ANÁLISIS ESTADÍSTICO DE DATOS CUALITATIVOS TEXTUALES: EL ENFOQUE LEXICOMÉTRICO

Javier Gil, Eduardo García, Gregorio Rodríguez y Ana Corrales
Universidad de Sevilla

El propósito de este trabajo es revisar el modo en que es analizada la información textual desde la perspectiva lexicométrica desarrollada por la escuela francesa de análisis de datos, con el fin de hacer una valoración acerca de los procedimientos incluidos bajo este enfoque.

1. ANÁLISIS CUANTITATIVO DE DATOS CUALITATIVOS

Con frecuencia, el investigador educativo se encuentra ante datos de naturaleza textual, es decir, datos expresados en forma de cadenas verbales de cierta extensión y que generalmente acaban siendo registrados como textos escritos. El análisis cualitativo de este tipo de datos se ha realizado a menudo a través de procesos de naturaleza intuitiva y creativa, que han dado pie a las críticas de quienes caracterizan el análisis de los datos textuales puesto en práctica por los investigadores cualitativos como oscuro, difuso, falto de rigor, poco definido.

A pesar de que son cada vez más habituales los esfuerzos por afrontar la indefinición de los procedimientos y pautas por los que se lleva a cabo el análisis, tratando de clarificar, guiar, dar rigor y exhaustividad a las técnicas analíticas cualitativas (Taylor y Bogdan, 1986; Lofland y Lofland, 1984; Miles y Huberman, 1984), son muchos los investigadores que asumen una perspectiva cuantitativa al analizar los datos cualitativos textuales (Berelson, 1952), recurriendo en algún momento de su proceso analítico a las técnicas estadísticas, con la intención de complementar o contrastar las conclusiones obtenidas por otras vías. Incluso este rasgo ha sido visto como característico de los enfoques de investigación interpretativos o cualitativos, los cuales recurren, sobre todo en el análisis de datos, a procedimientos cuantitativos característicos del enfoque positivista (De Miguel, 1988). Las ventajas del análisis de datos numéricos ha contribuido a un aumento en el uso del análisis estadístico en la investigación cualitativa, que ha quedado reflejado en la literatura a través de un incremento sostenido, durante las últimas décadas, de los trabajos que incluyen análisis estadísticos (Pelto y Pelto, 1991).

Parece evidente que, aunque las palabras aportan mayor significado, los números resultan menos ambiguos y son procesados con mayor economía. Por esta razón, buena parte de los investigadores prefiere trabajar con números y trasladan los textos a números lo más rápidamente posible. Generalmente, los valores numéricos surgen a partir del cómputo de los elementos diferenciados —frecuencias de códigos empleados en la categorización— al analizar el corpus de datos cualitativos textuales. Para

los investigadores que toman este camino, el terreno está bien marcado por las teorías de la medición, reglas de decisión, niveles de confianza, algoritmos para el cálculo, etc. (Miles y Huberman, 1984).

La cuantificación y el análisis estadístico son, por tanto, herramientas analíticas con las que cuenta el investigador en su trabajo con datos cualitativos, y pueden ser utilizadas conjuntamente con otras herramientas no cuantitativas. La formación del investigador, sus objetivos y sus concepciones epistemológicas le llevan a configurar un proceso de análisis orientado en mayor o menor medida hacia uno de estos tipos de técnicas. Dependiendo del estudio, la extensión de la cuantificación de datos varía ampliamente, desde ninguna hasta el uso de técnicas estadísticas de variado grado de complejidad (Wilcox, 1982). La cuantificación puede llegar a ser el aspecto central del análisis, e incluso a veces los datos cualitativos son, desde el primer momento, trasladados a índices numéricos y analizados cuantitativamente. En esta línea se inscriben los enfoques cuantitativos lexicométricos, apoyados en las técnicas estadísticas desarrolladas por la escuela francesa de análisis de datos.

2. EL ENFOQUE LEXICOMÉTRICO EN EL ANÁLISIS DE TEXTOS

Lexicometría es el término bajo el que se agrupan una serie de métodos que permiten reorganizar las unidades presentes en una secuencia textual y operar ciertos análisis estadísticos a partir de la cuantificación del vocabulario resultante de una segmentación operada sobre el texto (Lebart y Salem, 1988). Las posibilidades actuales en el campo de la lexicometría se han visto ampliamente incrementadas con el uso del ordenador; los primitivos análisis basados en la construcción de glosarios de términos y en el estudio de las concordancias (diferentes contextos verbales en los que ocurre una misma palabra) han pasado a constituir el material de partida para la aplicación de métodos estadísticos multidimensionales, basados en las técnicas de análisis de datos desarrolladas por Bénzecri (1973) —métodos factoriales y de clasificación, fundamentalmente—, que permiten extraer conclusiones acerca del sentido de un texto.

La aplicación de las técnicas del análisis de datos de la escuela francesa a datos de naturaleza textual (respuestas a cuestionarios de preguntas abiertas, entrevistas, textos literarios...etc.), apoyándose en las posibilidades ofrecidas por el ordenador, representa un intento de eludir la precodificación del texto y trabajar directamente con el corpus textual. De esta forma, resulta posible estudiar el sentido latente del texto a partir de los datos brutos, evitando el riesgo de la inconsistencia en la codificación o la pérdida de información que comporta la mediación de una categorización inicial.

Los estudios lexicométricos surgen, por tanto, de la necesidad sentida en el estudio de los textos de sobrepasar los enfoques tradicionales, marcados por la subjetividad, y abordar su análisis a partir del recuento y la localización de las unidades que contienen (Salem, 1982), considerándose como tales la palabra o una secuencia de palabras. La unidad considerada como base de los recuentos es generalmente la **palabra**, como ha ocurrido en otros enfoques del análisis de contenido de carácter cuantitativo realizado con ordenador (Mochmann, 1985; Mohler, 1985). La palabra cuenta con la ventaja de ser una unidad formal fácilmente identificable por el ordenador, al estar separada de otras unidades mediante un espacio o un signo de puntuación (Sánchez Carrión, 1985), y ello la hace especialmente adecuada para tratamientos automáticos. Lebart y Salem (1988), denominan *forma gráfica* a la unidad básica en que dividen a la cadena textual, y la definen del siguiente modo:

> «La unidad básica de los recuentos será la forma gráfica, definida como una sucesión de caracteres no delimitadores (en general, letras) rodeada por dos caracteres delimitadores (espacios, puntos, comas,...)» (p. 6).

Un riesgo implícito en el uso de formas gráficas como unidades radica en que diferentes palabras,

a veces con la misma raíz léxica, pueden expresar significados equivalentes, o podemos encontrarnos también ante palabras polisémicas, con más de un significado. Ambos aspectos deben ser tenidos en cuenta a efectos de enumeraciones posteriores. Por esta razón, Bolasco (1992:70) propone un tipo alternativo de unidad: la *forma textual*, identificada con una raíz léxica que corresponde a una sola forma gráfica o a sus diferentes inflexiones, siempre que sean portadoras de un significado indudablemente equivalente para las finalidades del estudio, o identificada con diferentes raíces léxicas que corresponden a diferentes formas gráficas con igual significado. Sin embargo, trabajar con formas textuales implica una intervención previa del analista, que debe definir las equivalencias entre formas gráficas o reducirlas a raíces léxicas comunes (lematización).

La unidad para el análisis puede ser un **grupo de palabras**. Lebart y Salem (1988), considerando que descender hasta el nivel de palabras puede conducir a un aislamiento que haga perder el sentido de lo expresado en el texto, utilizan también como unidad los *segmentos repetidos*, constituidos por dos o más palabras delimitadas por separadores tales como los signos de puntuación.

A partir de la segmentación, el corpus textual queda transformado en *tablas lexicales*, en las que se distribuyen las unidades para individuos o conjuntos de individuos con determinadas características. Estas tablas recogen ordenadamente la información sobre ocurrencia de las formas gráficas, situando en las columnas las partes consideradas en el texto (sobre la base de momentos diferentes, emisores distintos, etc.) y en las filas las formas del corpus textual o sólo aquéllas que rebasan una determinada frecuencia. Los términos son colocados en orden lexicométrico: frecuencias o longitudes decrecientes y/u orden alfabético. La aplicación de procedimientos estadísticos multivariables de carácter descriptivo, exploratorio, a este tipo de tablas arroja como resultado índices, agrupamientos, representaciones gráficas que permiten profundizar en la compresión de los datos y de la forma en que se estructuran. Del mismo modo, es posible realizar distintos análisis a partir de *tablas de segmentos repetidos*.

Las técnicas del análisis de datos de la escuela francesa son especialmente adecuadas para el tratamiento de valores numéricos procedentes de datos de naturaleza cualitativa. Precisamente, el tratamiento cuantitativo de lo cualitativo y el análisis simultáneo de un conjunto de variables definen las aportaciones fundamentales de tales métodos (Cornejo, 1988). Se trata de métodos, a la vez exploratorios y multidimensionales, que permiten descubrir la estructura de una gran tabla de números de varias dimensiones, traduciéndola a otra estructura más simple y que la resume mejor (Clapier, 1983).

3. EL PAQUETE ESTADÍSTICO SPAD.T

Dentro del enfoque lexicométrico desarrollado por la escuela francesa de análisis de datos (Lebart Salem, 1988), se ha diseñado un software específico capaz de trabajar con datos textuales procedentes de respuestas abiertas a cuestionarios, entrevistas, textos literarios, etc. El programa SPAD.T —Système Portable pour l'Analyse des Données Textuelles— (Lebart, Morineau y Bécue, 1989) permite extraer, de este tipo de datos, tablas de frecuencias construidas sobre las palabras o grupos de palabras presentes en una serie de textos para llevar a cabo diversos análisis lexicométricos, a los que añade métodos estadísticos como el análisis de correspondencias, la clasificación automática y el cáculo de las especificidades.

El **análisis de correspondencias**, en este caso, básicamente consiste en transformar la información de una tabla lexical en un conjunto de puntos, correspondientes a individuos o variables, posicionados en un espacio multidimensional, que pueden ser proyectados en los planos que mejor permiten visualizar la estructuración del conjunto. Las semejanzas y diferencias entre formas gráficas y/o puntos-columna quedan unívocamente reflejadas por las distancias entre los puntos que los representan en el espacio geométrico considerado. La interpretación de los gráficos factoriales resultantes consistirá en extraer conclusiones a partir del juego de proximidades y oposiciones puestas de relieve (Cibois, 1991).

Los **métodos de clasificación** permiten agrupar en clases, o jerarquías de clases, objetos o individuos sobre los cuales poseemos cierta información. Utilizan procedimientos basados en cálculos algorítmicos y producen representaciones gráficas generalmente en forma de árbol invertido.

El **cálculo de las especificidades** o formas características de los textos (Lafon, 1980) puede resultar de gran ayuda en la interpretación de las correspondencias reflejadas en los planos factoriales. Permite establecer qué formas se distribuyen uniformemente en el corpus de datos y cuáles no lo están y son características de una de las partes consideradas en éste. Se apoya en la obtención de un valor estadístico basado en la comparación de los porcentajes de frecuencia global e interno para cada forma gráfica. A cada forma se asocia un valor de contraste cuya magnitud aumenta a medida que disminuye la probabilidad de encontrar una determinada subfrecuencia interna. Este estadístico presenta una distribución normal. Por tanto, sus valores pueden ser trasladados a una curva normal para conocer la probabilidad asociada a la hipótesis nula que postula la repartición aleatoria de la forma entre los textos.

4. VALORACIÓN DEL ENFOQUE LEXICOMÉTRICO

El análisis lexicométrico permite enfrentarse a los datos textuales eludiendo la codificación de los mismos, y por tanto, el componente subjetivo, el tiempo y el esfuerzo asociados a esta operación. Se consigue así dar respuesta satisfactoria a la aspiración compartida por una buena parte de los investigadores de desarrollar el análisis de los datos de modo que cualquier analista siguiendo el mismo proceso llegue a los mismos resultados, objetivo que no resulta posible cuando se introducen en el proceso fases de codificación de los datos, incluso aunque se hagan esfuerzos por seguir un mismo patrón o esquema.

Considerado como lectura automática de textos, el análisis lexicométrico permite aproximarnos a la estructura de la información contenida en los textos, planteando interrogantes, abriendo vías de exploración y, en definitiva, contribuyendo eficazmente al planteamiento de hipótesis en las fases iniciales del análisis. Al apoyarse en técnicas descriptivas, resulta idóneo para ser aplicado cuando no existen hipótesis previas de trabajo y se requiere una exploración de la realidad estudiada con el fin de establecer los puntos de partida para posteriores análisis (García Santesmases, 1984).

Sin embargo, este tipo de procedimientos que recurren a la cuantificación es frecuentemente objeto de controversia, dado que no todos los investigadores están de acuerdo en considerar que los números puedan dar cuenta de las estructuras de significado contenidas en los datos textuales, sin que ello implique una pérdida excesiva de la riqueza informativa que les caracteriza. La segmentación del texto implica la pérdida de los significados contextuales de las palabras, limitando el estudio a la «superficie discursiva» del texto e ignorando el fondo significativo que subyace a ella. Implica un tratamiento indiferenciado de las palabras con una misma forma, que presupone una relación biunívoca entre el sentido con que son empleados los significantes objeto de recuento y sus significados, ignorando fenómenos como la polisemia y homonimia. Un problema inverso surge con la dispersión de formas que aluden a un mismo significado. Es el caso de los sinónimos y, con mayor amplitud, de las múltiples inflexiones de un verbo.

Se trata de un tipo de análisis subsidiario de las aplicaciones informáticas. Sin ellas, teniendo en cuenta que la unidad considerada en el análisis es la palabra, la tarea de extraer, ordenar y proceder al recuento de las apariciones de cada una de los vocablos presentes en un texto sería una tarea prácticamente inabordable cuando trabajamos con textos de cierta extensión. Del mismo modo, las técnicas estadísticas a que son sometidas las frecuencias precisan del soporte informático si queremos operar sobre grandes masas de datos. El programa SPAD.T reúne las características apropiadas para abordar este tipo de análisis.

A pesar del recelo que en algunos investigadores pudieran despertar el uso de la cuantificación o la rigidez del análisis automatizado, consideramos que la información textual puede ser analizada por

técnicas estadísticas como las representadas por el enfoque lexicométrico. La primera aproximación al significado de los datos que tales técnicas nos proporcionan, además de descubrirnos diferencias entre distintas partes del corpus textual, significados latentes, estructuras en los datos,..., resulta especialmente útil en la formulación de hipótesis o en el establecimiento de categorías para posteriores análisis cualitativos.

REFERENCIAS BIBLIOGRÁFICAS

BENZECRI, J. P. (1973): *L'Analyse des Données*. París: Dunod.

BERELSON, B. (1952): *Content Analysis in Communications Research*. Nueva york: Free Press.

BOLASCO, S. (1992): Sur différentes stratégies dans une analyse des formes textuelles: une expérimentation à partir de données d'enquête. En M. Becue, L. Lebart y N. Rajadell: *Jornades Internacionals d'Anàlisi de Dades Textuals*. Barcelona: Servicio de Publicaciones de la UPC. (pp. 69-88).

CIBOIS, PH. (1991): *L'analyse factorielle*. París: Presses Universitaires de France.

CORNEJO, J. M. (1988): *Técnicas de investigación social: El análisis de correspondencias*. Barcelona: PPU.

CLAPIER, P. (1983): *Análisis de Datos*. Vitoria: Publicaciones del Gobierno Vasco.

DE MIGUEL, M. (1988): Paradigmas de la Investigación Educativa Española. En I. Dendaluce (Coord.). *Aspectos Metodológicos de la Investigación Educativa. II Congreso Mundial Vasco*. Madrid: Narcea, (pp. 60-77).

GARCÍA SANTESMASES, J. M. (1984): Análisis factorial de correspondencias. En J. J. Sánchez Carrión, *Introducción a las técnicas de análisis multivariable aplicadas a las ciencias sociales*. Madrid: Centro de Investigaciones Sociológicas, (pp. 75-105).

LAFON, P. (1980): Sur la variabilité de la fréquence des formes dans un corpus. *Mots*, (1), 127-165.

LOFLAND, J. y LOFLAND, L. H. (1984): *Analyzing social settings. A guide to qualitative observation and analysis*. Belmont, CA: Wadsworth.

LEBART, L. y SALEM, A. (1988): *Analyse Statistique des Données Textuelles. Questions ouvertes et Lexicométrie*. París: Bordas.

LEBART, L., MORINEAU, A. y BECUE, M. (1989): *SPAD-T. Système portable pour l'analyse des données textuelles. Manuel de l'utilisateur*. París: CISIA.

MILES, M. B. y HUBERMAN, A. M. (1984): *Qualitative Data Analysis. A Sourcebook of New Methods*. Beverly Hills: Sage Publications.

MOCHMANN, E. (1985): Análisis de Contenido mediante Ordenador Aplicado a las Ciencias Sociales. *Revista Internacional de Sociología, 43* (1), 11-44.

MOHLER, P. Ph. (1985): Algunas Observaciones Prácticas sobre la Utilización del Análisis de Contenido en Ordenador. *Revista Internacional de Sociología, 43* (1), 45-57.

PELTO, P. J. y PELTO, G. H. (1991): *Anthropological research. The structure of inquiry*. Nueva York: Cambridge University Prees.

SALEM, A. (1982): Analyse factorielle et lexicométrie. Synthèse de quelques expériences. *Mots*, (4), 147-168.

SÁNCHEZ CARRIÓN, J. J. (1985): Técnicas de Análisis de los Textos mediante Codificación Manual. *Revista Internacional de Sociología, 43* (1), 89-118.

TAYLOR, S. J. y BOGDAN, R. (1988): *Introducción a los métodos cualitativos de investigación*. Buenos Aires: Paidós.

WILCOX, K. (1982): Ethnography as a methodology and its application to the study of schooling: a review. En G. Spindler (Ed.): *Doing the Ethnography of Schooling: Educational anthropology in action*. Nueva York: Holt, Rinehart and Winston, (pp. 456-488).