Universidad de Navarra (Pamplona), Noviembre 27, 2002

MÉTODOS DE LA ESTADÍSTICA TEXTUAL APLICACIONES A LAS PREGUNTAS ABIERTAS EN ENCUESTAS.

(Text Mining…)

Ludovic Lebart
CNRS - ENST

Bajo el nombre de « Text Mining » se encuentra una serie de técnicas de Análisis exploratorio de datos textuales.

Ambito del *Text Mining:*

WEB

Periodicos

Vigilancia technológica

(Information Retrieval)
Búsqueda documental

Preguntas abiertas

Discursos, Informes, Entrevistas libres

Cartas de reclamación

a l ’origen del enfoque « data Mining »...

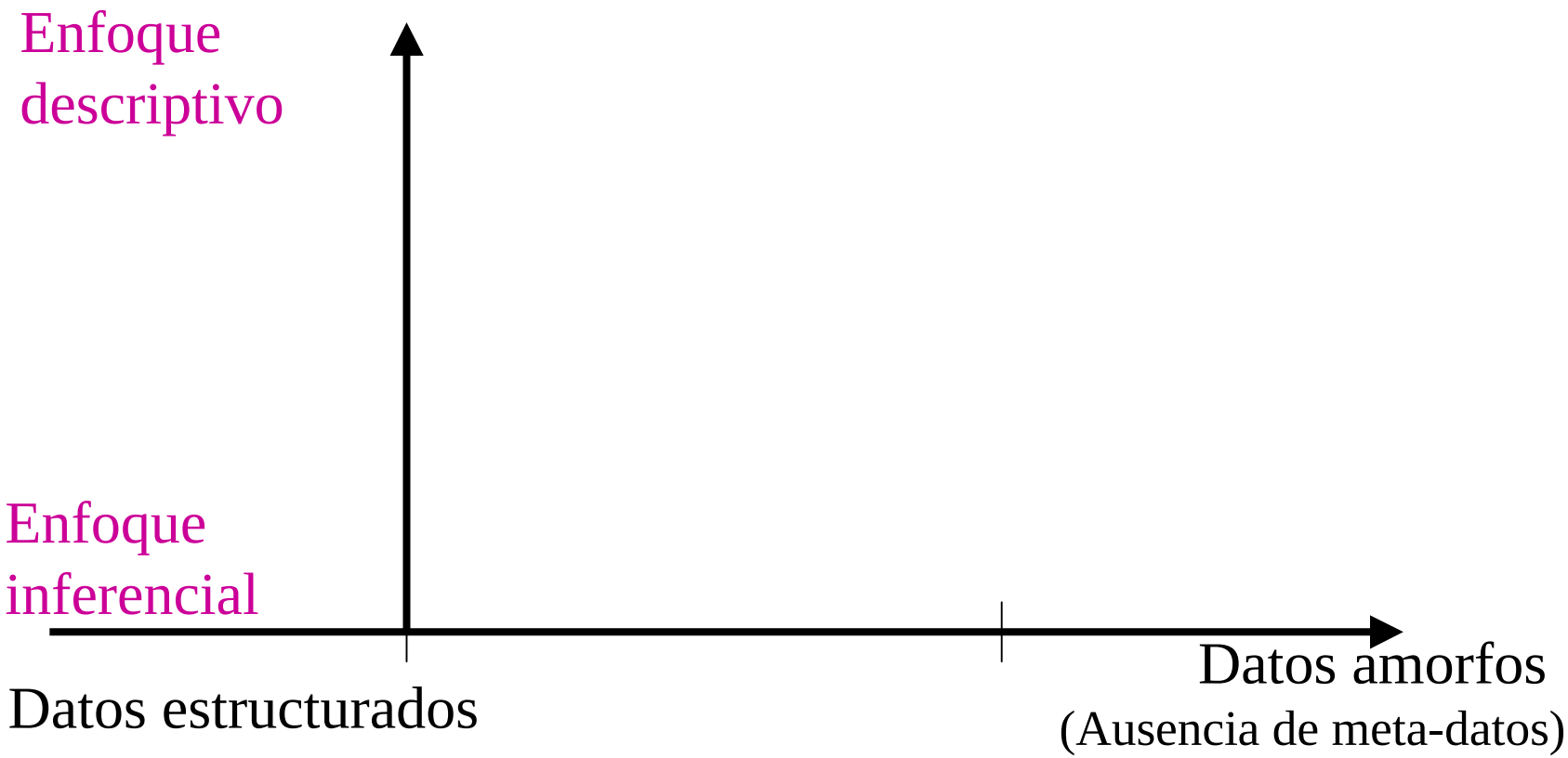
- 1- *Antiguas técnicas* en su principio son ahora faciles de emplear, y son mejoradas.
- 2- *Nuevas técnicas* son concebidas.
- 3- *Nuevos campos de investigación* pueden ser estudiados
- 4- *Nuevos productos* aparecen: "Los Paquetes".
- 5- Necesidad de selección de métodos y de estrategía de tratamiento.
- 6- Los bases de datos… numéricas y textuales

Tipo de situación práctica: el eje básico

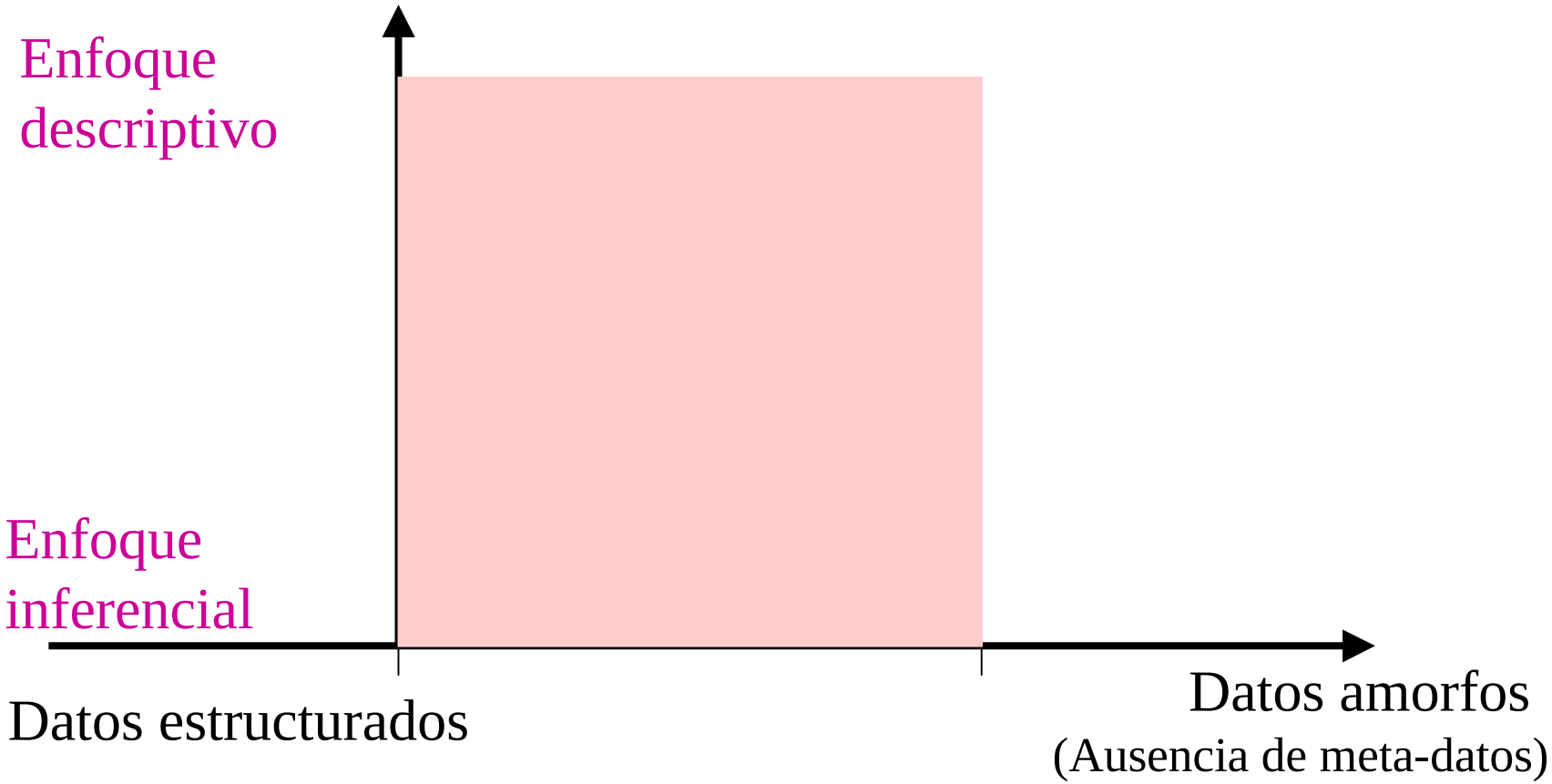
Datos estructurados

Datos amorfos
(Ausencia de meta-datos)

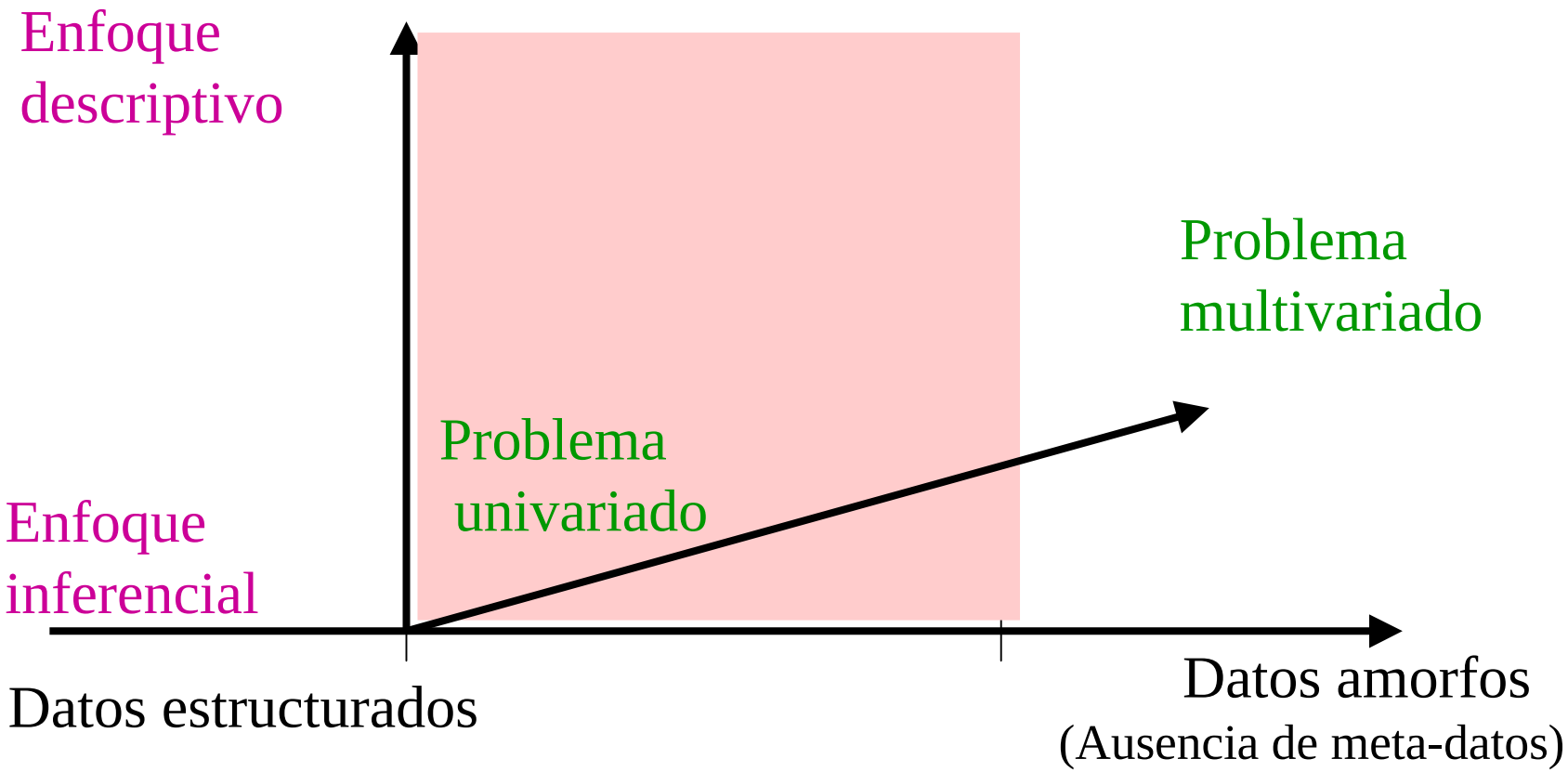
Tipo de situación práctica: el segundo eje



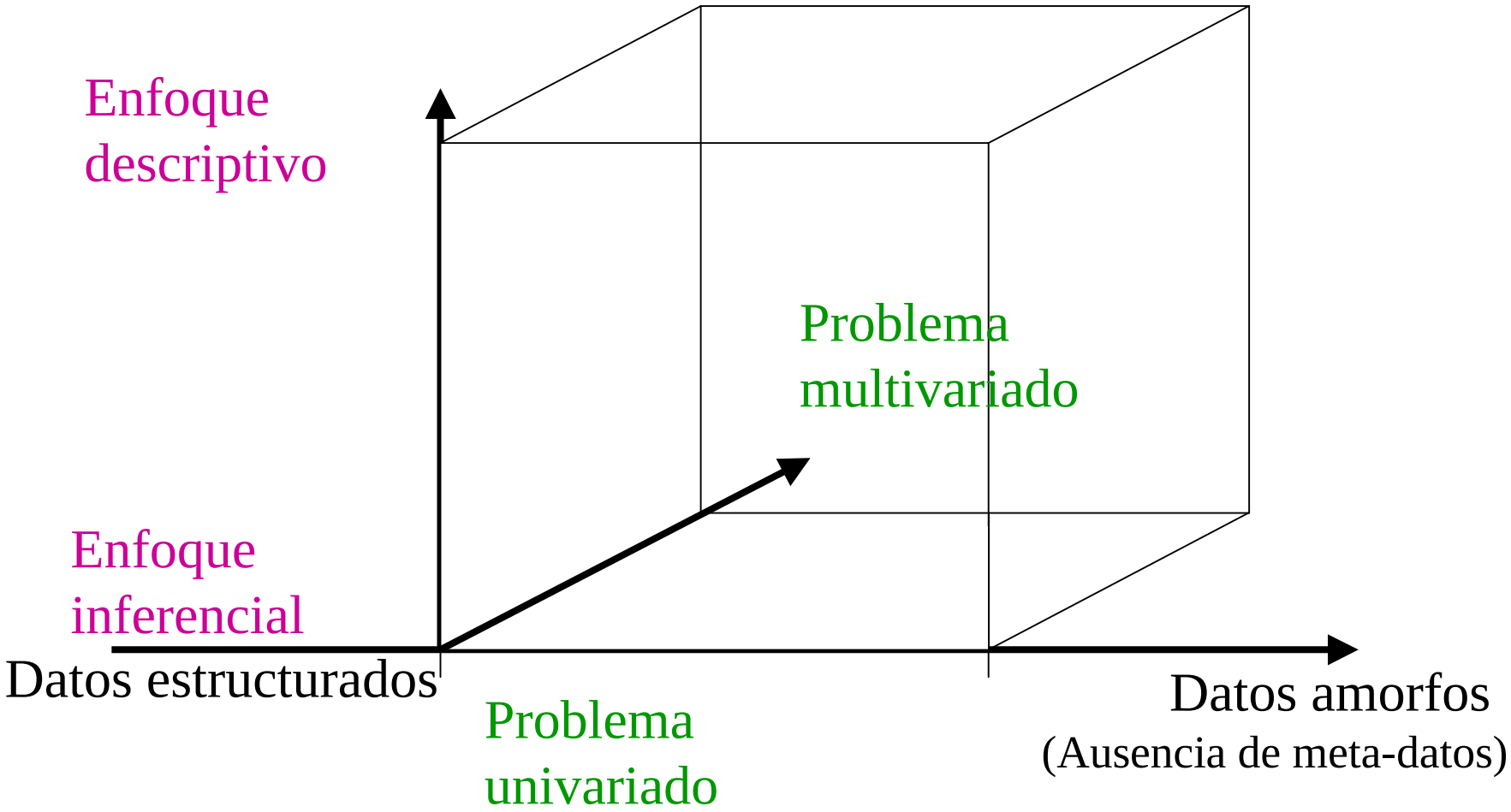
Tipo de situación práctica: el plano básico



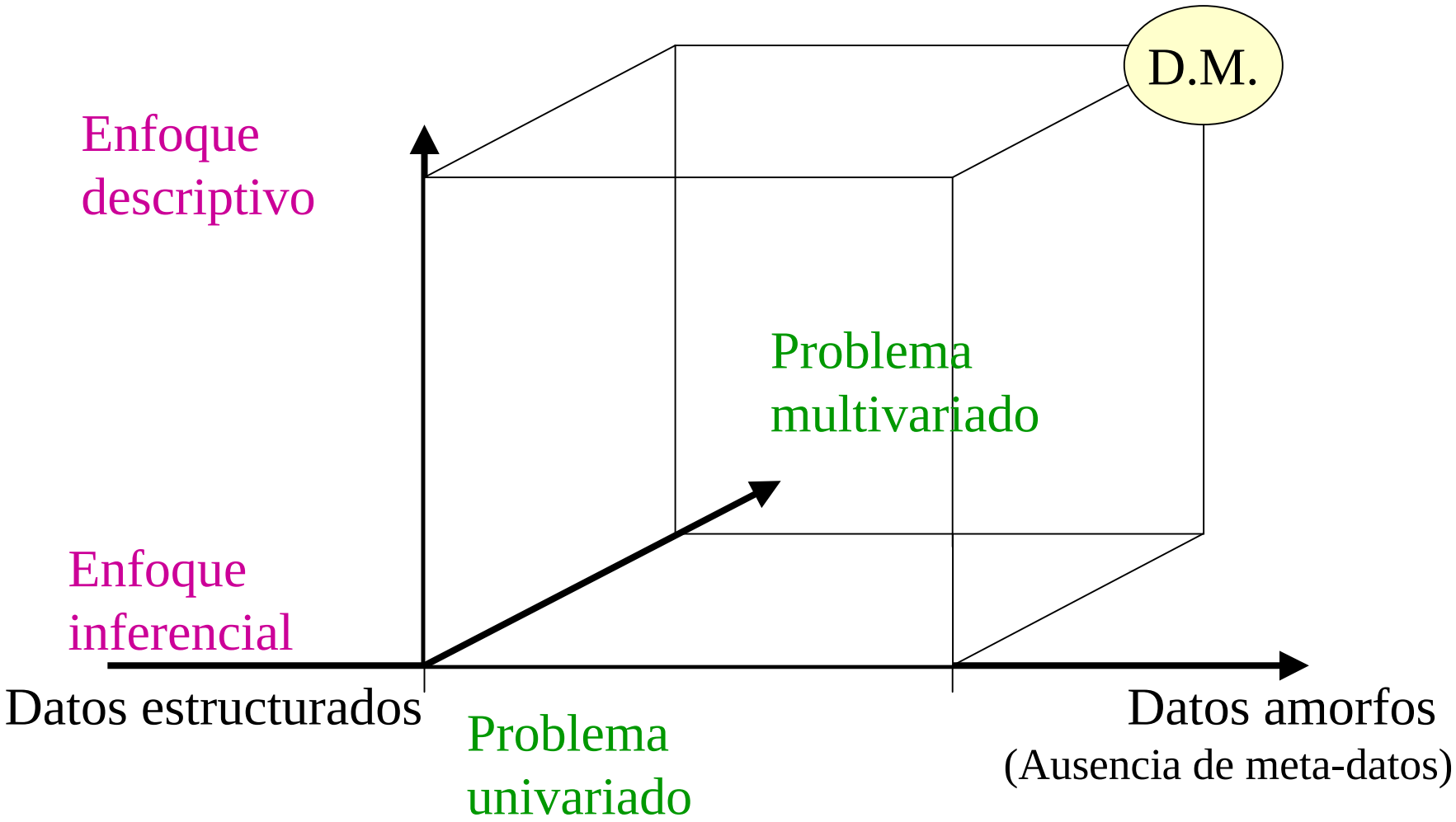
Tipo de situación práctica: el tercero eje



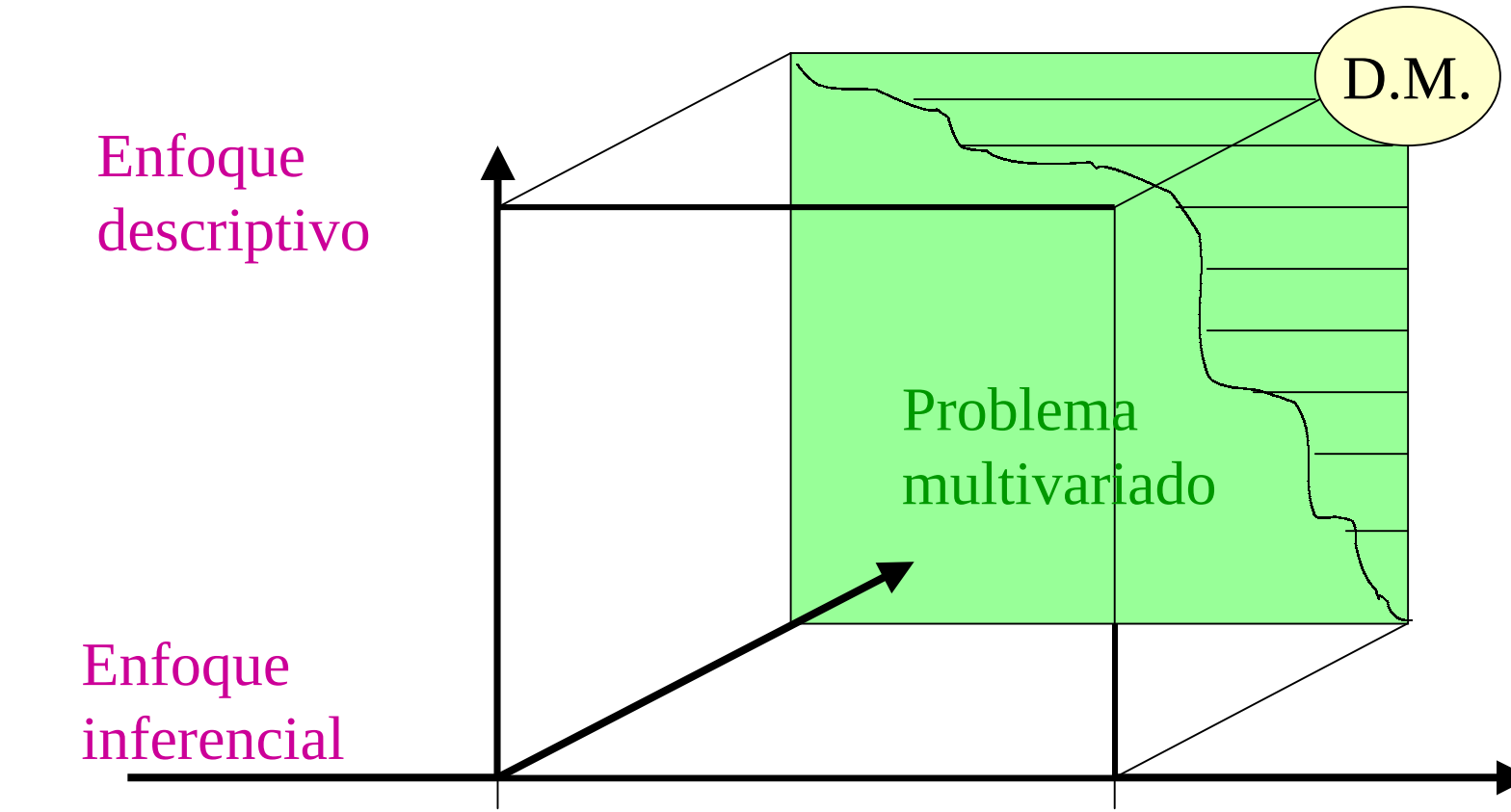
Tipo de situación práctica: el cúbico básico



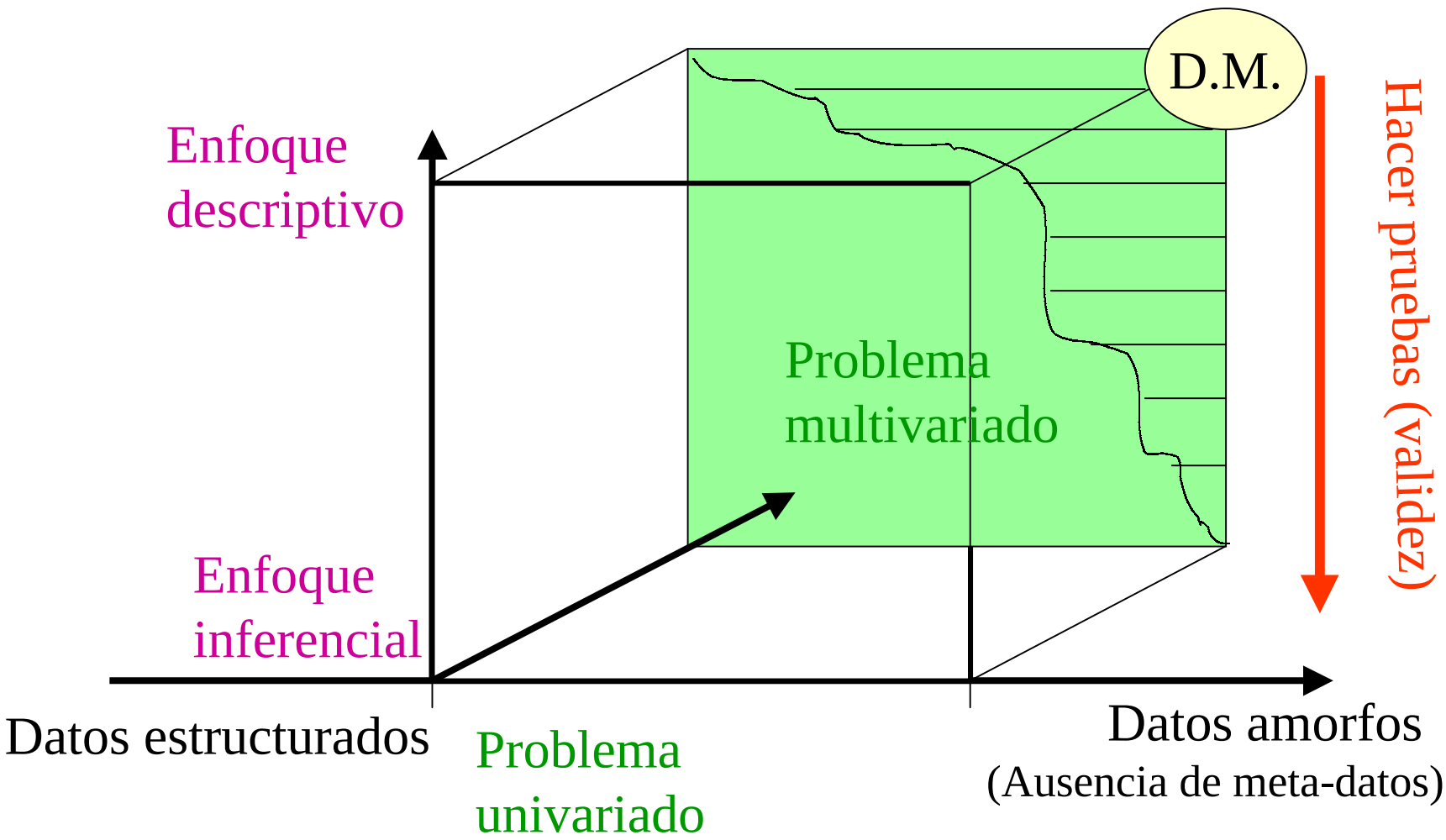
Tipo de situación práctica: el cúbico básico
Ubicación del análisis exploratorio de datos y del Data Mining



Tipo de situación práctica: el cúbico básico
El cuadrado de los problemas pendientes..



Tomar en cuenta la meta información



El “Text Mining” y el Análisis Exploratorio Multidimensional de Textos

- **« paradigma » inicial** (Benzécri, 1970):
- Extraer unidades de los textos
- Enriquecer la « lexicometria » con el « multidimensional»
- Aplicar las herramientas de visualisación a las tablas léxicas
- **evolución y diversificación de los métodos**

El “Text Mining” y el Análisis Exploratorio Multidimensional de Textos

- La presentacion se hara en el marco de las respuestas a preguntas abiertas.
- Entonces, tendremos que insistir sobre los rasgos particulares de ese tipos de datos.
- Definiremos tambien una estrategia de tratamiento propia a las preguntas abiertas.

Preguntas abiertas : Cuando?

◆ *Para reducir la duración de la entrevista:*
Las preguntas abiertas son más económicas en cuanto a tiempo de de entrevista, y generan menos cansancio y tension.
(caso de listas de items extensas)

◆ *Para recoger una información espontánea*
Ejemplo: Marketing.
(¿ Que recuerda de este anuncio publicitario ?)

Preguntas abiertas : Cuándo? (2)

◆ *Para entender y explicitar la respuesta a una pregunta cerrada*

La pregunta : « ¿ Por qué ? »

Las explicaciones relativas a una respuesta ya dada deben expresarse de manera espontánea.

◆ *Para obtener informaciones a priori no comparables*

- Medio ambiante,
- Costumbres alimenticias en un contexto internacional)

Preguntas abiertas: inconvenientes, ventajas...

INCONVENIENTES

Coste

Complejidad

Especificidad

VENTAJAS

Rapidez

Libertad

Riqueza

Comparación « abierto » - « cerrado »

Postcodificación manual des las respuestas

(principales inconvenientes)

Mediación del codificador Se añade a los sesgos de la entrevista (interpretación, decisiones discutibles)

Destrucción de la forma : se pierde la tonalidad de la entrevista y el estilo de la repuesta.

Empobrecimiento del contenido.
Les respuestas complejas, compuestas, vagas son laminadas por la codificación.

Problema de las respuestas raras eliminadas *a priori*.

Ejemplo de riqueza y de complejidad :

A la pregunta "¿por qué?", planteada después de una pregunta bastante general sobre "*los anhelos de las personas respecto a los próximos anos":*

"*que tengamos con buena salud y que nuestros hijos sean felices es lo que más nos importa; cuanto al resto, no creemos mucho en ello*".

Los ítems *salud personal*, *felicidad de los hijos* se pueden identificar y conservar.
Pero ¿cómo codificar la idea de exclusión de los otros ítems, lo que podria ser fundamental y bastante característico del contenido de esta respuesta en términos de estilo de vida?

A proposito de la importancia del enunciado de una pregunta

Los trabajos de Rugg (1941) muestran que la respuesta *"yes"* a la pregunta:

"Do you think the United States should forbid public speeches against democracy?"

obtiene 21 puntos (sobre 100) menos que las respuesta *"no"* a la pregunta:

"Do you think the United States should allow public speeches against democracy?

A proposito de la comparacion « abierto » - « cerrado »

Un ejemplo clásico (*ver: Schuman et al. , 1981*).
Interrogados a propósito del…

"*problema más importante que debe afrontar Estados Unidos*",

16% de los estadounidenses mencionan

"*crime and violence*"
(respuestas libres reagrupadas),

… mientras que este mismo ítem propuesto en respuesta a una pregunta cerrada recoge 33% de las respuestas.

A proposito de la comparacion « abierto » - « cerrado » (2)

La explicación dada por los autores es la siguiente:
la inseguridad se considera a menudo como un fenómeno local y no nacional, lo que hace que el ítem *crime and violence* se cite poco de forma espontánea.

El hecho de cerrar la pregunta indica que esta respuesta es pertinente o incluso *posible*;

→ de donde un más elevado porcentaje de respuestas.

Cerrar la pregunta modifica de hecho su redactado y su significación

Ejemplo (textos Ingles)

Pregunta abierta :

"What is the single most important thing in life for you?"
seguida por :
"What other things are very important to you?".

Preguntas incluidas en una encuesta internacional (7 paises)
(Japon, France, Allemagne, Italie, Hollande,
U K, USA) cerca de 1990 (Hayashi *et al*., 1992).

El ejemplo concierne la parte Ingles de la encuesta
(tamaño de la muestra : *1043*).

Ejemplos de repuestas en Inglès:

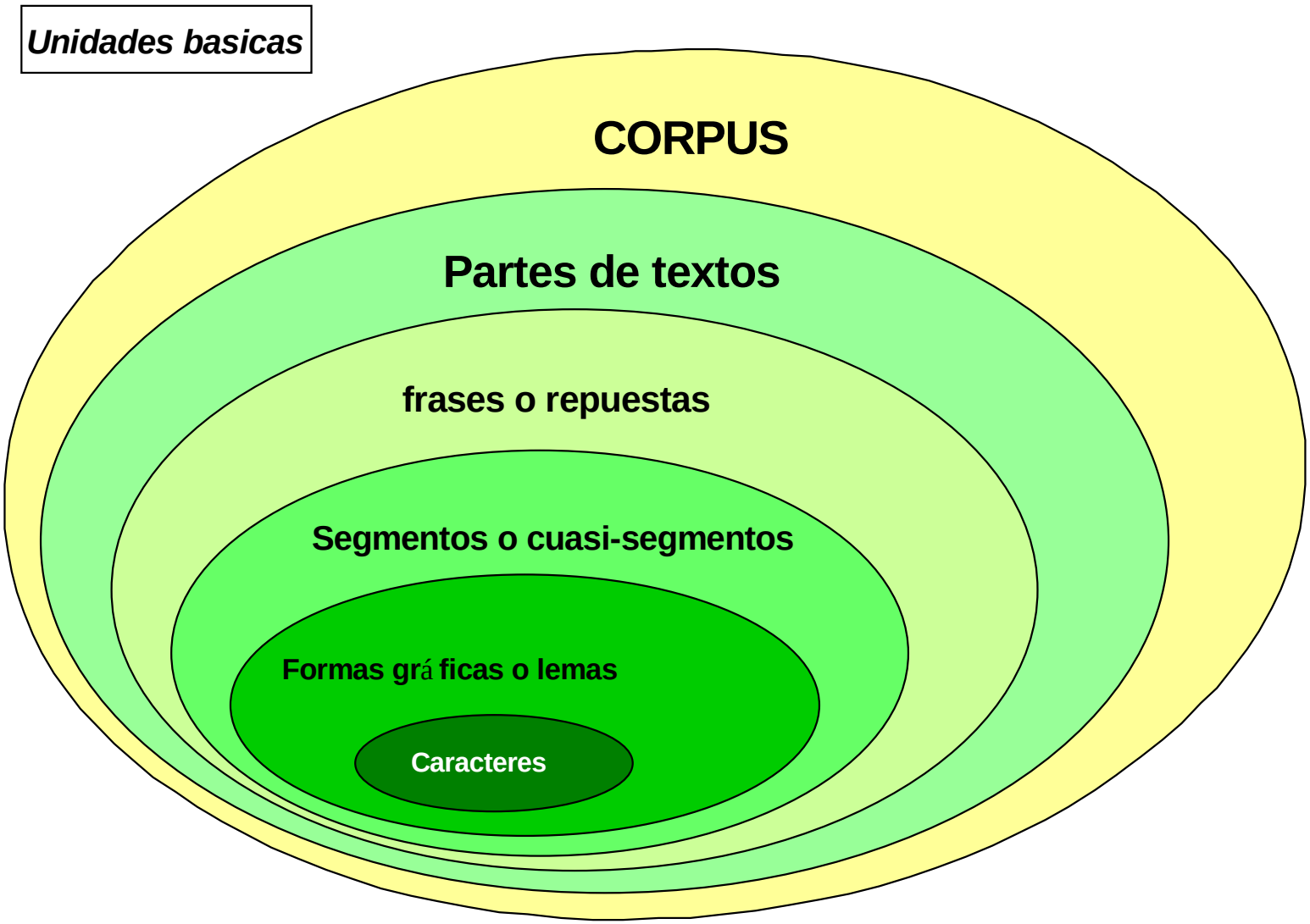
"Life question"

Gender	*Educ.*	*Age*	*Responses*
1	1	4	happiness in people around me, contented family, would make me happy
1	2	2	my own time, not dictated by other people
1	2	2	freedom of choice as to what I do in my leisure time
1	3	2	I suppose work
1	2	1	firm, my work, which is my dad's firm
2	1	6	just the memory of my last husband
2	2	6	well-being of my handicapped son
1	1	5	my wife, she gave me courage to carry on even in the bad times
2	2	3	my sons, my kids are very important to me, being on my own, I am responsible for their education
1	3	3	job, being a teacher I love my job, for the well-being of the children

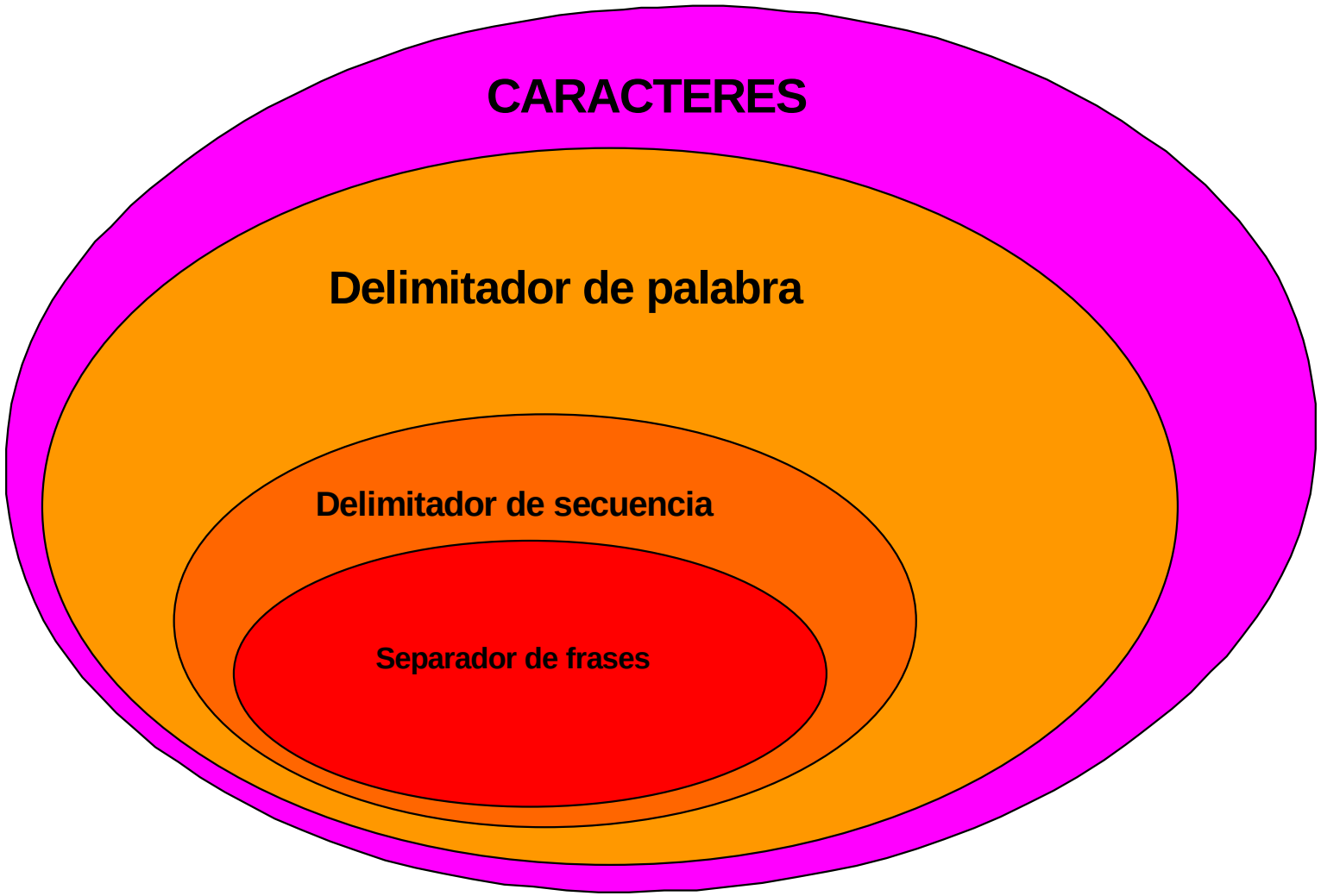
El “Text Mining” y el Análisis Exploratorio Multidimensional de Textos

- **« paradigma » inicial** (Benzécri, 1970):
- Extraer unidades de los textos
- Enriquecer la « lexicometria » con el « multidimensional»
- Aplicar las herramientas de visualisación a las tablas léxicas
- **evolución y diversificación de los métodos**

« Extraer unidades en los textos »



Los tipos de caracteres



Ejemplo 1, *continuacion*

El balance de la primera fase de segmentacion automatica :

Por *1043* respuestas, hay *13 669* ocurrencias (*tokens*),

con *1 413* palabras distintas (*types*).

Seleccionando las palabras que aparecen mas de *16* veces,
quedan: *10 357* ocurrenciass de esas palabras (*tokens*),

con *135* palabras distintas (*types*).

Cuidado con la polisemia de « *word* » y de « *palabra* »

Enriquecer la lexicometria con el multidimensional

- Herramientas de visualisación
  - ***Analisis en ejes principales***
  - ***Clasificación de las palabras y de los textos***
  - ***Mapas de Kohonen***
- Selección de las unidades y de las frases características
  - ***Unidades características (formas, segmentos, lemas)***
  - ***Selección de las respuestas modales***

Palabras (formas gráficas) apareciendo más de 16 veces (orden alfabético)

Words Appearing at Least Sixteen Times (Alphabetic Order) in the *1043* responses to the open question

Word	Frequency	Word	Frequency	Word	Frequency
I	*248*	go	*19*	of	*312*
I'm	*22*	going	*26*	on	*59*
a	*298*	good	*303*	other	*33*
able	*55*	grandchildren	*30*	others	*17*
about	*31*	happiness	*227*	our	*29*
after	*26*	happy	*137*	out	*34*
all	*86*	have	*99*	own	*16*
and	*504*	having	*70*	peace	*77*
anything	*19*	health	*609*	people	*61*
are	*65*	healthy	*45*	really	*28*

El tratamiento básico :

- Agrupamiento de las respuestas
- Ayuda por la lectura de los textos artificiales

Estrategia de análisis

- Agrupamiento *a priori*
- Yuxtaposiciones de agrupamientos
- Análisis directo de la tabla palabras - respuestas
- Uso de particiones instrumentales

¿ Como reagrupar las respuestas?

Se puede utilizar categorías o de combinaciones de categorías pertinentes en relación con la pregunta abierta analizada.

Al reagrupar las respuestas correspondientes a cada una de las categorías se obtienen "discursos artificiales" cuya significación es tanto más clara cuando las categorías se escogieron con cuidado.

Así, la lectura y la interpretación vienen considerablemente facilitadas; en efecto aparecen, para cada categoría, repeticiones y asociaciones de palabras significativas.
No obstante, esta reorganización de la información bruta se puede hacer de numerosas maneras.

Ejemplo de una tabla léxica

- *Primera partición:* 3 categorias de edad

- menos de 30 años [-30],
- entre 30 anos y 55 años [-55]
- mas de 55 años [+ 55] .

- *Segunda partición:* 3 niveles de educación

- nivel bajo [L],
- Medio [M],
- nivel alto [H]

El “Text Mining” y el Análisis Exploratorio Multidimensional de Textos

- **« paradigma » inicial** (Benzécri, 1970):
- Extraer unidades de los textos
- Enriquecer la « lexicometria » con el « multidimensional»
- Aplicar las herramientas de visualisación a las tablas léxicas
- **evolución y diversificación de los métodos**

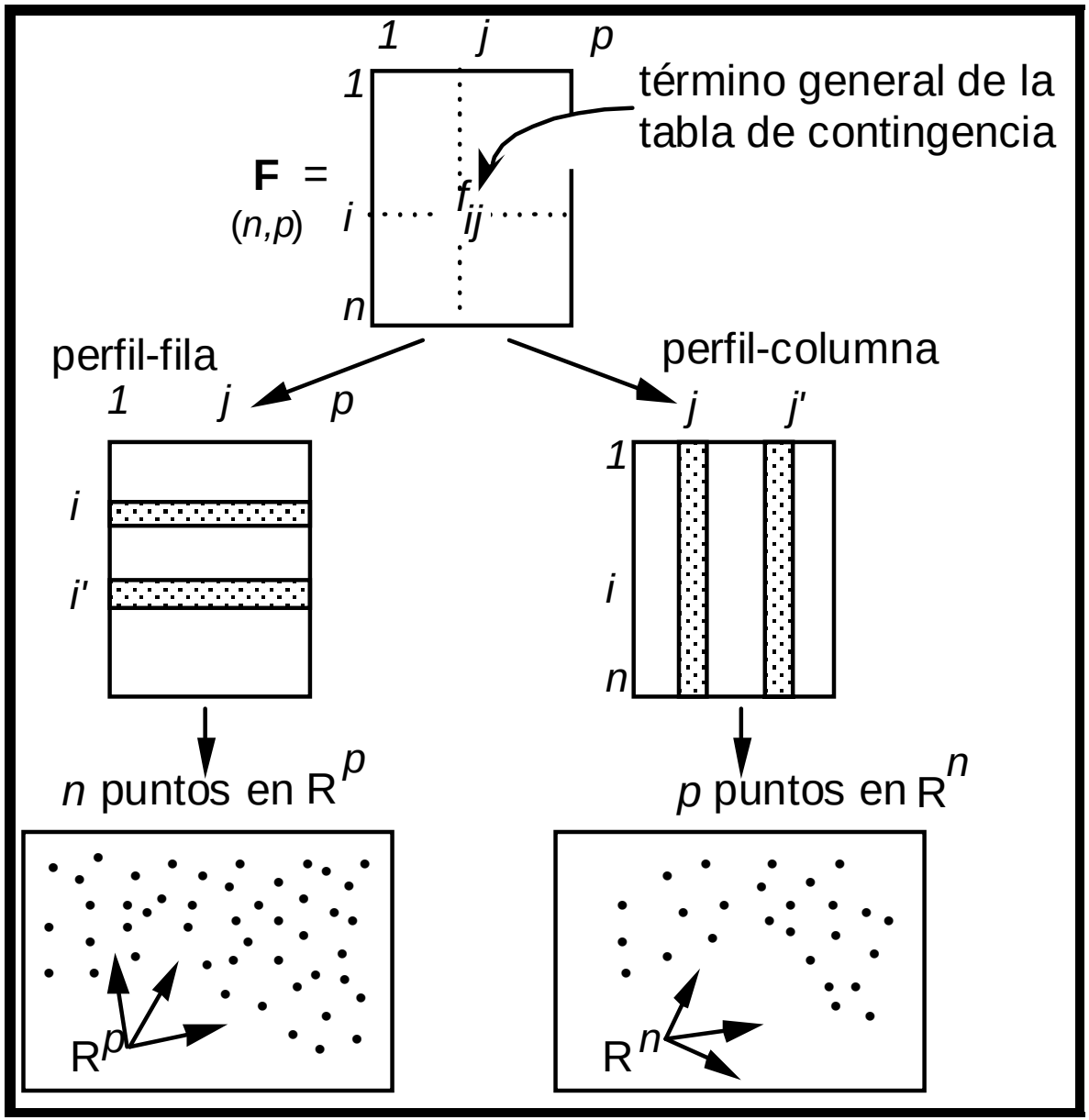
Ejemplo de tabla léxica

Partial listing of lexical table cross-tabulating 135 words of frequency greater than or equal to 16 with 9 age-education categories

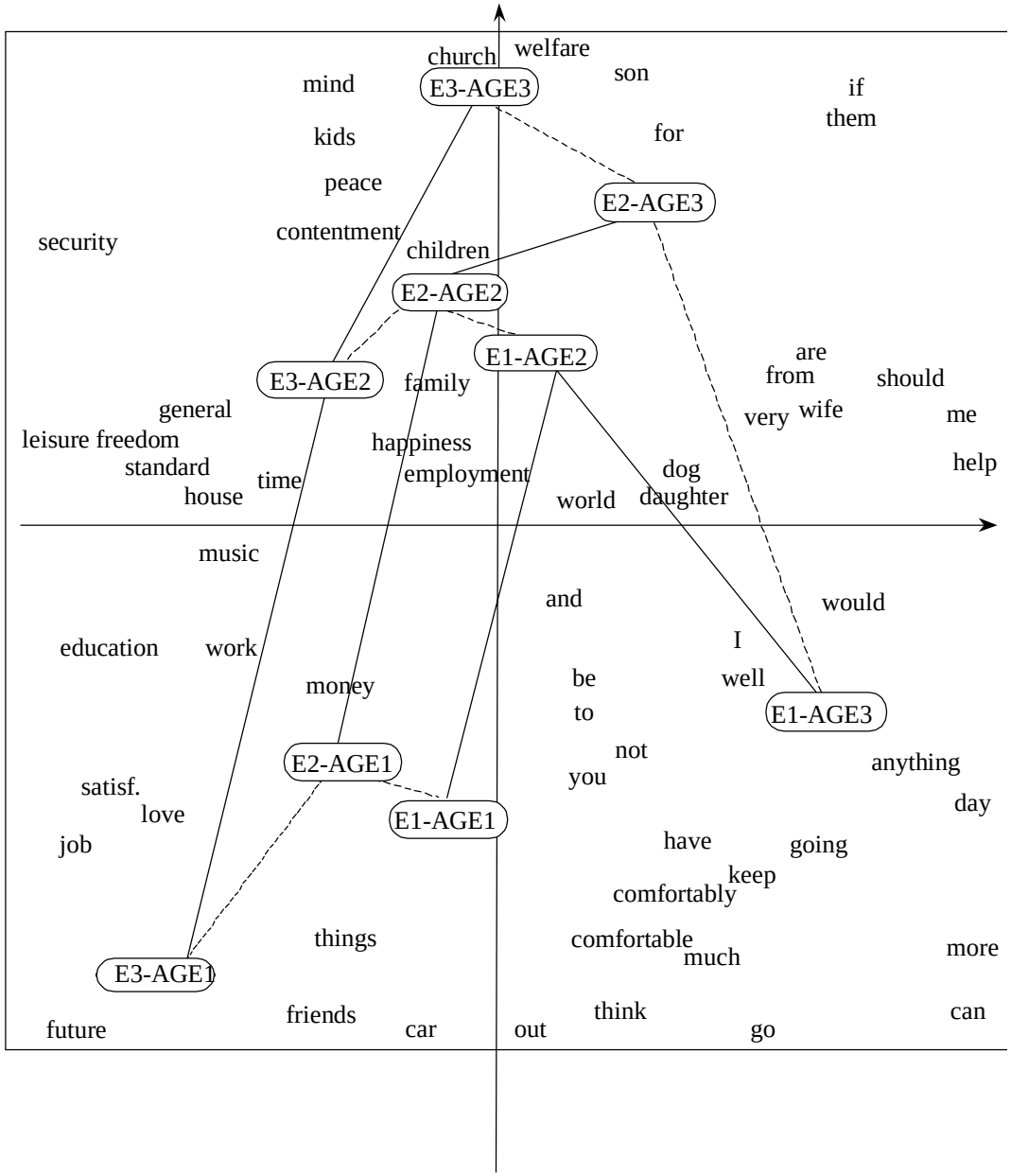
	L-30	L-55	L+55	M-30	M-55	M+55	H-30	H-55	H+55
I	*2*	*46*	*92*	*30*	*25*	*19*	*11*	*21*	*2*
I'm	*2*	*5*	*9*	*3*	*2*	*1*	*0*	*0*	*0*
a	*10*	*56*	*66*	*54*	*44*	*19*	*20*	*22*	*7*
able	*1*	*9*	*16*	*9*	*7*	*4*	*4*	*5*	*0*
about	*0*	*3*	*13*	*7*	*1*	*2*	*4*	*1*	*0*
after	*1*	*8*	*11*	*3*	*1*	*2*	*0*	*0*	*0*
all	*1*	*24*	*19*	*8*	*18*	*6*	*3*	*5*	*2*
and	*8*	*89*	*148*	*86*	*73*	*30*	*25*	*32*	*13*
anything	*0*	*4*	*9*	*1*	*3*	*0*	*1*	*1*	*0*

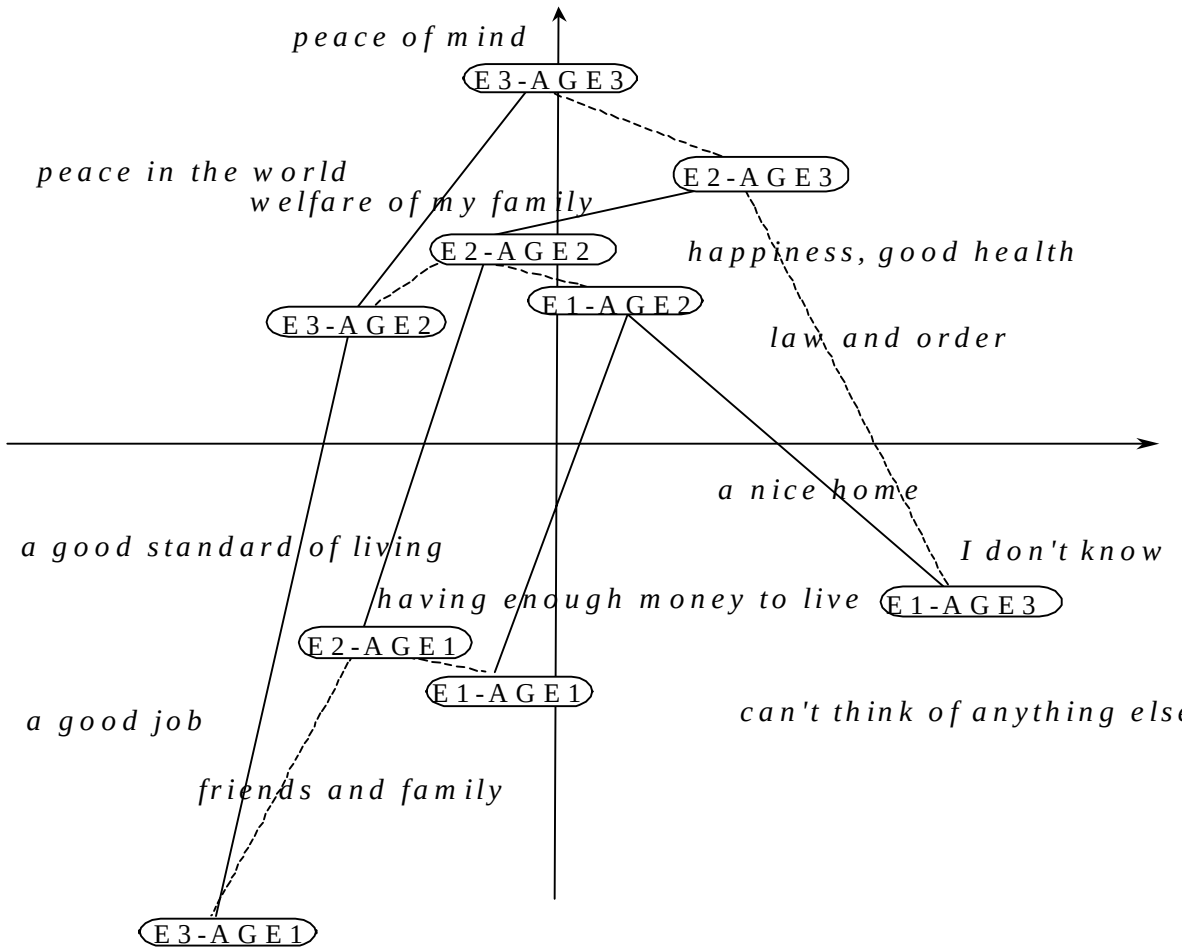
El “Text Mining” y el Análisis Exploratorio Multidimensional de Textos

- **« paradigma » inicial** (Benzécri, 1970):
- Extraer unidades de los textos
- Enriquecer la « lexicometria » con el « multidimensional»
- Aplicar las herramientas de visualisación a las tablas léxicas
- **evolución y diversificación de los métodos**


1
j
p
1
término general de la
tabla de contingencia
F =
(n,p)
i
f
ij
n
perfil-fila
perfil-columna
1
j
p
j
j'
i
i'
1
i
n
n puntos en R^p
p puntos en R^n
R^p
R^n

Correspondencia Palabras - Categorias





Ubicación de los Segmentos

welfare mind if H+55+55/high	peace	the getting M-5530-55/me	others of contentment	suppose general freedom H-5530-55/hi	security music love
which enjoy	with other husband content all	home	world time live is	people important	things food
for	son health daughter M+55+55/medi	that my life in happiness family enough children L-5530-55/lo	nothing and able	think comfortable	want friends do H-30-30/high
wife very ve them should on from are		up their s no get	it comfortably at	really out just healthy else about	what having car being

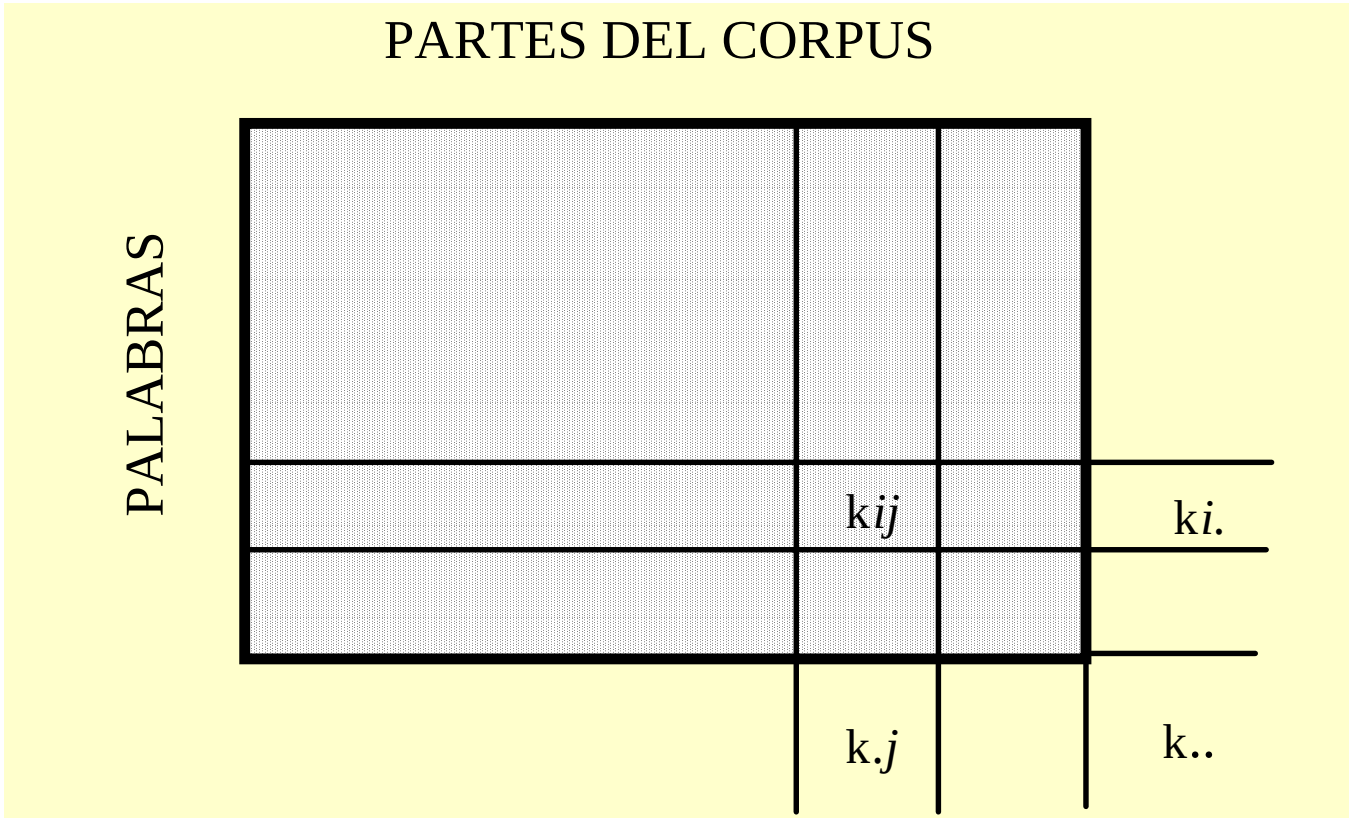
Kohonen map (parcial)

- Selección de las unidades características y de las respuestas modales
  - unidades características ***(palabras, segmentos, lemas)***
  - respuestas o frases modales

Palabras características *(especifidades)*

words	%W	%glob	FR.W	TestValue
H-30 = -30 * high				
1 friends	2.87	1.11	17	116	3.44
2 do	1.35	.45	8	47	2.60
3 want	1.01	.30	6	31	2.44
4 being	2.19	1.11	13	116	2.18
5 job	2.53	1.36	15	142	2.16
6 having	1.52	.67	9	70	2.11
7 things	.84	.27	5	28	2.06
---					
2 wife	.00	.65	0	68	-2.10
1 health	2.70	5.85	16	609	-3.59
H+55 = +55 * high				
1 mind	2.55	.45	5	47	2.91
2 welfare	1.53	.21	3	22	2.42
3 peace	2.55	.74	5	77	2.17

Palabras características



k..	tamaño del corpus
k*i.*	frecuencia de la palabra en el corpus
k*ij*	frecuencia de la palabra en la parte
k.*j*	tamaño de la parte

Palabras características

$$\text{Prob}(k, k_{i.}, k_{.j}, n) = \frac{\binom{k_{i.}}{n}\binom{k-k_{i.}}{k_{.j}-n}}{\binom{k}{k_{.j}}}$$

Formula de la ley hypergeometrica

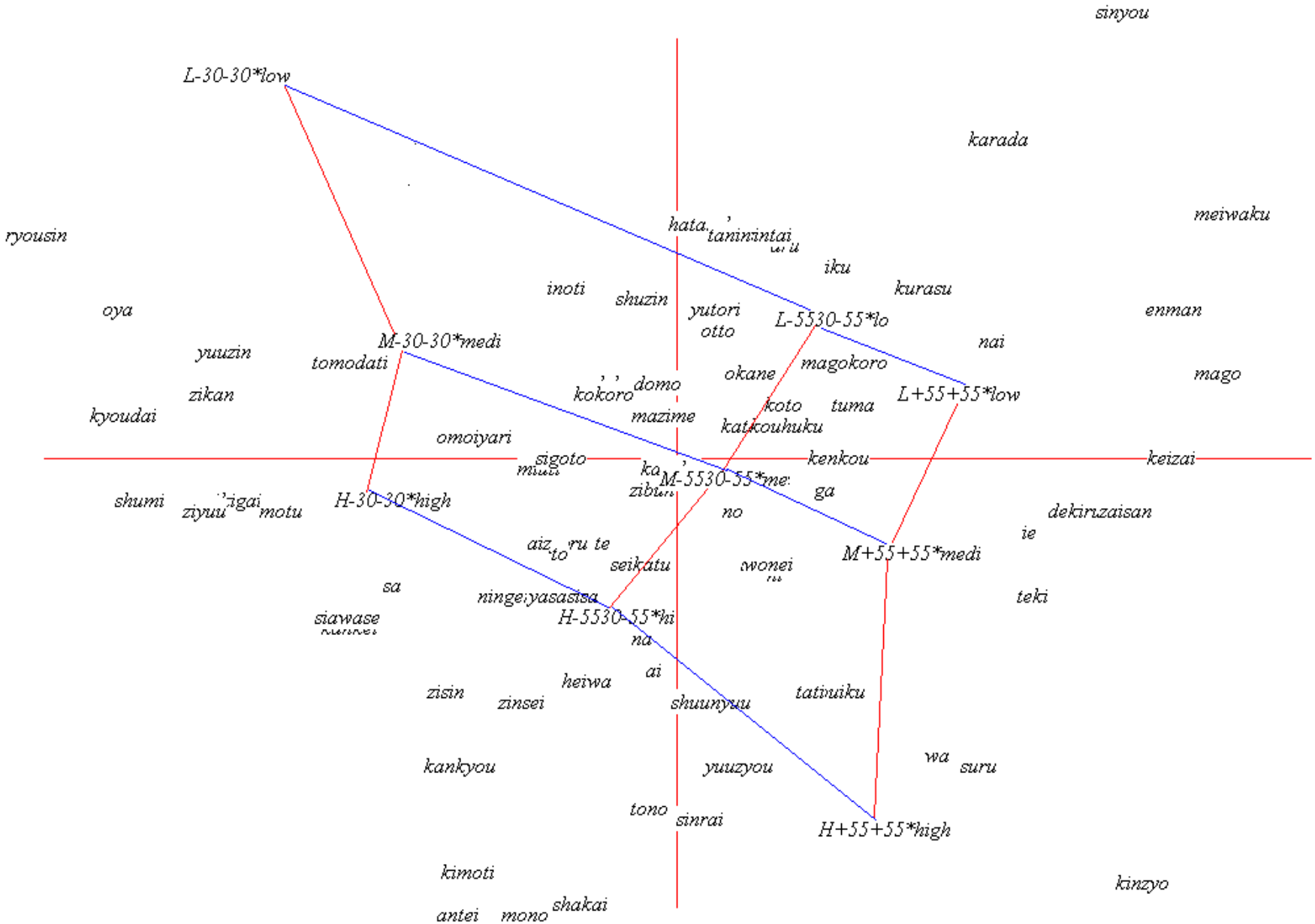
Respuestas modales

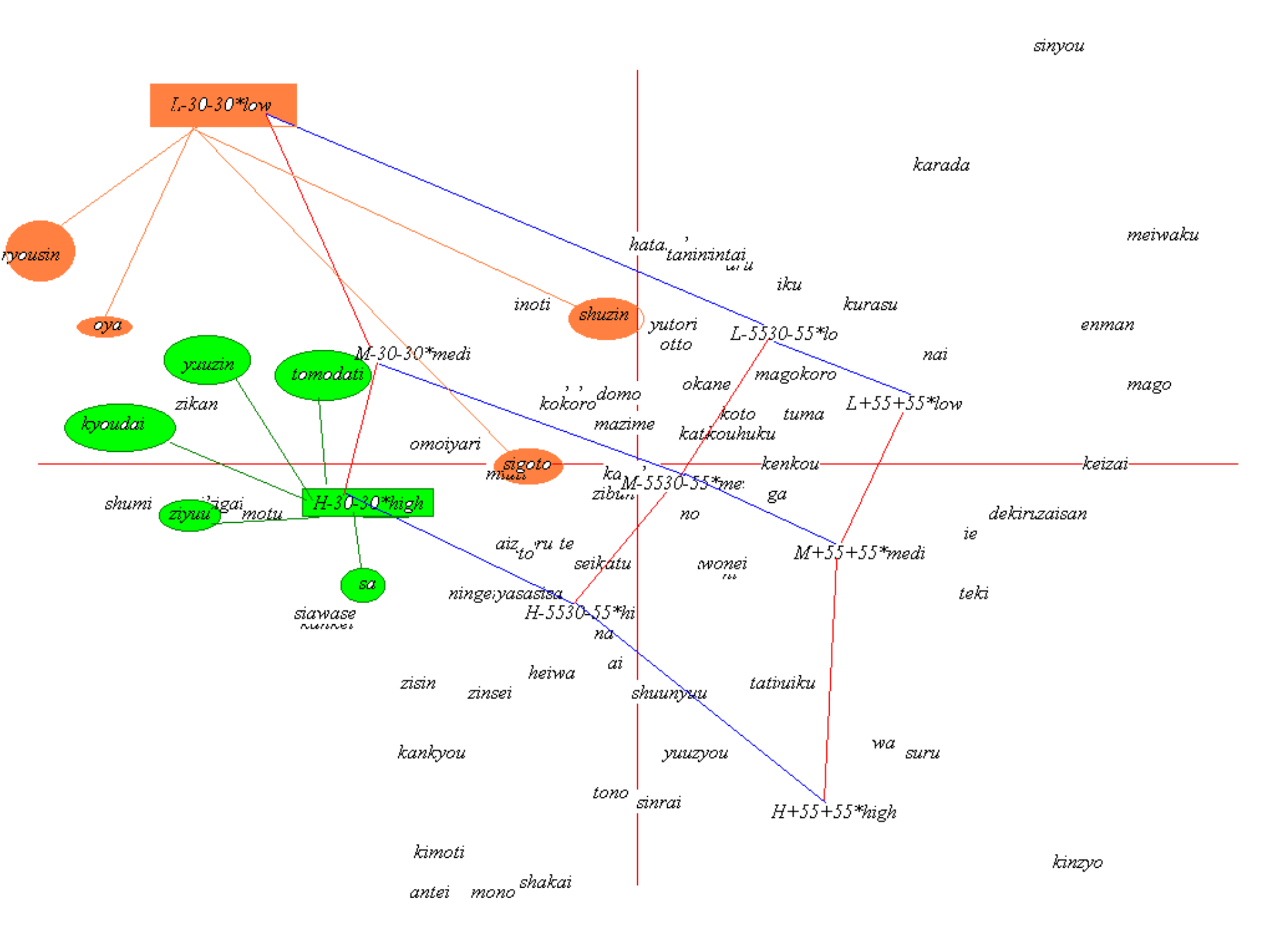
Category 7 Less than 30 years, high level of education

1.33	- 1	friends, friends, my homelife
1.12	- 2	being content having enough money to do what you want to do, within reason, having good friends, having a fulfilling job to do, having some idea of what you want to do and the freedom to choose, protection of the environment
1.05	- 3	to have good friends around having a good job, living in a good area, having lots of freedom to do the things you want to do
.93	- 4	good living education, good job, money

Category 9 Over 55 years, high level of education

.97	- 1	togetherness, peace of mind, good health, religion, no
.64	- 2	not to die, peace of mind, don't like people living envious of each other
.63	- 3	peace of mind good health, happiness, enough money to keep a standard of living
.38	- 4	welfare of my family work, satisfaction, good health, travel


sinyou
L-30-30*low
karada
meiwaku
ryousin
hata,
taninintai
iku
inoti
shuzin
kurasu
oya
yutori
otto
L-5530-55*lo
enman
M-30-30*medi
nai
yuuzin
tomodati
magokoro
okane
mago
zikan
kokoro
domo
L+55+55*low
kyoudai
mazime
koto
tuma
omoiyari
kenkou
keizai
sigoto
M-5530-55*me
ga
shumi
ziyuu
motu
H-30-30*high
no
dekiruzaisan
ie
aiz
to
ru te
seikatu
wonei
M+55+55*medi
sa
ninge;yasasisa
teki
siawase
H-5530-55*hi
na
ai
heiwa
zisin
zinsei
shuunyuu
tatiuiku
kankyou
yuuzyou
wa
suru
tono
sinrai
H+55+55*high
kimoti
kinzyo
antei
mono
shakai


sinyou
L-30-30*low
karada
yousin
meiwaku
hata
taninintai
iku
inoti
shuzin
kurasu
oya
yutori
otto
L-5530-55*lo
enman
yuuzin
M-30-30*medi
nai
tomodati
okane
magokoro
mago
zikan
kokoro
domo
L+55+55*low
kyoudai
mazime
koto
tuma
katkouhuku
omoiyari
sigoto
kenkou
keizai
ka
M-5530-55*me
ga
shumi
ziyuu
motu
H-30-30*high
no
dekirizaisan
ie
aiz
to
ru te
seikatu
wonei
M+55+55*medi
sa
ninge
yasasisa
teki
siawase
H-5530-55*hi
na
ai
heiwa
zisin
zinsei
shuunyuu
tatiuiku
kankyou
yuuzyou
wa
suru
tono
sinrai
H+55+55*high
kimoti
kinzyo
antei
mono
shakai

El “Text Mining” y el Análisis Exploratorio Multidimensional de Textos

- **« paradigma » inicial** (Benzécri, 1970):
- Extraer unidades de los textos
- Enriquecer la « lexicometria » con el « multidimensional»
- Aplicar las herramientas de visualisación a las tablas léxicas
- **evolución y diversificación de los métodos**

Evolución y diversificación de los métodos

- - Validez de las visualisacions (Bootstrap) (Demo)
- - Meta-data, nuevas variables.
- - Semantica.

Validez de los resultados

Algunos problemas planteados por el *Bootstrap*

Valores propios : poco interesantes...

Vectores propios : interversiones, rotaciones, cambios de direcciones de los ejes.

➡ « parcial Bootstrap »

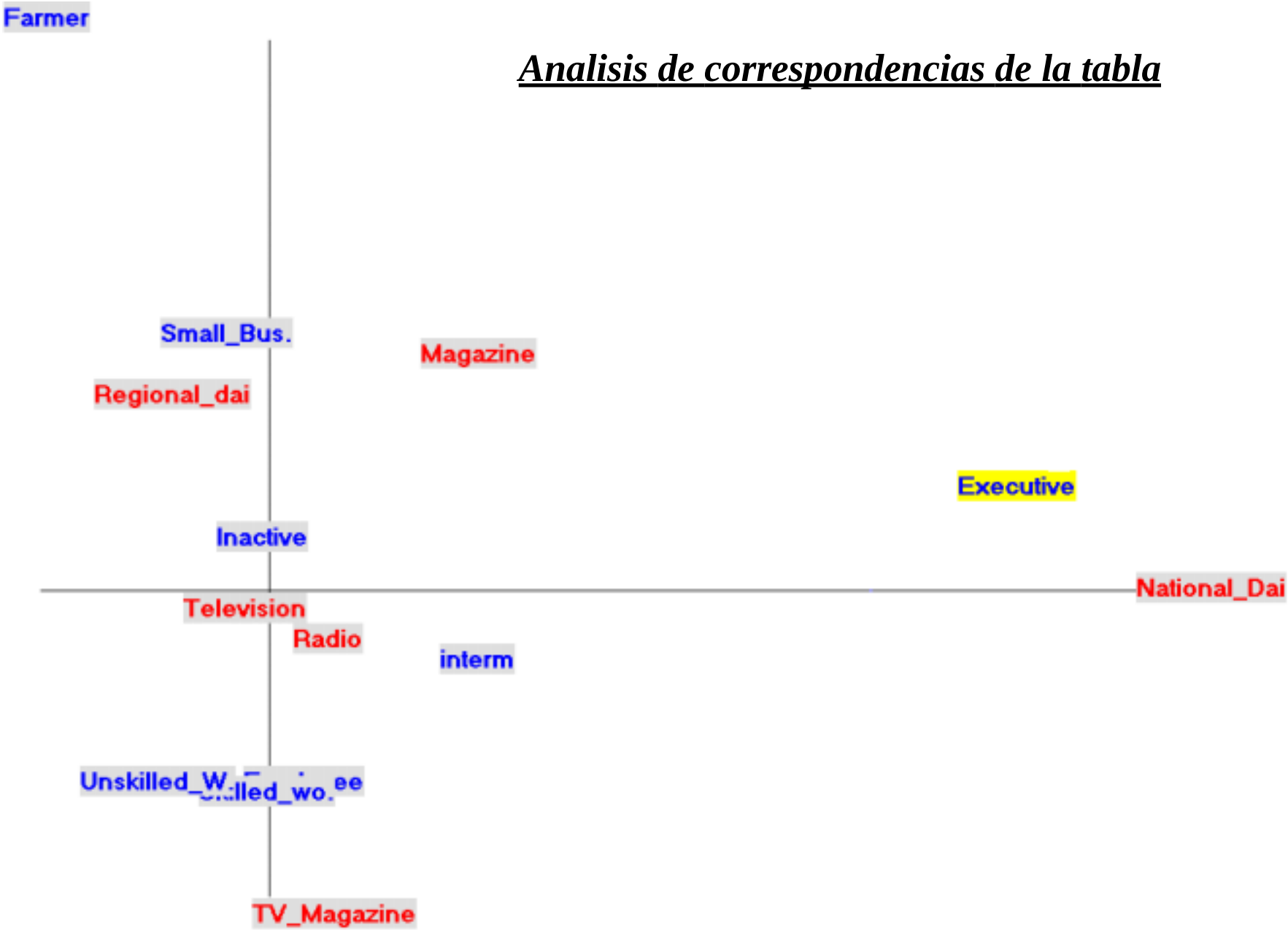
Ambigüedad de las frecuencias

Repaso : el bootstrap, Ejemplo : zonas de confianza

- "Tabla de contingencia" (Cross-tabulation) (CESP Multi-Media Survey, 1993).
- En cada celda: numero de contactos - media (el dia antes) .
- Columnas : **Media** [Radio, TV, National & Regional Daily N., Magazines].
- Filas : **Occupation groups**.

	Radio	Tele	Nat.	Reg.	Maga	TV_Mag
Farmer	96.	118.	2.	71.	50.	17.
Small Business	122.	136.	11.	76.	49.	41.
Executive	193.	184.	74.	63.	103.	79.
Intermediate	360.	365.	63.	145.	141.	184.
Employee	511.	593.	57.	217.	172.	306.
Skilled worker	385.	457.	42.	174.	104.	220.
Unskilled worker	156.	185.	8.	69.	42.	85.
Housewives, Ret.	1474.	1931.	181.	852.	642.	782.

Analisis de correspondencias de la tabla

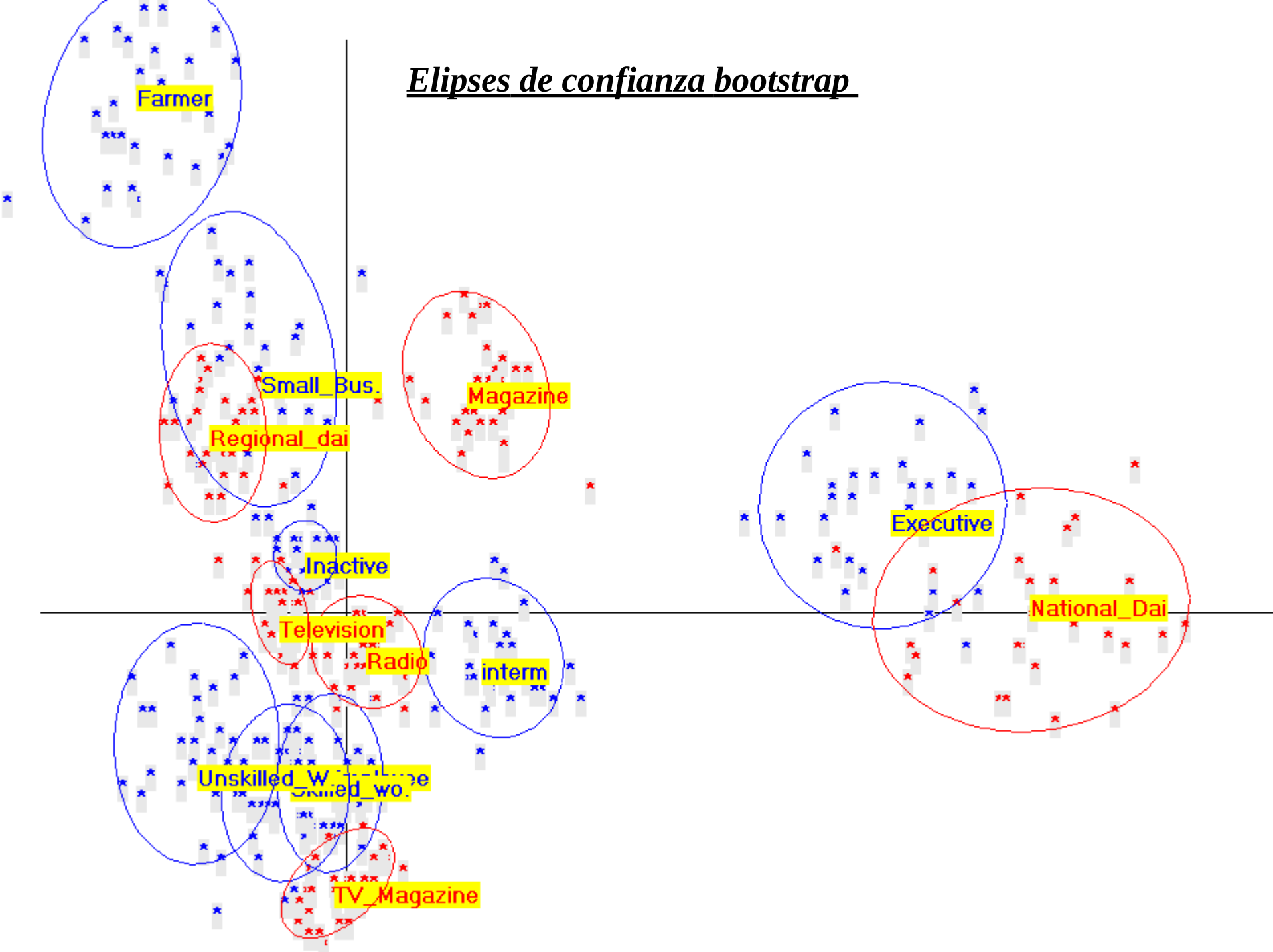


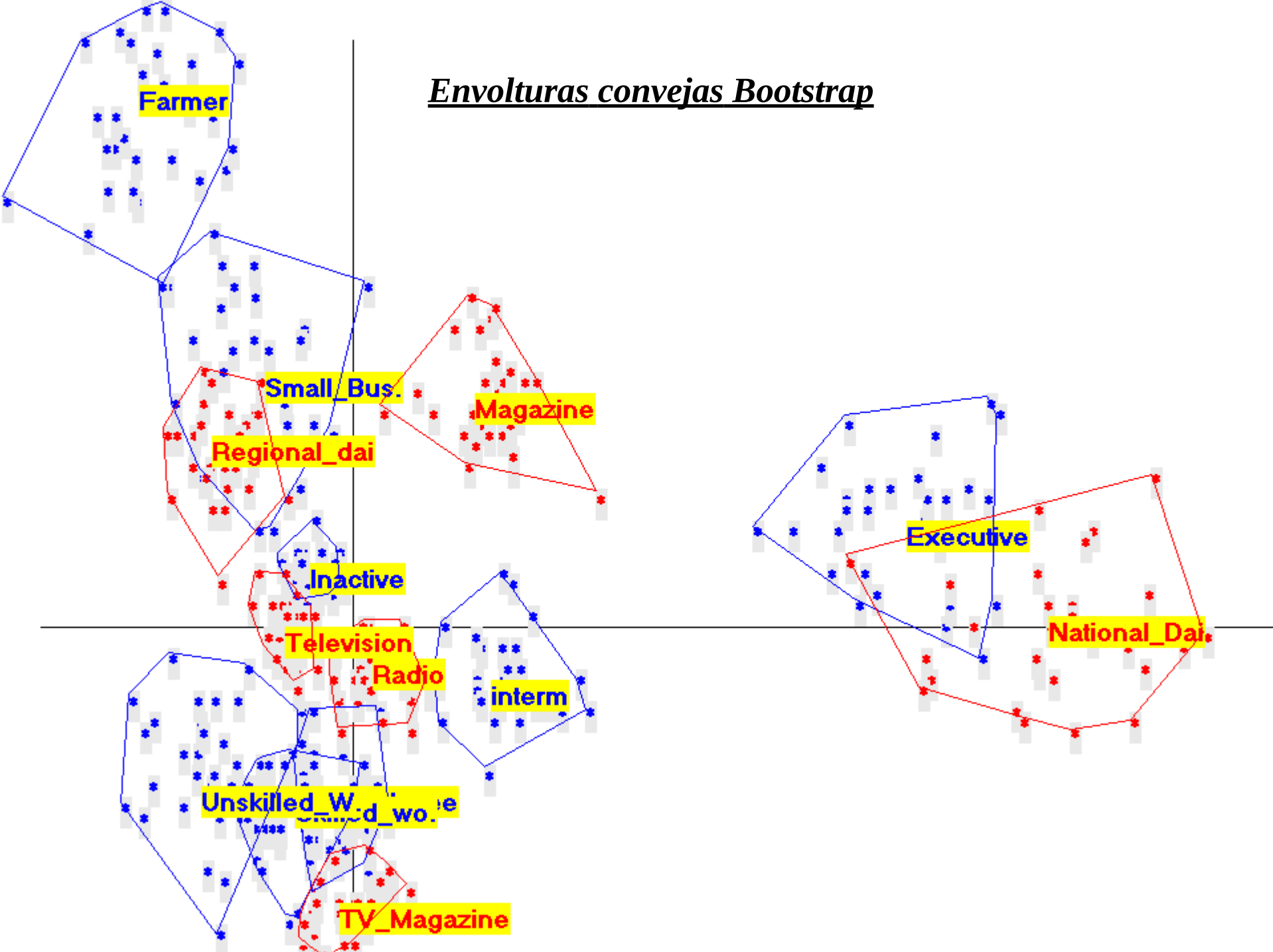
Repaso sobre el bootstrap, Ejemplo : zonas de confianza

- Ejemplo de una replicación de la tabla

	Radio	Tele	Nat.	Reg.	Maga	TV_M
Farmer	109.	120.	1.	78.	48.	20.
Small Business	126.	142.	8.	76.	53.	30.
Executive	196.	181.	80.	77.	109.	72.
Intermediate	384.	365.	60.	133.	138.	203.
Employee	514.	596.	59.	228.	172.	316.
Skilled worker	378.	467.	33.	171.	100.	223.
Unskilled worker	169.	188.	8.	79.	38.	81.
Housewives, Ret.	1519.	1961.	158.	893.	632.	764.

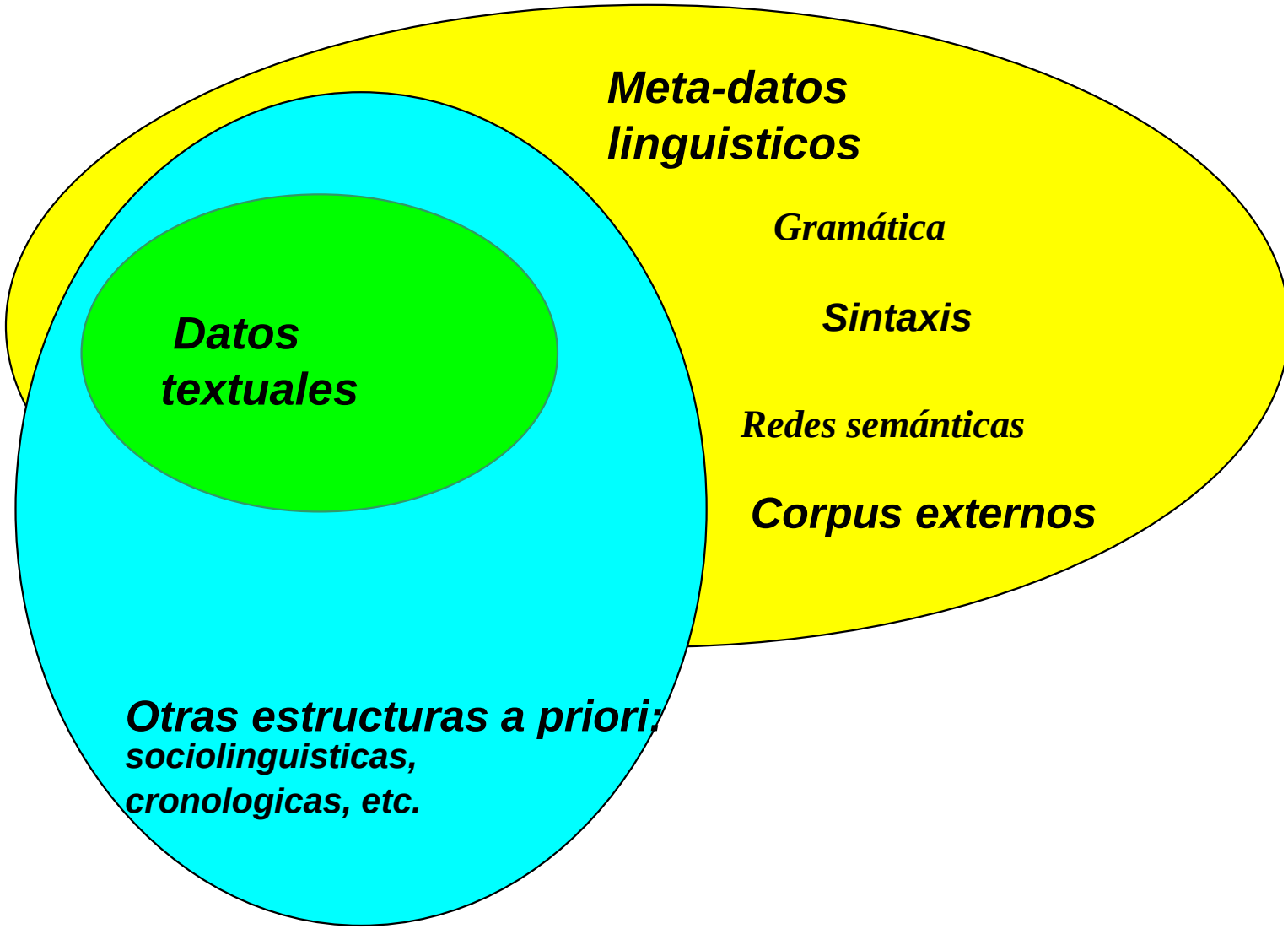
Elipses de confianza bootstrap




Envolturas convejas Bootstrap
Farmer
Small_Bus.
Magazine
Regional_dai
Executive
Inactive
Television
Radio
interm
National_Dai
Unskilled_W
TV_Magazine

Ambigüedad de las frecuencias

Meta-información



Las cuatro fases de un análisis lingüístico

(A bxg flower)

Morfología

A big flower A bug flower

A bag flower A bog flower

Sintaxis

The spoon speaks

(The speaks)

Semántica

A woman thinks

(A stone thinks)

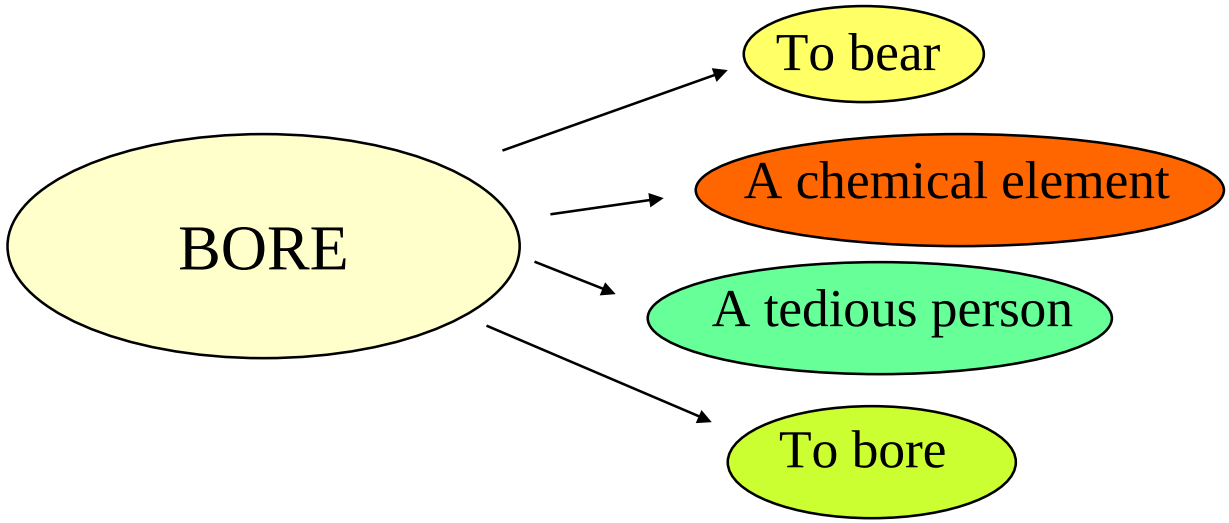
Pragmática

Un desafío (reto) para la I.A.

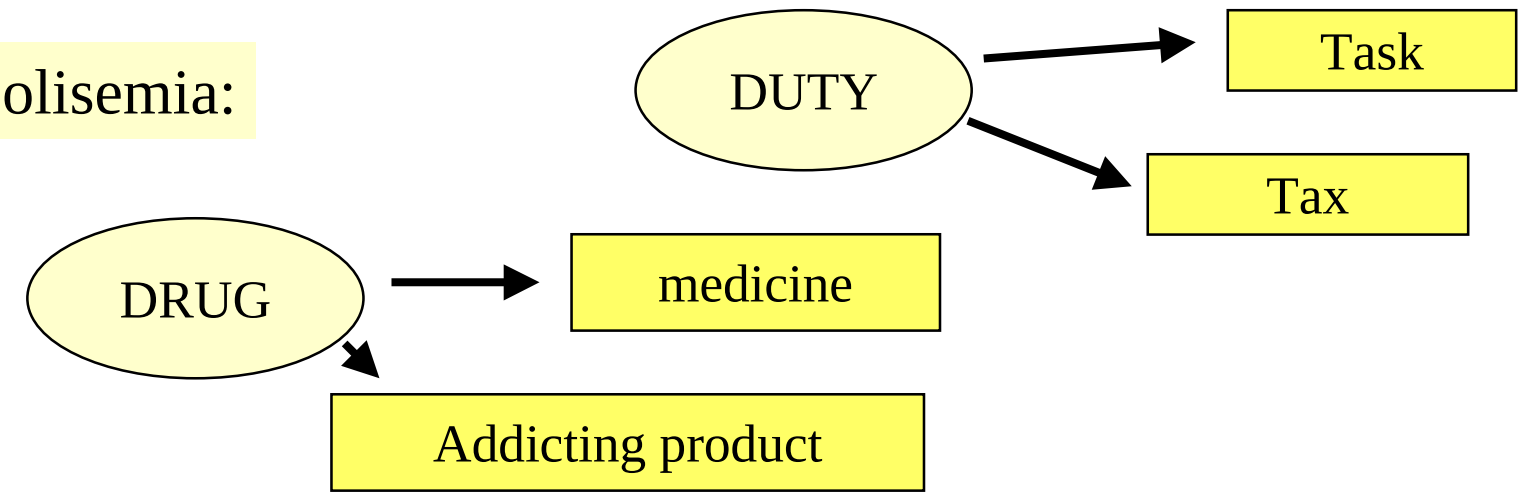
[Ejemplo del alumno]

Homógrafia, Polisemía, Sinonimía

Homógrafos:



Polisemia:



Etiquetado de las ocurrencias del corpus *"Life question"*

Gender	*Educ.*	*Age*	*Tagged responses*
1	1	4	happiness/NN in/IN people/NNS around/IN me/PRP ,/, contented/VBN family/NN ,/, would/MD make/VB me/PRP happy/JJ
1	2	2	my/PRP$ own/JJ time/NN ,/, not/RB dictated/VBN by/IN other/JJ people/NNS
1	2	2	freedom/NN of/IN choice/NN as/IN to/TO what/WP I/PRP do/VB in/IN my/PRP$ leisure/NN time/NN
1	3	2	I/PRP suppose/VBP work/NN
1	2	1	firm/NN ,/, my/PRP$ work/NN ,/, which/WDT is/VBZ my/PRP$ dad's/NNS firm/NN
2	1	6	just/RB the/DT memory/NN of/IN my/PRP$ last/JJ husband/NN
2	2	6	wellbeing/NN of/IN my/PRP$ handicapped/JJ son/NN
1	1	5	my/PRP$ wife/NN ,/, she/PRP gave/VBD me/PRP courage/NN to/TO carry/VB on/IN even/RB in/IN the/DT bad/JJ times/NNS

Nuevas variables (categorias gramaticales)

Substantivos

Verbos

Adjectivos

Pronombres

Determinantes

Adverbios

Preposiciones

Conjunciones

El contenido semántico de un perfil léxico:

Lingüística distributional
(Z. Harris)

X tiene una cola

X tiene bigotes

a X le gusta la leche

a X le gusta el pescado

X tiene miedo del agua

a X le encanta cazar los ratones

La semantica

La similaridad semantica no es una relacion transitiva.

Ejemplos de cadenas semanticas :

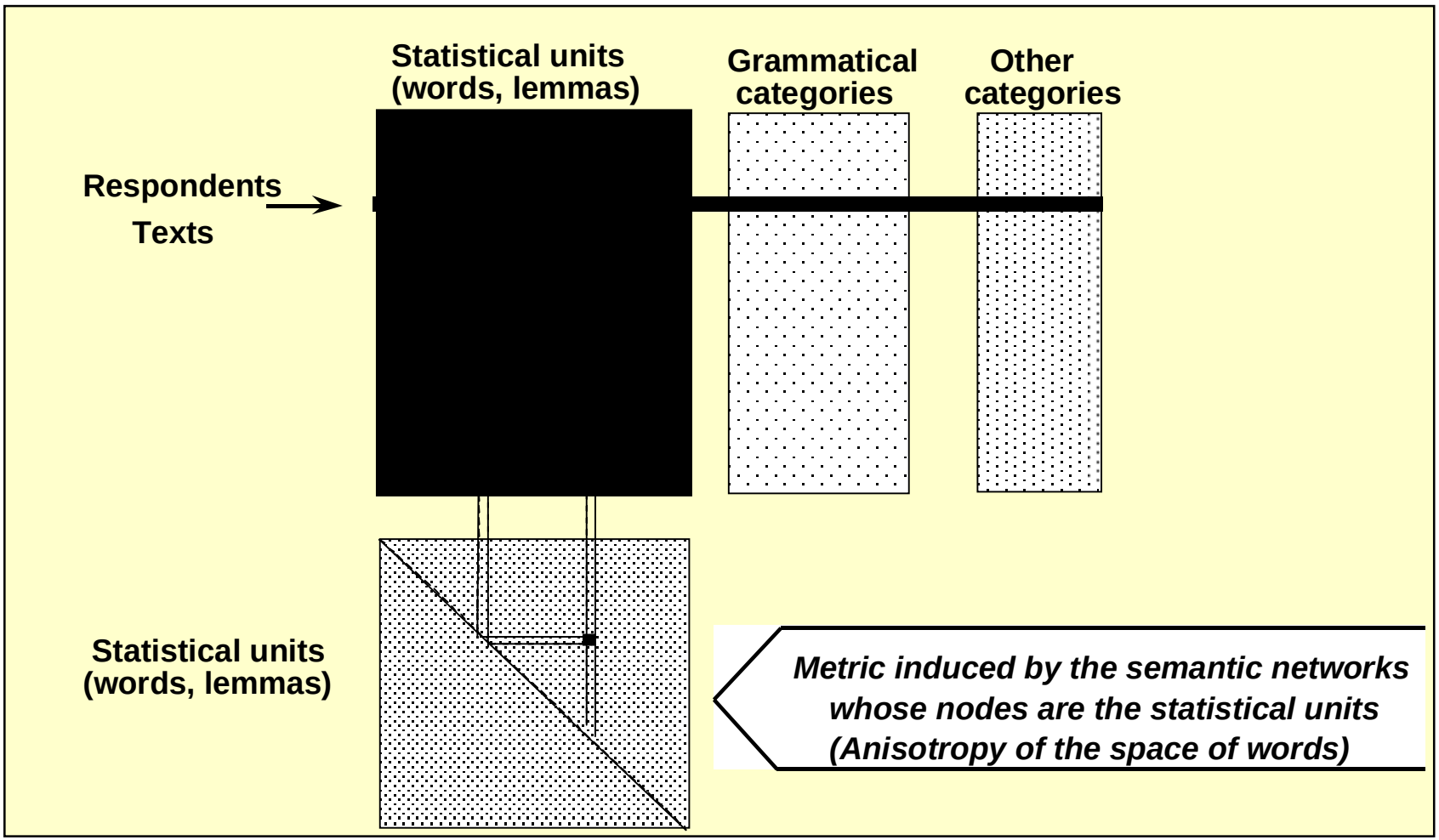
(1) *calm–wisdom–discretion–wariness–fear–panic*,

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

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

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

Nuevas variables, nuevas métricas



Procedimiento de las respuestas a preguntas abiertas

Estrategia de análisis

- Agrupamiento *a priori*
- Uso de particiones instrumentales
- Análisis directo de la tabla palabras - respuestas
- Yuxtaposiciones de agrupamientos

▪ Uso de particiones instrumentales

Se puede buscar una partición que sea la más universal posible teniendo en cuenta el tamaño de la muestra:
Es el principio que rige la elaboración de las *situaciones-tipo* (o *núcleos factuales*, *o tambien partición instrumental*).

Las principales características consideradas relevantes en función del objetivo (por ejemplo: *edad*, *categoría socio-profesional*, *sexo*, *nivel de instrucción*, *región*) se sintetizan en una partición única mediante una técnica de clasificación automática.

Esto equivale a sustituir uno o varios millares de individuos por una treintena de grupos lo más homogéneos posibles en cuanto a los criterios precitados.

- Análisis directo de la tabla palabras - respuestas

Se puede por el contrario, obtener una tipología directa, sin reagrupamiento previo, de las respuestas a partir de sus perfiles léxicos.

Esto tiene sentido únicamente cuando las respuestas no se reducen a 2 ó 3 formas.

Podemos después seleccionar las categorías que presentan un mayor grado de asociación con dicha tipología y utilizar éstas categorías para reagrupar las respuestas.

Es más una manera de detectar las categorias vinculadas a las respuestas que un analisis completo.

▪ Análisis directo de la tabla palabras - respuestas (2)

Se puede encontrar respuestas del tipo

- *Desarrollar el uso de transportes públicos*

o bien: - *Incitar a la gente a utilizar lo más que pueda*

los trenes, los autobuses

que son dos respuestas teniendo respectivamente 6 y 13 ocurrencias, sin ninguna palabra común, y cuyos contenidos son bastante vecinos

A la inversa, las dos respuestas siguientes a la misma pregunta:

- *respetar los límites de velocidad*
- *hacer respetar los límites de velocidad*

no se distinguen sino por una sola palabra y tienen, no obstante, contenidos sensiblemente distintos.

¿ Palabras o lemas ?

Dos grandes opciones: trabajar a partir del recuento de las formas gráficas (palabras) o a partir del recuento de los lemas.

La pertinencia de la elección depende en gran medida del dominio de aplicación y de los objetivos del estudio.

Por ejemplo, un investigador que explora un conjunto de artículos reunidos en una base de datos del dominio químico pedirá que se reagrupen el singular del sustantivo *ácido* con su plural *ácidos* para poder acceder a todos los textos que contengan una u otra de estas dos palabras en una misma búsqueda.

¿ Palabras o lemas ? (2)

Pero, por el contrario, cuando se trata de estudiar textos políticos, los investigadores han constatado que el singular y el plural de un mismo sustantivo aluden a nociones diferentes, incluso opuestas.
(Ejemplo, la oposición entre *defensa de la libertad/ defensa de las libertades*, que remiten a corrientes políticas distintas).

En este caso, será preferible indexar separadamente las dos palabras e incluir ambas en el análisis como unidades distintas.

Thank You

Gracias

Grazie

Obrigado

Merci