UNIVERSIDAD SANTO TOMAS
PRIMER CLAUSTRO UNIVERSITARIO DE COLOMBIA

Comunicaciones en Estadística
Diciembre 2014, Vol. 7, No. 2, pp. 139–156

Preguntas abiertas en encuestas ¿cómo realizar su análisis?

Open questions in surveys. How to perform the analysis?

William Arley Rincón Gómez[a]
williamrincon@usantotomas.edu.co

Resumen

El presente artículo, muestra algunas de las ventajas que tiene en una encuesta la utilización de preguntas abiertas, preguntas que debido a la complejidad de su análisis, son utilizadas con poca frecuencia, o cuando se utilizan, simplemente se dejan como parte en el cuestionario pero no se analizan las respuestas a este tipo de preguntas. Técnicas estadísticas como el análisis de datos textuales, desarrolladas inicialmente por Ludovic Lebart, son presentadas en este artículo como una alternativa a los métodos tradicionales de análisis de grandes masas de datos de tipo textual como la poscodificación y como una motivación al estudio de técnicas y software que permitan analizar datos de tipo textual.

***Palabras clave*:** datos, estadística, preguntas abiertas, textuales.

Abstract

This paper shows the advantages obtained by using open questions in a statistical survey, considering that they are rarely used today because of the complexity presented in their analysis. Therefore, these questions are often left as part of the questionnaire without any reasoning in their responses. Statistical techniques, such as the analysis of textual data, which was developed by Ludovic Lebart, are presented in this paper as an alternative for the traditional methods in analysis for large amounts of textual data as in poscodificacion and simultaneously serve as a motivation to study more techniques with the use of software, and thus improve the analysis of textual data type.

***Keywords*:** statistics, textual data, open questions..

[a]Docente. Facultad de Estadística. Universidad Santo Tomás. Colombia.

1. Introducción

Las respuestas a preguntas abiertas se utilizan en encuestas realizadas en economía, sociología, educación, politología, epidemiología, mercadotecnia, medicina, etc. Dichas respuestas constituyen una prolongación indispensable de los cuestionarios, cuando las encuestas van más allá de una simple búsqueda de sufragio, cuando se trata de explorar y profundizar sobre una tema complejo o poco conocido (Lebart et al. 2000).

Una de las disciplinas que tiene que ver con el estudio de la información textual es el análisis de contenidos, según Berelson & Lazarsfeld (1948)

> El análisis de Contenidos es una técnica que describe objetiva, sistemática y cualitativamente el contenido manifiesto en la comunicación.

El análisis de contenido se propone acceder directamente a la significación de diferentes segmentos que componen el texto. Es una técnica de investigación para la descripción objetiva, sistemática y cuantitativa del contenido manifiesto en la comunicación. Opera en dos fases: se empieza construyendo un conjunto de textos que serán sucesivamente analizados,y como segunda fase se hacen los conteos para cada uno de los temas previstos. Las unidades en un análisis de contenido pueden ser los temas, las palabras o los elementos de sintaxis o semántica. El análisis de contenido así definido tiene una dimensión estadística.

El análisis de las respuestas a esta clase de preguntas se ha venido realizando con métodos como: poscodificación, edición de índices y glosarios, ediciones de concordancias, entre otras.

En la actualidad se han desarrollado técnicas estadísticas como el análisis de datos textuales, que con ayuda de software estadístico se encargan de procesar y analizar esta clase de información. El análisis de datos textuales es una técnica que describe, sintetiza y analiza información contenida en las respuestas a preguntas abiertas. Si se utilizan de manera simultánea información de carácter textual y no textual (preguntas abiertas y cerradas), se pueden obtener las respuestas de los individuos por categorías (edad, sexo, nacionalidad, partido político, religión, etc.) y contrastar sus perfiles léxicos por categoría.

Este artículo presenta algunas de las ventajas que tiene el incluir preguntas de respuesta libre dentro de los cuestionarios, junto con una síntesis de algunos métodos de análisis de información textual; el método codificación y una primera aproximación a los métodos utilizados en el análisis estadístico de textos presentando con un ejemplo las primeras fases de este análisis, tales como el tratamiento inicial al corpus de datos, las unidades de la estadística textual y documentos lexicográficos además se realiza el planteamiento mas no la ejemplificación del análisis de correspondencias y clasificación de tablas léxicas. Este artículo pretende despertar en el lector el interés por el análisis de datos textuales lo cual permite, dado su amplio desarrollo, aprovechar de manera eficiente la información de tipo textual.

2. Preguntas abiertas en encuestas

Las preguntas abiertas utilizadas en encuestas proporcionan información de carácter textual; opiniones, explicaciones, justificaciones. La pregunta abierta no obliga a escoger entre un conjunto fijo de alternativas, es de respuesta libre, por eso, según la naturaleza de la pregunta y el interés de la persona, las repuestas varían mucho en cuanto a su extensión y profundidad; la utilización de este tipo de preguntas, aún a pesar de la dificultad en la codificación y en el análisis de sus respuestas, está justificada en muchas ocasiones gracias a las ventajas que ofrecen, algunas de ellas, según Pope (2012), son:

> recolección de información espontánea, enriquecimiento del informe definitivo (mediante la inclusión de cuotas reales de las respuestas que se consideren significativas), utilidad para explicar y comprender la respuesta a una pregunta cerrada; además proporciona información acerca de la opinión de un grupo de personas.

Las preguntas abiertas(de respuesta libre) utilizadas en encuestas sirven en la fase preparatoria de un estudio, como un elemento fundamental en la preparación y puesta a punto de una batería de ítems de respuestas para una pregunta cerrada; además existen situaciones en las cuales el interés se centra en reducir el tiempo de una entrevista, en recoger una información espontánea o en conocer y entender las explicaciones a una pregunta cerrada en las cuales la utilización de preguntas abiertas se impone.

El análisis de respuestas a preguntas abiertas es un problema de análisis de tipo textual. Distintos métodos han surgido con el objetivo de describir, sintetizar, clasificar y analizar la información textual; a esto se suma que el análisis de esta clase de información se ha venido realizando con técnicas de poscodificación, ordenación alfabética, ediciones de concordancias, entre otras. En la actualidad con la llegada del computador y con el desarrollo de técnicas estadísticas, dicha labor se ha facilitado y se han superado inconvenientes como: mediación del codificador, empobrecimiento del contenido, y eliminación *a priori* de respuestas confusas (Montenegro & Pardo 1998).

3. Codificación de respuestas a preguntas abiertas

A continuación se presenta una breve descripción de la codificación tradicional haciendo referencia a su definición uso, ventajas y desventajas. Las preguntas abiertas son preguntas de discusión, las cuales generan una gama tan amplia de respuestas que las posibles contestaciones son demasiado variadas y numerosas como para incluirlas en una lista en el cuestionario. Para estas preguntas se deja un espacio en el cuestionario a fin de que el entrevistador o el encuestado apunte las respuestas textualmente; posteriormente las respuestas son categorizadas o codificadas.

3.1. Codificación

El proceso de convertir las respuestas individuales en categorías se llama *codificación*. La codificación determina si los resultados constituyen información útil, se trata, en efecto de un problema de análisis de contenido que pretende presentar los resultados en forma simple, el propósito de la codificación es reducir toda la variedad de respuestas dadas para una pregunta, a pocos tipos de contestaciones que pueden ser tabuladas y luego analizadas. El primer paso en la codificación es determinar las clases de respuestas que se han dado a una pregunta, esto se hace, normalmente, tomando una muestra de los cuestionarios terminados, 25 % es típico, haciendo una lista de las respuestas y su frecuencia; luego, los comentarios en lista se organizan en agrupaciones lógicas a las cuales se les asigna un código; dichas agrupaciones están determinadas tanto por la frecuencia de las respuestas como por los objetivos de la prueba.

Por último, a estas categorías o códigos se les asignan números, de tal manera que los cuestionarios puedan ser tabulados después de haber sido codificados (Pope 2012). En resumen, el trabajo de Codificación clásico se realiza en dos etapas: La primera a través del análisis de una muestra del corpus que conduce a la elaboración del código mismo; la segunda consiste en hacer corresponder cada respuesta a uno o varios códigos. Un código, según Ghiglione (1989), debe reunir varias características para ser utilizado de manera satisfactoria:

- Primera, el número de categorías que lo constituyen no debe ser muy elevado, por razones de comodidad de manejo y, sobre todo por razones de estadísticas, lo que es válido, por otra parte para todas las preguntas; cuando se distinguen demasiadas respuestas, quedan pocos sujetos en cadacategoría y ya no es posible extraer conclusiones significativas. Cada vez que se trate de considerar una diferencia, no se debe preguntar si esta es interesante en sí misma (la respuesta tiene todas las posibilidades de ser afirmativa), sino si es verdaderamente necesaria para las metas perseguidas. De no serlo, el introducir categorías complementarias por tomar en cuenta complica el trabajo de los codificadores y esconde la información importante tras matices inútiles.
- En segundo lugar, es necesario asegurarse de que todas las categorías previstas se utilicen efectivamente y de que el número inevitable de respuestas inclasificables sea poco significativo.
- En tercer lugar, las reglas que definen la atribución de una respuesta a una categoría deben ser tan explícitas como sea posible y dejar el mínimo de campo al juicio del codificador. En la sucesión del trabajo, cuando las preguntas así codificadas se inserten en análisis más complejos, se deberá tener en mente estas reglas, que constituyen la verdadera definición de las categorías.
- En cuarto lugar, se debe precisar si cada respuesta debe atribuirse a una sola categoría o si se admite la posibilidad de codificar una respuesta en varias categorías: pero, esta posibilidad solo es útil si se emplea con frecuencia; si

las respuestas múltiples son la excepción, es preferible aceptar la pérdida de cierta información y realizar una elección entre las diversas categorías consideradas.

3.2. Ventajas y desventajas de la codificación

Los principales defectos de la poscodificación son: mediación del codificador, empobrecimiento del contenido y eliminación de las respuestas raras (Montenegro & Pardo 1998). Cuando la pregunta es abierta, resulta posible que se empobrezcan demasiado las respuestas al codificarlas una sola vez, pues la persona que realiza el proceso, debe interpretar y decidir, lo cual introduce un sesgo personal, ya que el codificador debe tomar decisiones difíciles y a veces discutibles para el profesional experto en el tema; o bien que se construya un código que tenga un gran número de ocurrencias que rápidamente se convierte en inutilizable, con la excepción evidente del caso en que se disponga de muchas respuestas (individuos).

Otra desventaja que se puede mencionar es que se realiza en una etapa preliminar en la cual no se ha analizado la base de datos, lo cual hace que muchas decisiones de asignación y de reagrupamiento se tomen sin un análisis global del corpus que tenga en cuenta toda su diversidad, complejidad y riqueza.

El método tradicional de la poscodificación de respuestas a preguntas abiertas en encuestas tiene únicamente la ventaja de que los resultados son fácilmente explotables, pues la técnica más habitual para el tratamiento de las respuestas abiertas consiste en construir una batería de ítems a partir de una respuesta (en general entre 100 y 200) a fin de codificar después el conjunto de las respuestas de manera que se sustituya la pregunta abierta por una o varias preguntas cerradas (Lebart et al. 2000).

4. Análisis de datos textuales

Se denomina estadística textual al estudio de textos mediante la aplicación de métodos estadísticos, las técnicas estadísticas que se utilizan corresponden a los métodos de análisis estadístico multivariante (análisis de correspondencias y análisis de conglomerados). El análisis factorial de correspondencias busca proyectar los datos sobre un espacio de dimensión reducida que guarden la mayor parte de la información original. Los métodos de clasificación agrupan a los individuos en clases homogéneos con respecto a las variables observadas, estos métodos son complementarios y se utilizan de manera simultánea; el análisis de datos textuales se realiza considerando una variable denominada variable léxica, cuyas modalidades son las formas gráficas del corpus tratado.

El análisis de datos textuales, a diferencia de los procesos de poscodificación, no busca una reducción *a priori* de la información bruta sino al contrario, busca una valoración de dicha información por medio de la utilización de los datos disponi-

bles sobre los encuestados o entrevistados, por ejemplo, en el cuestionario podemos indagar sobre: estrato socioeconómico, estado civil, profesión, nivel de escolaridad, género, entre otros datos que pueden ser empleados junto con la información textual, con el propósito de caracterizar tales segmentos de la población. Se puede por ejemplo obtener la opinión de los profesionales en relación con un determinado tema; en su fase inicial el análisis de datos textuales permite realizar una lectura global del vocabulario lo cual orienta los procesos posteriores.

El corpus textual "Demora en cumplir los requisitos para grado"servirá como ejemplo para ilustrar y realizar algunos procedimientos básicos de la estadística textual. Este corpus esta constituido por 971 respuestas dadas a pregunta abierta 32 (Tabla 1) dentro del marco de la encuesta realizada a los egresados de la Universidad Nacional De Colombia *Formulario de opinión de los graduandos sobre la carrera*[1], allí se pedía dar respuestas libres en las cuales debían expresar las principales causas asociadas con el trabajo o proyectos de grado, lo cual retrasó su graduación, teniendo en cuenta razones personales, sociales y universitarias.

El análisis se centra en 835 respuestas, debido a que algunos encuestados técnicamente no respondieron de forma acertada algunas preguntas, ya que las respuestas no aportaban nada nuevo al estudio pues el argumento dado al contestar la pregunta es similar a la pregunta dada, es decir son respuestas como "por el trabajo de grado", que no dan explicación de las causas o circunstancias que originaron retraso en el trabajo de grado (Rincón 2002).

Las primeras primeras respuestas del corpus "Demora en cumplir los requisitos para grado"son representadas a continuación en la forma como las procesa el programa SPAD T.[2]

- - -1

La monografía se convirtió en un trabajo que conlleva mucho tiempo.

- - -2

Demora en las correcciones por parte del director de la tesis.

- - -3

Muy mal enfoque en los exámenes preparatorios y con directores que tienen poco interés en ellos y desactivan muy fácilmente a los estudiantes.

- - -4

Extenuante trabajo de campo en empresas y con asesores.

- - -5

Deficiente asesoría por parte del director de la monografía, falta de motivación.

4.1. Unidades de la estadística textual

Para poder utilizar los métodos de análisis multidimensionales a datos textuales se construyen dos tablas de contingencia particulares, la tabla léxica (que contiene

[1]Esta encuesta tenía un total de 24 preguntas y fue respondida por 2303 egresados de las promociones de diciembre de 2000, abril y junio de 2001, pertenecientes a diferentes carreras en las sedes de Bogotá, Medellín, Palmira y Manizales.

[2]Programa desarrollado por Mónica Bécue, con las mismas técnica y filosofía del SPAD N.

la frecuencia relativa con la que cada forma léxica o palabra ha sido empleada por cada individuo) y la tabla léxica agregada (que contiene la frecuencia con la que se encuentra una forma en una parte del corpus cuando existe una o varias particiones del corpus). Los métodos de *análisis de datos textuales* se basan en mediciones y conteos realizados a partir de los objetos que se desean comparar (palabras, segmentos repetidos, unidades semánticas, etc.) para ello es necesario realizar un tratamiento preliminar al corpus de datos (Lebart et al. 2000), el proceso tiene las siguientes unidades estadísticas:

Tabla 1: *pregunta 13 formulario de opinión de los graduandos sobre la carrera. Fuente: Rincón 2002.*

13. Si usted se está graduando exactamente en el tiempo o en menor tiempo del establecido en el plan de estudios, por favor pase a la pregunta numero 14. En caso contrario señale todas las circunstancias que sean su caso personal. 1. Se retrasó por no haber inscrito la totalidad de las asignaturas del semestre correspondiente. 2. Se retrasó por cancelación de asignaturas 4. Se retrasó por haber perdido al menos una materia. 8. Se retrasó por haber tenido empleos simultáneos con sus estudios. 16.Se retrasó por circunstancias o problemas personales 32. Se retrasó por circunstancias ligadas al desarrollo del trabajo de grado (monografía, proyecto, pasantía, taller final, practica final, preparatorios,etc.). Estas circunstancias en síntesis fueron: _____ _____ _____ Se retrasó por otra razón que en síntesis fue: _____ _____ _____ _____

- Formas gráficas: unidad para la descomposición del texto y unidad mínima para los cálculos estadísticos; está definida como una sucesión de caracteres definidos entre dos delimitadores.
- Alfabeto: es el conjunto de caracteres del teclado del computador en el cual es registrado el Corpus.
- La forma gráfica y el vocabulario. El vocabulario de texto es el conjunto de formas distintas de un corpus (Tabla 1). Una ocurrencia es una cadena de caracteres acotada por dos delimitadores, dos cadenas idénticas, son por lo tanto, dos ocurrencias de una misma forma gráficas. La segmentación definida de esta forma permite considerar el texto como una sucesión de

ocurrencias separadas entre ellas por uno o más delimitadores. La longitud del texto es el número total de ocurrencias de formas gráficas.

- Segmentación del corpus: la segmentación del corpus es la operación en la cual se descompone el texto en unidades mínimas o formas gráficas. Lo cual permite realizar mediciones y conteos útiles para comparar estas unidades. Para realizar una segmentación automática de un texto es suficiente seleccionar del conjunto de caracteres un subconjunto que se denomina caracteres delimitadores; los demás se consideran caracteres no delimitadores. Se consideran dos tipos de delimitadores, fuertes y débiles; con frecuencia se define la coma (,) como delimitador débil y los dos puntos(:) y el punto(.), como delimitadores fuertes.

- Los segmentos repetidos: un segmento se define como una sucesión de formas gráficas que se encuentra entre dos delimitadores fuertes. Un segmento se denomina segmento repetido cuando se presenta al menos dos veces en el corpus. Los segmentos repetidos permiten indagar como se encuentran combinadas las palabras del vocabulario, se puede realizar sobre todo el vocabulario, sobre el vocabulario corregido o sobre el vocabulario recortado (cuasi segmentos).

- Numeración del texto: es el proceso de cambiar el texto original por números, lo cual facilita los cálculos y la aplicación del proceso informático. Los números que se asignan corresponden cada uno a una forma gráfica, la correspondencia se establece en un diccionario de formas gráficas del texto en el cual a cada forma gráfica le corresponde su número de orden alfabético (diccionario alfabético) o su número de orden de frecuencia, (diccionario de frecuencia).

4.2. Documentos lexicográficos

La lexicometría comprende les métodos que permiten operar las reorganizaciones formales de la secuencia textual y proceder a realizar los análisis estadísticos pertinentes sobre el vocabulario a partir de una segmentación.

Glosario de las formas gráficas. Si a cada forma se le asocian las coordenadas de sus ocurrencias en el corpus, se obtiene el índice de este. El índice puede ser en orden lexicográfico (índice lexicográfico) o en orden de frecuencia (índice jerárquico).

Concordancia. El conjunto de contextos en los cuales es utilizada una forma llamada forma polo se denomina concordancia de la forma. Las concordancias de una forma muestran bajo qué contextos es utilizada una palabra en el corpus.

Umbral de frecuencia. Para que el análisis; estadístico tenga sentido, será necesario que las formas aparezcan con una frecuencia mínima, pero ello normalmente se eliminan las formas poco frecuentes del corpus, escogiendo un umbral de frecuencias por encima del cual conservamos las formas, estas formas solo se eliminan

para los análisis estadísticos; es decir, las formas continúan en el corpus pero se eliminan de las tablas léxica y léxica agregada.

Medida y comparación de la riqueza del vocabulario. En cuanto más crece el corpus, más; aumenta el vocabulario, sin embargo el crecimiento marginal del número de formas tiende a disminuir. El tamaño del vocabulario no es proporcional a la longitud del corpus; con el propósito de comparar partes del corpus es conveniente que sean de tamaño similar. Los elementos de comparación son el tamaño del vocabulario, el numero de hápax (forma empleada solamente una vez) y las frecuencias máximas en cada parte.

Formas y segmentos característicos. Una forma gráfica se caracteriza por el número de sus ocurrencias o frecuencia y por las posiciones de la forma en el corpus, cuyo conjunto constituye la localización de la forma. El corpus se caracteriza por la frecuencia máxima, y por la distribución de las formas según su frecuencia, lo cual puede representarse por el histograma de efectivos por frecuencia. La detección de formas particularmente altas o particularmente bajas dentro de un corpus son usualmente de importancia para el investigador, pues representan características distintivas de los textos entre sí.

Tabla 2: *formas léxicas por orden de frecuencia. Fuente: elaboración propia.*

Num	Palabras empleadas	Frecuencia	Longitud
66	De	907	2
154	La	416	2
98	El	369	2
103	En	295	2
270	Trabajo	290	7
283	Y	254	1
134	Grado	231	5
72	Del	189	3
201	Para	156	4
211	Por	145	3
7	A	117	1
230	Que	109	3
275	Un	100	2
266	Tiempo	99	6
189	No	98	2
225	Proyecto	95	8
246	Se	94	2
163	Los	93	3
124	Falta	88	5
52	Con	85	3
265	Tesis	85	5
78	Desarrollo	76	10

4.3. Tratamiento de la base de datos

- Para realizar el proceso inicial de depuración de la base de datos se crea una aplicación en el programa SPAD T, lo cual permite una visualización global del vocabulario del corpus objeto de estudio. La longitud del corpus referido es de 9254 formas de las cuales 1629 son formas distintas, lo cual

equivale a un 17,60 % del total de palabras. Parte del vocabulario es presentado a continuación en la Tabla 2, junto con la frecuencia de cada forma, la cual es proporcionada en la salida del programa SPAD T y tiene especial importancia.[3]

- Reducción del vocabulario. Al realizar la lectura del vocabulario en la Tabla 2, se puede observar que las palabras consideradas herramienta son las de mayor frecuencia, por ejemplo: de, la, el, en, y, con frecuencias de 907, 416, 369, 295, y 254, respectivamente. Para una reducción inicial del vocabulario se eliminan estas formas junto con las formas porque, por, para, además; entre otras, las cuales también aparecen con frecuencia alta y son utilizadas como palabras herramienta en las diferentes respuestas.

  Como sinónimos dentro del contexto se pueden observar las formas proyecto y monografía; director y profesor, iniciar, inicie, inicial, inicio, las cuales serán representadas por una sola de las formas. Para observar si estas palabras son empleadas de la misma manera dentro de cada contexto, se realizó un análisis de concordancias, algunas de estas se muestran en el apéndice. La reducción de estas formas a una sola se realizó mediante el procedimiento CORTE del programa SPAD T, en el cual además es posible suprimir formas que se consideran innecesarias, y mediante equivalencia hacer más largas las formas cortas que se deseen conservar.

  Con este procedimiento, como se mencionó antes, se eliminaron las palabras: por, para, además, porque, etc y se dejó una sola forma para sinónimos. Mediante el procedimiento SETEX del mismo paquete, se eliminaron las formas de menos de cuatro letras, aumentando la longitud de algunas palabras que son cortas y que pueden cambiar el sentido de una frase, como son: no, sí, más, mal, con el objeto de que no sean eliminadas cuando se supriman del corpus las formas cortas o herramienta.

  Vocabulario reducido. Una segunda reducción del corpus se realizó también mediante el procedimiento SETEX, fue la reducción del número de palabras del vocabulario por frecuencia lo cual se hizo con un umbral de 4, con este umbral se retienen formas con frecuencia mayor o igual que 5. Estas formas son las que constituyen las respuestas características de los individuos, una forma pronunciada con poca frecuencia hace parte de respuestas aisladas de algunos individuos. Luego de este procedimiento se conservan 6039 ocurrencias y 151 palabras distintas, las cuales son retenidas para el análisis, parte del listado de las 151 formas retenidas es presentado en la Tabla 3.

  Para una descripción inicial del corpus de datos se realiza una lectura global del vocabulario. La lectura del vocabulario reducido en la tabla anterior muestra que formas como: demora, dificultad, falta, información, pasantía, investigación, tesis, tiempo, trabajar, trabajando, problemas, materias, entre otras, son las de mayor frecuencia, obedeciendo probablemente a respuestas que apuntan a justificar el retraso en el grado por: demora en la tesis, demora

[3]Los procedimientos estadísticos fueron realizados con el paquete estadístico SPAD T, y editados para su publicación en LaTeX.

probablemente del director, problema con el proyecto o personales, problemas con las pasantías, por estar trabajando, problemas de tiempo, etc.

Tabla 3: *diccionario de palabras después de la reducción. Fuente: elaboración propia.*

Num	Palabras	Fre.	Long.	Num	Palabras	Fre.	Long.
1	Acerca	5	6	16	Bastante	5	4
2	Alargó	6	6	17	Bibliografía	6	12
3	Algunas	9	7	18	Cambiar	6	7
4	Algunos	14	7	19	Cambio	18	6
5	Análisis	8	8	20	Campo	21	5
6	Anteproyecto	10	12	21	Cancelación	6	11
7	Anterior	5	8	22	Carrera	21	7
8	Aplicación	5	10	23	Circunstancias	5	14
9	Apoyo	9	5	24	Colombia	5	8
10	Aprobación	14	10	25	Comencé	5	7
11	Director	40	8	26	Compañero	7	9
12	Asesoría	11	8	27	Complejidad	12	11
13	Asignación	5	10	28	Consecución	9	11
14	Asignatura	5	11	29	Conseguir	11	9
15	Años	7	4	30	Cuao	16	4
33	Datos	7	5	39	Demasiado	6	9
40	Demora	37	6	42	Desarrollar	92	11
43	Después	8	7	106	Parte	29	5
44	Dificultad	65	10	107	Pasantía	34	8
67	Extenso	21	7	108	Perder	31	6
68	Falta	88	5	109	Personales	8	10
69	Final	13	5	70	Financiación	10	12
80	Información	42	11	142	Tesis	85	5
81	Iniciar	26	7	143	Tiempo	99	6
82	Investigación	32	13	145	Trabajar	29	8
83	Jurados	21	7	146	Trabajando	298	7
88	Mala	10	4	147	Trámites	9	8
89	Matemáticas	7	11	149	Universidad	15	11
90	Materias	51	8	150	Varias	7	6
97	Monografía	43	10	151	Veces	7	5

- Segmentos repetidos: se editaron en la Tabla 4 los segmentos repetidos construidos a partir del vocabulario recortado, utilizando un umbral de frecuencia de 3 para los segmentos de longitud dos, y un umbral de frecuencia de 0 para los de longitud 3 o más. Se puede observar que los de mayor frecuencia son: más tiempo, primer proyecto, problemas económicos, problemas personales, opción Colombia, trabajo dirigido, primer semestre.

  De longitud 3 podemos observar los segmentos: requería más tiempo, cuando terminé materias, cambiar *dos* veces, probablemente de director o de trabajo de grado; la edición de concordancias de segmentos como trabajo dirigido, y cambiar dos veces, dan idea del significado de estos segmentos poco claros. (Tabla 4), allí se puede apreciar que dos veces es utilizado por los egresados para afirmar que cambiaron dos veces el proyecto de grado, que tomaron dos veces seminario de grado,o que perdieron dos veces una materia; lo que genera retraso en el tiempo en terminar materias; con trabajo dirigido se refieren a el trabajo dirigido de grado. La lectura del vocabulario y de las formas de combinación de las palabras que lo constituyen junto con el análisis de los primeros planos factoriales dan idea general de las respuestas utilizadas por los egresados para justificar el retraso en la graduación, por circunstancias asociadas con el trabajo de grado.

Tabla 4: *segmentos repetidos por orden de frecuencia. Fuente: elaboración propia.*

Frecuencia	Segmento del texto
20	11- Más tiempo
8	15-Primer proyecto
7	18-Problemas personales
7	17-Problemas económicos
5	5-Décimo semestre
5	14-Opción Colombia
5	24- Trabajo dirigido
4	16-Primer semestre
4	10-Información necesaria
4	7-Dificultad econócica
4	8- Dos veces
4	9-Grado duro
4	1-Algunas materias
4	3-Conseguir información
3	19-Recursos económicos
3	13-Mismo tiempo
3	21-Requirió más
3	23-Terminé materias
3	12-Mayor tiempo
3	6-Demoró más
2	20-Requería más tiempo
2	22-Requirió más tiempo
2	2-Cambiar dos veces
2	4-Cuando terminé materias

Tabla 5: *concordancias de la forma Dos veces. Fuente: elaboración propia.*

Tomé	Dos veces	Seminario de trabajo de grado	1
Repetir	Dos veces	Un preparatorio	1
Perdí	Dos veces	Un preparatorio de la opción de grado	1
Perdí	Dos veces	Matemáticas y estuve enfermo en el 94 lo cual me obligó	1
Tuve que cambiar	Dos veces	De trabajo de grado debido a la complejidad de los mismos	1
Rechazo de	Dos veces	De la propuesta de grado	1

4.4. Particiones del corpus

- En respuestas individuales; esta partición se define en la entrada de los datos, puede corresponder a una realidad *a priori Caso de preguntas abiertas en encuestas* o ser definida en una forma arbitraria, por ejemplo frases o párrafos de un texto literario.

- La partición del corpus en textos puede también venir dada *a priori*, o puede ser el resultado de un agrupamiento de respuestas individuales según un criterio externo. En el primer caso están los textos literarios que son corpus divididos en frases, cada texto puede ser un párrafo, en el segundo caso están las respuestas a preguntas abiertas, el corpus se divide en textos según las características de los individuos, se pueden reagrupar por ejemplo según la categoría profesional y obtener respuestas, por ejemplo: de los profesores, de los abogados, de los médicos, de los ingenieros, etc.

Para la aplicación los métodos de análisis estadístico se construyen tablas de contingencia particulares y se considera una nueva variable, la variable léxica, cuyas modalidades serán las formas gráficas del corpus tratado.

- Tabla léxica: contiene la frecuencia relativa con la que cada forma gráfica forma léxica o palabra ha sido empleada por cada individuo; la tabla léxica es una tabla de contingencia que contiene los perfiles léxicos de los individuos. El objetivo de construir esta tabla es comparar los perfiles léxicos de las respuestas individuales. (Es la tabla de contingencia respuestas * formas). Esta tabla de contingencia se analiza mediante correspondencias simples y clasificación. El análisis de correspondencias de la tabla respuestas * formas hace una representación gráfica de las asociaciones entre filas y entre columnas, permitiendo visualizar de manera general el contenido de las respuestas, la clasificación de las respuestas nos permite obtener grupos de individuos que se parecen en cuanto al vocabulario que emplean para justificar su respuesta.

- Tablas léxicas agregadas: se construye cuando el corpus es particionado en textos que se desean comparar. El objetivo de construir esta tabla es comparar los perfiles léxicos de los textos en los cuales se ha particionado el corpus. En el caso de preguntas abiertas en encuestas se comparan los perfiles léxicos de cada grupo, según las categorías utilizadas para particionar el corpus. Esta tabla contiene las frecuencias de las formas en una parte del corpus de datos cuando este ha sido particionado una o varias veces. (Tabla de contingencia formas * textos). Tablas similares a las anteriores pueden ser construidas al sustituir las palabras por segmentos de frase repetidos. A estas tablas se les puede aplicar el análisis factorial de correspondencias simples (ACS) y los métodos de clasificación automática.

  Los métodos de análisis de datos aplicados a tablas léxicas permiten una aproximación diferenciadora de las respuestas individuales o de las partes del corpus. (Se procede por comparación de los perfiles léxicos). Estos métodos permitirán saber si un grupo de individuos dice lo mismo con respecto a un tema o no.

  El análisis de correspondencias aplicado a estas tablas da una visualización de las proximidades entre individuos y entre formas, permite observar que formas y/o expresiones diferencian a los individuos. Si se usa de manera simultánea en un análisis información textual y no textual, se puede observar cuáles son las características objetivas de los individuos asociadas a un tipo de vocabulario y también, la opinión respecto a un tema de un grupo de individuos. La clasificación automática de los individuos completa y enriquece el análisis. Se puede caracterizar cada clase en función de la información objetiva que se tiene sobre los individuos que la componen (Montenegro & Pardo 1998).

5. El modelo estadístico

El modelo estadístico utilizado usualmente para detectar las formas características en los textos, cuando el corpus ha sido particionado en textos es como sigue: (Montenegro & Pardo 1998). Se considera el texto como una posible muestra del corpus y se sitúa en conjunto de todas las muestras posibles de la misma longitud que pueden ser obtenidas. La variabilidad de la frecuencia se analiza con respecto a la totalidad de sus ocurrencias en el corpus. Con el propósito de establecer el modelo de probabilidad que servirá para detectar las formas características, se consideran equiprobables todas las muestras posibles, que se pueden construir a partir del corpus.

El modelo de probabilidad se establece así: sea X la variable aleatoria definida como el número de veces que la forma i (que tiene frecuencia total en el corpus $(f_{i.})$) aparece en una muestra de tamaño $f_{.j}$ entonces la probabilidad de que la variable aleatoria X tome el valor x está dada por:

$$prob(X = x) = \frac{\binom{f_{i.}}{x}\binom{f_{..} - f_{i.}}{f_{.j} - x}}{\binom{f_{..}}{f_{.j}}}$$

En donde: $f_{i.}, f_{.j}, f_{..}$ son respectivamente: frecuencia de la forma i en todo el corpus, tamaño de la parte j y longitud del corpus. Como puede verse la variable aleatoria X tiene una distribución hipergeométrica con parámetros $f_{i.}, f_{.j}, f_{..}$

Una forma característica positiva. En un texto es aquella con frecuencia interna alta en relación con su frecuencia en todo el corpus. Se usa la notación PSUP(f_{ij}) para la probabilidad de encontrar por lo menos (f_{ij}) ocurrencias de la forma i en el texto j bajo las hipótesis de una extracción al azar sin reposición de $f_{.j}$ entre las $f_{..}$ ocurrencias del corpus. Nótese que si PSUP(f_{ij}) es inferior que un cierto umbral, (normalmente 0.025) definido previamente, se declara la forma característica de especificidad positiva. Para facilitar la lectura se asocia a PSUP(f_{ij}) el valor de prueba (*V. test*) correspondiente a la distribución normal reducida. Un valor *test* se considera en general significativo si es mayor que 1.96.

Una forma característica es negativa. En un texto cuando presenta una frecuencia dentro del texto (frecuencia interna) significativamente baja en relación con su frecuencia en todo el corpus. La notación Pinf(f_{ij}) para denotar la probabilidad de que se encuentren a lo más (f_{ij}) ocurrencias de la forma i en el texto j, bajo las mismas hipótesis anteriores. Como antes si Pinf(f_{ij}) es inferior que un cierto umbral, usualmente 0,025 se declara la forma característica de especificidad negativa. Para este caso el valor de prueba se considera significativo si es menor que -1.96 . El razonamiento para asociar un modelo probabilístico a la aparición de un segmento i de longitud l en la parte j del texto es similar al seguido para las formas características.

Respuestas características. Estas no son respuestas artificiales construidas a partir de las formas características, sino respuestas reales, escogidas según un criterio como representantes del texto.

Criterio del Ji-cuadrado. Cada respuesta puede considerarse como un vector fila cuyas componentes son las frecuencias de cada una de las formas en esta respuesta. Un texto es un conjunto de vectores fila. El perfil léxico promedio del texto es la media de los perfiles de las respuestas del texto. Es legítimo Calcular distancias entre respuestas y textos. La distancia seleccionada entre textos y respuestas es precisamente la distancia Ji-cuadrado. La respuesta más característica será aquella más cercana al perfil medio del texto lo que se hace es ordenar las respuestas en orden decreciente de distancia al perfil medio; dicho criterio tiende a favorecer las respuestas largas.

Criterio del valor medio. Cuando se calcularon las formas características se ha asociado a cada par *forma, texto* un valor *test*, que puede ser positivo o negativo. Según la pertenencia de una respuesta a un texto, se le puede atribuir la media de los valores *test* correspondiente a las formas que componen la respuesta. La respuesta más característica será aquella cuya media sea más alta. Este criterio tiende a favorecer a las respuestas cortas.

Las respuestas características son respuestas originales pronunciadas por los individuos entrevistados. En general se extraen varias respuestas características para cada texto (10 a 20 según el caso).

6. Conclusiones

- Las respuestas libres a preguntas abiertas en encuestas proporcionan información muy valiosa, la cual debe ser tenida en cuenta para enriquecer los análisis cuantitativos, en especial cuando se abordan temas que generan polémica en donde las preguntas de este tipo permiten que el entrevistado o encuestado se exprese libremente, además son de gran ayuda cunado se trata de explicar la respuesta dada a una pregunta cerrada, al incluir la pregunta "¿Por qué?"

- La poscodificación presenta algunos defectos como la mediación del codificador, empobrecimiento del contenido y eliminación de las respuestas raras, sin embargo el método tradicional de la poscodificación de respuestas a preguntas abiertas en encuestas tiene la ventaja de que los resultados son fácilmente explotables.

- Al utilizar de manera conjunta información de carácter textual e información no textual, podemos caracterizar las respuestas dadas por los diferentes grupos de individuos al reagrupar las respuestas por género, edad, profesión, filiación política, por ejemplo. Para el corpus presentado podemos identificar las respuestas características que dan a la pregunta los estudiantes en las

diferentes sedes o segmentarlos de acuerdo a la carrera que cursaron en la Universidad Nacional.

- El método de análisis de datos textuales permite realizar una lectura global del corpus, al realizar una lectura preliminar del diccionario o del diccionario recortado.
- Con los métodos de análisis de datos textuales se puede obtener una tipología directa sin reagrupamiento previo de las respuestas a partir de sus perfiles léxicos.

Recibido: 16 de febrero de 2014
Aceptado: 27 de junio de 2014

Referencias

Berelson, B. & Lazarsfeld, P. F. (1948), *The analysis of communication content*, Universitetets studentkontor.

Ghiglione, R. (1989), *Las encuestas sociológicas: teoría y práctica*, Editorial Trillas, México.

Lebart, L., Salem, A. & Bécue, M. (2000), *Análisis estadístico de textos*, Editorial Milenio, San Salvador.

Montenegro, A. & Pardo, C. E. (1998), *Introducción al análisis Estadístico de datos textuales*, Unidad de Extensión, Departamento de Matemáticas, Universidad Nacional de Colombia, Bogotá.

Pope, J. (2012), *Investigación de mercados. Guía maestra para el profesional*, Norma, Bogotá.

Rincón, W. (2002), Comparación del análisis de datos textuales y el método de las palabras asociadas en el análisis de preguntas abiertas en encuestas, Master's thesis, Universidad Nacional de Colombia.

A. concordancias de las formas de trabajo y monografía y concordancias de la forma proyecto

Tabla 6: *concordancias de las formas de trabajo y monografia.*

LA MONOGRAFIA SE CONVIRTIO EN UN	TRABAJO	QUE CONLLEVA MUCHO TIEMPO	1
EXTENUENTE	TRABAJO	DE CAMPO EN EMPRESAS Y CON ASESORES	1
PARTE APLICADA REQUERIA MUCHO TIEMPO YA QUE SE	TRABAJO	EN 4 EMPRESAS	1
DURE 1 AÑO ELABORANDO MI	TRABAJO	DE	1
SE DEBIO REALIZAR UN	TRABAJO	DE INVESTIGACION QUE DEMANDO MAS TIEMPO	1
A TODOS LOS ANTROPOLOGOS NOS TOCA HACER EL	TRABAJO	DE GRADO PRACTICAMENTE SOLOS	1
EL	TRABAJO	NECESITABA MAS TIEMPO PARA LLEGAR A UN BUEN TERMINO	1
ME LLEVE MAL CON EL DIRECTOR DE MI PRIMER	TRABAJO	DE GRADO Y ME TOCO DEJARLO Y	1
EL	TRABAJO	DE GRADO FUE ALGO MUY NOVEDOSO	1
EL	TRABAJO	DE GRADO FUE MUY LARGO	1
EL	TRABAJO	DE GRADO TUVO PROBLEMAS DE FUNDAMENTACION Y DIRECCION	1
DISTANCIA DEL SITIO DE	TRABAJO		1
CAMBIO DE PROYECTO DE	TRABAJO	DE GRADO	1
EL	TRABAJO	FUE BASTANTE LARGO Y FUE NECESARIO UTILIZAR HASTA	1
IMPOSIBLE DE HACER	TRABAJO	DE CAMPO	1
	TRABAJO	DE CAMPO PREVIO A LA ELABORACION DE LA MONOGRAFIA	1
TOME DOS VECES SEMINARIO DE	TRABAJO	DE GRADO PORQUE	1
MI	TRABAJO	DE GRADO SE DESARROLLO EN LA GUAJIRA Y EL APOYO DE LA	1
DIFICULTAD EN LA CONCEPTUALIZACION Y REALIZACION DEL	TRABAJO	DE GRADO	1
EL DESARROLLO DE UN BUEN	TRABAJO	DE MONOGRAFIA EN MI CONCEPTO REQUIERE DE UN TIEMPO	1
NO HABER DESAROLLADO A TIEMPO EL	TRABAJO	DE GRADO	1
	TRABAJO	COMO AUXILIAR DE INVESTIGACION	1
POR TENER UN	TRABAJO	QUE ME GUSTO MUCHO ME DEMORE	1
REINICION DE	TRABAJO	DE GRADO	1
DEMASIADO	TRABAJO	PARA UNA SOLA PERSONA	1
RETRASO CON EL	TRABAJO	Y ADEMAS TUVE QUE TRABAJAR TODA LA CARRERA	1
LA	MONOGRAFIA	SE CONVIRTIO EN UN TRABAJO QUE CONLLEVA MUCHO TIEMPO	1
DEFICIENTE ASESORIA POR PARTE DEL DIRECTOR DE LA	MONOGRAFIA		1
RETASO DE LOS JURADOS DE LA	MONOGRAFIA	PARA LA SUSTENTACION Y ELABORACION DEL C	1
FALTA DE RECURSOS ECONOMICOS PRA COSTEAR LA	MONOGRAFIA	Y SOSTENIMIENTO PROPIO	1
NO PUDE CULMINAR MI	MONOGRAFIA	Y AL PASARME A PREPARATORIOS NO HAY LA SUFICIENTE	1
DIFICULTADES PARA LA CONSECUCION DE DIRECTOR PARA LA	MONOGRAFIA	?	1
EL DESARROLLO DE LA	MONOGRAFIA	FUE MUY EXTENSO	1
SE FUE MAS TIEMPO DEL PROYECTADO EN ELABORAR LA	MONOGRAFIA		1
IMPOSIBILIDAD DE HACER TRABAHO DE CAMPO EN LA	MONOGRAFIA		1
ES IMPOSIBLE DEDICARSE A LA	MONOGRAFIA	MIENTRAS SE TERMINAN MATERIAS	1

Tabla 7: *concordancias de la forma proyecto.*

PARA EL	PROYECTO	
INEFICIENCIA DE QUIENES FUERON LOS EVALUADORES DEL	PROYECTO	DE GRADO 1
EL PRIMER	PROYECTO	DE GRADO NO OBTUVO FINANCIACION 1
CAMBIO DE	PROYECTO	DE TRABAJO DE GRADO 1
TUVE QUE CAMBIAR	PROYECTO	Y BUSCAR NUEVO DIRECTOR PORQUE EL ANTERIOR NO ESTUVO 1
DE RECURSOS PARA EL TRABAJO DE RETRASO UN POCO EL	PROYECTO	1
REPROBE EL PRIMER	PROYECTO	1
EN EL DESARROLLO DEL TRABAJO	PROYECTO	DE MONOS 1
LA EXTENSION Y DIMENSION DEL	PROYECTO	1
DIFICULTADES EN EL DESARROLLO DEL	PROYECTO	DE GRADO 1
EN EL CASO DEL TRABAJO DE GRADO QUE QUERIA PRESENTAR UN	PROYECTO	DE CALIDAD 1
DIFICULTADES EN ACTIVADES PROPIAS DEL	PROYECTO	1
EN PRINCIPIO EL	PROYECTO	ERA MUY AMBICIOSO 1
FALTA DE ALGUNOS ELEMENTOS NECESARIOS PARA EL	PROYECTO	LOS CUALES SON DE DIFICIL 1
DIFICULTAD AL ESTABLECER UN	PROYECTO	DE INVESTIGACION 1
TRABAJE DURANTE UN AÑO EN UN	PROYECTO	PARA CUYO DESARROLLO FINALMENTE NO SE COMPLETO 1
EL DESARROLLO DEL	PROYECTO	DE GRADO DURO UN AÑO RAZON POR LA CUAL CURSE 11 SEMESTRES 1
MALA DIRECCION DEL	PROYECTO	DE GRADO 1
EL DIRECTOR DEL	PROYECTO	NUNCA SE PREOCUPO 1
EL	PROYECTO	DE GRADO NO FUE REALIZADO EN UN SOLO SEMESTRE 1
EL	PROYECTO	DE GRADO RESULTO SER BASTANTE DISPENDIOSO 1
INICIE UN	PROYECTO	DE GRADO QUE NO FUE EXITOSO Y TUVE QUE REINICIAR 1
DESARROLLO DEL	PROYECTO	DE GRADO 1
EL	PROYECTO	DEMORO DOS AÑOS PARA SU REALIZACION 1
EL	PROYECTO	QUE DESARROLLE FUE MUY COMPLEJO Y REQUIRIO DE MUCHO TRABAJO 1
EL PLANEO DEL	PROYECTO	1
VENCIMIENTO DE TERMINOS DEL	PROYECTO	1
EL DESARROLLO DE UN BUEN	PROYECTO	DE GRADO REQUIERE DE POR LO MENOS UN SEMESTRE 1