Revista Colombiana de Estadística
Nº 25 - 26, págs. 93 - 104

ANÁLISIS DE CORRESPONDENCIA BINARIA EN RESPUESTAS ABIERTAS

Leonardo Bautista S.[1] y Héctor O. Amaya T.[2]

Resumen

Este artículo hace una propuesta metodológica para el análisis de respuestas abiertas, consistente en un proceso conformado por un análisis sintáctico (automatizado por el SISCRE), un análisis semántico de carácter subjetivo y la aplicación de la correspondencia binaria a tablas de contingencia muy especialmente diseñadas para el estudio de contenidos linguísticos.

1. El Análisis de preguntas abiertas

Numerosos linguistas han mostrado que el lenguaje no es solamente un medio de comunicación entre los hombres, sino que dota de la posibilidad de organizar y de explicar su experiencia sensible. Las lenguas como conceptos de categoría varían según las culturas y los países. La forma de emplear la lengua del propio país varia fuertemente según los medios sociales, los niveles de educación, la edad, el sexo, la religión, la moda, la poesía, la época, etc. Estas diferencias no se reducen a estrictos aspectos formales sino que reflejan variaciones de percepción y de aprehensión de la realidad con correcciones de significación y con distinciones de contenidos.

Las preguntas abiertas constituyen elementos de información más específica que las preguntas cerradas. En éstas el respondiente tiene oportunidad

---

[1] Profesor Asociado del Departamento de Matemáticas y Estadística de la Universidad Nacional de Colombia

[2] M. S. en Estadística, Universidad Nacional de Colombia.

de contestar en sus propios términos y en su propio marco de referencia sin tener el problema de una lista de alternativas en las que las respuestas pueden forzarse en una categoría a la que no pertenecen propiamente. Rara vez son utilizadas en razón a que el aprovechamiento de las respuestas recogidas es a la vez difícil y costosa.

2. Propuesta metodológica

Si el lenguaje quiere ser sometido a un análisis formal, éste será necesariamente un análisis de contenidos, de pensamientos, de la forma de expresar la comprensión de la realidad. El considerar el lenguaje tan solo como un medio de comunicación ha conducido a que su análisis se haga en forma equivocada, por cuanto la información textual es mantenida en estado bruto, sin selección ni codificación a lo largo del análisis y cada forma léxical mantiene su sentido en un triple registro: aquel que lo pronuncia, aquel que le da el sentido y aquel que le confiere el lugar que ocupa en el espacio designado por todas las otras formas lexicales enunciadas por los otros locutores. Se impone entonces la necesidad de una metodología de estudio que permita, respetando en su totalidad el conjunto de propósitos, poner en evidencia aquellas relaciones entre las formas lexicales utilizadas y las determinantes socioeconómicas de quien responde.

El primer paso en la labor de analizar las respuestas abiertas es el proceso de organización en forma tal que se facilite su lectura e interpretación. Sin embargo esta reorganización se puede efectuar de numerosas formas.

Si por medio del lenguaje se puede explicar la experiencia sensible de los individuos, es de esperar que quien escucha o lee puede estar dándole una significación diferente a la que quiere dar quien emite la idea. Por esto el hecho de agrupar respuestas abiertas no puede ser simplemente un trabajo realizado por un grupo de personas, las cuales entre otras características presentan diferentes puntos de vista para comprender una determinada respuesta. La categorización de respuestas abiertas, realizada por diferentes individuos conduce a grupos temáticos de los que finalmente no se sabe si determinan la expresión de quienes respondieron, la comprensión de quienes clasificaron o una unión de estos dos efectos (más interacciones).

La metodología que aquí se propone para el análisis de respuestas a preguntas abiertas consta de cuatro puntos. Un análisis sintáctico, un análisis semántico, un análisis de Correspondencia Binaria y un análisis e interpretación de los mapas obtenidos en el análisis de correspondencia binaria.

2.1. Análisis sintáctico

Para someter una serie de objetos a comparaciones estadísticas es necesario definir unidades que puedan ser claramente reconocidas y uniones sistemáticas que conduzcan a categorías más generales. Las unidades gramaticales de primer orden serán entonces oraciones, en donde cada una de las palabras que la componen es calificada según su papel sintáctico, es decir si es un sustantivo, verbo, adjetivo, pronombre, etc. o simplemente si se trata de una palabra sin interés de contenido y se llamará entonces PALABRA IGNORADA. La tarea consiste en determinar la combinación gramatical a la que pertenece la oración es decir si es de la forma [verbo-sustantivo],[verbo-verbo-sustantivo], [sustantivo-verbo-sustantivo-adjetivo], [verbo-sustantivo-adjetivo], [verbo adjetivo-sustantivo], etc.

Estas combinaciones gramaticales son entonces las categorías dentro de las cuales vale la pena hacer comparaciones. Dentro de una misma combinación gramatical se organizan las respuestas por orden de similaridad de palabras. En este análisis sintáctico se logra mantener la información en su estado natural de significación, puesto que el investigador solo ha determinado palabras que pueden ser ignoradas.

El resultado de éste proceso es un orden textual que no ha variado los contenidos, pero que por su estructura gramatical y su similitud de palabras, se acerca a un ordenamiento semántico que facilita la siguiente fase.

2.2. Categorización semántica

El anterior análisis sintáctico genera una organización del material textual que permite la realización por parte del investigador de la categorización semántica o de contenidos. Esta categorización es naturalmente subjetiva,

está influenciada por la percepción y opinión que tiene el investigador sobre el tema. Pero es inevitable, puesto que en un momento dado tiene que intervenir el investigador a través de la lectura de las respuestas abiertas con el fin de organizar ideas o conceptos que sobre la pregunta emiten las personas. Este proceso de "convertir" palabras u oraciones para interpretarlas como ideas o conceptos puede ser realizado únicamente por el cerebro humano. De otro lado el hecho de contar con un ordenamiento sintáctico significa que se ha logrado una ganancia sustancial; las respuestas se tienen organizadas en forma tal que la categorización semántica puede ser realizada por un único individuo en un tiempo relativamente corto, es decir, se convierte en un método factible.

2.3. Análisis de correspondencia binaria

En particular se trata de establecer las relaciones que se presentan entre las variables socio demográficas, socioeconómicas, u otras, las cuales generalmente se miden en forma categórica o cualitativa, y palabras o categorías semánticas. Se establezca para esto, tablas de contingencia que informan sobre la frecuencia de uso de categorías semánticas (o palabras) según las variables socioeconómicas. Estas tablas son entonces sometidas al análisis de correspondencia binaria.

2.4. Análisis e interpretación de mapas

Se pretende finalmente establecer las hipótesis que relacionen la forma de comprender el mundo expresado por la respuesta abierta y las características que determinan su proceso de aprehensión de la realidad por parte de los individuos, es decir, si las palabras o las ideas expresadas a través de las respuestas a preguntas abiertas presentan patrones de comportamiento de tipo social, de género, aspectos demográficos, culturales, etc., e incluso detectar cambios sociales extralinguísticos, todo esto en términos del grado de correspondencia entre categorías semánticas utilizadas y modalidades socioeconómicas.

3. El tratamiento con el computador

Para hacer viable este procedimiento. Se requiere de un software capaz de organizar la información de la manera prevista, mediante el análisis sintáctico. El desarrollo de dicho software se hizo en forma separada y consiste en crear una base de conocimientos, un vocabulario, que crece automáticamente a medida que el sistema se alimenta. El sistema aproxima a la creación de categorías conceptuales de acuerdo a la similitud semántica de las oraciones mismas, permitiendo identificar significados literales y la manera como se representan en una base de conocimiento. Para la identificación del significado de las oraciones, se analizan los métodos de conformación y su representación en una base de conocimiento, mediante los conceptos fundamentales del cálculo de predicados.

4. Ejercicio de aplicación de la propuesta metodológica

4.1. Recolección de información

Con el objetivo de realizar un análisis estadístico a respuestas obtenidas de preguntas abiertas, se diseñó un formulario sobre la conceptualización, que poseen las personas, de Santafé de Bogotá acerca del trabajo. Se pretendía conocer, con la primera pregunta abierta, qué opinaban las personas que tenían un empleo, con respecto a si el contar de repente con mucho dinero los motivaría a seguir trabajando ó no. Obteniendo que la gran mayoría está de acuerdo en seguir laborando, se preguntó entonces el por qué. Como una segunda pregunta, se les propuso escoger entre contar con suficiente dinero y no tener poder de decisión, ó contar con poco dinero pero mucho poder de decisión. En esta pregunta se encontró que gran parte de los entrevistados prefieren el poder de decisión, entonces se les indagó la razón de su preferencia.

Para poder relacionar el lenguaje utilizado por los entrevistados, desde el punto de vista semántico, se incluyeron en el formulario algunas preguntas cerradas relacionadas con su status sociodemográfico. Se registraron las variables: sexo, estado civil, grado de escolaridad y estrato social.

El formulario se aplicó a trescientas personas, en diferentes horas y sitios de la ciudad de Santafé de Bogotá. Los sitios seleccionados trataron de obtener entrevistas con personas de las diferentes clases sociales. La entrevista se realizó en forma directa y grabada. La grabación de las respuestas obedeció a recomendaciones que hacen los sociolingüistas al respecto. Ellos aseguran que el método primordial para recopilar un conjunto de datos relevantes acerca del habla de una persona es la entrevista individual grabada. Esto con el fin de evitar que los métodos empleados para la recolección de otros tipos de datos interfieran en éstos mismos. Cuando la entrevista no se graba, sino que se registra en un formulario, se puede observar el triple registro; la respuesta emitida por el entrevistado, lo que el entrevistador registra, es decir, lo que cree haber entendido y el análisis que le confiere el investigador. Esta interferencia es la que en el estudio del lenguaje se presenta como el problema metodológico clásico.

La información recopilada se llevó a una base de datos, y se procedió de acuerdo a los objetivos del estudio a no considerar las siguientes entrevistas:

1. Las personas que manifestaron no tener un empleo.
2. Para los pocos que er la pregunta dos preferían dinero y a la vez poder de decisión.
3. Las que se manifestaron indecisas.
4. Quienes respondieron lo mismo en las dos preguntas abiertas.

4.2. Análisis sistemático

En esta etapa se toma, con el programa SISCRE, cada una de las respuestas a la pregunta abierta que se desea analizar, se examina sintácticamente para: a) eliminar las palabras ignoradas, b) generar combinaciones gramaticales y asignar el código respectivo a la respuesta. La mayor dificultad de esta etapa radica en la necesidad de determinar si las palabras tienen la función de sustantivo, adjetivo, verbo, preposición, conjunción, adverbio o pronombre indefinido. Este vocabulario es el que conforma la "Base de Conocimientos" del SISCRE.

4.3. Análisis semántico

Esta etapa es realizada por el investigador. El SISCRE produce dos archivos, en el primero el orden es el siguiente: Número de la respuesta, el nombre asociado al grupo al que pertenece, las respuestas cerradas y las respuestas en forma reducida. El segundo denominado palabras, que contiene cada una de las palabras de las respuestas reducidas, indicando el número de la respuesta, las respuestas cerradas y el grupo al que pertenecen.

Para la respuesta a la primera pregunta, al consultarse el archivo REDUCIDO se observa por ejemplo:

```
11   BA29   F C B B    trabajo es salud
127  BA29   M S U M    trabajo es terapia
-------------------------------------------------------
14 BA39 F C U M trabajo es vital.
```

En la primera columna se presenta el número de la entrevista (11), en la segunda el grupo asociado según la combinación gramatical de la primera pregunta abierta que es la que se está categorizando (BA29), en la tercera las respuestas obtenidas de las cuatro preguntas cerradas realizadas (F = femenino, C = casado, B = bachillerato, B = estrato bajo), y en la cuarta la respuesta reducida, es decir, las palabras que están incluidas en la combinación gramatical asociada a dicha respuesta. En las dos primeras líneas se transcribe lo referente a las entrevistas 11 y 127. Como se puede apreciar, el grupo asociado a estas respuestas es también el BA29, por presentar una combinación gramatical semejante, sustantivo-verbo-sustantivo.

Al consultarse el archivo PALABRAS, con respecto a la entrevista en mención, aparece lo siguiente:

```
trabajo   11 F C  B B BA29
es        11 F C  B B BA29
salud     11 F C  B B BA29
```

Este archivo muestra en la primera columna todas las palabras que contiene el archivo REDUCIDO, en la segunda el número de la entrevista que contiene cada palabra, en la tercera las respuestas a las preguntas cerradas, y en la cuarta el grupo al que pertenece cada una de las palabras.

El proceso de análisis semántico consiste en analizar las respuestas reducidas, con el fin de realizar la categorización de contenidos. Este se realiza subjetivamente y consiste en clasificar las respuestas de contenido similar en grupos. Adicionalmente a este análisis es posible obtener los sinónimos o formas equivalentes.

Se generan finalmente dos tablas de contingencia: en una la primera variable de clasificación corresponde a las preguntas cerradas y la segunda, a las formas o palabras obtenidas en las respuestas. En la otra tabla de contingencia, la primera variable de clasificación corresponde también a las preguntas cerradas y la segunda a los grupos generados en el proceso de análisis semántico

5. Análisis estadístico

Comprende el análisis de la pregunta "Si contara con suficiente dinero, ¿seguiría trabajando? ¿Por qué?", para las personas que contestaron afirmativamente (n=229).

5.1. Análisis de las palabras

Del análisis semántico realizado a las respuestas en esta pregunta, se obtuvo una tabla de contingencia con doscientas ochenta y una filas, es decir, palabras y once columnas, que corresponden a las categorías de las variables sociodemográficas. Las palabras con una frecuencia inferior a dos no fueron tenidas en cuenta para el Análisis de Correspondencia Binaria, el que finalmente fue realizado a una tabla de contingencia con ciento diez filas o palabras y once columnas o categorías de las variables sociodemográficas.

5.2. Análisis estadístico de las categorías temáticas (frases)

De otra parte, el Análisis de Correspondencia Binaria se aplicó a la tabla de contingencia de los GRUPOS TEMÁTICOS como primera variable de clasificación y las variables sociodemográficas como la segunda. La tabla obtenida incluía sesenta y tres filas, es decir, categorías semánticas y once co-

lumnas o categorías de las variables demográficas. Debido a que la tabla presentó frecuencias marginales inferiores a dos, éstas no fueron tenidas en cuenta para el análisis de correspondencia binaria. De los resultados obtenidos de éste análisis, se conforman tres grupos relacionados con el estrato social de las personas entrevistadas. Las CATEGORÍAS TEMÁTICAS que identifican a cada uno de estos grupos son:

CATEGORÍA TEMÁTICA UNO. En este grupo, relacionado con el estrato bajo, de las personas con un grado de escolaridad de primaria, algunas están de acuerdo en trabajar su dinero o en administrar su propio negocio y el resto, creen que el dinero no lo es todo. Quienes son bachilleres, piensan en incrementar el dinero obtenido. Se podría decir que el trabajo es considerado por este grupo, como un objeto. El trabajo es útil en la medida que les permite HACER o emprender negocios que tal vez ahora no pueden desarrollar.

CATEGORÍA TEMÁTICA DOS. Caracterizado por la clase media, piensan que no pueden estar desocupados, que están acostumbrados a trabajar o que deben trabajar, otros hablan de invertir el dinero ya no solamente en sus propios negocios sino creando empresas, y piensan en alcanzar lo que han deseado.

GRUPO NUMERO TRES. Este grupo está relacionado con la clase alta y un nivel de escolaridad universitaria, opinan que el trabajo les agrada, que es algo vital, saludable física y mentalmente, y por el hecho de ser una actividad continuarían laborando. El trabajo es entonces una actividad necesaria.

Clasificación de individuos según grupo temático y estado social.

Estrat. / G. Temát.	Bajo	Medio	Alto	Total
I	9	9	1	19
II	26	38	6	70
III	16	40	27	83
Total	51	87	34	172

La tabla anterior, corresponde a una tabla de contingencia, obtenida al clasificarse por estratos las respuestas contenidas en cada uno de los grupos

temáticos. Al realizar la prueba de independencia de la JI-CUADRADO, se encontró un estadístico de prueba de T= 19.4, con el cual se podría afirmar que existe suficiente evidencia estadística para concluir, con un nivel de significancia de 0.001, que los grupos temáticos y el estrato social, no son independientes.

6. Discusión de resultados

1. La conformación de las categorías temáticas con un contenido definido es factible a partir de la unión sistemática de los grupos generados en un análisis de la estructura semántica.

   Aunque existen varias formas de categorización semántica, la propuesta metodológica que se presenta puede tener algunos inconvenientes. El lenguaje español presenta como característica que una palabra puede pertenecer a más de un atributo gramatical. Por ejemplo, cuando se dice, "el trabajo" y "yo trabajo", en la primera expresión [trabajo] es un sustantivo, mientras que en la segunda corresponde a una forma verbal.

   Este inconveniente implica que se requiere por momentos ser un especialista del lenguaje para poder definir el atributo gramatical de cada palabra no contenida en la base de conocimientos.

2. La opinión de los individuos se "obtiene" cuando se analizan contenidos semánticos.

   Los franceses proponen un análisis de palabras, pero al poderse enmarcar una misma palabra en más de dos atributos gramaticales hace que su interpretación sea muy difícil, al igual que una misma palabra en dos contenidos diferentes. Este análisis podría ser útil en el estudio de respuestas semiabiertas, en las que el entrevistado se limita a completar frases.

3. El proceso descrito como análisis semántico debe tenerse en cuenta en la interpretación de resultados, debido a que al ser realizada por el investigador y de acuerdo a la concepción que él tenga del tema, puede variar de persona a persona.

Esa diferencia de opiniones viene a corroborar las diferencias que existen en la lectura de las respuestas abiertas, pero son más de matiz, por cuanto las respuestas han sido mantenidas hasta el final en su estado natural.

4. El Análisis de Correspondencia Binaria utilizado en el ejemplo permitió formular hipótesis que pueden considerarse consistentes con los planteamientos de los lingüistas acerca del carácter social del lenguaje.

   En los mapas resultantes se observó que existen diferencias por genero, estado civil, grado de escolaridad, y estrato social, en cuanto a la concepción que tienen las personas acerca del trabajo y de la preferencia del poder de decisión. Es así como el análisis realizado a las tablas de contingencia de las categorías semánticas y las variables sociodemográficas, evidenció una relación entre estas dos variables de clasificación.

Bibliografía

Bautista, L. Ramos, J., 1988,*Análisis de Datos·de Encuestas y tabulados.* Universidad Nacional de Colombia, Facultad de Ciencias Departamento de Matemáticas y Estadística, Bogotá.

Bautista, L., 1991, UN/DTCD-NIC/85/018 - *Programas para el desarrollo de las Naciones Unidas.* Bogotá.

Bautista, A., 1991, *Desarrollo de una Interfaz entre una base de datos de conocimiento y una base de datos alfanumérica,* Universidad Nacional de Colombia Facultad de Ingeniería, Bogotá.

Diday, E. Jambu, M. Lebart, L. Pages, J. Tomassone, R., 1983, *Data Analysis and Informatics,*III. Elsevier Science Publishers,Versailles Francia.

Labov, W., 1983, *Modelos Sociolingüísticos,* Ediciones Catedra, S.A.Madrid.

Lebart, L. Morineau, A. Kenneth, M. Warwick., 1984, *Multivariate Descriptive Statistical Analysis,* John Wiley & sons. Inc. New York.

Lions, J., 1985, *Introducción en la Lingüística Teórica,* Edit. Tiede, Barcelona,.

Lebart, L. Salem. A., 1988, *Analyse Statistique des donnes Textualles*, París, Bordas.