ANÁLISIS ESTADÍSTICO DE DATOS DE ENCUESTAS. TRATAMIENTO CONEXO DE RESPUESTAS A PREGUNTAS ABIERTAS Y CERRADAS

Mónica Bécue Bertaut
Departament d'Estadística i Investigació Operativa
Facultat d'Informàtica de Barcelona (UPC)

Resumen

El sociólogo se encuentra muchas veces enfrentado , en la recogida de datos con información textual, sea con las preguntas abiertas de una gran encuesta, sea con entrevistas, sea con otros textos de fuentes de datos secundarios. Tanto con fines exploratorios o de clasificación previa, como a la hora de la comprobación de determinadas hipótesis el método, y el programa informático SPAD.T correspondiente, constituye un útil importante para el tratamiento de textos. En particular permite confrontar los resultados obtenidos del análisis estadístico de textos con otras variables estructurales provenientes de las grandes encuestas interviniendo como variables ilustrativas.

Resum

El sociòleg es troba moltes vegades enfrontat, a la recollida de les dades amb informació textual, bé sigui amb les preguntes obertes d'una gran enquesta, bé sigui amb entrevistes, bé sigui amb altres textos de fonts de dades secundàries. Tant amb finalitat exploratòries o de classificació prèvia com a l'hora de la comprovació de determinades hipòtesis, el mètode i el programa informàtic SPAD.T corresponent, constitueixen una eina important pel tractament de textos. Especialment, permet confrontar els resultats obtinguts de l'anàlisi estadística de textos amb altres variables estructurals provinents de les grans enquestes intervenint com a variables il·lustratives.

Abstract

The sociologist can be confronted with textual information, during data gathering, in large surveys, in interviews or in other secondary sources. The SPAD.T is a useful and important tool in dealing with texts, in exploratory aims or in a priori classifications, as well as the verifying moment of the hypothesis of the method. In particular, it allows to confront results obtained from text statistical analysis and other structural variables coming from large surveys, being introduced as ilustrative variables.

Los métodos estadísticos multidimensionales de análisis de grandes tablas de datos permiten tratar las respuestas a preguntas cerradas y abiertas de forma simultánea.

Entre estos métodos, los de análisis de correspondencias, correspondencias simples y múltiples, cumplen un papel privilegiado. Métodos de descripción de las tablas de contingencia (o tablas cruzadas) y de ciertas tablas binarias (matriz de respuestas a cuestionarios, por ejemplo), proporcionan una representación gráfica de las asociaciones entre las líneas y las columnas.

Son métodos muy difundidos desde los trabajos de J. P Benzecri (Benzécri 73), J.P. Fénelon, L. Lebart, A. Morineau (Lebart, Morineau, Fénelon 84), Y. Escoufier (Escoufier 85), J.M. Bouroche, G. Saporta (Bouroche, Saporta 80), de Volle (Volle 80) que los han estudiado de forma sistemática como herramientas flexibles para el tratamiento de datos multidimensionales.

Los métodos de clasificación automática pueden complementar la descripción gráfica obtenida. La reagrupación de los individuos en clases homogéneas efectuadas permite simplificar y sintetizar las representaciones gráficas.

Más recientemente, J.P Benzecri (Benzecri 81), L. Lebart y A. Salem (Lebart, Salem 88) han introducido la aplicación de estos métodos en el terreno textual. Esta aplicación ofrece una nueva aproximación a los datos textuales. Es una aproximación esencialmente diferencial que describe los contrastes entre los textos o las respuestas de encuesta (respuestas individuales o grupos de respuestas) y que confronta las respuestas abiertas y cerradas.

El sistema informático *SPAD.T* (Lebart, Morineau, Bécue 89) automatiza la aplicación de estos métodos, integrándolos en un mismo entorno y constituye, así, una herramienta para el tratamiento comparativo de textos. Permite trabajar a partir de los datos brutos, sin precodificación ni manipulación previas y analizar conjuntamente respuestas abiertas y cerradas.

APORTACIÓN ESPECÍFICA DE LAS PREGUNTAS ABIERTAS

Las respuestas abiertas son, todavía, poco utilizadas. El análisis de la información que dichas respuestas proporcionan es a la vez complejo y costoso. No obstante, en ciertos casos es imposible sustituir una pregunta abierta por una pregunta cerrada ya que estos dos tipos de preguntas aportan informaciones de naturaleza muy distinta y, por lo tanto, difícilmente comparables.

Por esta y otras razones, el uso de preguntas abiertas puede ser indispensable. Facilita la exploración de dominios mal conocidos, para los cuales todas las posibles respuestas cerradas no son conocidas a priori. Reduce el tiempo de la entrevista, cuando una sola pregunta abierta sustituye a varias listas de ítems. Permite explicitar las respuestas a preguntas cerradas (con la cues-

tión complementaria ¿Por qué?) y, así, saber si los entrevistados han entendido los valores de la pregunta cerrada de la misma forma.

EJEMPLO DE APLICACIÓN

El ejemplo utilizado en este artículo proviene de una encuesta efectuada sobre 2000 individuos representativos de los franceses mayores de edad residentes en Francia continental (Lebart , Houzel 80). El cuestionario intenta cubrir varios dominios de estudio tales como la familia, las condiciones de vida y las aspiraciones de los franceses. Comporta 410 preguntas cerradas y 6 preguntas abiertas.

La pregunta abierta escogida en este ejemplo tiene el libelado siguiente: «Le nombre de divorces augmente actuellement en France; à quoi est due, selon vous, cette augmentation?».

Las respuestas abiertas de los 2.000 individuos se graban sobre un soporte magnético sin otra codificación previa que la marca de fin de respuesta.

UNIDADES DE TRATAMIENTO ESTADÍSTICO

El sistema SPAD.T considera dos unidades para el tratamiento estadístico del corpus: la *forma gráfica* y el *segmento repetido.*

La forma gráfica será la unidad de base. Viene definida como sucesión de caracteres no delimitadores (en general letras) comprendidos entre dos delimitadores (blancos y signos de puntuación). A una misma forma léxica (unidad de lengua definida en el diccionario) pueden corresponder varias formas gráficas —femenino y masculino de un mismo adjetivo, flexiones distintas de un verbo. Inversamente, una misma forma gráfica puede referirse a formas léxicas distintas.

La conservación del genero de un adjetivo, del tiempo de un verbo no sólo facilita la completa automatización del tratamiento sino que, sobre todo, mantiene una información no despreciable, la información contenida en la propia utilización del género, de un tiempo pasado en lugar del presente, etc.

La no identificación de homógrafos (*como* del verbo comer y *como* conjunción, por ejemplo) no es, en general, un inconveniente grave porque las palabras no están siendo tratadas de forma aislada. No obstante, el usuario puede utilizar la edición de concordancias —edición de todos los contextos de todas las ocurrencias de una forma— para distinguirlos y introducir esta información en la cadena de tratamiento.

La segunda unidad estadística considerada es el *segmento* de frase *repetido.*

Es una unidad de recuento compuesta por varias formas contiguas. Esta nueva unidad estadística, introducida por A. Salem (Salem 82, Salem 87), permite tomar en cuenta el contexto de las formas.

GLOSARIO DE FORMAS Y SEGMENTOS REPETIDOS

El tratamiento preliminar del corpus constituido por las 2000 respuestas a la pregunta abierta sobre el divorcio consiste en identificar estas unidades, contarlas y construir las tablas léxicas o segmentales que serán sometidas al análisis de correspondencias.

Se obtiene así una primera información sobre el corpus: su longitud y el número de formas distintas que lo componen, respectivamente 18.620 ocurrencias y 1.667 formas distintas en el ejemplo.

La tabla 1 representa las *159* formas empleadas al menos *15* veces en todo el corpus. La forma más frecuente es *de*, empleada *1.180* veces. La primera forma plena es *vie*, pronunciada *341* veces. Se puede notar que las formas *femme* y *femmes*, pronunciadas respectivamente *220* y *205* veces, aparecen como las tercera y cuarta formas significativas mientras que las formas *homme* y *hommes* se encuentran empleadas solamente *32* y *18* veces respectivamente.

En la tabla 2, se muestran los segmentos repetidos del corpus. Están seleccionados en función de umbrales de frecuencia distintos según la longitud del segmento y listados en orden lexicográfico.

Tabla 1

Glosario de las formas más frecuentes

DICTIONNAIRE DES MOTS

NUM.	MOTS EMPLOYÉS	FRÉQUENCES	LONGUEURS	NUM.	MOTS EMPLOYÉS	FRÉQUENCES	LONGUEURS
1	A	476	1	12	AVEC	19	4
2	ACTUELLE	67	8	13	BEAUCOUP	42	8
3	AMOUR	21	5	14	C	125	1
4	ARGENT	65	6	15	CA	33	2
5	ASSEZ	69	5	16	CAUSE	46	5
6	AU	145	2	17	CE	29	2
7	AUCUNE	15	6	18	CELA	20	4
8	AUSSI	17	5	19	CHACUN	40	6
9	AUTRE	40	5	20	CHANGE	15	6
10	AUX	38	3	21	CHANGEMENT	28	10
11	AVANT	76	5	22	CHÔMAGE	55	7

NUM.	MOTS EMPLOYÉS	FRÉQUENCES	LONGUEURS
23	COMME	35	5
24	CONCESSIONS	24	11
25	CONDITIONS	25	10
26	CONTRAINTES	16	11
27	COUPLE	74	6
28	COUPLES	50	7
29	D	269	1
30	DANS	101	4
31	DE	1180	2
32	DES	407	3
33	DEUX	37	4
34	DIFFICILE	27	9
35	DIFFICULTÉS	33	11
36	DIVORCE	63	7
37	DIVORCER	45	8
38	DONC	19	4
39	DU	143	2
40	EFFORT	21	6
41	EGOISME	21	7
42	ELLE	35	4
43	ELLES	34	5
44	EMANCIPATION	23	12
45	EN	137	2
46	ENFANTS	19	7
47	ENGAGEMENT	25	10
48	ENSEMBLE	31	8
49	ENTENDENT	16	9
50	ENTENTE	19	7
51	ENTRE	34	5
52	ÉPOUX	15	5
53	EST	287	3
54	ET	305	2
55	ÊTRE	43	4
56	ÉVOLUTION	70	9
57	FACILE	62	6
58	FACILEMENT	17	10
59	FACILITÉ	52	8
60	FAIRE	29	5
61	FAIT	68	4
62	FAMILLE	23	7
63	FEMME	220	5
64	FEMMES	205	6
65	FINANCIÈRE	39	10
66	FINANCIERS	17	10
67	FONT	25	4
68	GENS	228	4
69	GRANDE	36	6
70	HOMME	32	5
71	HOMMES	18	6
72	IL	97	2
73	ILS	142	3
74	INCOMPATIBILITÉ	15	15
75	INDÉPENDANCE	107	12
76	INDÉPENDANTE	17	12
77	INDÉPENDANTES	27	13
78	JE	53	2
79	JEUNE	36	5
80	JEUNES	146	6
81	L	322	1
82	LA	760	2
83	LE	328	2
84	LES	585	3
85	LEUR	49	4
86	LIBÉRATION	44	10
87	LIBERTÉ	166	7
88	LIBRES	28	6
89	MAINTENANT	27	10
90	MA	18	3
91	MANQUE	190	6
92	MARIAGE	123	7
93	MARIAGES	27	8
94	MARIÉ	28	5
95	MARIENT	88	7
96	MARIER	17	6
97	MAUVAISE	17	8
98	MÊME	25	4
99	MESENTENTE	34	10
100	MODE	33	4
101	MODERNE	32	7
102	MOEURS	126	6
103	MOINS	133	5
104	N	128	1
105	NE	282	2
106	NON	20	3
107	ON	177	2
108	ONT	77	3
109	OU	36	2
110	PAR	32	3
111	PARCE	96	5
112	PAS	284	3
113	PATIENCE	24	8
114	PERSONNE	15	8

NUM.	MOTS EMPLOYÉS	FRÉQUENCES	LONGUEURS	NUM.	MOTS EMPLOYÉS	FRÉQUENCES	LONGUEURS
115	PEUT	33	4	138	SON	40	3
116	PLUS	533	4	139	SONT	122	4
117	POUR	94	4	140	SOUVENT	21	7
118	PROBLEME	20	8	141	SUR	21	3
119	PROBLEMES	34	9	142	TEMPS	31	5
120	QU	150	2	143	TOT	21	3
121	QUAND	21	5	144	TOUT	40	4
122	QUE	169	3	145	TRAVAIL	99	7
123	QUI	114	3	146	TRAVAILLE	39	9
124	REFLECHIR	19	9	147	TRAVAILLENT	67	11
125	RESPECT	17	7	148	TROP	330	4
126	RESPONSABILITÉS	21	15	149	UN	117	2
127	RIEN	20	4	150	UNE	86	3
128	RYTHME	16	6	151	VA	20	2
129	S	97	1	152	VALEURS	20	7
130	SA	25	2	153	VEULENT	46	7
131	SAIS	22	4	154	VEUT	38	4
132	SAIT	31	4	155	VIE	341	3
133	SANS	52	4	156	VIS	18	3
134	SE	280	2	157	VITE	43	4
135	SENS	17	4	158	VIVRE	47	5
136	SEXUELLE	16	8	159	Y	83	1
137	SOCIETÉ	53	7				

Tabla 2

Segmentos repetidos del corpus

TABLEAU DES SEGMENTS REPETÉS

SEUILS MINIMUM DE FRÉQUENCE DE RÉPÉTITION:

SEUIL GENERAL	9
SEGMENTS DE LONGUEUR 2	99
SEGMENTS DE LONGUEUR 3	19

SEG.	FREQ.	LONG.	TEXTE DU SEGMENT
			A
1	13	4	A CAUSE DE LA
2	101	2	A LA
3	25	3	A LAVIE
4	20	3	A PLUS DE
			AU
5	21	3	AU FAIT QUE
6	16	4	AU FAIT QUE LES
			C
7	121	2	C EST
8	22	3	C EST PLUS
9	15	4	C EST PLUS FACILE
			DANS
10	20	3	DANS LE COUPLE

SEG.	FREQ.	LONG.	TEXTE DU SEGMENT
			DE
11	254	2	DE LA
12	119	3	DE LA FEMME
13	58	3	DE LA VIE
14	12	4	DE LA VIE ACTUELLE
15	19	4	DE MOINS EN MOINS
16	48	4	DE PLUS EN PLUS
			EMANCIPATION
17	13	4	EMANCIPATION DE LA FEMME
			EST
18	24	3	EST PLUS FACILE
19	2	4	EST PLUS FACILE DE
			EVOLUTION
20	20	3	ÉVOLUTION DE LA
21	23	3	ÉVOLUTION DES MOEURS
			FAIT
22	23	3	FAIT QUE LES
23	10	4	FAIT QUE LES FEMMES
			FEMMES
24	20	3	FEMMES QUI TRAVAILLENT
			GENS
25	38	3	GENS SE MARIENT
26	21	4	GENS SE MARIENT TROP
27	0	5	GENS SE MARIENT TROP JEUNES
			IL
28	21	3	IL N Y
29	20	4	IL N Y A
30	14	5	IL N Y A PLUS
31	10	6	IL N Y A PLUS DE
32	43	3	IL Y A
			ILS
33	22	3	ILS SE MARIENT
34	15	4	ILS SE MARIENT TROP
35	12	5	ILS SE MARIENT TROP JEUNES
			INDEPENDANCE
36	13	4	INDÉPENDANCE DE LA FEMME
37	11	4	INDÉPENDANCE FINANCIÈRE DES FEMMES
			JE
38	13	4	JE NE SAIS PAS
			JEUNES
39	12	4	JEUNES SE MARIENT TROP
			LA
40	209	2	LA FEMME
41	20	3	LA FEMME TRAVAILLE
42	11	5	LA LIBERATION DE LA FEMME
43	189	2	LA VIE
44	34	3	LA VIE ACTUELLE
45	21	3	LA VIE MODERNE
			LE
46	12	4	LE MANQUE D ARGENT
47	13	4	LE MARIAGE N EST
48	12	4	LE TRAVAIL DES FEMMES
			LES
49	101	2	LES FEMMES
50	22	3	LES FEMMES SONT
51	10	4	LES FEMMES SONT PLUS
52	188	2	LES GENS
53	48	3	LES GENS NE
54	49	3	LES GENS SE
55	37	4	LES GENS SE MARIENT
56	20	5	LES GENS SE MARIENT TROP
57	24	3	LES GENS SONT
58	15	4	LES JEUNES SE MARIENT
59	11	5	LES JEUNES SE MARIENT TROP
			LIBÉRATION
60	22	4	LIBÉRATION DE LA FEMME
			LIBERTÉ
61	16	4	LIBERTÉ DE LA FEMME
62	24	3	LIBERTÉ DES MOEURS
			MANQUE
63	28	3	MANQUE D ARGENT
64	124	2	MANQUE DE
			MARIAGE
65	10	4	MARIAGE N EST PLUS
			MODE
66	24	3	MODE DE VIE
			N
67	23	3	N EST PLUS
			NE
68	10	4	NE S ENTENDENT PAS
69	30	3	NE SAIT PAS
			PARCE
70	26	3	PARCE QUE LES
71	13	4	PARCE QUE LES GENS
			PLUS
72	49	3	PLUS EN PLUS
73	13	4	PLUS FACILE DE DIVORCER
			QU

SEG.	FREQ.	LONG.	TEXTE DU SEGMENT
74	14	4	QU IL Y A
			----QUE
75	12	4	QUE LA FEMME TRAVAILLE
76	23	3	QUE LES FEMMES
77	24	3	QUE LES GENS
78	10	4	QUE LES GENS NE
			----SE
79	55	3	SE MARIENT TROP
80	31	4	SE MARIENT TROP JEUNES
81	14	4	SE MARIENT TROP VITE
			----TRAVAIL
82	13	4	TRAVAIL DE LA FEMME
83	29	3	TRAVAIL DES FEMMES
			----TROP
84	26	3	TROP DE LIBERTÉ
			----Y
85	21	3	Y A PLUS
86	16	4	Y A PLUS DE

TABLAS LÉXICAS Y SEGMENTALES

Para aplicar el análisis de correspondencias a las respuestas abiertas, se construyen tablas de contingencia particulares:

1. La tabla léxica contiene la frecuencia con la cual una forma gráfica es empleada por cada uno de los individuos. El análisis de correspondencias, aplicado a esta tabla de frecuencias, llamada tabla léxica, procede por comparación de las distribuciones de las formas en los individuos, es decir compara los perfiles léxicos de los individuos.
2. Si existen una o varias particiones pertinentes del corpus —partición del corpus en grupos de respuestas según la clase de edad del individuo, según el sexo, etc.— se puede construir la tabla de contingencia que contiene la frecuencia de cada forma en cada parte del corpus. Esta tabla se llama tabla léxica agregada.
3. Tablas similares se obtienen sustituyendo las formas por los segmentos repetidos.

ANÁLISIS DE LA TABLA LÉXICA Y ASOCIACIÓN ENTRE EL VOCABULARIO Y LAS CARACTERÍSTICAS DE LOS INDIVIDUOS

En una tabla de contingencia, las filas y las columnas representan dos particiones de una misma población y ambas particiones juegan un papel análogo: para analizar el contenido de la tabla tiene sentido considerar tanto la nube de puntos-fila como la nube de puntos-columna. El análisis de correspondencias ofrece una representación gráfica conjunta de ambas; para ello efectúa la proyección de las nubes sobre subespacios de dimensión reducida pero manteniendo la máxima dispersión posible.

El análisis de correspondencias de la tabla Respuestas*Formas proporciona una visualización de la dispersión del vocabulario sobre los primeros ejes. Dos formas próximas habrán sido pronunciadas frecuentemente por los mismos individuos. Las formas alejadas del centro de gravedad, que están en la periferia sobre las gráficas de los planos factoriales, son formas cuyo empleo o cuya frecuencia de empleo diferencian a los individuos. Se puede, así, detectar asociaciones entre formas. En el gráfico 1, se representa el plano factorial principal del análisis de la tabla léxica.

El análisis de dicho gráfico presenta rasgos particulares: las respuestas cortas se distinguen más por la presencia o ausencia de formas gráficas que por la diferencia de sus perfiles de frecuencia. Esto hace que las distancias interindividuos sean difíciles de interpretar. Además, la información se reparte sobre numerosos ejes, lo que dificulta su aprehensión global. Se puede decir que, en este primer análisis, se reagrupan e interpretan las respuestas idénticas o similares repetidas con una cierta frecuencia, dejando para otro tipo de análisis las respuestas más originales. Se trata de efectuar un trabajo preparatorio, encaminado a establecer un criterio de agrupamiento de las respuestas.

Este tratamiento exploratorio puede ser completado y guiado con la utilización de dos informaciones suplementarias: las respuestas a las preguntas cerradas y los segmentos repetidos. La primera proporciona una herramienta poderosa para detectar relaciones entre las características de los individuos y su lenguaje; la segunda contextualiza el empleo de las formas y precisa los argumentos empleados por los individuos y como son expresados.

En el ejemplo tratado, se proyectan sobre las gráficas factoriales las modalidades de 8 preguntas cerradas (véase gráfico 3). Dichas modalidades son consideradas columnas suplementarias del análisis anterior. Los indicadores estadísticos, llamados valores-test, calculados por SPAD.T miden, en desviaciones-tipo, cuán lejos del centro de gravedad se situa una modalidad sobre un eje dado: dicho valor-test está normado de tal forma que se puede leer como una realización de una variable norma centrada y reducida, bajo la hipótesis de repartición al azar de las modalidades sobre el eje. Por lo tanto, se considera relacionada con el eje una modalidad cuyo valor-test asociado es mayor que 1.96 o menor que -1.96. En efecto, bajo la hipótesis de repartición aleatoria de las modalidades, la probabilidad de que el valor-test esté entre estos dos valores es del 95% (véase tabla 3).

Los gráficos 2 y 3 muestran el posicionamiento de las características de los individuos y de los segmentos repetidos sobre el plano factorial del gráfico 1. Formas, segmentos y características individuales son puntos de un mismo espacio, lo que legitima interpretar la proximidad entre dos puntos. La lectura simultánea de las tres figuras permite ver las características de los individuos que emplean un cierto argumento, con qué palabras y con qué cons-

Tabla 3

Coordenadas y Valores-test de modalidades sobre los ejes factoriales

COORDOONNÉES ET VALEURS-TEST DES MODALITÉS SUR LES AXES 1 A 3

NOTE : LES DISTANCES A L'ORIGINE SONT DIVISÉES PAR 10.

MODALITÉS					COORDONNÉES					VALEURS-TEST				
IDEN -	LIBELLE	EFF.	P.ABS	DISTO	1	2	3	0	0	1	2	3	0	0
4 .	*Sexe de l'enquété*													
AD01 -	masculin	933	6344.00	0.13	-0.03	0.04	0.01	0.00	0.00	-3.6	4.8	1.5	0.0	0.0
AD02 -	feminin	1067	8022.00	0.08	0.03	-0.04	-0.01	0.00	0.00	3.6	-4.8	-1.5	0.0	0.0
10 .	*A2a: avez-vous eu des enfants?*													
AI01 -	oui	1378	9881.00	0.05	-0.01	-0.01	-0.01	0.00	0.00	-1.3	-1.3	-1.6	0.0	0.0
AI02 -	non	622	4485.00	0.22	0.02	0.02	0.02	0.00	0.00	1.3	1.3	1.6	0.0	0.0
15 .	*C1: la famille est le seul endroit où l'on se sent bien*													
AL01 -	fam. endroit bien	1225	8617.00	0.07	0.08	-0.02	0.02	0.00	0.00	11.3	-2.9	3.5	0.0	0.0
AL02 -	famm. non endroit bien	774	5748.00	0.15	-0.12	0.03	-0.04	0.00	0.00	-11.3	3.0	-3.5	0.0	0.0
AL03 -	famille n.s.p.	1	1.00	1436.50	-1.10	-0.42	0.50	0.00	0.00	-1.1	-0.4	-0.5	0.0	0.0
16 .	*C2: opinion sur le mariage*													
AM01 -	indissoluble	440	3048.00	0.37	0.10	-0.03	0.03	0.00	0.00	6.4	-1.8	2.1	0.0	0.0
AM02 -	dissout si pb. grave	714	5311.00	0.17	0.03	0.01	-0.06	0.00	0.00	2.5	1.1	-5.1	0.0	0.0
AM03 -	dissout si accord	764	5499.00	0.16	-0.10	0.02	0.04	0.00	0.00	-9.7	1.9	3.4	0.0	0.0
AM04 -	mariage n.s.p.	82	508.00	2.73	0.21	-0.18	-0.01	0.00	0.00	4.8	-4.1	-0.2	0.0	0.0

MODALITÉS					COORDONNÉES					VALEURS-TEST				
IDEN -	LIBELLE	EFF.	P.ABS	DISTO	1	2	3	0	0	1	2	3	0	0
433.	*age * sexe de l'enquété*													
OU01 -	24 et moins, homme	130	780.00	1.74	0.12	0.08	0.29	0.00	0.00	3.4	2.3	8.2	0.0	0.0
OU02 -	25 @ 39 ans, homme	339	2356.00	0.51	-0.04	0.06	0.01	0.00	0.00	-2.1	3.1	0.6	0.0	0.0
OU03 -	40 @ 59 ans, homme	268	1794.00	0.70	-0.13	0.05	-0.05	0.00	0.00	-5.8	2.1	-2.3	0.0	0.0
OU04 -	60 et plus, homme	196	1414.00	0.92	0.01	0.00	-0.05	0.00	0.00	0.5	0.1	-1.9	0.0	0.0
OU05 -	24 et moins, femme	153	1169.00	1.13	0.07	0.02	0.09	0.00	0.00	2.5	0.6	3.1	0.0	0.0
OU06 -	25 @ 39 ans, femme	362	2778.00	0.42	-0.07	-0.05	-0.04	0.00	0.00	-4.0	-2.7	-2.3	0.0	0.0
OU07 -	40 @ 59 ans, femme	291	2227.00	0.55	0.00	0.00	0.00	0.00	0.00	0.0	0.0	-0.2	0.0	0.0
OU08 -	60 et plus, femme	261	1848.00	0.68	0.17	-0.09	-0.04	0.00	0.00	8.0	-4.2	-1.8	0.0	0.0
510.	*age*diplome*													
E-30 -	- 30 ans, et.element	171	1119.00	1.18	0.21	-0.08	0.19	0.00	0.00	7.4	-2.8	6.6	0.0	0.0
B-30 -	- 30 ans, baccalaure	266	1923.00	0.65	-0.03	0.04	0.05	0.00	0.00	-1.3	1.8	2.5	0.0	0.0
S-30 -	- 30 ans, et.superie	71	576.00	2.39	-0.26	-0.04	-0.07	0.00	0.00	-6.4	-0.9	-1.8	0.0	0.0
E-50 -	- 50 ans, et.element	325	2298.00	0.53	0.05	0.07	0.03	0.00	0.00	2.6	3.4	1.5	0.0	0.0
B-50 -	- 50 ans, baccalaure	297	2148.00	0.57	-0.11	0.01	-0.03	0.00	0.00	-5.7	0.3	-1.5	0.0	0.0
S-50 -	- 50 ans, et. superie	106	789.00	1.72	-0.30	0.07	-0.09	0.00	0.00	-8.7	1.9	-2.7	0.0	0.0
E+50 -	+ 50 ans, et.element	544	3817.00	0.28	0.13	-0.07	-0.01	0.00	0.00	9.6	-4.8	-0.8	0.0	0.0
B+50 -	+ 50 ans, baccalaure	173	1339.00	0.97	-0.12	0.06	-0.10	0.00	0.00	-4.6	2.3	-3.7	0.0	0.0
S+50 -	+ 50 ans, et. superie	47	357.00	3.92	-0.05	0.00 -	0.08	0.00	0.00	-0.9	0.1	1.5	0.0	0.0

GRÁFICO 1

Plano factorial principal de la tabla léxica

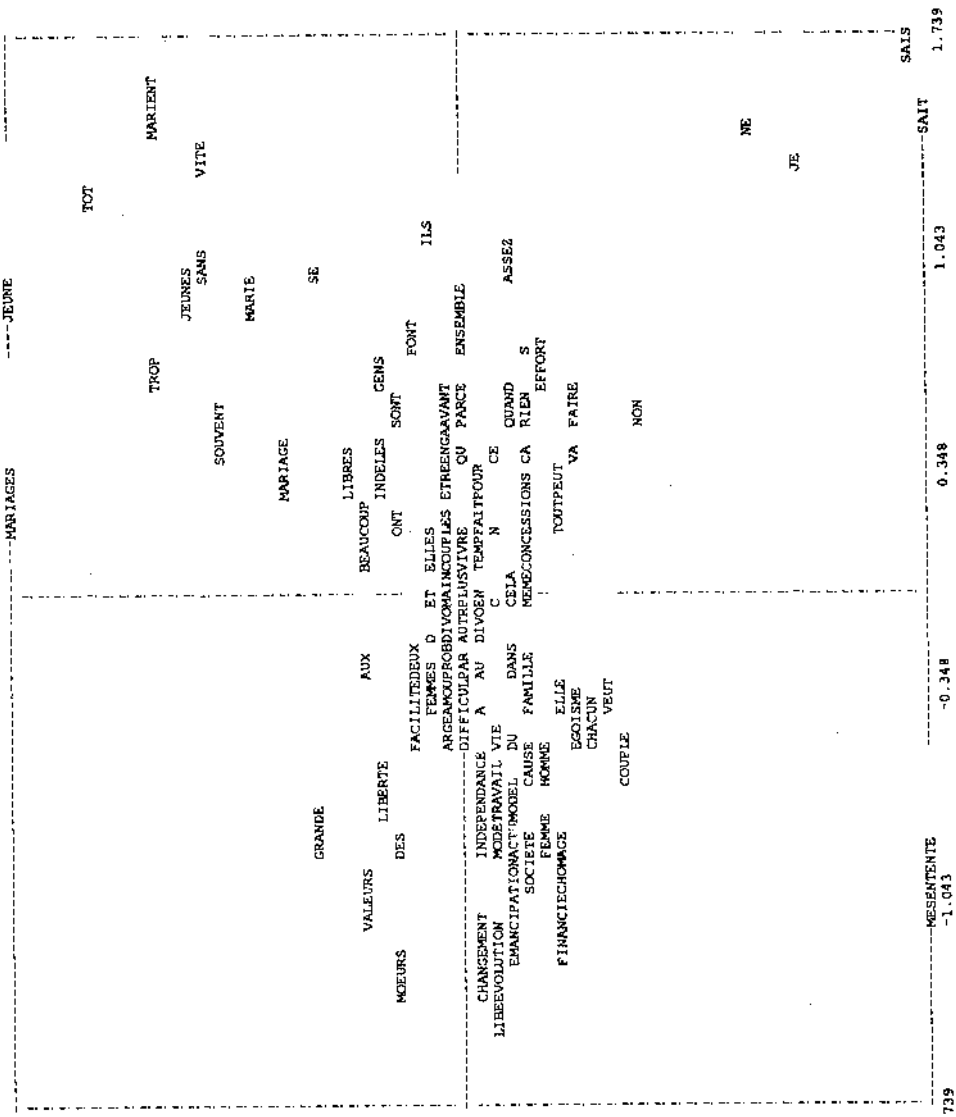


GRÁFICO 2

Proyección de los segmentos como elementos suplementarios

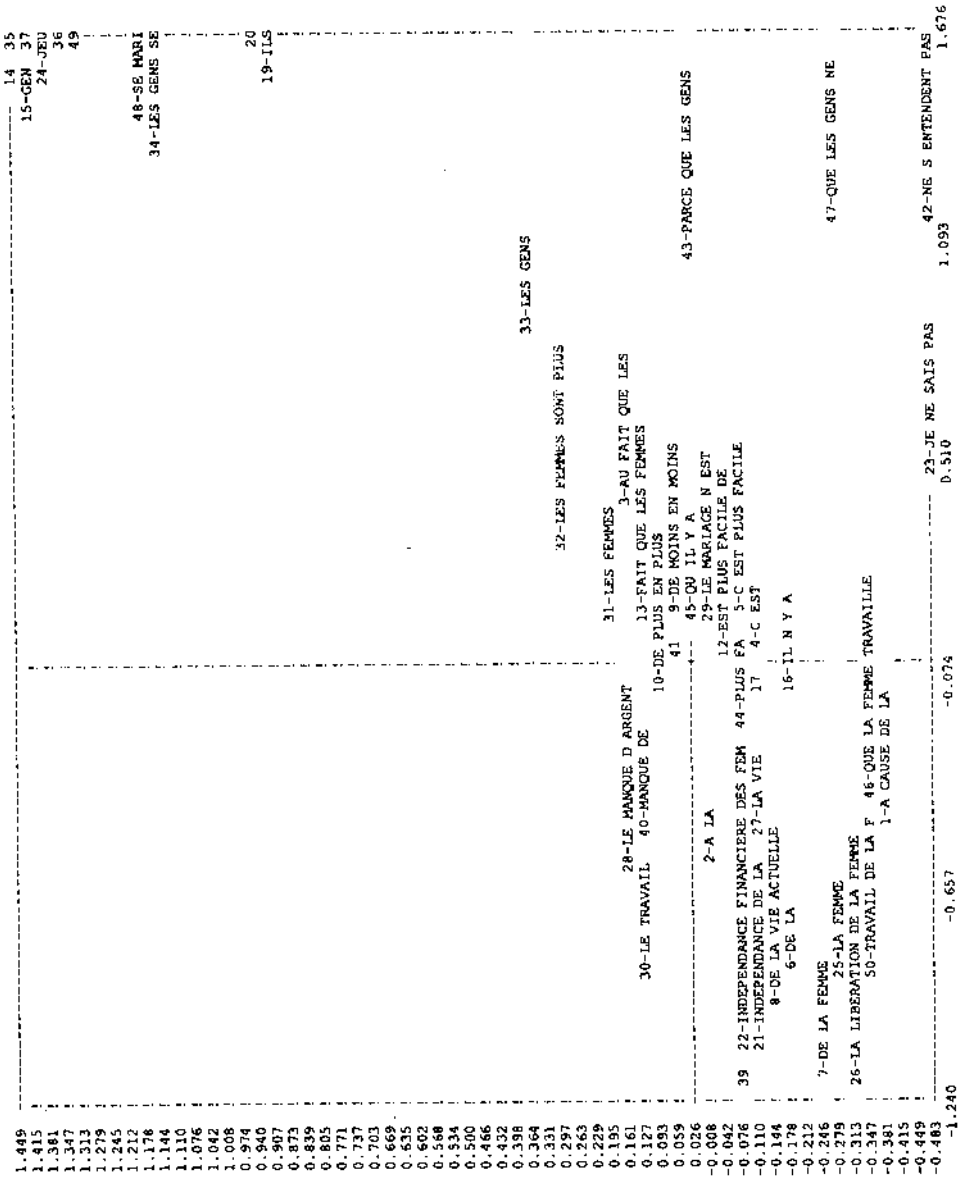


GRÁFICO 3

Modalidades proyectadas sobre el plano factorial Forma*Individuos

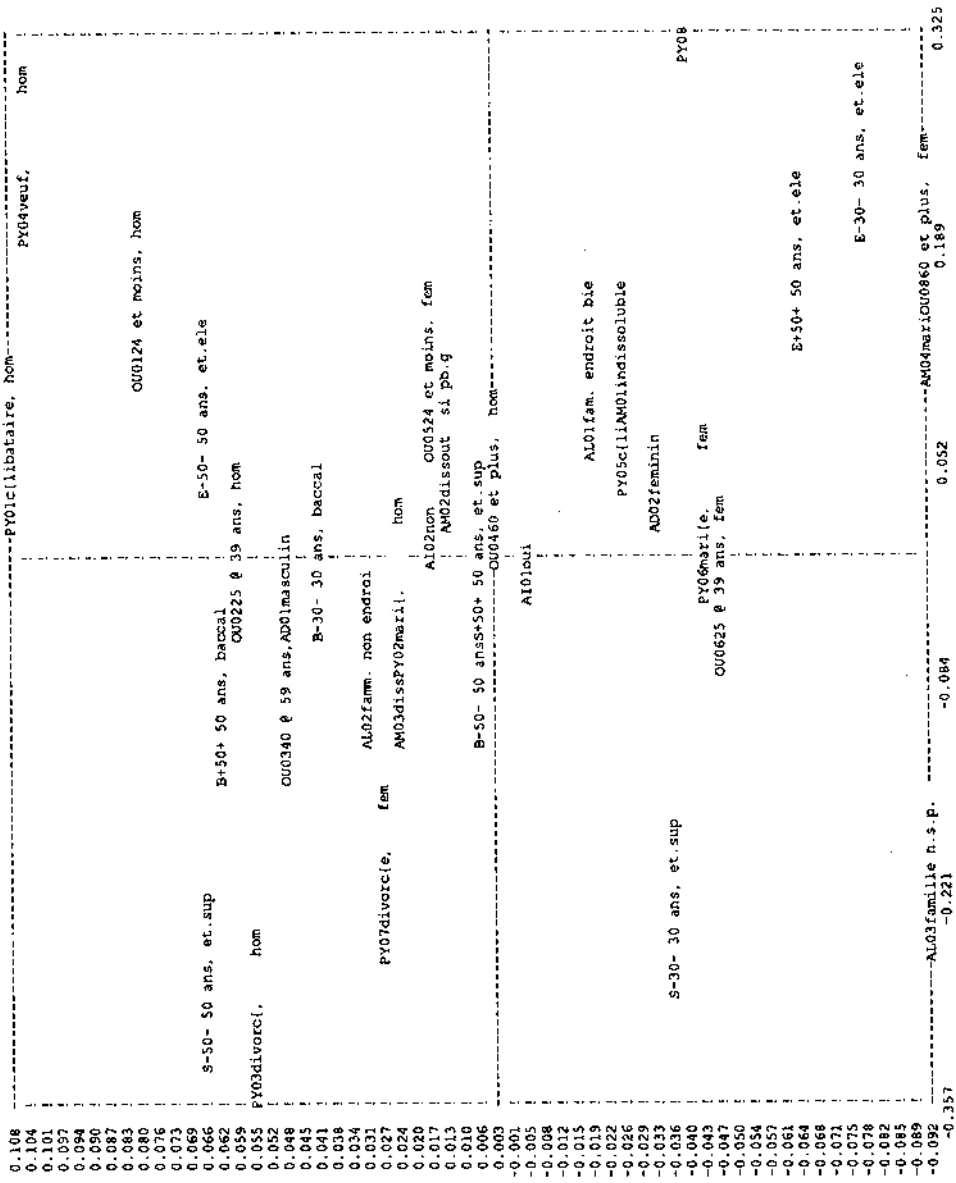
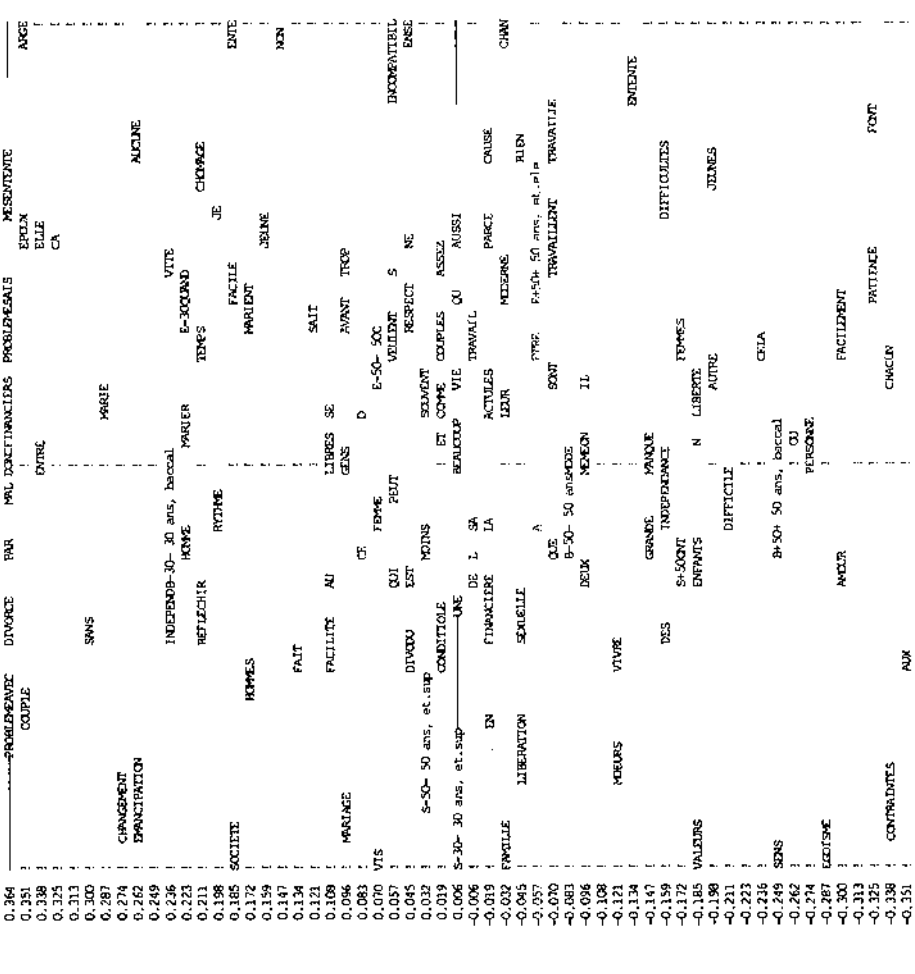


GRÁFICO 4

Plano factorial principal de la tabla léxica agregada



trución sintáctica lo expresan, es decir permite ver «quién dice qué y como lo dice». Por ejemplo, se pone de manifiesto la asociación entre un nivel de estudios bajo y la no respuesta dada por *Je ne sais pas*, entre un nivel de estudios alto y las respuestas que hablan de la liberación de la mujer, del cambio de las costumbres.

ANÁLISIS DE UNA TABLA LÉXICA AGREGADA Y FORMAS CARACTERÍSTICAS

El tratamiento estadístico de grupos de respuestas tiene mucho más interés que el de las respuestas individuales. Los análisis presentados en los párrafos anteriores constituyen una ayuda para la selección de un criterio de agrupamiento pertinente.

Entre todas las variables cualitativas consideradas, se puede escoger una de ellas y construir la tabla léxica agregada correspondiente. Escoger una variable supone escoger un punto de vista; las estructuras que se observarán habrán sido, en parte, suscitadas por esta elección: la dispersión del vocabulario en función de la pertenencia a una u otra categoría de esta variable constituirá la trama de fondo sobre la cual se superpondrán otras estructuras. Será, en cierto sentido, una trama de referencia.

Un criterio pertinente para el agrupamiento de las respuestas individuales es la modalidad de la variable Título*Edad. La tabla léxica agregada obtenida contiene las frecuencias con la cual cada una de las formas ha sido empleada por cada una de las 9 categorías de individuos.

El análisis de esta tabla consiste en comparar los perfiles lexicales de las distintas categorías de individuos. Se proponen tres herramientas para efectuar este análisis: el análisis de correspondencias, las listas de formas características y las listas de respuestas características.

ANÁLISIS DE CORRESPONDENCIAS DE LA TABLA LÉXICA AGREGADA

En el gráfico 4, se presenta el plano factorial principal del análisis de la tabla léxica agregada. Para facilitar la lectura de la gráfica, se puede unir con líneas continuas las modalidades que indican una edad parecida, y por líneas discontinuas las modalidades que indican un mismo nivel de título

Se puede interpretar el primer eje como un eje de titulación: la progresión del nivel de la titulación de izquierda a derecha del eje define una trayectoria extraordinariamente próxima al eje en casi toda su extensión. El segundo eje

parece ser un eje de edad. Opone los más jovenes —arriba— a los mayores. Más que proyectar como elemento suplementario la variable «Edad», es preferible proyectar la variable «Edad*Sexo». La notable similaridad de las dos trayectorias de edad en el caso de los jovenes tiende a desaparecer en el caso de los mayores. Existen de hecho correlaciones entre las variables utilizadas, en particular entre el sexo, la edad y el título: la mujer mayor suele tener menor titulación que el hombre mayor. Por esto el punto-modalidad «Mujer mayor» está atraido por el nivel de estudios bajo. Recordemos que el punto de la gráfica asociado a una modalidad se encuentra en el centro de gravedad de las formas utilizadas por el grupo de individuos que poseen esta modalidad.

SELECCIÓN DE FORMAS CARACTERÍSTICAS

Se puede completar la representación gráfica obtenida por la selección de las formas más características de cada uno de los 9 grupos determinados por la variable Edad-Título. Esta selección, apoyada sobre criterios probabilistas, detecta las formas «anormalmente» frecuentes en las respuestas de un grupo de individuos. Para facilitar la lectura de la caracterización de un grupo por una forma, el sistema *SPAD.T* asocia a cada forma un valor-test que mide la diferencia entre la frecuencia de la forma en el grupo y la frecuencia de la misma forma en la población.

De la misma forma que antes, dicho valor-test está normado de tal forma que se puede leer como una realización de una variable norma centrada y reducida, bajo la hipótesis de repartición aleatoria de la forma considerada en las clases. Por lo tanto, se consideran características de una clase las formas cuyo valor-test asociado es mayor que 1.96 (formas sobrerepresentadas en la clase) o menor que -1.96 (formas subrepresentadas en la clase). En la Tabla 4, se muestran las formas características de 4 de los 9 textos formados según las 9 modalidades de la variable Título*Edad.

SELECCIÓN DE RESPUESTAS CARACTERÍSTICAS

Dado un grupo de individuos, se puede calcular el perfil léxico medio del grupo, a partir de los perfiles léxicos de los individuos que lo componen. Se puede considerar como características de un grupo, las respuestas más próximas a este perfil medio, próximas en el sentido de la distancia de Chi-2, distancia entre distribuciones de frecuencias ya utilizada en el análisis de correspondencias.

TABLA 4

Formas características de 4 textos:
Menores de 30 años sin estudios y con estudios superiores.

SÉLECTION DES FORMES LEXICALES CARACTÉRISTIQUES

TEXTE NUMERO 1 E-30 = - 30 ans, et.element

	Libelle de la forme graphique	*Pourcentage*		*Fréquence*		*V. test*	*Proba*
		Interne	*Global*	*Interne*	*Globale*		
1	SAIS	0.78	0.15	9.	22.	4.097	0.000
2	ARGENT	1.29	0.44	15.	65.	3.676	0.000
3	PROBLEMES	0.69	0.23	8.	34.	2.650	0.004
4	COUPLE	1.12	0.50	13.	74.	2.602	0.005
5	JE	0.86	0.36	10.	53.	2.436	0.007
5	ON	0.69	1.20	8.	177.	-1.579	0.057
4	AU	0.52	0.98	6.	145.	-1.586	0.056
3	IL	0.26	0.66	3.	97.	-1.666	0.048
2	DIVORCE	0.00	0.43	0.	63.	-2.538	0.006
1	ONT	0.00	0.52	0.	77.	-2.918	0.002

TEXTE NUMERO 3 S-30 = - 30 ans, et.superie

	Libelle de la forme graphique	*Pourcentage*		*Fréquence*		*V. test*	*Proba*
		Interne	*Global*	*Interne*	*Globale*		
1	DEUX	1.02	0.25	6.	37.	2.734	0.003
2	MARIAGE	2.04	0.83	12.	123.	2.686	0.004
3	EN	2.04	0.93	12.	137.	2.388	0.008
4	RESPONSABILITÉS	0.68	0.14	4.	21.	2.381	0.009
5	EST	3.23	1.94	19.	287.	2.036	0.021
5	TROP	1.02	2.23	6.	330.	2.032	0.021
4	PARCE	0.00	0.65	0.	96.	-2.063	0.020
3	JEUNES	0.17	0.99	1.	146.	-2.089	0.018
2	FEMMES	0.34	1.39	2.	205.	2.302	0.011
1	ILS	0.00	0.96	0.	142.	2.752	0.003

Formas características de 4 textos:
Menores de 50 años sin estudios y con estudios superiores.

SÉLECTION DES FORMES LEXICALES CARACTÉRISTIQUES

TEXTE NUMERO 1 E-30 = - 30 ans, et.element

	Libelle de la forme graphique	*Pourcentage*		*Fréquence*		*V. test*	*Proba*
		Interne	*Global*	*Interne*	*Globale*		
1	ILS	1.56	0.96	61.	142.	4.186	0.000
2	PAS	2.50	1.92	98.	284.	2.966	0.002
3	NE	2.48	1.91	97.	282.	2.914	0.002
4	POUR	0.97	0.64	38.	94.	2.854	0.002
5	CHANGE	0.23	0.10	9.	15.	2.495	0.006
5	ÉVOLUTION	0.23	0.47	9.	70.	-2.597	0.005
4	COUPLE	0.23	0.50	9.	74.	-2.840	0.002
3	DE	6.90	7.98	270.	1180.	-2.904	0.002
2	SOCIETÉ	0.10	0.36	4.	53.	-3.284	0.001
1	MARIAGE	0.41	0.83	16.	123.	-3.509	0.000

TEXTE NUMERO 9 S+50 = + 50 ans, et.superie

	Libelle de la forme graphique	*Pourcentage*		*Fréquence*		*V. test*	*Proba*
		Interne	*Global*	*Interne*	*Globale*		
1	L	4.29	2.18	16.	322.	2.417	0.008
2	RESPECT	0.80	0.11	3,	17.	2.395	0.008
3	MARIER	0.80	0.11	3.	17.	2.395	0.008
4	QUE	2.68	1.14	10.	169.	2.302	0.011
5	RIEN	0.80	0.14	3.	20.	2.221	0.013
5	COUPLE	0.00	0.50	0.	74.	-1.046	0.148
4	LA	3.75	5.14	14.	760.	-1.119	0.131
3	ILS	0.27	0.96	1.	142.	-1.162	0.123
2	FEMME	0.54	1.49	2.	220.	-1.398	0.081
1	TROP	1.07	2.23	4.	330.	-1.421	0.078

TABLA 5

Respuestas características de 4 textos:
Selección según el criterio de frecuencia de las formas

TEXTE NUMERO 1 E-30 = - 30 ans, et.element

Critère de classification		Réponse ou individu caractéristique
2.626 —	1	NE SAIS PAS
2.626 —	2	NE SAIS PAS
2.579 —	3	JE NE SAIS PAS
2.579 —	4	JE NE SAIS PAS
2.488 —	5	JE SAIS PAS, NON JE SAIS PAS

TEXTE NUMERO 3 S-30 = - 30 ans, et.superie

Critère de classification		Réponse ou individu caractéristique
1.243 —	1	L'ÉVOLUTION DES MOEURS
1.237 —	2	MOINS DE RESPONSABILITÉS DES DEUX PARTIES VIS A VIS DU MARIAGE, LIBÉRATION DES MOEURS, MORALE MOINS STRICTE
1.108 —	3	MANQUE DE MATURITÉ AU MOMENT DU MARIAGE, ÉVOLUTION DIFFERENTE DES DEUX PARTENAIRES
1.020 —	4	ÉVOLUTION DES MOEURS LA FEMME AU TRAVAIL
1.010 —	5	A LA DÉGRADATION DES VALEURS MORALES

TEXTE NUMERO 7 E+50 = + 50 ans, et.element

Critère de classification		Réponse ou individu caractéristique
2.369 —	1	ILS NE S'ENTENDENT PAS
1.960 —	2	NE SAIT PAS
1.960 —	3	NE SAIT PAS
1.960 —	4	NE SAIT PAS
1.960 —	5	NE SAIT PAS

TEXTE NUMERO 9 S+50 = + 50 ans, et.superie

Critère de classification		Réponse ou individu caractéristique
1.024 —	1	LIBÉRALISATION DES MOEURS
1.024 —	2	LIBÉRALISATION DES MOEURS
1.013 —	3	AU NON RESPECT DES ENGAGEMENTS
0.845 —	4	L'ENGAGEMENT EST SUBJECTIF ET LIÉ A L'ATTRAIT DE L'UN POUR L'AUTRE ET NON PLUS A L'AMOUR QUE L'UN DOIT AVOIR POUR L'AUTRE
0.806 —	5	L'AFFAIBLISSEMENT DE L'INSTITUTION FAMILIALE

Se pueden, también, seleccionar las respuestas características siguiendo otro criterio, el criterio del valor-test medio. Como lo hemos visto en el párrafo anterior, se afecta a cada forma y para cada grupo un valor-test que califica la significación de su frecuencia en el grupo comparada a su frecuencia en la población. Se puede atribuir a cada respuesta la media de los valores-test de las formas que la componen. Las respuestas con valor medio más alto serán las más características del grupo (ver table 5).

CONCLUSIÓN

Los tratamientos posibles son más numerosos que los aquí propuestos, pero se ha querido explicitar sobre todo la especificidad de los métodos empleados: la aproximación estadística a los datos textuales presentada en este artículo ofrece una nueva lectura de los textos, lectura esencialmente distinta pero complementaria de la lectura humana. Dicha lectura proporciona una descripción cuantitativa, sistemática y exhaustiva del vocabulario. Ofrece una aproximación comparativa: se describen, analizan e interpretan las diferencias entre los textos.

Los datos de encuesta constituyen el terreno de elección de estos métodos. Pero se puede analizar con provecho otro tipo de textos —textos literarios, discursos políticos, entrevistas no directivas, etc. El corpus constituido debe presentar un cierto grado de homogeneidad y de exhaustividad. Los resultados obtenidos facilitan entonces la construcción de hipótesis y orientan los análisis posteriores.

La integración del conjunto de los métodos en un mismo entorno informático, disponible tanto para microcomputadores como para grandes computadores, permite su utilización por todo tipo de usuarios. El coste de la grabación de los datos textuales sobre soporte magnético queda compensado por la calidad del instrumento de observación ofrecido. La codificación mínima requerida facilita el tratamiento de textos grabados prealablemente para otros fines.

BIBLIOGRAFÍA

Bécue M. *Un sistema informático para el análisis estadístico de datos textuales.* Tesis doctoral. Facultat d'Informàtica de Barcelona. UPC. 1989

Benzécri J.P. La taxinomie, vol. I, *L'Analyse de Correspondances,* vol. II, Dunod. París, 1973.

Benzécri J.P. «Pratique de l'Analyse des Données», tomo 3, *Linguistique & Lexicologie.* Dunod. París, 1973.

Bouroche J.M., Saporta G. *L'Analyse des Données,* «Que sais-je», n°1.854, P.U.F. París, 1980.

Brian E. *Analyse des Données Lexicométriques.* Rapport Credoc/D.G.T., 1984.

Escouffier Y. «L'Analyse des Correspondances, ses Propriétés.ses Extensions». *Bull. of the Inst. Stat. Inst.*, 4, 28-2. 1985.

Haeusler L. *Analyse Lexicale de Réponses Libres: Le Coût de l'Electricité.* Rapport Crédoc-EDF, 1984.

Lafon P., Salem A. «L'Inventaire des Segments Répétés d'un Texte», en *Mots* n° 6, pp. 161-177, 1983.

Lebart L., Houzel van Effenterre Y. «Le Système d'Enquêtes sur les Aspirations des Français, Une Brève Présentation». en *Consommation* n°1, pp. 3-25. Dunod, París, 1980.

Lebart L. «L'Analyse Statistique des Réponses Libres dans les Enquêtes Socio-économiques». *Consommation,* n°1, pp. 39-62, Dunod. París, 1982.

Lebart L., Morineau A., Fénelon J.P. *Traitement des Données Statistiques.* Dunod. París, 1979.

Lebart L., Morineau A., Warwick *Multivariate Descriptive Statistical Analysis.* J. Wiley and Sons. Nueva York, 1984.

Lebart L., Salem A. *Analyse Statistique des Données Textuelles.* Dunod. París, 1988.

Lebart L., Morineau A., Bécue M., (con la col. de P. Pleuvret et L. Haeusler) *SPAD.T, Système Portable pour l'Analyse des Données Textuelles.* Manuel de Référence. CISIA. París, 1989.

Morineau A. «Computational and Statistical Methods of Exploratory Analysis of Textual Data». COMPSTAT, Physica Verlag, Viena, pp 372-377, 1984.

Reinert M. «Un logiciel d'Analyse Lexicale». en *Les Cahiers de l'Analyse des Données,* 4, pp 471-484. Dunod. París, 1986.

Salem A. «Analyse Factorielle et Lexicométrie. Synthèse de Quelques expériences», *Mots* n°4, pp 147-168. 1982.

Salem A. *Pratique des Segments Répétés, Essai de Statistique Textuelle.* Klincksieck. París, 1987.

Volle M. *Analyse des Données.* Economica. París, 1980.