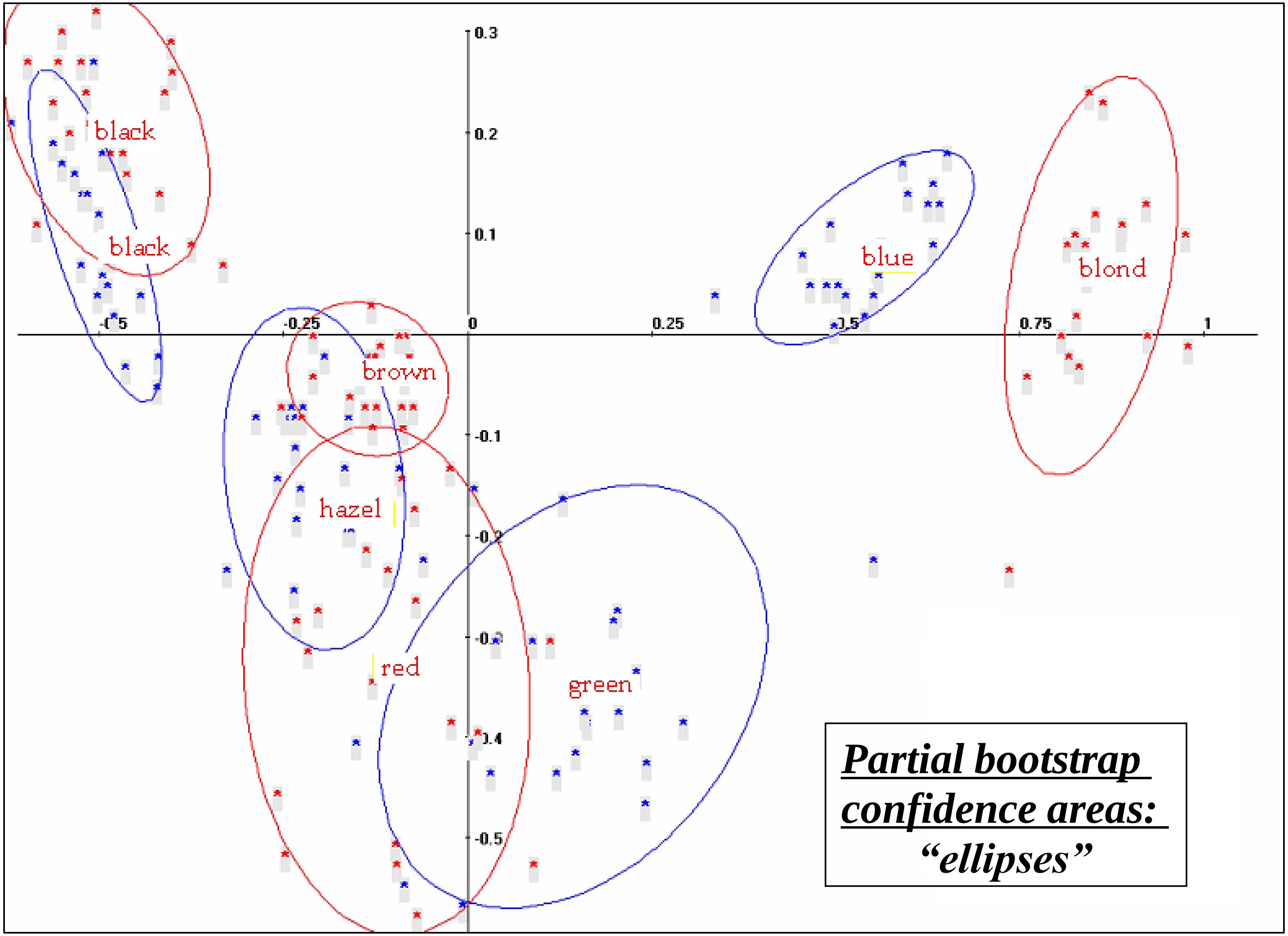
The partial bootstrap, makes use of simple *a posteriori* projections of replicated elements on the original reference subspace provided by the eigen-decomposition of the observed covariance matrix.

From a descriptive standpoint, this initial subspace is better than any subspace undergoing a perturbation by a random noise. In fact, this subspace is the expectation of all the replicated subspaces having undergone perturbations (however, the original eigenvalues are not the expectations of the replicated values).

The plane spanned by the first two axes, for instance, provides an optimal point of view on the data set.

4) Basic statistical tools: Visualization, Characteristic words, Bootstrap.

Principal plane (1, 2) ***Snee data. Hair - Eye***



Partial bootstrap confidence areas:

"ellipses"

Total bootstrap...

Total bootstrap type 1

Total bootstrap type 2

Total bootstrap type 3

Total bootstrap total type 1

Total Bootstrap type 1 (very conservative) : simple change (when necessary) of signs of the axes found to be homologous (merely to remedy the arbitrarity of the signs of the axes). The values of a simple scalar product between homologous original and replicated axes allow for this elementary transformation.

This type of bootstrap ignores the possible interchanges and rotations of axes.
It allows for the validation of stable and robust structures.
Each replication is supposed to produce the original axes with the same ranks (order of the eigenvalues).

Total bootstrap type 2

Total Bootstrap type 2 (rather conservative) : correction for possible interversions of axes. Replicated axes are sequentially assigned to the original axes with which the correlation (in fact its absolute value) is maximum. Then, alteration of the signs of axes, if needed, as previously.

Total bootstrap type 2 is ideally devoted to the validation of axes considered as latent variables, without paying attention to the order of the eigenvalues.

Total bootstrap type 3

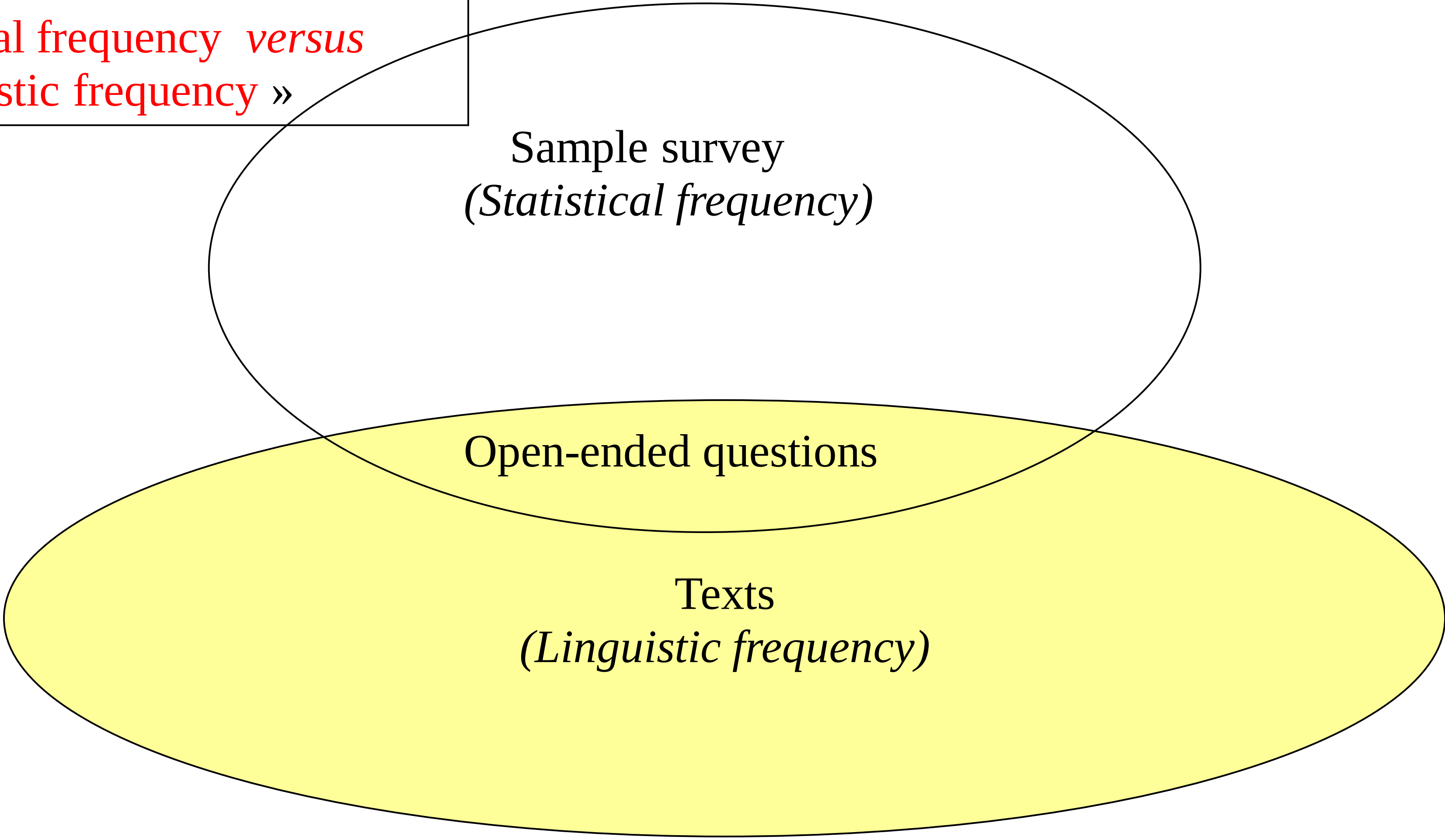
Total Bootstrap type 3 (could be lenient if the procrustean rotation is done in a space spanned by many axes) : a procrustean rotation (see: Gower and Dijksterhuis, 2004) aims at superimposing as much as possible original and replicated axes.Total bootstrap type 3 allows for the validtion of a whole subspace.

If, for instance, the subspace spanned by the first four replicated axes can coincide with the original four-dimensional subspace, one could find a rotation that can put into coincidence the homologous axes.

The situation is then very similar to that of partial bootstrap.

Specific (or: hierarchical) bootstrap

Textual data
Statistical frequency *versus*
« linguistic frequency »



Sample survey
(Statistical frequency)

Open-ended questions

Texts
(Linguistic frequency)

Text Mining and Open-ended Questions in Sample Surveys

Summary / Outline

1) Principles of Data Mining and Text mining: A reminder

2) Open-ended Questions: Why? How?

3) From texts to numerical data

4) Basic statistical tools: Visualization, Characteristic words, Bootstrap.

5) Applications: Open questions, sample surveys, texts

6) About textual data in general

7) Conclusions

Example 1: Comments about wines

The forthcoming diapositives show the principal plane produced by a correspondence analysis of lexical contingency table.

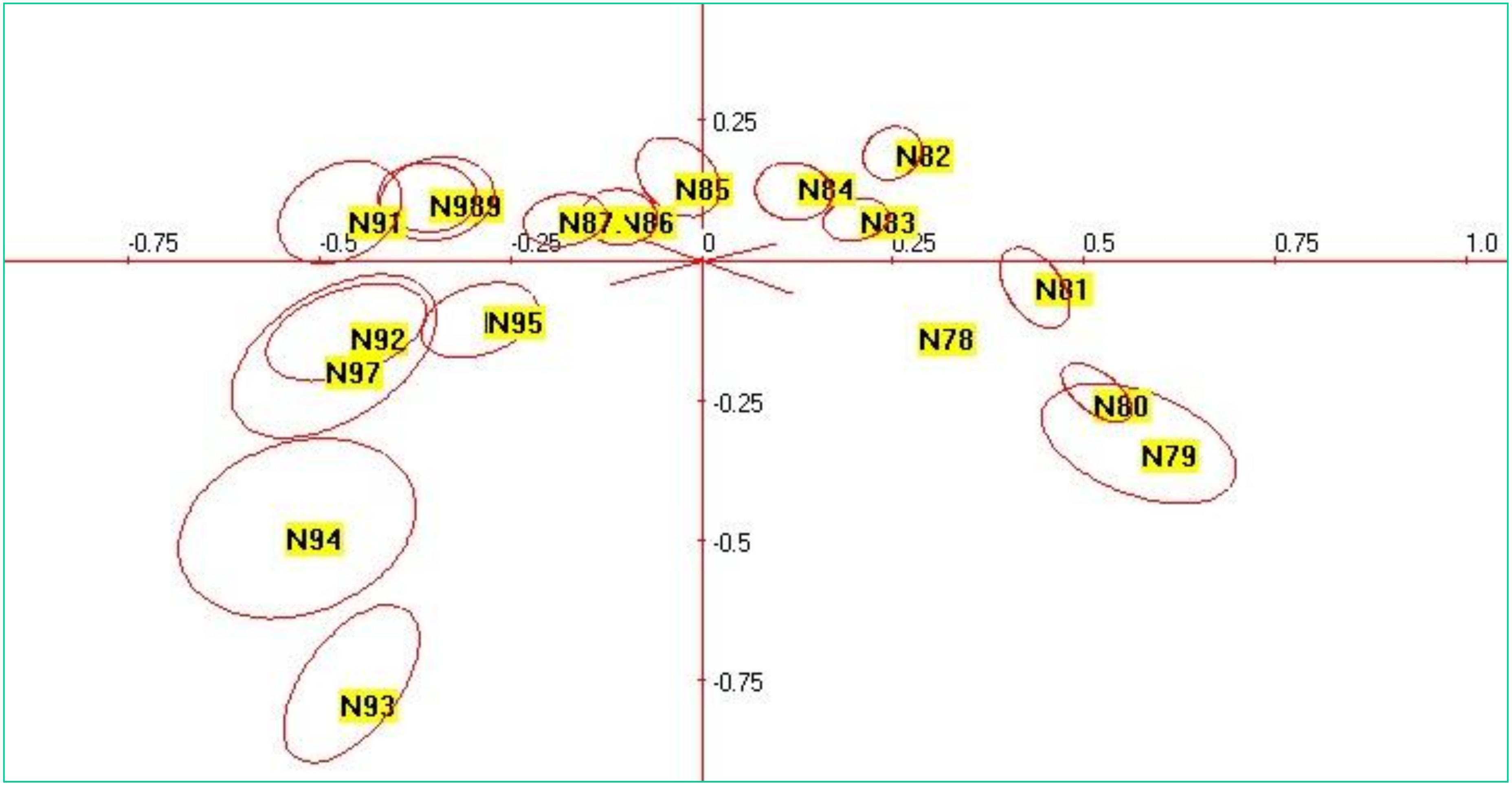
Proximity between 2 category-points (columns) means similarity of lexical profiles of the 2 categories.

Proximity between 2 word-points (rows) means similarity of lexical profiles of these words.

Example 1: Comments about wines

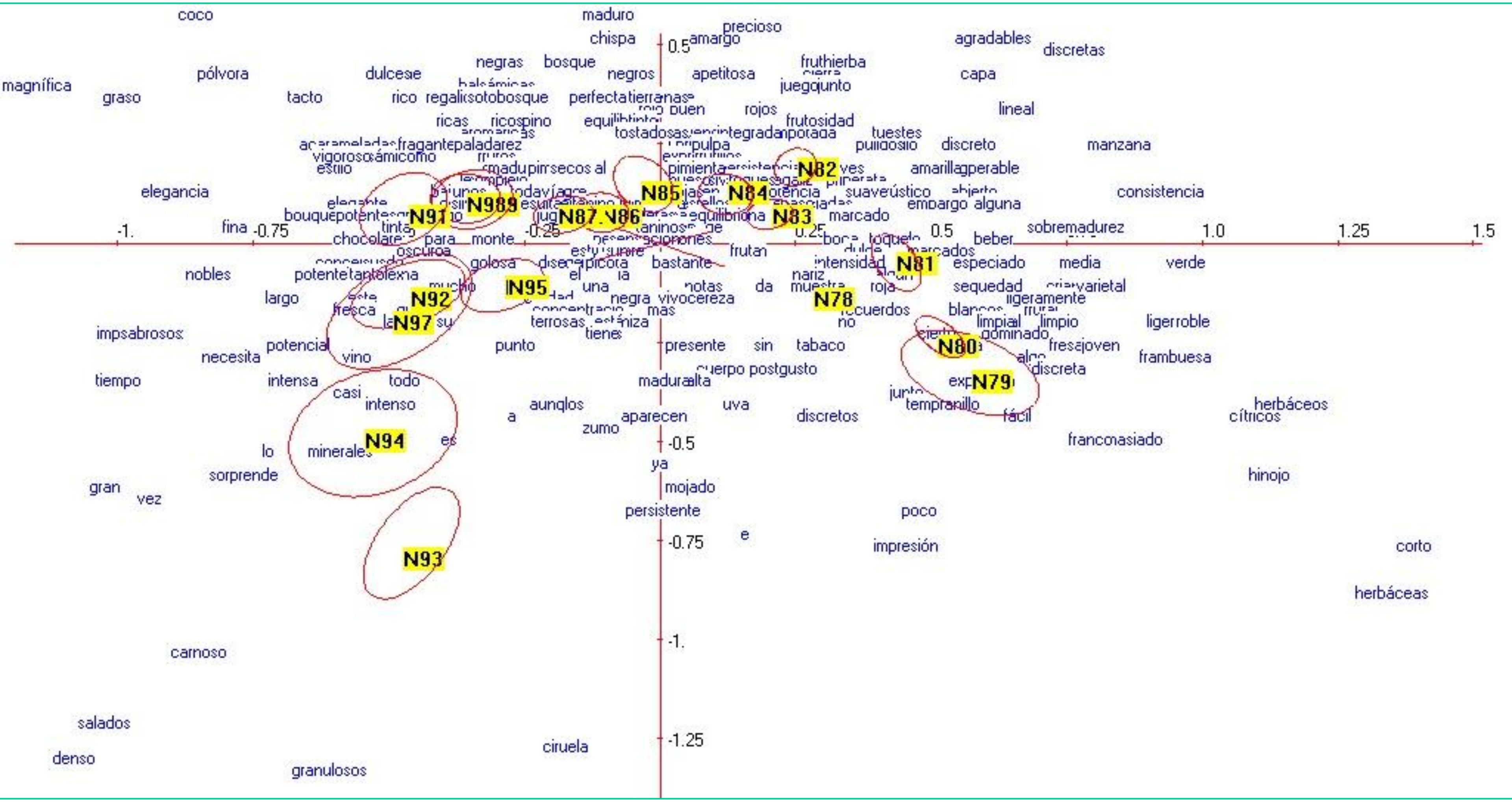
Principal plane of the CA of the contingency table crossing 395 words and 19 score groups (N79 -> N97).

Partial bootstrap confidence ellipIses.



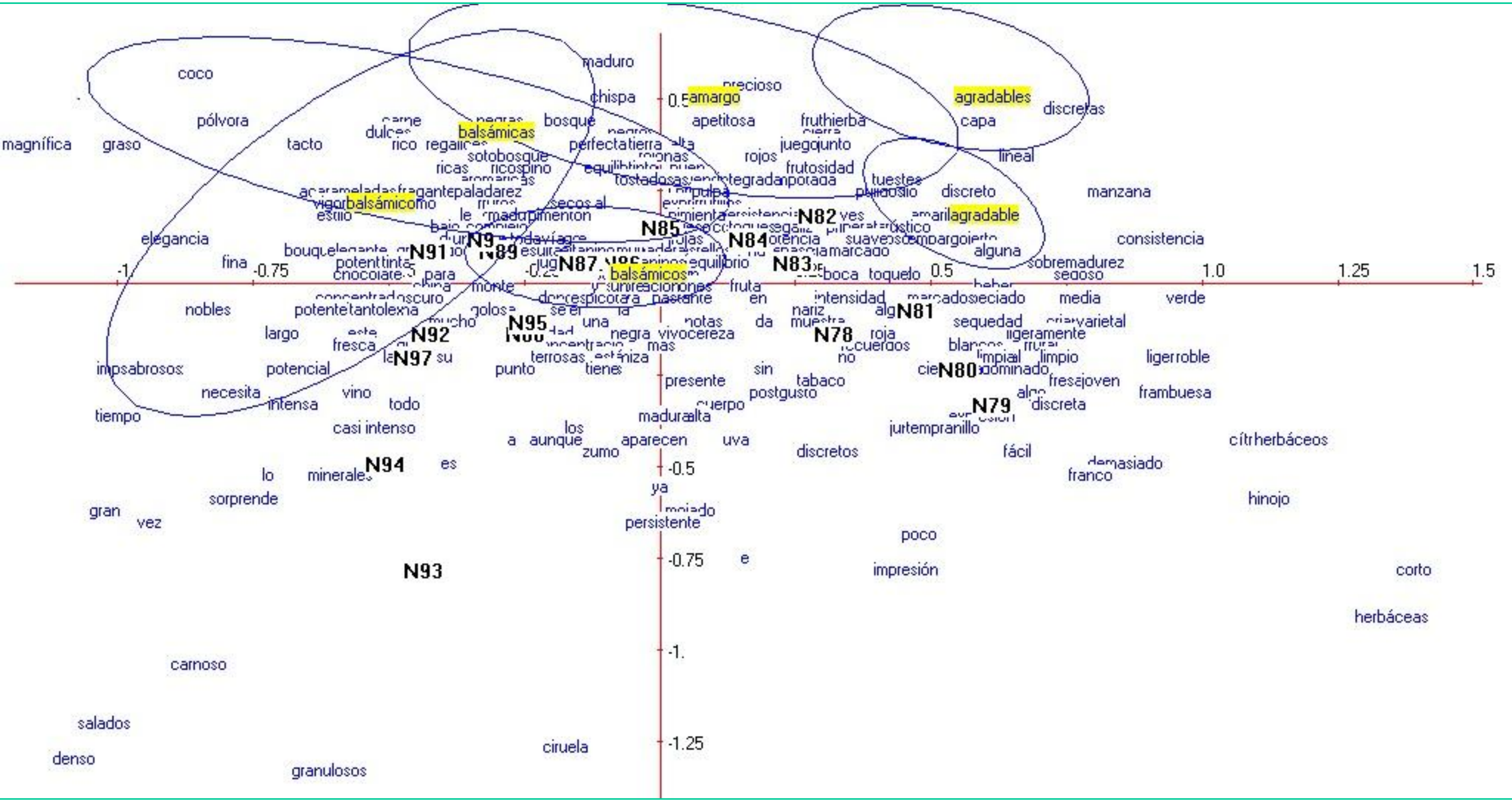
Example 1: Comments about wine

Same first plane with the 395 words.



Example 1: Comments about wine

Same first plane with the 395 words and some confidence ellipses for words.



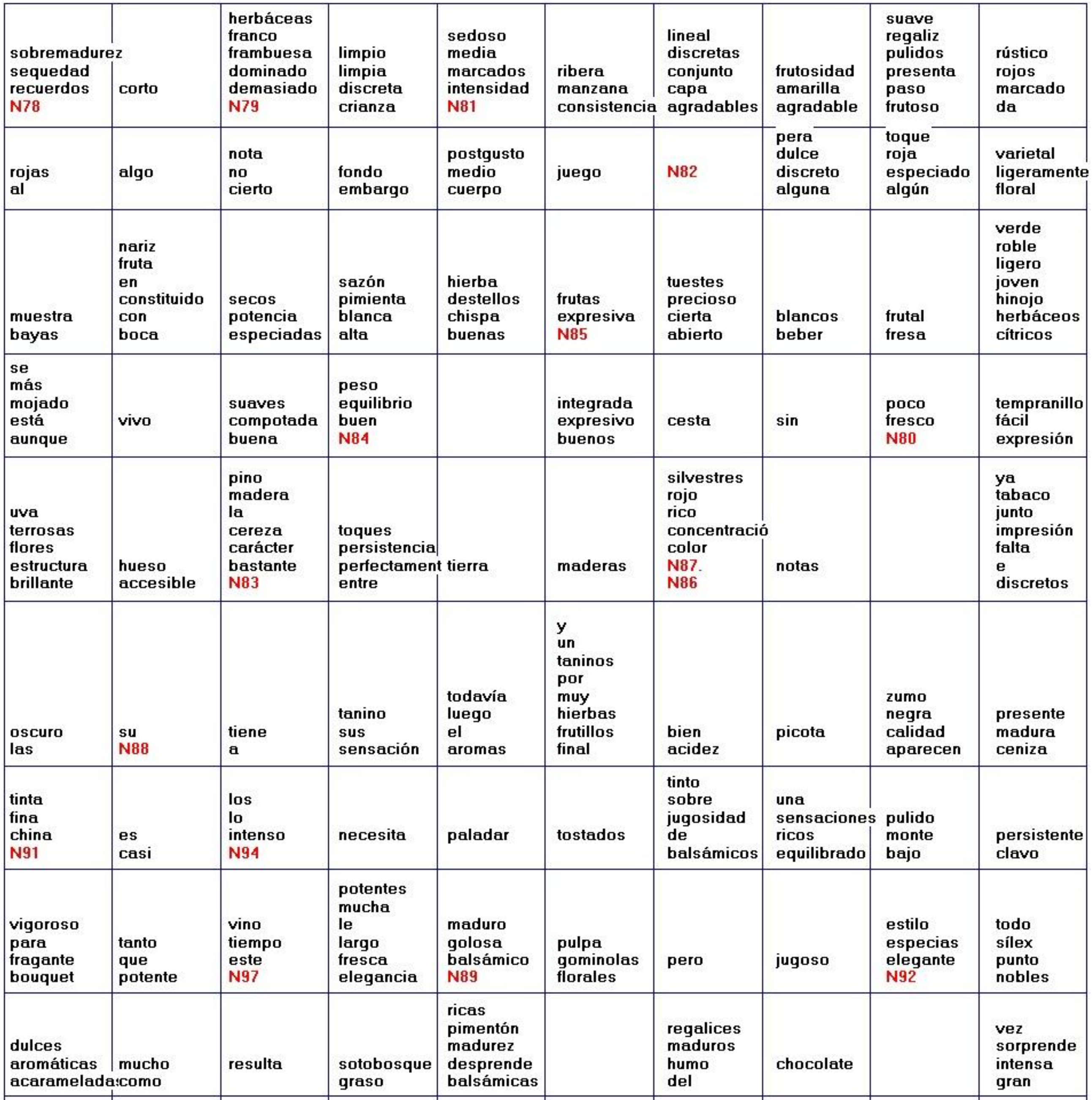
5) Applications: Open questions, sample surveys, texts

Example 1: Comments about wine

S.O.M.

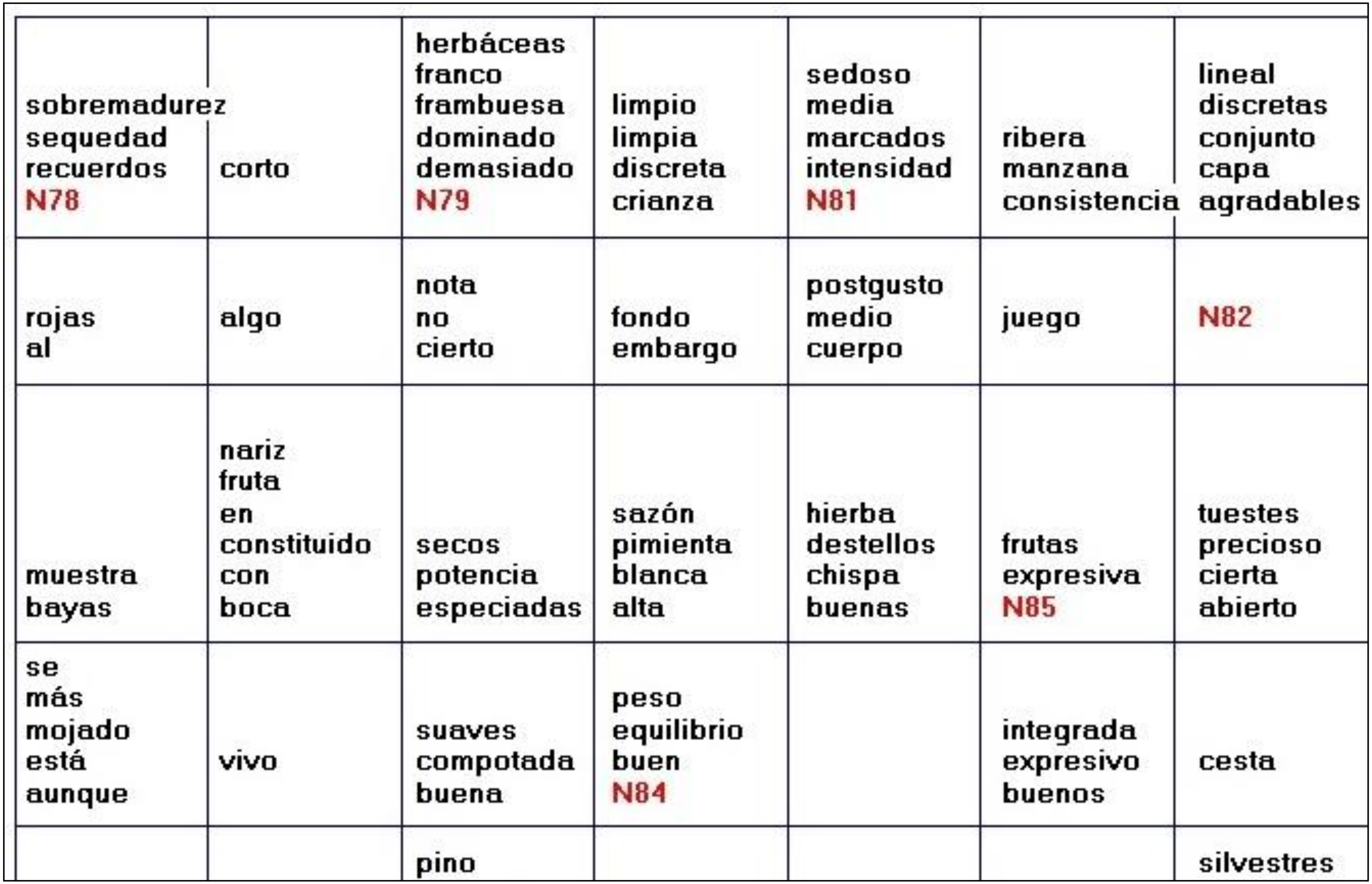
Self Organizing Map (Kohonen Map)

395 words, 19 categories



sobremadurez sequedad recuerdos N78	corto	herbáceas franco frambuesa dominado demasiado N79	limpio limpia discreta crianza	sedoso media marcados intensidad N81	ribera manzana consistencia	lineal discretas conjunto capa agradables	frutosidad amarilla agradable	suave regaliz pulidos presenta paso frutoso	rústico rojos marcado da
rojas al	algo	nota no cierto	fondo embargo	postgusto medio cuerpo	juego	N82	pera dulce discreto alguna	toque roja especiado algún	varietal ligeramente floral
muestra bayas	nariz fruta en constituido con boca	secos potencia especiadas	sazón pimienta blanca alta	hierba destellos chispa buenas	frutas expresiva N85	tuestes precioso cierta abierto	blancos beber	frutal fresa	verde roble ligero joven hinojo herbáceos cítricos
se más mojado está aunque	vivo	suaves compotada buena	peso equilibrio buen N84		integrada expresivo buenos	cesta	sin	poco fresco N80	tempranillo fácil expresión
uva terrosas flores estructura brillante	hueso accesible	pino madera la cereza carácter bastante N83	toques persistencia perfectament entre	tierra	maderas	silvestres rojo rico concentració color N87. N86	notas		ya tabaco junto impresión falta e discretos
oscuro las	su N88	tiene a	tanino sus sensación	todavía luego el aromas	y un taninos por muy hierbas frutillos final	bien acidez	picota	zumo negra calidad aparecen	presente madura ceniza
tinta fina china N91	es casi	los lo intenso N94	necesita	paladar	tostados	tinto sobre jugosidad de balsámicos	una sensaciones ricos equilibrado	pulido monte bajo	persistente clavo
vigoroso para fragante bouquet	tanto que potente	vino tiempo este N97	potentes mucha le largo fresca elegancia	maduro golosa balsámico N89	pulpa gominolas florales	pero	jugoso	estilo especias elegante N92	todo sílex punto nobles
dulces aromáticas acaramelada	mucho como	resulta	sotobosque graso	ricas pimentón madurez desprende balsámicas		regalices maduros humo del	chocolate		vez sorprende intensa gran

Example 1: Comments about wines (Zoom on the S.O.M.)



sobremadurez sequedad recuerdos N78	corto	herbáceas franco frambuesa dominado demasiado N79	limpio limpia discreta crianza	sedoso media marcados intensidad N81	ribera manzana consistencia	lineal discretas conjunto capa agradables
rojas al	algo	nota no cierto	fondo embargo	postgusto medio cuerpo	juego	N82
muestra bayas	nariz fruta en constituido con boca	secos potencia especiadas	sazón pimienta blanca alta	hierba destellos chispa buenas	frutas expresiva N85	tuestes precioso cierta abierto
se más mojado está aunque	vivo	suaves compotada buena	peso equilibrio buen N84		integrada expresivo buenos	cesta
		pino				silvestres

Example 1 («Wine » question) Characteristic words, score = 80

```
    words        %W     %glob         Fr.W      Fr.glob     TestValue    Prob.

text number     3        score = 80
----------------
 1 typical     .56     .11         7.        11.        3.803     .000
 2 light      1.13     .38        14.        40.        3.664     .000
 3 short       .64     .16         8.        17.        3.385     .000
 4 mouth      3.94    2.45        49.       256.        3.306     .000
 5 citrus      .80     .25        10.        26.        3.299     .000
 6 herbal      .40     .10         5.        10.        2.690     .004
 7 notes      2.82    1.81        35.       189.        2.576     .005
 8 discreet    .72     .28         9.        29.        2.570     .005
------------------------------
 8 and        6.36    7.82        79.       816.       -2.029     .021
 7 that        .16     .64         2.        67.       -2.323     .010
 6 fine        .00     .35         0.        37.       -2.362     .009
 5 wine        .1      .67         2.        70.       -2.435     .007
 4 long        .0      .41         0.        43.       -2.633     .004
 3 elegant     .00     .46         0.        48.       -2.842     .002
 2 good        .56    1.49         7.       156.       -3.051     .001
 1 powerful    .00     .54         0.        56.       -3.164     .001
---------------------------------------------------------------
```

Example 1 («Wine » question) Characteristic (or modal) responses

```
---------------------------------------------------------------------------
 text number    3       N80
 ----------------

  1.35 -  1   nice fruity nose. in the mouth the tannins are somewhat hard fruit.

  1.31 -  2   red fruit, some earthy and herbal notes. light on the palate, timidly
              fruity.

  1.19 -  3   nose citrus, hay, white berries. soft in the mouth without much
              expression.

  1.07 -  4   young tempranillo red clean and typical, with stone fruit on the nose.
               tannins in the mouth are somewhat discreet.
 ---------------------------------------------------------------------------
```

Reminder: Supervised and unsupervised approaches

In the statistical learning theory:

"Unsupervised approach" (exploratory or descriptive).
"Supervised approach (confirmatory or explanatory approach).

Factor analysis and classification are unsupervised,
Discriminant analysis or regression methods are supervised.

External validation is the standard procedure in the case of supervised learning.

Once the model parameters were estimated (learning phase),
external validation is used to evaluate the model (generalization phase),
usually with cross validation methods.

Reminder (continuation)

External validation in the context of correspondence analysis (CA).

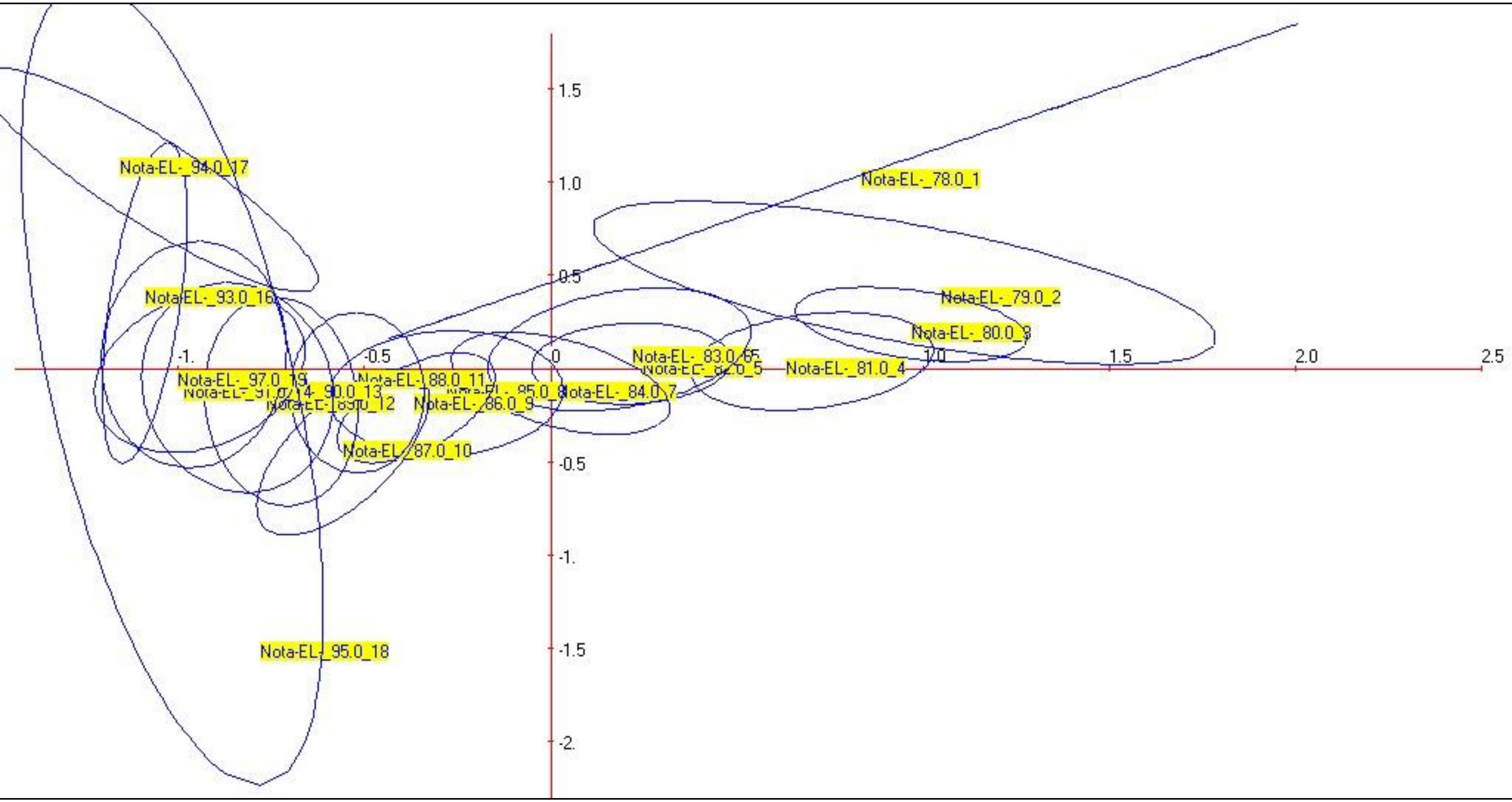
Two practical circumstances:

a) when the data set may be divided into two or more parts, one part being used to estimate the model, the other part used to verify the suitability of this model,

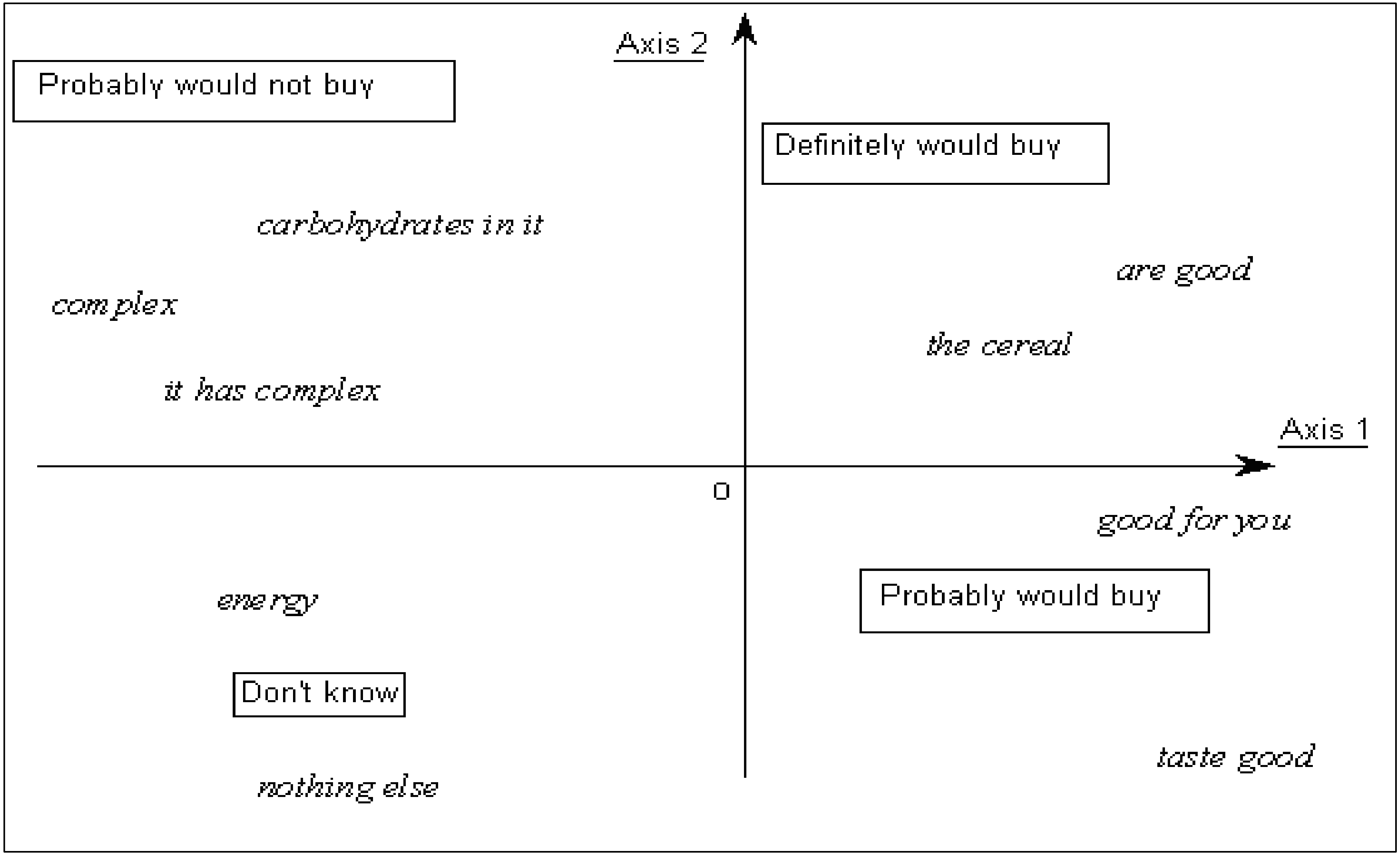
b) where certain metadata or external information are available to supplement the description of items.

We assume that external information in the form of “supplementary elements”.

Example 1 («Wine » question) Direct CA of responses, the score groups are projected afterwards on the principal plane. Bootstrap ellipses drawn after *bootstrapping the respondents*



Example 2: Open Questions / Copy-Test



Example 2: Open Questions / Copy-Test

Purchase intent and responses to open question

TEXT 1 : Probably would not buy

-- 1 to tell you about how long people have eaten them.
-- the complex carbohydrate that are in this cereal.
-- the people who eat this cereal and the product. that's all.

-- 2 it's supposed to be healthy
-- it has good carbohydrates in it.

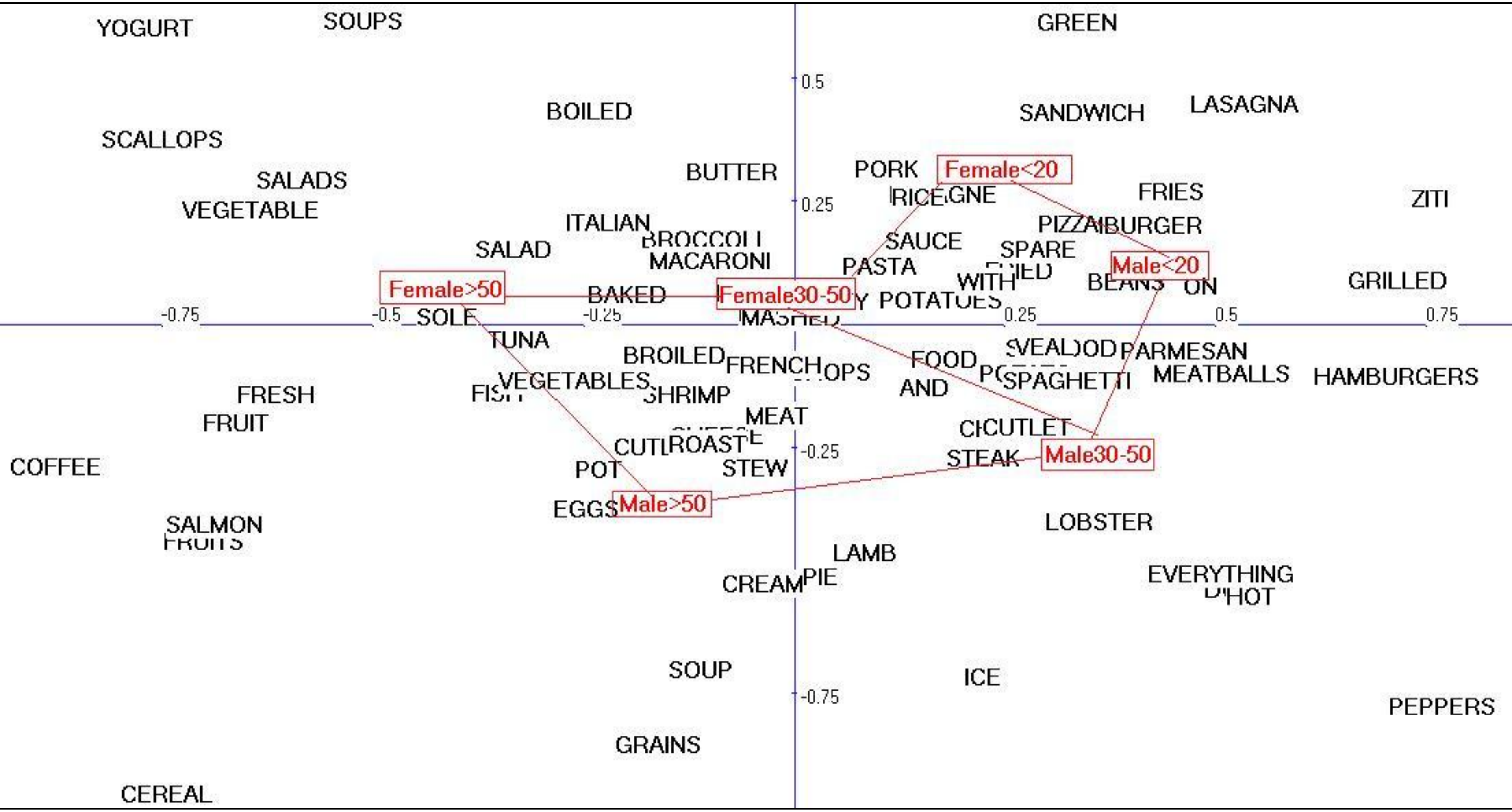
-- 3 that it has complex carbohydrate, to keep you going all morning, that people have eaten it a
-- long time, the years people have eaten this cereal and some didn't know about
-- the complex carbohydrate.

TEXT 3 : Probably would buy

-- 1 it's nutritious for you.
-- nothing else.

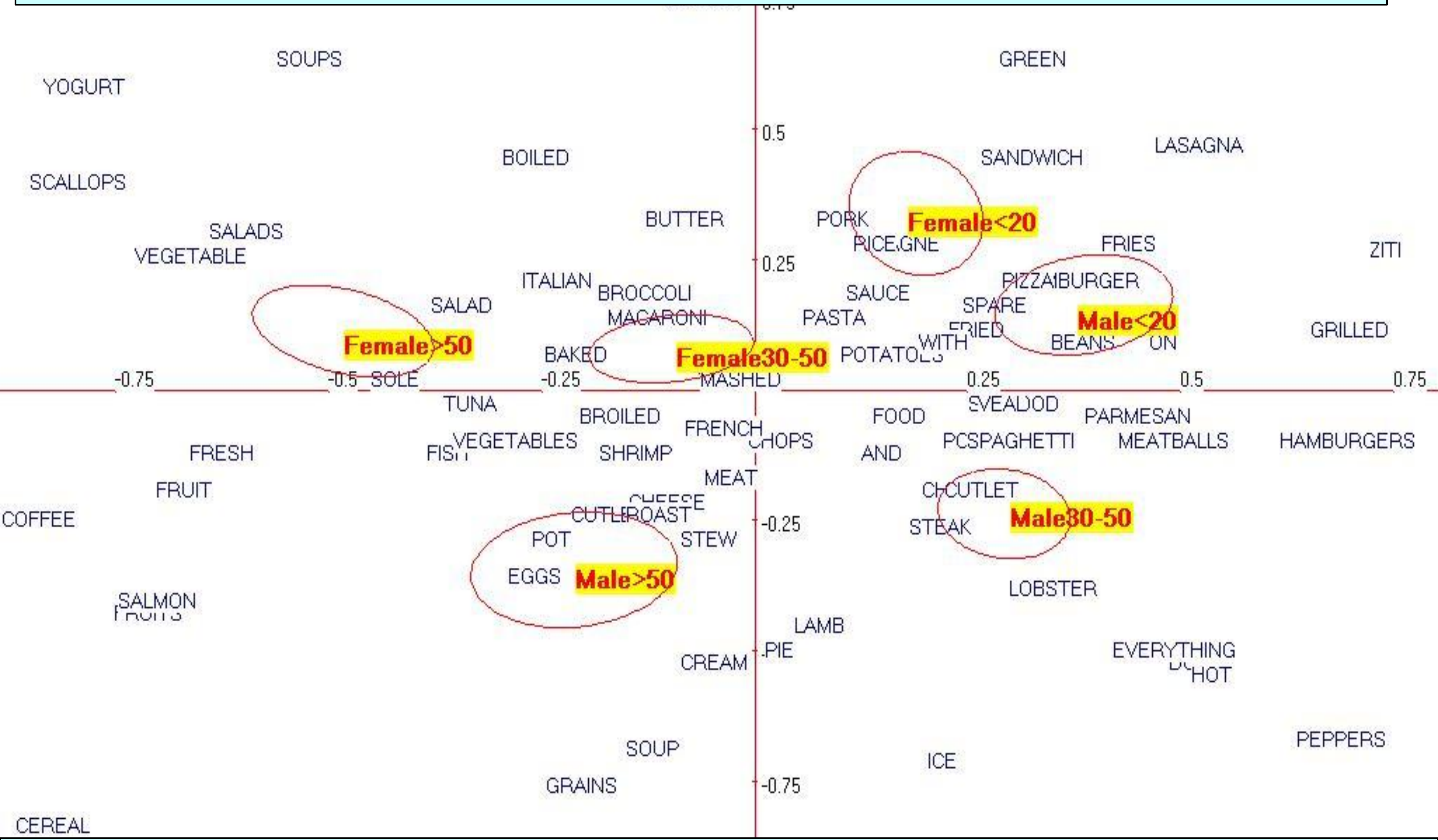
-- 2 that,is good for you
-- that,s all it said to me

Example 3: International survey (Tokyo Gas Company) about dietary habits. Open question: "*What dishes do you like and eat often?*



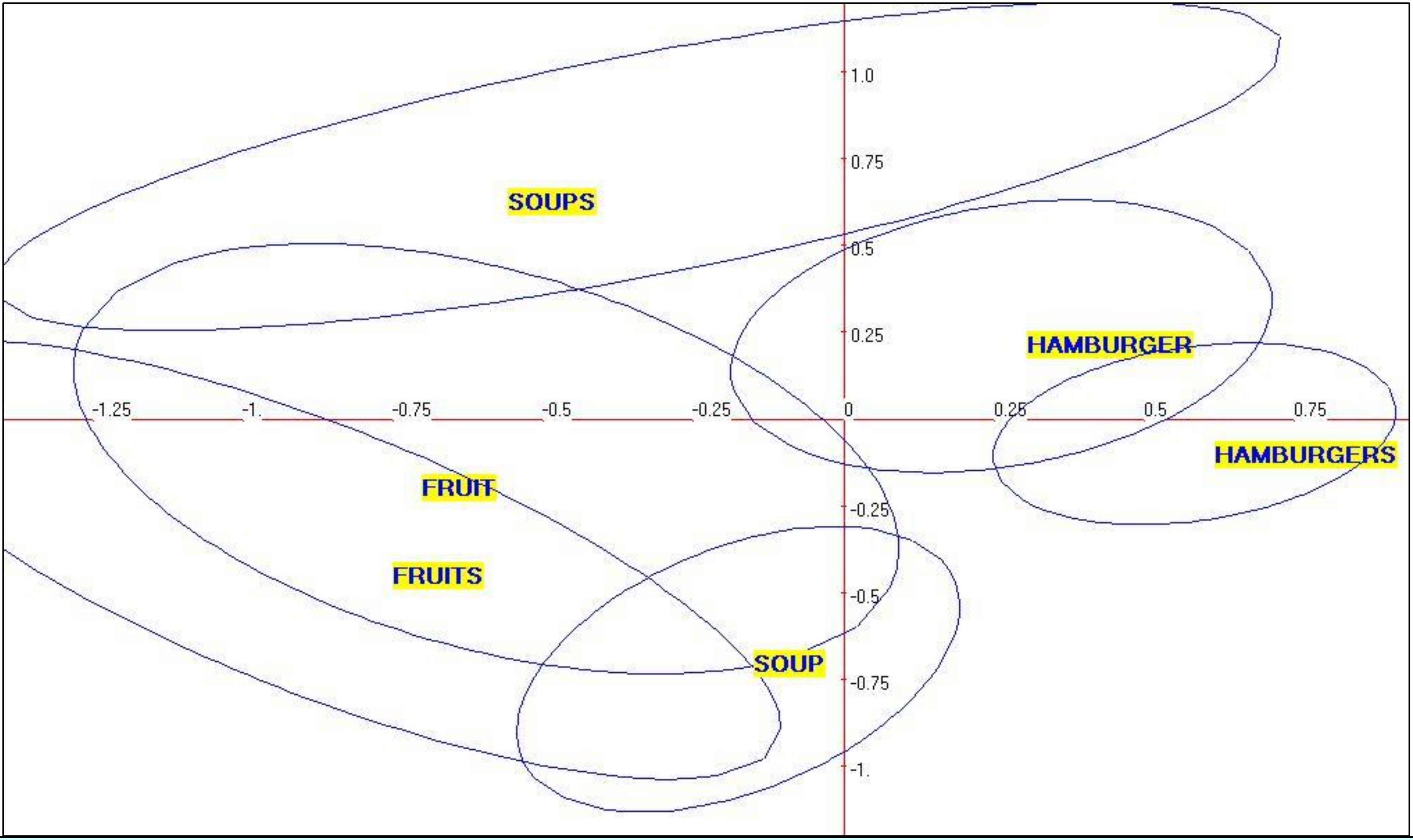
New York: First principal plane. Table crossing words and age x gender categories

Example 3: International survey (continuation). Question: "*What dishes do you like and eat often?*



New York: First principal plane. Example of confidence areas for categories (Bootstrap)

Example 3: International survey (continuation). Question: "*What dishes do you like and eat often?*



New York: First principal plane. Example of confidence areas for words (Bootstrap)

Example 3: International survey (continuation). *"What dishes do you like and eat often?*

ICE CREAM PEPPERS	PIE	SEAFOOD	POT CUTLET		SCALLOPS	FRUIT FRESH VEGETABLE	SALMON		FRUITS COFFEE
HOT DOGS	LOBSTER		FRENCH	BUTTER			**Female>50**		YOGURT SOUPS
	Male30-50	POTATO CHINESE	CHEESE	**Female30-50**			TUNA SALAD		SALADS
		STEAK FOOD		SHRIMP	VEGETABLES FISH	BAKED	ITALIAN		SOLE BROCCOLI
	VEAL MEATBALLS	SPAGHETTI	CHOPS		BROILED		MACARONI CHICKEN BEEF	TURKEY PASTA	
ON GRILLED	PARMESAN	AND		ROAST	CUTLETS				RICE BEANS
ZITI			LAMB		MEAT	LASAGNE BREAD		SAUCE PORK	**Female<20** FRIED
PIZZA	WITH HAMBURGERS		STEW	**Male>50**					SPARE RIBS
LASAGNA GREEN FRIES	**Male<20** HAMBURGER	POTATOES MASHED	EVERYTHING	GRAINS	SOUP CEREAL	EGGS	BOILED	GARLIC	SANDWICH

New York: First principal plane. Example of Kohonen Map (Self Organizing map).

Text Mining and Open-ended Questions in Sample Surveys

1) Principles of Data Mining and Text mining: A reminder

2) Open-ended Questions: Why? How?

3) From texts to numerical data

4) Basic statistical tools: Visualization, Characteristic words, Bootstrap.

5) Applications: Open questions, sample surveys, texts

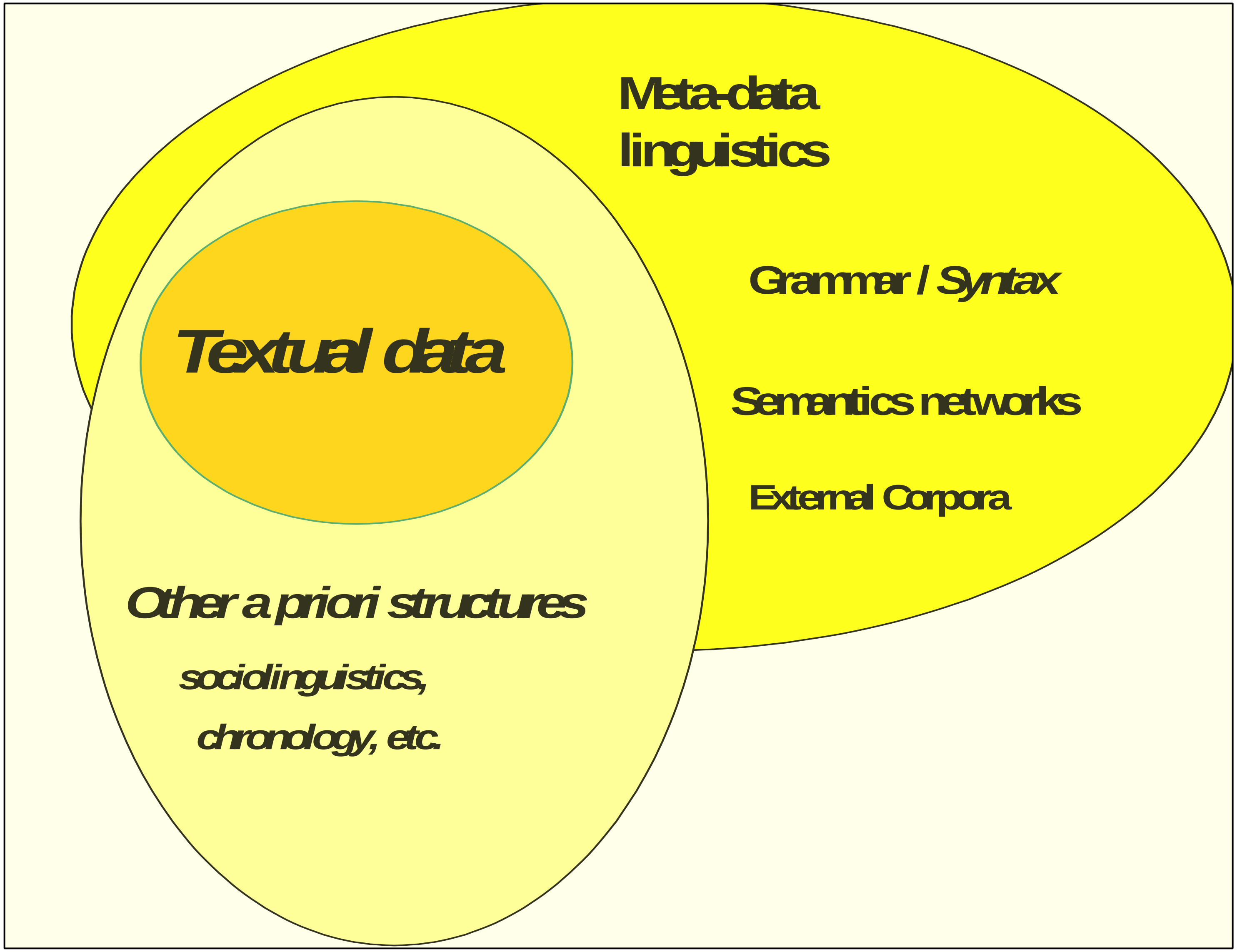
6) About textual data in general

7) Conclusions

Processing Strategy

- A priori Grouping (Lexical contingency table)
- Juxtaposition of Lexical contingency tables
- Direct Analysis of the sparse Lexical table

Importance of Meta-data



The four phases of a linguistic analysis

(A bxg flower)

Morphology

A big flower A bug flower

A bag flower A bog flower

Syntax

The spoon speaks

(The speaks)

Semantics

A man thinks

(A stone thinks)

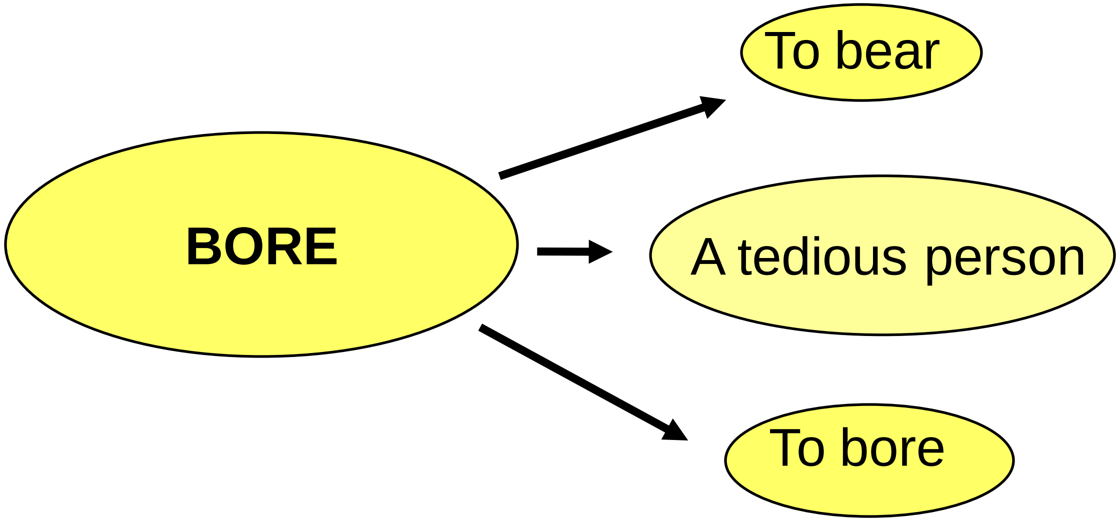
Pragmatics

A challenge to I.A.

6) About textual data in general

Homography, Polysemy, Synonymy

Homographs:



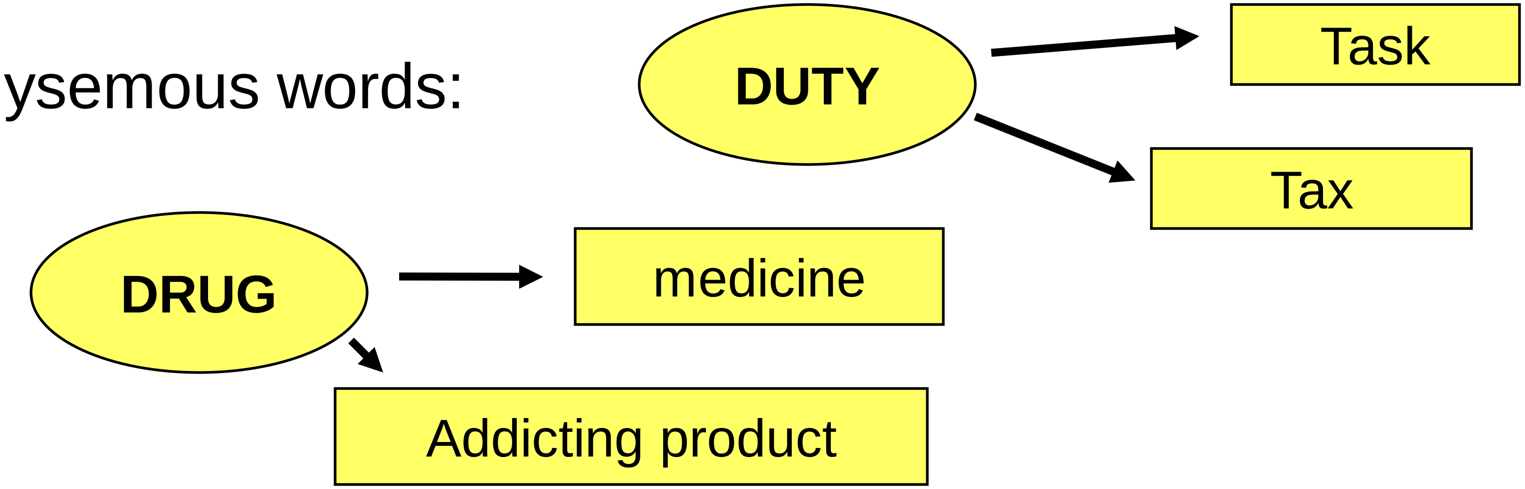
BORE

To bear

A tedious person

To bore

Polysemous words:



DUTY

Task

Tax

DRUG

medicine

Addicting product

Semantic content of a lexical profile

Distributional linguistics (Z. Harris)

A is sometimes purring

A mews

A has whiskers

A likes milk

A likes chasing mice

→ At the end, the point « A » will be superimposed with the point « CAT»

But semantic similarity is not a transitive relationship

(1) calm–*wisdom–discretion–wariness–fear*–*panic*,

(2) *fact*–*feature –aspect–appearance*–*illusion*

Text Mining and Open-ended Questions in Sample Surveys

Summary / Outline

1) Principles of Data Mining and Text mining: A reminder

2) Open-ended Questions: Why? How?

3) From texts to numerical data

4) Basic statistical tools: Visualization, Characteristic words, Bootstrap.

5) Applications: Open questions, sample surveys, texts

6) About textual data in general

7) Conclusions

As a conclusion...

For each open-ended question,

and for each partition of the sample of respondents,

we obtain, without any preliminary coding or other intervention:

- **A visualization of proximities between words and categories.**
- **Characteristic elements or words for each category .**
- **Modal responses for each category** (a kind of automatic summary).

[Remember also that the open question “Why” following a closed question provides an indispensable assessment of the real understanding of the question].

As a conclusion... (continuation)

All these processing are carried out under the supervision of robust assessment procedures:

- **Non-parametric statistical tests,**
- **Bootstrap validation.**

We are not dealing here with a novel sophisticated modeling involving complex hypotheses.

We use **simple instruments of observation** to get acquainted with the real concerns of the respondent, i.e.: the customer, the user, the client.

With the rapid development of online surveys, the spreading of e-mails and blogs, the presented set of tools could be a valuable component of a new methodology for a better customer knowledge.

7) Conclusions – Short Bibliography

- Akuto H. (1992). *International Comparison of Dietary Culture*. Nihon Keizai Simbun, Tokyo.
- Becue-Bertaut M, Alvarez-Esteban R, Pages J. (2008). Rating of products through scores and free-text assertions: Comparing and combining both*. Food Quality and Preference* **19**, 122–134.
- Bécue M., Lebart L. (1996). Clustering of texts using semantic graphs. Application to open-ended questions *Proceedings of the IFCS 96 Symposium*, Kobe, Springer Verlag, Tokyo (in press).
- Belson W.A., Duncan J.A. (1962): A Comparison of the check-list and the open response questioning system, *Applied Statistics,* 2, 120-132.
- Benzécri J.-P. (1992). *Correspondence Analysis Handbook.* Marcel Dekker, New York.
- Biber D. (1995). *Dimensions of register variation.* Cambridge Univ. Press, Cambridge.
- Bradburn N., Sudman S., and associates (1979): *Improving Interview Method and Questionnaire Design,* Jossey Bass, San Francisco.
- Greenacre M. (1993). *Correspondence Analysis in Practice*. Academic Press, London.
- Deerwester S., Dumais S.T., Furnas G.W., Landauer T.K., Harshman R. (1990). Indexing by latent semantic analysis, *J. of the Amer. Soc. for Information Science,* 41 (6), 391-407.
- Lebart L. (1982). Exploratory analysis of large sparse matrices, with application to textual data, *COMPSTAT,* Physica Verlag, 67-76.
- Lebart L., Salem A., Bécue M., (2000), *Análisis estadístico de textos*, Editorial Milenio, Lleida.
- Lebart L., Salem A., Berry E. (1998). *Exploring Textual Data.* Kluwer, Dordrecht.
- Lebart L., Morineau A., Warwick K. (1984). *Multivariate Descriptive Statistical Analysis.* John Wiley. N.Y.
- Ritter H., Kohonen T. (1989). Self Organizing Semantic Maps. *Biol. Cybern*. 61, 241-254.
- Salem A. (1984). La typologie des segments répétés dans un corpus, fondée sur l'analyse d'un tableau croisant mots et textes, *Cahiers de l'Analyse des Données,* 489-500.
- Schuman H., Presser F. (1981): *Question and Answers in Attitude Surveys*, Academic Press, New York.
- Sudman S., Bradburn N. (1974): *Response Effects in Survey,* Aldine, Chicago.

Software note: All the preceding computations (Multidimensional analysis of texts and images, Self organizing maps, various Bootstrap procedures) can be performed with the Software Dtm-Vic (Data and text Mining, Visualization, Inference, Classificaiton) freely downloadable from www.dtm-vic.com.

Software note: All the preceding computations (Multidimensional analysis of texts and images, Self organizing maps, Bootstrap) can be carried out with the Software Dtm-Vic (Data and text Mining, Visualization, Inference, Classificaiton) freely downloadable from the website: **www.dtm-vic.com.**

Thank You

Gracias

Grazie

Obrigado

Merci

Choukrane

Danke