*Anuario de Psicología*
1992, nº 55, 7-22


# El análisis estadístico de datos textuales. La lectura según los escolares de enseñanza primaria

Mónica Bécue
*Universidad Politécnica de Cataluña*
Ludovic Lebart
*École Nationale Supérieure de Télécommunications*
Núria Rajadell
*Universidad de Barcelona*

*Los investigadores se encuentran a menudo enfrentados en la recogida de datos con información textual, sea a través de las preguntas abiertas de una amplia encuesta, sea con entrevistas, sea con otro tipo de textos de fuentes de datos secundarios. Tanto con finalidad exploratoria o de clasificación previa como a la hora de comprobación de determinadas hipótesis, los métodos estadísticos constituyen una herramienta importante para el tratamiento de textos. Especialmente, permiten confrontar los resultados obtenidos del análisis estadístico de textos con otras variables estructurales procedentes de las grandes encuestas interviniendo como variables ilustrativas. En este artículo presentamos un ejemplo sobre la opinión de 895 escolares del nivel de Educación Primaria (10-11 años) del Área Metropolitana de Barcelona, generada a partir de la pregunta abierta «Para mí leer es...». Se muestran algunos resultados proporcionados a través del tratamiento estadístico de análisis de datos textuales, mediante el sistema informático SPAD-T.*
Palabras clave: *Análisis estadístico de textos, análisis de correspondencias, análisis multivariante.*

*The investigators can be confronted with textual information, during data gathering, in large surveys, in interviews or in other secondary sources. The Statistical Methods are useful and important tools in dealing with texts, in exploratory aims or in a priori classifications, as well as the verifying of certain hypothesis. 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. In this article we present an example about the opinion of 895 students of the*

*Dirección de los autores:* M. Bécue, Departamento de Estadística e Investigación Operativa, Facultad de Informática, Pau Gargallo 5. 08028 Barcelona. Ludovic Lebart, École National Supérieure de Télécommunications, Département d'Economie. 46, rue Barrault, 75013 Paris. Núria Rajadell, Departamento de Didáctica y Organización Escolar, Facultad de Pedagogía, Baldiri Reixac s/n, D, 4º. 08028 Barcelona.

*Primary School (10-11 years old) on the Metropolitan Area of Barcelona, generated for the open question «For me read is...». We present differents results obtained through SPAD-T system.*

Key words: *Statistical Textual Analysis, Analysis of Correspondances, Multivariate Analysis.*

## El análisis de datos textuales

Los métodos de la estadística textual han surgido del encuentro entre el estudio cuantitativo de los textos literarios, por una parte, y la corriente de la estadística moderna llamada análisis de datos, por otra.

Las distribuciones lexicales, inicialmente descubiertas como leyes empíricas para mejorar la transcripción estenográfica, han sido posteriormente estudiadas bajo el lema de «psicobiología del lenguaje» (Zipf, 1946).

En un segundo tiempo (Yule, 1944; Guiraud, 1960; Muller, 1968) la estadística lexical se enfrentó con problemas planteados por estilistas preocupados por el estudio comparativo del vocabulario de los «grandes autores»: comparación de la riqueza del vocabulario, análisis de la evolución del vocabulario de un mismo autor, etc. Esta corriente se vio reforzada por la difusión de los computadores, y enriquecida por los análisis de tipo morfosintáctico (según especifican por ejemplo Bourques y cols., 1988).

Los métodos de la estadística textual se aplican ahora a todo tipo de textos transcritos sobre soporte informático. Por lo tanto se pueden utilizar entre otros métodos de aproximación a los textos (lingüística/análisis del discurso, análisis de contenido, indexación automática, inteligencia artificial), en las distintas disciplinas que entran en relación con el texto (historia, sociología, psicología, etc.), teniendo en cuenta en cada caso, evidentemente, las perspectivas de investigación propias de dichas disciplinas. Una parte importante de los trabajos de investigación —que comporta aplicaciones industriales considerables— se dedica a la comprensión de los lenguajes naturales (obsérvese una síntesis en Coulon y cols., 1986).

Los métodos del análisis de datos han probado, en lo que concierne a los estudios textuales (Benzecri, 1981), su aptitud para elaborar tipologías mediante el recuento de las formas gráficas. Dichos métodos presentan la ventaja de estudiar los perfiles lexicales en su conjunto, y por lo tanto, tomar en cuenta redes de autocorrelaciones bastante finas. Así consiguen llegar bastante lejos en el estudio de los textos, a la vez que guardan una total independencia de la lengua tratada. Con el sistema informático SPAD.T (Lebart y cols., 1989) utilizado para analizar el ejemplo presentado en este artículo, se han tratado además textos en castellano, catalán, croata, francés, griego, inglés, italiano y provenzal.

## Las preguntas abiertas

Las preguntas abiertas son todavía poco utilizadas en las encuestas, porque la explotación de las respuestas recogidas es difícil y costosa. No obstante

la información obtenida mediante dichas preguntas puede ser muy distinta de la obtenida mediente preguntas cerradas (Schuman y Presser, 1981); por lo tanto puede ser necesario mantener una pregunta abierta en razón de la información buscada.

Citemos dos situaciones-tipo bastante corrientes para las cuales la utilización de un cuestionario abierto se impone. Para economizar el tiempo de entrevista (una pregunta abierta puede sustituir una larga lista de opciones), y para explicitar las respuestas a preguntas cerradas mediante el uso de la clásica pregunta abierta «¿Por qué?». Las explicaciones relativas a una respuesta anterior se deben dar de forma espontánea: proponer una batería de ítems podría ofrecer nuevos argumentos y falsear la sinceridad de la explicación.

La utilidad de este tipo de preguntas ha sido subrayada por numerosos autores y sólo las dificultades y el coste de explotación limitan su uso. Sin embargo, sólo una pregunta abierta permite saber si las distintas categorías de personas interrogadas han entendido la pregunta cerrada de la misma forma; hecho particularmente importante en las encuestas internacionales, porque permite detectar eventuales divergencias semánticas introducidas por el enunciado de la pregunta en función de la lengua utilizada.

No podemos más que resaltar la importancia que posee la postcodificación. Las técnicas clásicas de postcodificación operan construyendo una batería de ítems a partir de una muestra de respuestas. Después se codifica el conjunto de las respuestas de tal forma que se sustituye la respuesta abierta por una o más respuestas cerradas.

Para respuestas simples, muy tipificadas y poco numerosas (dicho de otra forma: para respuestas a una pregunta que se habría podido cerrar sin dificultad...) este procedimiento presenta pocos inconvenientes. Para otros casos se pueden mencionar rápidamente algunos de los defectos de este tipo de tratamiento: subjetividad de la codificación, empobrecimiento de la forma y mutilación del contenido.

## ¿Qué unidades estadísticas?

Cuando la recogida de textos en soportes informáticos constituía la parte más importante del tratamiento estadístico de los textos en ordenador, las polémicas sobre la naturaleza de las unidades estadísticas fueron bastante vivas. ¿Era indispensable, como afirmaban algunos, trabajar sólo con textos fuertemente sobrecodificados en categorías gramaticales, reduciendo los plurales al singular, los verbos flexionados al infinitivo, etc., cosa que entorpecía de manera notable la entrada de textos en computadores? ¿Se podían empezar a tratar recuentos obtenidos de manera totalmente automática teniendo en cuenta sólo las formas gráficas (relación de caracteres gráficos delimitados por blancos o signos de puntuación), dejando para las etapas ulteriores del análisis los problemas de falta de ambigüedad y de lematización?

*Una aproximación más pragmática*

Sin profundizar en estas divergencias teóricas, el desarrollo reciente de la informática aclara estas cuestiones con una nueva luz. En estos momentos la grabación sobre soporte magnético se transforma en la forma más natural de almacenar la información textual. El desarrollo de los lectores ópticos facilita, cada día más, la recogida de textos impresos.

Así pues, al tratamiento informatizado de textos brutos —considerados como una sucesión de formas gráficas— se le atribuyen nuevos objetivos, como la verificación de la corrección de entradas automáticas, las tipologías de textos realizadas en una primera fase con carácter exploratorio o la localización de las unidades que funcionan en los textos agrupados en corpus, entre otras.

La definición de la unidad de base debe merecer una reflexión específica. En efecto, para efectuar los recuentos utilizables por los algoritmos de análisis de datos, es posible definir de formas muy distintas las unidades de la cadena textual.

*Formas gráficas y segmentos repetidos*

La unidad de base será la forma gráfica definida como una sucesión de caracteres no delimitadores (normalmente letras) junto a caracteres delimitadores (blanco, puntos, comas...). Una misma palabra podrá, en general, dar lugar a numerosas formas gráficas, según su caso o género gramatical en el texto; una misma forma gráfica puede igualmente reflejar numerosas palabras. Esto no supone un gran inconveniente, pues las formas gráficas no serán tratadas aisladamente.

El tratamiento integrado constará de dos aspectos: un aspecto multidimensional clásico, que se interesará por los perfiles de frecuencias de formas gráficas, es decir, por los vectores cuyos componentes equivalen a las frecuencias de cada una de las formas utilizadas por un individuo o un grupo de individuos; conteniendo estos perfiles una información extremadamente rica. El segundo aspecto, al que se le califica de contextual, consiste en tomar en cuenta nuevas unidades estadísticas, los segmentos repetidos (Salem, 1987). Se trata de secuencias de formas simples que aparecen con cierta frecuencia y que enriquecen los perfiles de formas y ayudan a aclarar ciertas ambigüedades de interpretación, en la cual interviene el contexto de estas formas.

Los algoritmos particulares de cálculo permitirán descubrir dichos segmentos repetidos. Precisando todavía más, las técnicas evidenciarán las diferencias entre perfiles de formas gráficas y de segmentos.

Mientras que la interpretación de un perfil puede ser delicada (p. ej. ¿por qué esta categoría de entrevistados utiliza estas palabras con estas frecuencias?), la interpretación de las diferencias es todavía más sencilla: sin especular sobre el significado de los perfiles, se puede observar claramente que, por ejemplo, dos categorías de entrevistados tienen unos perfiles próximos, alejados de los de las otras categorías. Simplificando al máximo, se puede resumir esta aproximación

«por contraste o por diferencia» a través de la fórmula siguiente: no es necesario saber lo que han dicho dos categorías para saber si han expresado o no la misma opinión o concepto.

Para seleccionar formas y segmentos es imprescindible utilizar umbrales de frecuencia, los cuales permitirán efectuar filtros a diferentes niveles sobre la información de base. Esta fase de tratamiento preliminar consiste en destinar a cada nueva forma gráfica un número de orden que será asociado a todas las ocurrencias de esta misma forma. Estos números serán almacenados en un diccionario de formas, o vocabulario, propio de cada explotación, el cual permitirá, a la salida de los cálculos o de las impresiones, reconstituir el grafismo de las formas evidenciadas a través de los cálculos estadísticos.

Una etapa intermediaria de tratamiento podría consistir en «lematizar» el vocabulario (p. ej. declarar equivalentes las formas gráficas correspondientes a una misma palabra) o depurar este vocabulario de palabras-herramientas (artículos, conjunciones, etc., véase por ejemplo Reinert, 1986).

La experiencia obtenida en el análisis de preguntas abiertas demuestra que esta etapa no es del todo indispensable, o que no debe intervenir demasiado pronto. Las formas gráficas diferentes de una misma palabra pueden estar relacionadas con un contexto y un contenido particular, y algunas palabras-herramientas pueden caracterizar de una manera concreta actitudes u opiniones.

## Tablas de contingencia léxicas

Los métodos de análisis de datos suelen tratar grandes tablas de datos creadas a partir de variables nominales, ordinales o cuantitativas. Para aplicar estos métodos —en particular el análisis de correspondencias y los métodos de clasificación— 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, aplicada 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...— 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. Se obtienen tablas similares sustituyendo las formas por los segmentos repetidos.

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 aplicado a las tablas léxicas proporciona una visualización de las similitudes entre perfiles de frecuencias de formas. El análisis de las tablas segmentales permite, además, tener en cuenta el orden en el cual aparecen las formas.

## El análisis de datos textuales aplicado a la investigación educativa

Con el fin de conocer la opinión sobre la lectura que poseen los escolares del nivel de educación primaria, se ha realizado una amplia investigación en el área metropolitana de Barcelona. Para medir las actitudes lectoras a través de las cuatro facetas consideradas configuradoras de dichas actitudes —personales, familiares, escolares y ambientales—, se pueden aplicar diferentes instrumentos con enfoques desde lo más cuantitativo hasta lo más cualitativo. Pero nos interesan de una manera especial dos de los cuestionarios elaborados en el curso de dicha investigación (Rajadell, 1990), para poder caracterizar las actitudes lectoras; uno de ellos pretende de manera específica conocer las actitudes a partir de una escala tipo Likert con cinco posibilidades de respuesta; el otro, mucho más amplio, pretende recoger, de manera general, la máxima información en torno a los hábitos, intereses y realidades relacionados con la lectura. En este último se encuentran dos preguntas de carácter abierto con los siguientes enunciados:

1. *Para mí leer es...*
2. *Creo que leer es importante porque...*

La primera pregunta facilita el conocimiento sobre el concepto de lectura que poseen los escolares, mientras que la segunda cuestión nos informa sobre la importancia que otorgan al acto y efecto de leer.

La muestra estudiada está formada por 895 alumnos y alumnas que están cursando quinto curso de EGB, con una edad de 10-11 años, cuya proporción está configurada por un 51.2 % de niños y un 49.6 % de niñas, asegurando la presentación de la variable sexo con un respetable equilibrio. Este alumnado se encuentra ubicado en centros escolares de variada tipología (públicos 56 % y privados 34 %).

Por otra parte, los niños contestan a un amplio cuestionario que incluye preguntas sobre su actitud hacia la lectura. La consulta de las fichas escolares ha permitido obtener, además, variables-indicadoras de la situación socioeconómica de sus familias.

En este artículo presentaremos algunos resultados proporcionados por el tratamiento estadístico de la primera pregunta abierta, utilizando métodos estadísticos de análisis de datos textuales.

## Métodos de análisis

Mediante el tratamiento del cuestionario presentado se pretenden ilustrar las principales etapas del análisis de una encuesta que incluye preguntas abiertas y cerradas.

Una etapa preliminar permite reagrupar a los escolares en clases homogéneas en cuanto a las características socioeconómicas de sus familias. Para dicho reagrupamiento se emplea la técnica llamada de «Núcleos factuales» (Lebart y Salem, 1989). No es posible extenderse aquí sobre este método que permite obtener reagrupamientos operativos de centenares o millares de individuos en un número reducido de clases, teniendo en cuenta las respuestas a un grupo de variables así como sus interrelaciones. En este ejemplo, se tienen en cuenta las respuestas a las variables indicadoras de la situación socioeconómica.

Una vez eliminados del estudio los escolares de los cuales no se poseen estos indicadores —restando un global de 857 individuos— se han obtenido seis clases de escolares. La Tabla 1 describe las clases, de forma precisa, mediante la comparación de los porcentajes de respuestas internas en cada clase y de los porcentajes globales con el fin de seleccionar las modalidades más características. Se puede observar que ciertas modalidades ilustrativas (es decir, no utilizadas para la construcción de las clases) son repartidas de forma diferenciada en las clases.

TABLA 1. CARACTERIZACIÓN DE LAS SEIS CLASES SOCIOECONÓMICAS DE ESCOLARES

| Modalidades características | | IDEN | Porcentajes CLA/MOD | MOD/CLA | GLOBAL | Peso | V. Test. Prob. | |
|---|---|---|---|---|---|---|---|---|
| Clase 1/6 | | aa1a | | | 45.04 | 386 | | |
| Estudios madre | Elementales | ES02 | 68.56 | 100.00 | 65.69 | 563 | 21.69 | 0.000 |
| Estudios padre | Elementales | ER02 | 71.27 | 97.67 | 61.73 | 529 | 21.40 | 0.000 |
| Trabajo madre | Ama de casa | TM09 | 55.95 | 98.70 | 79.46 | 681 | 13.99 | 0.000 |
| Tipo de escuela | Pública | PP01 | 60.04 | 88.34 | 66.28 | 568 | 12.77 | 0.000 |
| Lengua familiar | Castellano | FU02 | 56.78 | 92.23 | 73.16 | 627 | 11.90 | 0.000 |
| Libros texto escuela | Me gustan | TX02 | 47.38 | 82.12 | 78.06 | 669 | 2.53 | 0.006 |
| En casa tenemos | Pocos libros | HH01 | 67.65 | 5.96 | 3.97 | 34 | 2.53 | 0.006 |
| Número hermanos | Tiene 2 hermanos | NG03 | 52.74 | 27.46 | 23.45 | 201 | 2.42 | 0.008 |
| Número hermanos | Tiene 3 hermanos | NG04 | 56.67 | 13.21 | 10.50 | 90 | 2.23 | 0.013 |
| Clase 2/6 | | aa2a | | | 15.64 | 134 | | |
| Estudios padre | Medios | ER03 | 70.21 | 49.25 | 10.97 | 94 | 12.92 | 0.000 |
| Estudios madre | Medios | ES03 | 55.88 | 28.36 | 7.93 | 68 | 7.98 | 0.000 |
| Estudios padre | Universitarios | ER04 | 69.44 | 18.66 | 4.20 | 36 | 7.31 | 0.000 |
| Tipo de escuela | Privada | PP02 | 28.03 | 60.45 | 33.72 | 289 | 6.83 | 0.000 |
| Trabajo madre | Ama de casa | TM09 | 19.09 | 97.01 | 79.46 | 681 | 6.20 | 0.000 |
| Lengua familiar | Catalán | FU01 | 38.83 | 29.85 | 12.02 | 103 | 6.09 | 0.000 |
| Lengua familiar | Catalán y castellano | FU03 | 36.59 | 11.19 | 4.78 | 41 | 3.22 | 0.001 |
| En casa tenemos | Muchos libros | HH03 | 17.85 | 76.87 | 67.33 | 577 | 2.51 | 0.006 |
| Número hermanos | Tiene 1 hermano | NG02 | 19.75 | 47.01 | 37.22 | 319 | 2.43 | 0.007 |
| Cualquier lengua | Excelente | QL05 | 23.33 | 20.90 | 14.00 | 120 | 2.29 | 0.011 |

| Modalidades características | | IDEN | Porcentajes CLA/MOD | MOD/CLA | Peso GLOBAL | | V. Test. | Prob. |
|---|---|---|---|---|---|---|---|---|
| Clase 3/6 | | aa3a | | | 9.45 | 81 | | |
| Estudios madre | Sin | ES01 | 88.46 | 85.19 | 9.10 | 78 | 18.70 | 0.000 |
| Estudios padre | Sin | ER01 | 92.98 | 65.43 | 6.65 | 57 | 16.08 | 0.000 |
| Tipo de escuela | Pública | PP01 | 13.38 | 93.83 | 66.28 | 568 | 6.03 | 0.000 |
| Lengua familiar | Castellano | FU02 | 12.28 | 95.06 | 73.16 | 627 | 5.15 | 0.000 |
| Trabajo madre | Ama de casa | TM09 | 11.45 | 96.30 | 79.46 | 681 | 4.35 | 0.000 |
| Cualquier lengua | Suspenso | QL01 | 23.08 | 14.81 | 6.07 | 52 | 2.88 | 0.002 |
| Número hermanos | Tiene 4 hermanos | NG05 | 25.81 | 9.88 | 3.62 | 31 | 2.52 | 0.006 |
| Leo con | Alguna dificultad | LA02 | 12.72 | 53.09 | 39.44 | 338 | 2.50 | 0.006 |
| Leo cuando | trabajo | PL01 | 14.13 | 32.10 | 21.47 | 184 | 2.23 | 0.013 |
| Número hermanos | Tiene 7 hermanos | NG08 | 50.00 | 3.70 | 0.70 | 6 | 2.22 | 0.013 |
| Número hermanos | Tiene 6 hermanos | NG07 | 66.67 | 2.47 | 0.35 | 3 | 1.96 | 0.025 |
| Clase 4/6 | | aa4a | | | 13.54 | 116 | | |
| Estudios madre | Missing estudios de la madre | ESMI | 75.19 | 86.21 | 15.52 | 133 | 19.28 | 0.000 |
| Estudios padre | Missing estudios del padre | ERMI | 73.88 | 85.34 | 15.64 | 134 | 18.96 | 0.000 |
| Lengua familiar | Missing lengua | FUMI | 61.63 | 45.69 | 10.04 | 86 | 11.22 | 0.000 |
| Tipo de escuela | Privada | PP02 | 29.07 | 72.41 | 33.72 | 289 | 9.11 | 0.000 |
| Número hermanos | Missing nº hermanos | NGMI | 42.40 | 45.69 | 14.59 | 125 | 8.83 | 0.000 |
| Clase 5/6 | | aa5a | | | 9.10 | 78 | | |
| Trabajo madre | Oficios | TM08 | 93.35 | 52.56 | 5.02 | 43 | 99.99 | 0.000 |
| Trabajo madre | Comer. | TM06 | 100.00 | 44.87 | 4.08 | 35 | 99.99 | 0.000 |
| Nivel socioecon. | Medio | NS02 | 12.11 | 70.51 | 52.98 | 454 | 3.18 | 0.001 |
| Porque leo en la escuela los libros propuestos por el maestro | No tienen acción | QU08 | 50.00 | 5.13 | 0.93 | 8 | 2.71 | 0.003 |
| Asignatura rechazada | Plástica | AR07 | 18.29 | 19.23 | 9.57 | 82 | 2.62 | 0.004 |
| Leo con | Mucha dificultad | LA01 | 25.00 | 8.97 | 3.27 | 28 | 2.34 | 0.010 |
| Tipo de escuela | Privada | PP02 | 12.46 | 46.15 | 33.72 | 289 | 2.27 | 0.012 |
| Número hermanos | Ningún hermano | NG01 | 17.11 | 16.67 | 8.87 | 76 | 2.18 | 0.015 |
| Clase 6/6 | | aa6a | | | 7.23 | 62 | | |
| Trabajo madre | Profesión industria | TM04 | 100.00 | 38.71 | 2.80 | 24 | 11.36 | 0.000 |
| Trabajo madre | Adm. Banc. Empr. | TM01 | 100.00 | 24.19 | 1.75 | 15 | 8.72 | 0.000 |
| Trabajo madre | Profesiones liberales | TM03 | 100.00 | 24.19 | 1.75 | 15 | 8.72 | 0.000 |
| Estudios madre | Universitarios | ES04 | 71.43 | 16.13 | 1.63 | 14 | 5.93 | 0.000 |
| Estudios madre | Medios | ES03 | 29.41 | 32.26 | 7.93 | 68 | 5.75 | 0.000 |
| Trabajo madre | Funcionaria | TM02 | 100.00 | 8.06 | 0.58 | 5 | 4.65 | 0.000 |
| Lengua familiar | Catalán | FU01 | 20.39 | 33.87 | 12.02 | 103 | 4.61 | 0.000 |
| Tipo de escuela | Privada | PP02 | 13.15 | 61.29 | 33.72 | 289 | 4.48 | 0.000 |
| Estudios padre | Universitarios | ER04 | 30.56 | 17.74 | 4.20 | 36 | 4.15 | 0.000 |
| Lengua familiar | Catalán y castellano | FU03 | 21.95 | 14.52 | 4.78 | 41 | 2.94 | 0.002 |
| Número hermanos | Ningún hermano | NG01 | 15.79 | 19.35 | 8.87 | 76 | 2.54 | 0.006 |
| Estudios padre | Universitarios sup. | ER05 | 42.86 | 4.84 | 0.82 | 7 | 2.32 | 0.010 |
| Estudios padre | Medios | ER03 | 13.83 | 20.97 | 10.97 | 94 | 2.24 | 0.012 |

### *Glosario de formas y segmentos repetidos*

La Tabla 2 muestra las 109 formas repetidas al menos 8 veces en el corpus formado por las respuestas a la primera de las dos preguntas abiertas de los 857 individuos que configuran la población. La forma más frecuente es «*y*». La forma llena más frecuente es «*divertido*». Las formas «*importante*» y «*aprender*» aparecen a continuación.

Las formas-herramientas tienen el mismo tratamiento que las formas llenas. Si su distribución en las respuestas es aleatoria no perturban los resultados. Si, por el contrario, su distribución no es debida al azar aportan una información interesante. De la misma forma, si dos formas gráficas referidas a la misma palabra —las diferentes formas del verbo «*aprender*», por ejemplo—, tienen un comportamiento similar, se pueden sustituir por la misma palabra, si no se refieren a usos diferenciados de la misma palabra.

La Tabla 3 muestra los segmentos observados en las respuestas abiertas, segmentos seleccionados por umbrales de frecuencia: los segmentos formados por dos palabras empleados al menos 30 veces, por tres palabras empleados al menos 10 veces y los más largos empleados al menos 5 veces.

### *Construcción y análisis de correspondencias de una tabla léxica agregada*

Las respuestas de los escolares son reagrupadas en función de la clase de pertenencia del escolar obtenida anteriormente. Se construye la tabla léxica agregada que contiene la frecuencia con la cual cada grupo emplea cada una de las 109 formas conservadas. Para el análisis de correspondencias, se considera la clase 4 como un elemento ilustrativo: en efecto, los escolares de esta clase no han contestado a numerosas preguntas cerradas y sus respuestas abiertas son extremadamente estereotipadas. En la Figura 1, se presenta el primer plano obtenido mediante el análisis de correspondencias de dicha tabla.

El primer eje opone las clases 5 y 6, caracterizadas por el trabajo de la madre, a las otras clases. El segundo eje opone las familias de clase baja (clases 3 y 1 sobre todo, con padres sin estudios o con estudios elementales) a las familias de clase media (clases 2 y 6, padres con estudios medios o superiores): configura un eje socioeconómico. Que las diferentes actitudes hacia la lectura, expresadas en la respuesta abierta, estén relacionadas con la pertenencia a una u otra clase socioeconómica no es sorprendente. Puede parecer más inesperada la relevancia que tiene el trabajo de la madre en cuanto a la actitud lectora: el primer eje está constituido por la oposición entre las clases de escolares cuya madre trabaja y las otras clases.

### *Selección de formas y segmentos característicos*

Se puede completar la representación gráfica obtenida por la selección de las formas más características de cada una de las 5 clases. Esta selección, apoya-

*Balance del tratamiento*
Número total de respuestas = 857.
Número total de palabras = 5692.
Número de palabras distintas = 628.
Porcentaje, palabras distintas = 11.0

*Selección de las palabras*
Umbral de frecuencia = 7.
Total de palabras retenidas = 4749.
Palabras distintas retenidas = 109.

Formas lexicales por orden de frecuencia

| Núm. | Palabras empleadas | Frec. | Longitud. |
|---|---|---|---|
| 108 | y | 305 | 1 |
| 77 | muy | 279 | 3 |
| 45 | es | 225 | 2 |
| 33 | divertido | 214 | 9 |
| 91 | que | 152 | 3 |
| 104 | un | 147 | 2 |
| 57 | importante | 144 | 10 |
| 105 | una | 137 | 3 |
| 71 | me | 129 | 2 |
| 10 | aprender | 125 | 8 |
| 88 | porque | 124 | 6 |
| 25 | cosas | 120 | 5 |
| 27 | de | 111 | 2 |
| 62 | leer | 111 | 4 |
| 1 | a | 91 | 1 |
| 82 | para | 83 | 4 |
| 78 | no | 80 | 2 |
| 39 | en | 67 | 2 |
| 18 | bonito | 67 | 6 |
| 59 | la | 62 | 2 |
| 101 | te | 60 | 2 |
| 22 | como | 58 | 4 |
| 38 | el | 57 | 2 |
| 52 | gusta | 56 | 5 |
| 24 | cosa | 54 | 4 |
| 75 | mucho | 50 | 5 |
| 26 | cuando | 45 | 6 |
| 73 | mí | 44 | 2 |
| 31 | diversión | 43 | 9 |
| 70 | más | 42 | 3 |
| 66 | libro | 41 | 5 |
| 68 | lo | 40 | 2 |
| 15 | aventuras | 40 | 9 |
| 11 | aprendes | 40 | 8 |
| 58 | interesante | 40 | 11 |
| 98 | si | 38 | 2 |
| 79 | nuevas | 37 | 6 |
| 106 | veces | 35 | 5 |
| 96 | se | 34 | 2 |
| 34 | divertirme | 34 | 10 |
| 17 | bien | 32 | 4 |
| 67 | libros | 31 | 6 |
| 86 | pero | 30 | 4 |
| 69 | los | 29 | 3 |
| 60 | las | 29 | 3 |
| 64 | leo | 28 | 3 |
| 92 | rato | 28 | 4 |
| 3 | aburrido | 28 | 8 |
| 83 | pasar | 28 | 5 |
| 74 | muchas | 27 | 6 |
| 87 | poco | 26 | 4 |
| 8 | algo | 24 | 4 |
| 48 | estudiar | 24 | 8 |
| 44 | entretenimiento | 23 | 15 |
| 80 | o | 23 | 1 |
| 65 | leyendo | 22 | 7 |
| 81 | palabras | 21 | 8 |
| 14 | aventura | 21 | 8 |
| 99 | son | 20 | 3 |
| 12 | aprendo | 19 | 7 |
| 37 | e | 18 | 1 |
| 72 | mejor | 18 | 5 |
| 43 | entretenido | 18 | 11 |
| 40 | enseña | 17 | 6 |
| 89 | puedes | 17 | 6 |
| 28 | del | 17 | 3 |
| 76 | mundo | 16 | 5 |
| 13 | así | 16 | 3 |
| 102 | tiempo | 16 | 6 |
| 16 | bastante | 14 | 8 |
| 49 | etc. | 14 | 3 |
| 23 | con | 14 | 3 |
| 94 | saber | 14 | 5 |
| 47 | estoy | 14 | 5 |
| 100 | también | 14 | 7 |
| 51 | forma | 13 | 5 |
| 85 | paso | 13 | 4 |
| 6 | además | 13 | 6 |
| 35 | divertirse | 13 | 10 |
| 50 | fantasía | 13 | 8 |
| 32 | divertida | 12 | 9 |
| 84 | pasas | 12 | 5 |
| 41 | entrar | 12 | 6 |
| 97 | según | 11 | 5 |
| 93 | rollo | 11 | 5 |
| 21 | cada | 11 | 4 |
| 20 | bueno | 11 | 5 |
| 53 | hacer | 10 | 5 |
| 55 | imaginación | 10 | 11 |
| 95 | sabes | 10 | 5 |
| 4 | aburrimiento | 10 | 12 |
| 36 | divierto | 10 | 8 |
| 103 | todo | 9 | 4 |
| 46 | escribir | 9 | 8 |
| 54 | hay | 9 | 3 |
| 29 | depende | 9 | 7 |
| 9 | algunas | 9 | 7 |
| 63 | lees | 9 | 4 |
| 30 | distracción | 9 | 11 |
| 90 | pues | 8 | 4 |
| 61 | lectura | 8 | 7 |
| 42 | entretenerse | 8 | 12 |
| 7 | al | 8 | 2 |
| 107 | vez | 8 | 3 |
| 5 | aburro | 8 | 6 |
| 19 | buen | 8 | 4 |
| 56 | imaginar | 8 | 8 |
| 2 | aburrida | 8 | 8 |
| 109 | yo | 8 | 2 |

TABLA 3. SEGMENTOS REPETIDOS EN EL CORPUS

aprender
1 4 4 aprender cosas nuevas y
2 5 4 aprender cosas que no
aprendes
3 4 4 aprendes muchas cosas y
cosas
4 10 3 cosas que no
divertido
5 41 2 divertido y
es
6 59 2 es muy
7 10 3 es muy bonito
8 21 3 es muy divertido
9 7 4 es muy divertido y
10 19 3 es muy importante
11 8 4 es muy importante porque
12 11 3 es una cosa
13 5 4 es una cosa que
leer
14 30 2 leer es
15 10 3 leer es muy
16 4 5 leer es muy divertido y
me
17 55 2 me gusta
18 12 3 me gusta leer
19 18 3 me gusta mucho
20 4 4 me gusta mucho leer
21 5 5 me lo paso muy bien
muy
22 35 2 muy bonito
23 10 3 muy bonito porque
24 5 4 muy bonito porque aprendes
25 93 2 muy divertido
26 20 3 muy divertido y
27 4 4 muy divertido y entretenido
28 5 4 muy divertido y me
29 91 2 muy importante
30 20 3 muy importante porque
31 4 4 muy importante porque aprendes
32 13 3 muy importante y
33 4 4 muy importante y divertido
no
34 12 3 no me gusta
para
35 32 2 para mi
36 24 4 para mi leer es
37 9 5 para mi leer es muy
pasar
38 4 4 pasar un buen rato
que
39 5 4 que me gusta mucho
si
40 4 4 si es de aventuras
una
41 52 2 una cosa
42 16 3 una cosa muy
43 8 4 una cosa muy importante
44 19 3 una cosa que
45 8 4 una cosa que me
46 6 5 una cosa que me gusta
47 4 6 una cosa que me gusta mucho
48 31 2 una diversión

da 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, se 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. Dicho valor-test está normalizado de tal forma que se pueda leer como una realización de una variable normal centrada y reducida, bajo la hipótesis de repartición aleatoria de la forma considerada en las clases. Por lo tanto, se declaran características de una clase de 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 positivas de las clases 1 y 5 que son las clases situadas al mismo nivel sobre el eje 2 y opuestas sobre el eje 1. En la Tabla 5 se muestran los segmentos característicos de esas mismas clases, ampliados a través de la Tabla 6.

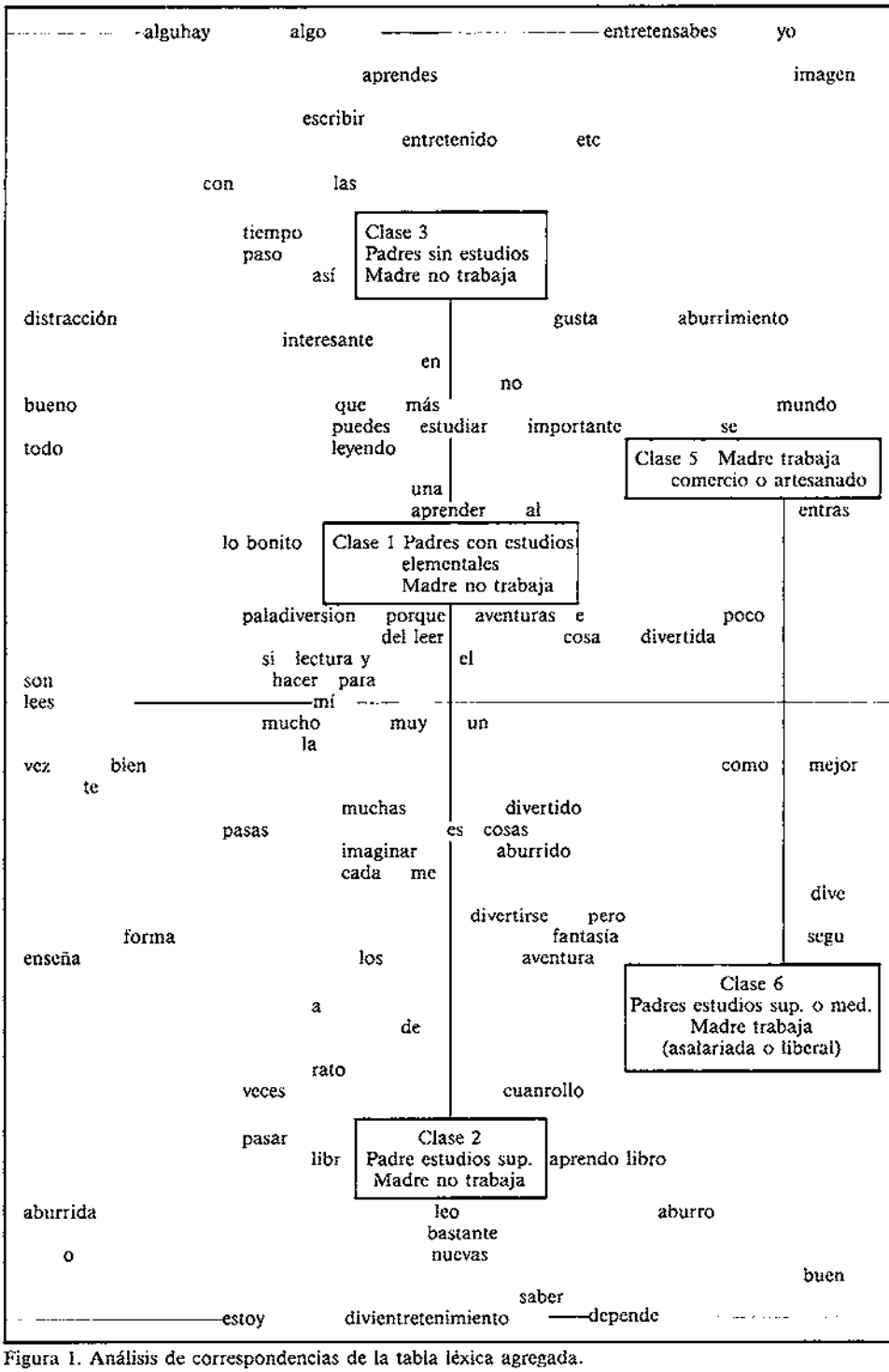


Figura 1. Análisis de correspondencias de la tabla léxica agregada.

TABLA 4. FORMAS CARACTERÍSTICAS DE LAS CLASES 1 Y 5

*Texto número 1* aala = Clase 1/6

| Expresión de la forma gráfica | Porcentaje Interno | Porcentaje Global | Frecuencia Interna | Frecuencia Global | V. Test | Probabilidad |
|---|---|---|---|---|---|---|
| 1 interesante | 1.27 | 0.84 | 27. | 40. | 2.751 | 0.003 |
| 2 entretenido | 0.66 | 0.38 | 14. | 18. | 2.611 | 0.005 |
| 3 también | 0.52 | 0.29 | 11. | 14. | 2.300 | 0.011 |
| 4 pasas | 0.42 | 0.25 | 9. | 12. | 1.826 | 0.034 |
| 5 algunas | 0.33 | 0.19 | 7. | 9. | 1.666 | 0.048 |

*Texto número 5* aa5a = Clase 5/6

| Expresión de la forma gráfica | Porcentaje Interno | Porcentaje Global | Frecuencia Interna | Frecuencia Global | V. Test | Probabilidad |
|---|---|---|---|---|---|---|
| 1 importante | 5.19 | 3.03 | 22. | 144. | 2.402 | 0.008 |
| 2 como | 2.59 | 1.22 | 11. | 58. | 2.251 | 0.012 |
| 3 mejor | 1.18 | 0.38 | 5. | 18. | 2.100 | 0.018 |
| 4 divertirme | 1.65 | 0.72 | 7. | 34. | 1.914 | 0.028 |
| 5 se | 1.65 | 0.72 | 7. | 34. | 1.914 | 0.028 |

TABLA 5. SEGMENTOS CARACTERÍSTICOS DE LAS CLASES 1 Y 5

*Texto número 1* aala = Clase 1/6

| Expresión de la forma gráfica | Porcentaje Interno | Porcentaje Global | Frecuencia Interna | Frecuencia Global | V. Test | Probabilidad |
|---|---|---|---|---|---|---|
| 1 28-divertido y entretenido | 0.98 | 0.43 | 7. | 7. | 2.708 | 0.003 |
| 2 27-divertido y | 3.63 | 2.54 | 26. | 41. | 2.308 | 0.011 |
| 3 71-muy divertido y entretenido | 0.56 | 0.25 | 4. | 4. | 1.763 | 0.039 |
| 4 108-una cosa que me gusta | 0.56 | 0.25 | 4. | 4. | 1.763 | 0.039 |
| 5 46-es una diversión | 0.70 | 0.37 | 5. | 6. | 1.511 | 0.065 |

*Texto número 5* aa5a = Clase 5/6

| Expresión de la forma gráfica | Porcentaje Interno | Porcentaje Global | Frecuencia Interna | Frecuencia Global | V. Test | Probabilidad |
|---|---|---|---|---|---|---|
| 1 104-una cosa muy importante | 1.97 | 0.50 | 3. | 8. | 1.848 | 0.032 |
| 2 73-muy importante | 9.21 | 5.65 | 14. | 91. | 1.739 | 0.041 |
| 3 44-es una cosa | 1.97 | 0.68 | 3. | 11. | 1.425 | 0.077 |
| 4 31-en un mundo | 1.32 | 0.43 | 2. | 7. | 1.101 | 0.135 |
| 5 103-una cosa muy | 1.97 | 0.99 | 3. | 16. | 0.893 | 0.186 |

TABLA 6. SEGMENTOS AMPLIOS CARACTERÍSTICOS DE LAS CLASES 1 Y 5

SELECCIÓN DE INDIVIDUOS O RESPUESTAS CARACTERÍSTICAS (CRITERIO DE FRECUENCIA DE PALABRAS)

*Texto número 1* aala = Clase 1/6

| Criterio de clasificación | Respuesta o individuo característico |
|---|---|
| 2.751 | 1 interesante |
| 2.611 | 2 entretenido |
| 1.375 | 3 muy interesante |
| 1.375 | 4 muy interesante |
| 1.375 | 5 muy interesante |

SELECCIÓN DE INDIVIDUOS O RESPUESTAS CARACTERÍSTICAS (CRITERIO DE CHI-2)

*Texto número 1* aala = Clase 1/6

| Criterio de clasificación | Respuesta o individuo característico |
|---|---|
| 0.882 | 1 es muy importante porque alguien compra un periódico y se entera lo que 1 anuncia, para no ser un analfabeto toda la vida |
| 0.902 | 2 a mí me gusta mucho leer porque como dice el refrán: si te gusta la 2 aventura, lánzate a la lectura, leyendo me lo paso muy bien |
| 0.908 | 3 es aprender leyendo otras cosas nuevas y aventurarte a leer y aprendes 3 cosas que en el cole no. para mí es muy importante |
| 0.909 | 4 para mí leer es muy divertido y me lo paso muy bien leyendo porque aprendo 4 mejor a poner los puntos y las comas |
| 0.914 | 5 muy interesante, muy gracioso y es muy divertido |

## *Selección de las respuestas modales*

La selección de las respuestas modales de las distintas clases (Lebart, 1982) permite extraer respuestas reales tales que su vocabulario sea representativo de vocabulario específico de dicha clase. 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 pueden 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. Se pueden, también, seleccionar las respuestas características siguiendo otro criterio, el criterio del valor-test medio. Según hemos visto en el párrafo anterior, se asigna 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. La Tabla 6 muestra las respuestas modales de la clase 1 obtenidas mediante la utilización de los dos criterios.

## Discusión

Los tratamientos posibles son más numerosos que los que hemos propuesto a partir de este ejemplo, sin embargo hemos pretendido a través de él explicitar básicamente la especificidad de los métodos empleados. La aproximación estadística al análisis estadístico de los datos textuales presentado a través de este artículo ofrece una nueva lectura de los textos, lectura esencialmente distinta y a su vez complementaria con la lectura analizada desde un enfoque mucho más manual. 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, entre los grupos de individuos.

Los datos de encuesta constituyen el terreno de elección de estos métodos. Ante una pregunta abierta concreta y dados diferentes grupos de individuos se pueden obtener, sin ninguna precodificación previa, las características principales de las diferencias entre los grupos. La visualización de las proximidades entre formas y categorías, mediante el análisis de correspondencias de la tabla léxica agregada y/o de la tabla segmental agregada, proporciona un resumen de las similitudes entre los grupos y una descripción de la asociación entre palabras.

También se pueden analizar con provecho otro tipo de textos —textos literarios, discursos políticos, entrevistas no directivas...—. El corpus constituido, sin embargo, debe presentar un cierto grado de homogeneidad y de exhaustividad. Los resultados obtenidos facilitan entonces la construcción de hipótesis y orientan los análisis posteriores.

## Referencias

*Actes de les Jornades d'Anàlisi de Dades Textuals*, Barcelona, 10-12. Diciembre 1990, Facultat d'Informàtica de Barcelona, Universitat Politècnica de Catalunya. Editores: Bécue, M., Lebart, L. y Rajadell, N., Servei de Publicacions de la UPC. Barcelona, 1992.

Bécue, M. (1991). *Análisis Estadístico de Datos Textuales: Métodos de Análisis y Algoritmos*. Paris: CISIA.

Benzécri, J.P. (1973). *La taxinomie*, Vol. I, *L'Analyse de Correspondances*, Vol. II. Paris: Dunod.

Benzécri, J.P. (1981). *Pratique de l'Analyse des Données*, tome 3, Linguistique et Lexicologie. Paris: Dunod.

Bourques, G. & Duchastel, J. (1988). *Restons Traditionnels et Progressifs. Pour une Nouvelle Analyse du Discours Politique*. Montreal: Boréal.

Coulon, D. & Kayser, D. (1986). Informatique et Langage Naturel: Présentation Générale des méthodes d'interprétation des Textes. *Techniques et Sciences Informatiques*. Vol. 5, *2*, 103-128.

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

Guiraud, P. (1960). *Problèmes et Méthodes de la Statistique Linguistique*. Paris: PUF.

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

Lafon, P. & Salem, A. (1983). «L' Inventaire des Segments Répétés d'un Texte», *Mots, 6*, 161-177.

Lebart, L. (1982). L' Analyse Statistique des Réponses Libres dans les Enquêtes Socio-économiques. *Consommation, 1*, pp. 39-62, Paris: Dunod.

Lebart, L., Morineau, A. & Warwick (1984). *Multivariate Descriptive Statistical Analysis*. New York: J. Wiley and Sons.

Lebart, L. & Salem, A. (1989). *Analyse Statistique des Données Textuelles*. Paris: Dunod.
Lebart, L., Morineau, A. & Bécue, M. (avec la coll. de P. Pleuvret et L. Haensler) (1989). *SPAD.T, Système Portable pour l'Analyse des Données Textuelles*. Manuel de Références., Paris: CISIA.
Muller, C. (1968). *Initiation à la Statistique Linguistique*. Paris: Larousse.
Rajadell, N. (1990). *Les Actituds envers la lectura. Un model d'Anàlisi per a l'Educació Primària*. Tesis doctoral no publicada. Universitat de Barcelona.
Rajadell, N. (1991). El análisis de datos en la investigación educativa. *Lectura y Vida, 12, 4*, 31-40.
Reinert, M. (1986). Un Logiciel d'Analyse Lexicale. *Les Cahiers de l'Analyse des Données, 4*, 471-484. Paris: Dunod.
Salem, A. (1982). «Analyse Factorielle et Lexicométrie. Synthèse de Quelques expériences», *Mots, 4*, 147-168.
Salem, A. (1987). *Pratique des Segments Répétés, Essai de Statistique Textuelle*. Paris: Klincksieck.
Schuman, H. & Presser, F. (1981). *Questions and Answers in Attitude Surveys*. New York: Academic Press.
Yule, G.U. (1944). *A Statistical Study of Vocabulary*. Cambridge University Press.
Zipf, G.K. (1935). *The Psychobiology of Language, an Introduction to Dynamic Philology*. Boston: Houghton-Mifflin.