Metodología de Encuestas
Vol 5, Núm 1, 2003, 55-66

ISSN: 1575-7803

ANÁLISIS TEXTUAL: GENERACIÓN Y APLICACIONES[1]

Karmele Fernández Aguirre
Universidad del País Vasco

RESUMEN

El presente trabajo surge como comentario al artículo de Elena Abascal y M. de los Ángeles Franco titulado "Análisis textual de encuestas: aplicación al estudio de las motivaciones de los estudiantes en la elección de su titulación". Este artículo presenta una síntesis muy interesante del análisis estadístico textual aplicado al caso de respuestas libres a cuestiones abiertas dentro de un cuestionario.

En las investigaciones mediante encuestas se plantean, cada vez más a menudo, este tipo de cuestiones junto con las cerradas entre las que se encuentran las referidas a las características de los encuestados. Estas últimas juegan un papel importante debido a la relación que puede establecerse entre los dos tipos de cuestiones. La síntesis efectuada en el artículo constituye un claro ejemplo de integración de los aspectos lexicales y estadísticos. Así, en cuanto a la metodología de encuestas, las autoras combinan de manera fructífera, métodos de estadística descriptiva multivariante con aspectos lexicales de términos, segmentos repetidos, formas...y sus contextos.

En esta nota, vamos a introducir una serie de comentarios, de tipo histórico, sobre el desarrollo del análisis textual dentro de la escuela francesa de análisis de datos, incidiendo, al mismo tiempo, en ciertos aspectos técnicos reseñables de la metodología estadística subyacente, para finalizar con una visión sucinta del panorama actual dentro de los análisis estadísticos textuales.

Palabras clave: *análisis textual, análisis de correspondencias, clasificación automática.*

[1] Este trabajo ha sido financiado por el grupo de investigación consolidado UPV 038.321-13631/2001.

Nacimiento y primeros pasos del análisis de datos textual

La técnica del Análisis Factorial de Correspondencias, como método de análisis descriptivo multivariante, fue planteada por J.P. Benzécri en un curso de lingüística matemática publicado en 1964 e impartido en la Facultad de Ciencias de Rennes (Francia) desde el comienzo de los años 60 (Benzécri y col., 1981). En esta misma facultad B. Escofier defendió, en 1965, su tesis doctoral titulada L´Analyse des Correspondances, donde se resaltan las principales propiedades del método.

El profesor Benzécri, padre de la escuela francesa de análisis de datos, es un científico de una talla excepcional influenciado por las corrientes lingüísticas del siglo XX. Estas corrientes pueden asociarse a dos nombres fundamentalmente, el lingüista con formación matemática y filosófica Noam Chomsky y el que fue su maestro Zellig Harris.

Para Chomsky la lingüística es una ciencia deductiva tal que partiendo de ciertos axiomas se engendren las lenguas reales. En su obra *Syntactic Structures* (Chomsky, 1956) defiende la existencia de una gramática universal bajo la cual se adecuan las gramáticas particulares, bajo el cumplimiento de ciertos supuestos. Así, el lingüista debe formular un modelo teórico, como un modelo ideal, que no surgirá jamás de los datos. Chomsky obtiene frases escritas en un vocabulario natural mediante fórmulas de tipo deductivo, partiendo de símbolos algebraicos que forman lo que él llama vocabulario no terminal.

Harris, por el contrario, parte del discurso y mediante un procedimiento inductivo sistemático descubre la estructura que expresa, también, mediante símbolos algebraicos. Harris entiende por distribución de una palabra el conjunto de todos sus contextos posibles: *"The distribution of an element will be understood as the sum of all its environments"* (Harris, 1954). La lingüística estructural americana se había caracterizado por el estudio de las lenguas amerindias, de las cuales algunas habían sido recopiladas, antes de su extinción, de labios de personas incapaces de expresarse claramente en ninguna lengua conocida. Este hecho llevó al desarrollo de métodos originales que permitían utilizar un material o corpus limitado absteniéndose de considerar el sentido. En estas investigaciones, el análisis distribucional de Harris llega al máximo. Así, en una serie de artículos de gran influencia, afirma que todo o casi todo de una lengua puede obtenerse, sin recurrir al sentido, por el análisis de los hechos distribucionales.

En opinión de Benzécri ambos lingüistas se mueven sobre un mismo eje en sentidos opuestos y, sin duda alguna, elige el sentido de Harris. Benzécri considera idealistas las tesis de Chomsky debido a que tienden a separar el espíritu de los hechos de su inspiración y de su objeto. Afirma que: "a falta de un algoritmo universal que permita sobrepasar las 10.000 páginas de texto de una lengua con sintaxis provista de una semántica, pretendemos ofrecer al lingüista, por medio de la estadística, un método inductivo eficaz para tratar de modo útil las tablas de datos que podrían ser objeto de recuento inmediato con el horizonte de investigaciones sucesivas escalonadas, no dejando nada en la sombra de las formas, de los sentidos o del estilo" (Benzécri, 1982).

Benzécri aborda así un nuevo método, inductivo y algebraico, al que denomina Análisis de Correspondencias, como método de estadística multivariante para el tratamiento de grandes tablas de datos (en principio lingüísticos) en base a las posibilidades

abiertas por el ordenador en los años 60. Benzécri, que tomó el término mismo de distribución de Harris, definió la distancia distribucional entre distribuciones condicionadas conocida como distancia P^2 (chi-cuadrado).

El Análisis de Correspondencias (en adelante AC), aplicado actualmente en múltiples disciplinas, reposa en una noción de lingüística, la equivalencia distribucional, propiedad fundamental de AC, tal como veremos más adelante.

En una construcción inductiva de la lingüística a partir de los datos, sin supuestos *a priori*, hay que observar las palabras, las frases y el discurso como una sucesión de elementos. Una palabra o un segmento (sucesión de palabras) estará caracterizado por el conjunto de todos los contextos en los que pueda insertarse para obtener una frase correcta. Si nos remitimos al estudio de un corpus cerrado, una frase correcta es la que tiene sentido en el contexto de ese corpus. En su estudio, conviene proceder a la búsqueda de frases correctas en etapas sucesivas, de modo que las primeras sean las más accesibles y se avance lentamente y sin interrupciones. Se considerarán, primero, las frases más cortas que sigan un modelo como sujeto-verbo-complemento... Así, la noción extensible de contexto permite que una palabra sea tratada en los contextos que comprenden únicamente dos palabras.

En los primeros tiempos de la estadística lingüística se comienza a investigar con tablas como la siguiente (Benzécri, 1982). Sea una tabla de datos rectangular con *I* nombres en fila y *J* verbos (o adjetivos) en columnas, en la intersección de la fila *i* y de la columna *j* tenemos el número k_{ij} de veces que el nombre *i* se asocia al verbo *j*, la asociación entre *i* y *j* será tanto más lícita cuanto mayor sea k_{ij} . Si consideramos el nombre *i*, conviene conocer la importancia relativa de esta asociación entre *i* y *j* medida por el cociente $k_{ij} /k_{i\cdot}$ es decir, del número k_{ij} de veces que i se ha empleado con *j* entre el número total de veces, k_{ij} , que se ha empleado *i* (suma de la fila *i*). La sucesión de valores $k_{ij} /k_{i\cdot}$ para todos los verbos *j, j´, j´´*... será el perfil del nombre *i*. Dos nombres serán sinónimos (desde el punto de vista de su asociación con los verbos) si tienen el mismo perfil, esta sinonimia es aceptable en el sentido de que dos seres que corren, ríen, cantan... con la misma frecuencia no pueden más que parecerse. La definición de perfil de un verbo *j* es análoga. En la práctica no es verosímil que dos nombres *i* e *i´* o dos verbos *j* y *j´* tengan el mismo perfil, pero la similitud de perfiles puede ser más o menos grande lo que lleva al problema de la representación espacial del conjunto de perfiles.

El problema de la representación espacial, sugiere el recurso al Análisis Factorial. Al comienzo de los años 60 este método estaba reservado a la práctica de psicólogos y biómetras y el problema de la reducción de variables aparecía ligado a hipótesis de normalidad multivariante. Para Benzécri como matemático formado en el álgebra lineal, en la geometría euclidiana multidimensional y en el cálculo tensorial, no existía más que un único problema: los individuos (filas de la tabla) son puntos o vectores de un espacio; las variables (columnas de la tabla) son formas lineales, o vectores de un espacio dual. Los coeficientes de correlación de los estadísticos, identifican a los productos escalares de los geómetras y el espacio y su dual son isomorfos si se ha fijado una métrica euclídea (o fórmula de distancia). Para aplicar las fórmulas clásicas únicamente quedaba por resolver el problema de fijar una métrica.

Principio de equivalencia distribucional

Es en este estadio en el que surge el principio de equivalencia distribucional: la distancia *d(i, i´)* entre dos nombres i e *i´* no debe modificarse si se identifican dos verbos *j* y *j´* que son sinónimos en cuanto a su distribución (tienen el mismo perfil); es decir, si se reemplazan las columnas *j* y *j´* que son proporcionales entre sí por la columna *j´´* suma de las anteriores y proporcional a ambas.. Así, por ejemplo, si *decir* y *hablar* tienen los mismos sujetos en las mismas proporciones, se pueden identificar ambos verbos. El principio de equivalencia distribucional unido a la exigencia de la geometría euclidiana multidimensional de que la distancia sea cuadrática, lleva a la formulación de la distancia distribucional.

La distancia entre los perfiles de dos filas, entre el perfil fila *i* y el perfil fila *i´*, será:

$$d^2(i,i') = \sum_j \frac{k}{k_{.j}} \left(\frac{k_{ij}}{k_{i.}} - \frac{k_{i'j}}{k_{i'.}} \right)^2 ,$$

siendo k el efectivo total de la tabla.

Posteriormente la distancia distribucional se ha denominado distancia chi-cuadrado debido a su parentesco con la prueba P^2 clásica.

La equivalencia distribucional confiere gran estabilidad a los resultados del AC, el caso de la proporcionalidad perfecta entre dos columnas (o filas) no se dará en la práctica, pero pueden ser próximas. Si éstas se agregan los resultados del análisis no se alteran de modo sensible.

A partir de la tabla *(I x J)* se construyen dos nubes de puntos, *N(I)* y *N(J)*, o conjuntos de puntos previstos de masas y distancias (en el ejemplo nombres y verbos) Cada una de estas nubes se sitúa en un espacio euclidiano en el que se buscan los ejes principales de inercia, con el criterio de ajuste mínimo cuadrático. Para poder observar las nubes en un espacio accesible a los sentidos, éstas se proyectan en planos formados por dos de sus ejes principales. En un principio, las dos nubes de puntos *N(I)* y *N(J)*, construidas de forma simétrica a partir de la misma tabla de datos, flotaban en espacios diferentes y no se relacionaban los ejes de ambas nubes.

Representación simultánea de los dos conjuntos puestos en correspondencia

La correspondencia entre los datos, nombres y verbos, sugiere una identificación paralela de los ejes factoriales; así, si en un eje se obtuviera una graduación desde nombres inanimados en el extremo izquierdo hasta nombres animados en el derecho, este hecho tendría que tener un reflejo en las afinidades entre nombres y verbos. Resulta muy interesante la lectura (Benzécri, 1982) de la construcción geométrica que llevó al descubrimiento de la coincidencia de los ejes de la nube de parejas de puntos *(i, j), N(I x J),* del espacio suma directa de los espacios de las dos nubes, con los ejes de las nubes *N(I)* y *N(J)* de los subespacios separados de ambas.

En la tesis doctoral de B. Escofier se recoge el fruto de la experimentación con un ordenador IBM 1620, instalado en 1963 en un laboratorio de cálculo de Rennes, y de las demostraciones que constituyen las propiedades inigualables del AC (Escofier, 1965). Así, se encuentra la equivalencia de los factores obtenidos de las tres nubes *N(I x J), N(I), N(J)* con las fórmulas de transición que permiten pasar, de modo muy simple, de los factores de un espacio a los de otro. En consecuencia, se cumple el deseo de que en la representación simultánea un elemento *i* (resp. *j*) esté rodeado de los elementos *j* (resp. *i*) con los que más se asocia.

Los factores obtenidos del análisis de las nubes N(I), N(J) corresponden a los mismos valores propios 8, lo que permite analizar una sola nube. Las fórmulas o relaciones de transición, también llamadas baricéntricas se escriben para el factor de orden ":

$$F_\alpha(i) = \frac{1}{\sqrt{\lambda_\alpha}} \sum_j \frac{k_{ij}}{k_{i.}} G_\alpha(j) \qquad (1)$$

$$G_\alpha(j) = \frac{1}{\sqrt{\lambda_\alpha}} \sum_i \frac{k_{ij}}{k_{.j}} F_\alpha(i) \qquad (2)$$

Estas fórmulas permiten precisar las relaciones entre los puntos que representan por un lado las filas (nombres) y por otro las columnas (verbos) de la tabla de contingencia. Los elementos que intervienen en las mismas son:

- *F "(i)*: proyección de la fila i sobre el eje de rango " de la nube de perfiles fila.
- *G "(j):* proyección de la columna j sobre el eje de rango " de la nube de perfiles columna.
- *8"*: valor común de la inercia asociada a cada uno de los ejes F" y G"

En la representación simultánea, las relaciones entre la posición de los puntos fila y de los puntos columna pueden ser descritas así (véase (1)): prescindiendo del factor $(1/\sqrt{\lambda\alpha}$, la proyección de la fila *i* sobre el eje " es el centro de gravedad (baricentro) de las proyecciones de las columnas *j*. Este centro de gravedad se calcula como una media ponderada de esas proyecciones, siendo los valores $k_{ij}/k_{i.}$ las ponderaciones. La segunda relación de transición se describe de modo análogo.

Al atraer los elementos "pesados" al baricentro, una columna *j* atrae tanto más a una fila *i* cuanto más elevado sea el valor $k_{ij}/k_{i.}$. En la práctica, los puntos alejados del origen llaman particularmente la atención, debido a que son los que tienen un perfil muy diferente del perfil medio, situado en el origen de coordenadas. Como vemos, es posible interpretar la posición de una fila (un nombre) con respecto a un conjunto de columnas (verbos) y este hecho es de gran interés para el estudio de las asociaciones entre las modalidades en filas y en columnas de una tabla.

El isomorfismo entre un espacio vectorial y su dual, unido a la perfecta simetría del papel que juegan los dos conjuntos I, J puestos en correspondencia, había permitido unificar los dos puntos de vista precedentes de la Estadística Multivariante clásica; el del Análisis Factorial de un conjunto de variables y el de escalas multidimensionales o Reescalado Multidimensional: representación de una nube de puntos. Cada i (nombre), es a

la vez un punto caracterizado por sus asociaciones con los elementos j (conjunto de verbos) y lo mismo para los j. Quedaba así probado que los dos puntos de vista conducen a los mismos factores, a condición de elegir de modo adecuado los coeficientes de ponderación.

Evolución desde los años sesenta hasta mediados de los ochenta

El análisis de correspondencias, ignorado por el mundo científico anglosajón hasta bien entrados los ochenta, se manifestó rápidamente como una técnica nueva y poderosa a la hora de estudiar la estructura de grandes corpus, extrayendo las principales características y en definitiva, la principal información contenida en los mismos. Desde un principio fueron desarrollándose programas de ordenador que permiten obtener el recuento de formas, segmentos,... de los corpus introducidos en el mismo. Al mismo tiempo se desarrollaron métodos de Análisis de Clasificación Automática y se comenzaron a aplicar como complementarios del Análisis de Correspondencias. Además de la técnica de proyección de elementos suplementarios consistente en proyectar filas o columnas de la tabla de datos, sobre los ejes o planos factoriales, sin haber intervenido en la formación de los mismos.

Benzécri había propuesto un método estadístico para la resolución de los problemas fundamentales que interesaban al lingüista. Con este método, se efectúa una abstracción cuantitativa partiendo de tablas de datos diversos, para obtener por medio del cálculo, nuevos parámetros que miden entidades situadas en un nivel de abstracción superior al de los hechos recensados en primer lugar.

En 1973 se celebró en Grenoble un primer coloquio consagrado al análisis de datos lingüísticos, donde se dio cuenta de los trabajos de AC realizados entre 1963 y 1973. Posteriormente, se celebró una jornada en la Ecole Normale Supérieure de Fontenay-Saint-Cloud y un segundo coloquio en Montpelier en 1976. Es este mismo año el del comienzo de la publicación de la revista *Les cahiers de l´analyse des données* fundada y dirigida por J-P. Benzécri hasta 1997. En ella se recogen interesantes artículos de análisis de datos textual; además, el libro ya citado (Benzécri y col. 1981) se dedica íntegramente a las aplicaciones de lingüística y lexicología después de unas introducciones metodológicas. Esta obra recoge trabajos bajo cuatro epígrafes correspondientes a otros tantos tipos de tablas: de contingencia, de presencia-ausencia o lógicas, de cuestiones o tablas de Burt, y de secuencia o lógicas en yuxtaposición. Todos estos estudios responden a distintos objetivos, bien sea el establecimiento de parentesco entre textos de autores diferentes, aproximando los que tienen un vocabulario similar, o bien el establecimiento de tipologías entre los capítulos de una misma obra, o bien el establecimiento de los posibles cambios de vocabulario de una organización política o sindical en el curso de un período dado.

En comparación con los AC realizados sobre otro tipo de datos, el tratamiento de datos lingüísticos presenta peculiaridades derivadas de la multidimensionalidad propia de la materia. Así, el problema del léxico es más complejo que el de otros elementos del lenguaje ya que está formado por un conjunto de unidades que sin ser infinito, en el sentido matemático del término, no da la impresión de ser estrictamente finito.

Características observadas en las aplicaciones

Durante estas primeras décadas se pusieron de manifiesto ciertas características, de las cuales algunas son comunes con otro tipo de aplicaciones, como pueden ser la búsqueda empírica de la estabilidad de resultados introduciendo modificaciones en las tablas y otras son más específicas de las aplicaciones a tablas lexicales. En cuanto a las primeras, existía una preocupación sobre la disparidad de pesos de las partes del corpus en columnas de una tabla lexical. Así, estos pesos suelen ser parecidos, a lo sumo una columna suele cuadruplicar el resto. Sin embargo, trabajos de J.P. Benzécri y L. Lebart señalan que no es necesario recurrir a un análisis ponderado aunque los pesos de las columnas sean muy dispares (Benzécri, 1979), (Lebart, 1979), y conclusión segunda en (Fernández Aguirre, 1988).

Un tema muy debatido en esta época fue el establecimiento o no de un umbral mínimo de frecuencia para las formas gráficas, a menudo en filas de la tabla lexical. Existía la creencia, por otra parte lógica, de que una forma poco utilizada pudiera serlo debido al azar. Los términos utilizados una sola vez, denominados "hapax", constituyen un caso extremo y siempre se prescinde de los mismos.

El establecimiento de un umbral mínimo fue una práctica habitual desechada, con posterioridad, por los equipos del laboratorio de lexicología y textos políticos de la Ecole Normale Supérieure de Fontenay-Saint-Cloud. A. Salem, estadístico de este centro publicó dos trabajos en los que mostraba que la reducción del umbral de frecuencia no hacía sino favorecer la discriminación, en el primer plano factorial, entre los diarios de Hébert, Roux y Leclerc, aspirantes a la sucesión de Marat en 1973, empeorando únicamente cuando se introducen los "hapax" (Salem, 1981) y (Salem, 1982). Asimismo, J.P. Benzécri afirma que la solución de considerar todas las formas que se encuentren en al menos dos textos del corpus estudiado, puede llevar a excelentes resultados (Benzécri, 1984), la conclusión tercera en (Fernández Aguirre, 1988) apuntaba también en la misma dirección.

En los años ochenta subsistieron dos formas de proceder en cuanto a las reglas de segmentación que fijan los elementos fila de la tabla lexical. Ciertos trabajos siguen una forma de proceder clásica (Reiner, 1983), (Ait-Hamlat, 1984) y no retienen entre estos elementos más que lo que denominan *"mots pleins"*, es decir, sustantivos, verbos, adjetivos, ... prescindiendo de lo que consideran categorías gramaticales vacías como pronombres, preposiciones, artículos,... Esta forma de proceder fue perdiendo fuerza. Así, en el laboratorio de St-Cloud no se prescindió de estas formas funcionales a las que llaman *"mots utils"*, denominados términos instrumentales o "palabras herramienta" en el artículo objeto de esta nota. A. Salem, que colabora con los lexicólogos A. Geffroy, J. Guilhaume, P. Lafon y M. Tournier, afirma (Salem, 1982) que en la práctica la distinción entre las formas funcionales y formas lexicales parece muy difícil de realizar por lo que a partir de 1976 abandonan dicha práctica. Dos años después, A. Salem critica el proceder de A. Ait-Hamlat y defiende el del laboratorio de St-Cloud más explícitamente aduciendo además otras dos razones: ciertas formas funcionales juegan un papel de primer orden en el análisis de textos políticos y pone por ejemplo el paradigma *la liberté, les libertés* (Salem, 1984). Numerosos lingüistas se interesan particularmente en las

categorías gramaticales que constituyen según E. Benveniste el aparato formal de la enunciación (pronombres personales, pronombres y adjetivos posesivos, demostrativos...) de manera que fijan los trazos explícitos de la presencia del locutor en el texto que ha producido.

En el laboratorio de St-Cloud, como puede verse de gran transcendencia, surge la técnica de "Inventario de Segmentos Repetidos", en un trabajo de P. Lafon en colaboración con A. Salem (Lafon y Salem, 1983). En síntesis el método consiste en la obtención de todas las secuencias de formas, denominadas segmentos, de longitud 2, 3,... hasta 25, que se encuentran varias veces en un corpus textual. El método denominado de "vecinos recurrentes" permite documentar los análisis a partir de formas simples, mediante la reasignación de los segmentos repetidos en el corpus. En este mismo trabajo se afirma, que la tabla de datos analizada contendrá el conjunto de formas gráficas, como elementos principales (activos en el análisis) y el conjunto de segmentos repetidos, como elementos suplementarios. Estos, colaboran a mejorar la descripción de los ejes factoriales, así como a obtener resultados relativos a los segmentos, contextos cortos de formas en los que su distribución es aclarada por el AC de la tabla lexical.

Desarrollo del análisis de datos textual

A partir de mediados de los ochenta el Análisis de Correspondencias se ha ido haciendo cada vez más popular. A este respecto, la publicación en inglés de las obras de L. Lebart (Lebart, Morineau y Warwick, 1984) y de M. Greenacre (Greenacre, 1984) ha sido probablemente decisiva. El AC y en general los métodos de análisis multivariantes se han ido desarrollando conjuntamente en múltiples áreas.

En particular, la conjunción entre el análisis textual y el análisis de datos multivariante sufre un continuo desarrollo. A partir de 1990 comienzan a celebrarse las *Journées Internationales d´Analyse Statistique des Données Textuelles (JADT)*. A la primera celebrada en Barcelona le siguió la segunda celebrada en Montpelier en 1993, en 1995 se celebró la tercera en Roma y continúan de modo regular hasta nuestros días.

En estos años, el desarrollo de los programas informáticos se produce de forma paralela al de los métodos y sus aplicaciones a grandes conjuntos de datos procedentes de encuestas socioeconómicas, entrevistas, investigaciones litararias, de textos políticos, archivos históricos, bases de datos documentales, etc. L. Lebart, que años antes había comenzado a investigar en el campo de respuestas libres a cuestiones abiertas en encuestas, desarrolla junto a A. Morineau un primer módulo de tratamiento de textos en el sistema SPAD (Lebart y Morineau, 1984). Posteriormente M. Bécue Bertaut presenta en la Facultd de Informática de Barcelona su tesis doctoral titulada: Un sistema informático para el Análisis de Datos Textuales; así, en 1988 se presenta el programa SPAD.T (Bécue Bertaut, 1991). La autora investiga de forma sistemática los métodos lexicométricos y estadísticos utilizados en el análisis textual y se centra en un campo privilegiado: el del tratamiento de respuestas abiertas de encuestas y de su relación con respuestas e información cerrada. El sistema informático desarrollado facilita la experimentación que lleva a la evolución y perfeccionamiento de los métodos.

Por otra parte, A. Salem inició, a finales de los ochenta, la serie Léxico Software (Salem, 1987) y comenzó una estrecha colaboración con L. Lebart (Lebart y Salem,

1988) y (Lebart y Salem, 1994). Esta obra va incorporando nuevos métodos e introduciendo mejoras en los anteriores. En consecuencia, publican dos nuevos títulos en colaboración con L. Berry (Lebart, Salem y Berry, 1998) y M. Bécue Bertaur (Lebart, Salem y Bécue Bertaut, 2000) respectivamente.

Perfeccionamiento de los análisis textuales

Actualmente se progresa en los análisis de respuestas libres a cuestiones abiertas y su relación con el resto de la información recogida en la encuestas. La aplicación de métodos de clasificación automática como análisis complementarios es cada vez más frecuente. Las técnicas de visualización, por importantes que puedan ser, se limitan a planos factoriales y en el caso de grandes ficheros de datos lexicales resultan insuficientes. Así, la utilización conjunta de análisis factoriales y de clasificación sobre los primeros factores es de gran interés. Los algoritmos empleados en las técnicas de agrupamiento permiten emplear únicamente la dimensión real de la nube de puntos, al tomar la información proyectada en los primeros factores, prescindiendo de lo que puede considerarse ruido debido a las fluctuaciones del muestreo.

El método de Clasificación Ascendente Jerárquica comúnmente utilizado (criterio de Ward generalizado) se basa en cálculos de distancia entre elementos de base tomados dos a dos. En general la distancia aplicada es la distancia P^2 (chi-cuadrado) entre perfiles al igual que en AC. La complementariedad entre las dos técnicas estudiadas en (Lebart, 1994) es muy recomendable en análisis descriptivos y exploratorios de datos procedentes de grandes encuestas (Fernández Aguirre, Gallastegui Zulaica, Modroño Herrán y Núñez Antón, 2003) o complejos como los lexicales. La clasificación puede versar sobre los términos o las formas textuales de una tabla lexical efectuando un análisis directo de las respuestas o documentos o puede efectuarse sobre las variables de caracterización de los individuos encuestados o incluso sobre otras cuestiones cerradas en el cuestionario. En todos los casos los elementos no activos en el análisis pueden proyectarse como suplementarios para ilustrar la descripción de los planos factoriales y clases. Las coordenadas de estas variables o modalidades suplementarias pueden valorarse mediante los *"valores-test"*, que proporcionan una medida de su significación estadística (Lebart et al., 2000).

El artículo que comentamos, utiliza perfectamente además de la técnica de segmentos repetidos los cursos desarrollados en estos años: los relativos a elementos característicos. Éstos pueden ser palabras, lemas (formas gráficas que tienen la misma raíz y un significado equivalente) o segmentos; asimismo, pueden ser respuestas modales o características. Los elementos característicos de una parte del corpus, de una respuesta o de un conjunto de ellas, lo son debido a su utilización muy superior (por exceso) o muy inferior (por defecto) a la global en todo el corpus. La evaluación se efectúa por medio de un *"valor-test"* asociado a una probabilidad obtenida mediante la comparación de proporciones, en cada parte del corpus y en el total, en el marco de la ley hipergeométrica (Lebart et al., 1998) o (Lebart et al., 2000).

Nuevas tendencias

En los párrafos anteriores hemos pretendido dar una visión de la estadística textual de la escuela francesa de análisis de datos. Debemos también mencionar que en los últimos tiempos, tiempos de globalización, de internet, de multiculturalidad, se interrelacionan distintas tendencias, distintos idiomas y los campos de aplicación se multiplican. Como ejemplo cabe mencionar que en las *Journées Internationales d´Analyse Statistique des Données Textuelles (JADT´2000)* las dos conferencias invitadas fueron: *Word Frecuency Distribution* de R. H. Baayen y *Fut-il travailler pour etre heureux?* de C. Baudelot y M. Gollac, las áreas temáticas asimismo fueron bilingües francés/inglés.

Recientemente, además de continuarse con ámbitos de aplicación ya clásicos, se consideran particiones longitudinales de corpus, series de tiempo textuales y análisis discriminante textual. En cuanto a series de tiempo textuales pueden considerarse particiones de respuestas abiertas según grupos de edad, rentas mensuales, número de hijos, nivel de educación... pueden considerarse discursos de algún personaje político a lo largo del tiempo o la progresión del discruso de un fiscal (Lebart et al., 2000); pero, también existe un campo que constituye un auténtico reto: el tratamiento de datos textuales para predicciones en el campo de los mercados financieros. En cuanto al análisis discriminante entra dentro de las técnicas estadísticas decisionales empleadas para atribuir un texto a un autor o a una fecha, o seleccionar un documento en base a la respuesta a una pregunta y codificar información expresada a diario en modo textual. La idea es extraer los aspectos invariantes del autor o del período que pueden permanecer ocultos al lector. Se trata de análisis discriminante basado en reconocimiento de patrones o estilometría, un ejemplo clásico constituye el trabajo (Mosteller y Wallace, 1964) sobre la autoría de 12 de los *Federalist Papers*. El corpus lo constituyen 77 textos políticos anónimos de los que 12 eran de autoría difícilmente atribuible. Análisis estadísticos basados en la frecuencia de ciertos términos identificaron al autor más probable de los dos posibles. Los métodos usados en la mayoría de los trabajos se basaban en la construcción de índices en función de la longitud de las palabras, o de las frases, de la frecuencia de las palabras, de la riqueza de vocabulario,... El uso sistemático de las técnicas de análisis de datos (Análisis de Correspondencias y Clasificación Automática) ha supuesto un nuevo enfoque y un avance todavía medianamente reconocido.

Aunque no totalmente independiente del reconocimiento de patrones, existe otra área conocida como análisis discriminante global (Lebart et al, 1998) que incide sobre todo en el contenido, el significado y la esencia del texto. Este aspecto interesa en aplicaciones de recuperación de la información, codificación automática y análisis de respuestas libres en encuestas. En concreto, la recuperación de la información o *Information Retrieval* es hoy día una disciplina autónoma (Salton y Mc Gil, 1983) y (Salton, 1988) aplicada a grandes matrices de datos en múltiples contextos como lingüística computacional, caracterización de documentos por temas, identificación de tendencias en documentos... Las técnicas multivariantes más eficaces de acuerdo con los propios autores son similares a las debidas a Benzécri (Benzécri, 1977), (Benzécri y col., 1981) y (Lebart, 1982). Por ejemplo, (Deerwester, Dumais, Furnas, Landauer y Harshman, 1990) usan un método muy similar al análisis discriminante en los primeros ejes principales de

un AC al que llaman *Latent Semantic Indexing*. Asimismo, mucho autores usan la descomposición en valores singulares que está en la base tanto del Análisis de Correspondencias como del Análisis de Componentes Principales, como técnica de minería de datos aplicada a textos (*Textual Data Mining*).

En cuanto al análisis de respuestas libres de encuestas L. Lebart presenta en (Lebart et al., 1998) un interesante trabajo de comparación de respuestas abiertas en distintas lenguas, en el marco del análisis discriminante global. Se trata de una encuesta sobre hábitos de alimentación en tres grandes metrópolis: Paris, New York y Tokyo. Se obtienen seis grupos demográficos al cruzar las dos categorías de género con tres grupos de edad. El autor procede a la comparación entre las tres ciudades, a priori muy heterogéneas entre sí, y muestra que es posible predecir la pertenecia de un individuo a un grupo en base a las respuestas a una cuestión abierta.

La posibilidad de comparaciones múltiples en base a textos en diferentes idiomas, o en otros contextos, apunta como una posibilidad de avances en futuras investigaciones.

Referencias

Ait-Hamlat, A. (1984) Analyses des repetitions et indexation automatique des documents. *Les Cahiers de l´Analyse des Données*. IX-2, 173-204.

Bécue Bertaut, M. (1991 Análisis de Datos Textuales. Métodos Estadísticos y algoritmos. Paris, CISIA.

Benzécri, J.P. (1977) Analyse discriminante et analyse factorielle. *Les Cahiers de l´Analyse des Données*. II-4, 369-406.

Benzécri, J.P. (1979) Sur l´analyse d´un tableau dont l´une des colomnes a un poids prédominant. *Les Cahiers de l´Analyse des Données*.

Benzécri, J.P. (1982) Historie et Préhistorie de l´analyse des données. Paris. Dunod.

Benzécri, J.P. (1984) Description des textes et analyse documentaire. *Les Cahiers de l´Analyse des Données*. IX-2, 205-211.

Benzécri, J.P. y col. (1981) Pratique de l´analyse des données, T.3, Linguistique et lexicologie. Paris. Dunod.

Chomsky, N (1956) *Syntactic Structures*. La haye. Mouton.

Deerwester, S., Dumais, S.T., Furnas, G., Laundauer, T.K. y Harshman, R. (1990) Indexing by latent semantic analysis. *J. Of the Amer. Soc. For Information Science* 41-6. 391-407.

Escofier, B. (1965) L´Analyse des Correspondances. PhD thesis. Facultad de Ciencias de Rennes. Publicada en 1969 en Cahiers de Bureau Universitaire de Recherche Opérationnelle, nº 13.

Fernández Aguirre, K. (1988) Análisis Multivariante de Tablas de gran inercia. Aplicación al corpus con terminología económica en Euskara. PhD Thesis. Facultad de Ciencias Económicas y Empresariales. Bilbao.

Fernández Aguirre, K., Gallastegui Zulaica, I., Modroño Herrán, J.I. y Núñez Antón, V. (2003) Clients characteristics and marketing of products: some evidence from a financial institution. *The International Journal of Bank Marketing*. En prensa.

Greenacre, M. (1984) Theory and Applications of Correspondence Analysis. London. Academic Press.

Harris, Z. (1954) Distributional Structure. *Word* 10-2-3. 146-162

Lafon, P. y Salem, A. (1983) L´Inventaire des segments répétes d´un texte. *M.O.T.S.*

Lebart, L. (1979) Example d´analyse d´un tableau dont l´une des colomnes a un poids prédominant. *Les cahiers de l´analyse des Données*.

Lebart, L. (1982) Exploratory anályis of large sparse matrices, with application to textual data. *COMPSTAT*. 67-76.

Lebart, L. (1994) *Complementary use of Correspondence Analysis and Cluster Analysis*. Correspondence Analysis in the Social Sciences. Greenacre, M. And Blasius, J.

Lebart, L. y Morineau, A. (1984) SPAD. Système Portable pour l´Analyse des Données. T. III. Paris. CESIA.

Lebart, L. y Salem, A. (1988) *Analyse Statistique des Données Textuelles*. Paris. Dunod.

Lebart, L. y Salem, A. (1994) *Statistique Textuelle*. Paris. Dunod.

Lebart, L., Morineau, A. y Warwick, K. (1984) *Multivariate Descriptive Statistical Analysis*. New York. John Wiley and sons.

Lebart, L., Salem, A. y Bécue Bertaut, M. (2000) *Análisis Estadístico de Textos*. Lleida. Milenio.

Lebart, L., Salem, A. y Berry, E. (1998) *Exploring Textual Data*. Dordrecht. Kluwer Academic Publisher.

Mosteller, F. y Wallace, D.L. (1964) *Inference and disputed authorship: The Federalist*. Reading Mass. Addison Wesley.

Reinert, A. (1983) Une méthode de classification descendante hiérarchique: Application a l´analyse lexicale par contexte. *Les cahiers de l´analyse des données*. VIII-2. 187-198.

Salem, A. (1981) *Signalement et Inventaire Lexical: Textes Politiques FranÇais de 1973*. Pratique de l´analyse des données. T. 3. Lingüistique et Lexicologie. Paris. Dunod. Chapter LC1-5.

Salem, A. (1982) Analyse factorielle et lexicométrie, synthese de quelques experiences. *M.O.T.S*.

Salem, A. (1984) La typologie des segments répétes dans un corpus, fondée sur l´analyse d´un tableau croissant mots et textes. *Les cahiers de l´analyse des données*. IX-4. 489-500.

Salem, A. (1987) *Le lexicloud. Programes pour le traitament lexicométrique des textes*. Ecole Normale Supérieure de Fontenay-Siant-Cloud.

Salton, G. (1988) Automatic Text Processing: Transformation, Analysis and Retrieval of Information by Computer. New York. Addison-Wesley.

Salton, G. y Mc Gil, M.J. (1983) *Introduction to Modern Information Retrieval*. International Student Edition.