Metodología de Encuestas
Vol 5, Núm 1, 2003, 67-75

ISSN: 1575-7803

COMENTARIOS Y RESPUESTA A 'ANÁLISIS TEXTUAL DE ENCUESTAS: APLICACIÓN AL ESTUDIO DE LAS MOTIVACIONES DE LOS ESTUDIANTES EN LA ELECCIÓN DE SU TITULACIÓN'.

Ludovic Lebart
ENST

Mónica Bécue
Universitat Politècnica de Catalunya

Elena Abascal Fernández
María de los Ángeles Franco Manero
Universidad Pública de Navarra

Eduardo García
Universidad de Sevilla

Presentación

El pasado número de Metodología de Encuestas (volumen 4, número 2, Noviembre de 2002) se publicó un artículo de las profesoras Elena Abascal Fernández y María de los Ángeles Franco Manero, del área de Estadística de la Universidad Pública de Navarra. Su título: "Análisis textual de encuestas: aplicación al estudio de las motivaciones de los estudiantes en la elección de su titulación". Con él, apareció por primera vez en Metodología de Encuestas un trabajo sobre análisis textual. Se trata, sin duda, de un campo apasionante en las encuestas, con intenciones, logros y limitaciones similares a lo que encontramos en otros problemas actuales, como lo es la traducción automática.

El tradicional dilema entre respuestas abiertas o cerradas se ha venido salvando habitualmente a favor de las segundas, con el argumento de peso de que la impresionante masa de datos que generan las encuestas no permiten análisis de respuestas abiertas al ser poco propicios a la automatización. Pero esta decisión deja pasar de largo oportunidades de gran valor en la calidad de las informaciones que pueden recogerse mediante interrogación.

El análisis textual ha aparecido como respuesta a la demanda de realizar análisis de respuestas abiertas (o textos) en masas de datos. Requiere mucho esfuerzo, pero hace viable la aparición de alternativas a la clásica solución de ceñirse a las respuestas de opción múltiple.

La aparición del artículo mencionado en el número anterior ha generado suficiente respuesta, desde investigadores e investigadoras de indudable experiencia y prestigio en este campo, como para justificar la inclusión, en esta ocasión, de algunos comentarios, tanto al trabajo como al asunto o materia del que trata. Parte de los comentarios se incluyen en este artículo, y parte se publican como artículos independientes, si bien su origen sigue siendo la referencia al trabajo de las profesoras Abascal y Franco. Por este motivo, les hemos pedido que respondan en este mismo documento a todas las referencias.

Desde la revista es un placer y una oportunidad abundar en esta cuestión y estimular un mayor número de contribuciones que, como éstas, inciden en los interrogantes abiertos en la metodología de las encuestas.

Ludovic Lébart École Nationale Supérieure des Télécommunications

Cada análisis de respuestas a una pregunta abierta en una encuesta constituye un verdadero trabajo de investigación. En efecto, que sepamos, no existe una estrategia de tratamiento estándar y cada nuevo ejemplo aporta una nueva piedra al edificio del análisis textual. Esta investigación puede parecer pesada y laboriosa, pero no hay que olvidarse que aporta una calidad fundamental, comparada con los tratamientos clásicos de post-codificación: conserva el texto original de las respuestas lo más avanzado posible en el análisis y, como consecuencia, garantiza al usuario final que nada ha sido olvidado o descuidado. Por otro lado, este trabajo sugiere concebir un programa en el que todas las etapas que han sido realizadas estarían pre-programadas (análisis directo, análisis agregados, palabras y respuestas características) y en el que la parte manual del tratamiento estadístico sería el objeto de interfaces ágiles (por ejemplo, eliminación de palabras-herramientas).

¿Es posible ir más lejos en el análisis que han hecho las autoras? Probablemente, pero a costa de una cantidad desmesurada de trabajo, en comparación con las ventajas que se pueden obtener. En definitiva, los trabajos suplementarios que se pueden hacer servirían más para tranquilizar a los escépticos y para abundar en los mismos resultados que no a traer nuevos elementos. Se puede, por ejemplo, validar las representaciones gráficas (gráficos 1, 2 y 3) por intervalos de confianza de los puntos, calculados a partir de replicación *boot-strap*. Se puede introducir más información morfosintáctica, trabajando en un fichero *lematizado*, siguiendo el mismo espíritu, para verificar que la *lematización* no trastorne los resultados. Pero la experiencia que tenemos de este tipo de tratamientos nos lleva a pensar que las conclusiones de las autoras no serían diferentes con estos perfeccionamientos. Creo que este trabajo permite dar, de manera muy honrada y clara, una excelente idea del interés y, también, de las dificultades del análisis textual de las respuestas a las preguntas abiertas.

Original:

Chaque analyse de réponses à une question ouverte dans une enquête constitue un véritable travail de recherche. En effet, il n'existe pas de stratégie de traitement standard à notre connaissance, et chaque nouvel exemple apporte une petite pierre à l'édifice de l'analyse textuelle. Cette recherche peut sembler lourde et laborieuse, mais il ne faut pas oublier qu'elle a une qualité fondamentale par rapport aux traitements plus classiques par post-codage : elle conserve le texte original des réponses le plus loin possible dans l'analyse, et donc garantit à l'utilisateur final que rien n'a été oublié ou négligé. Ce travail suggère d'ailleurs de concevoir un logiciel où toutes les étapes réalisées seraient préprogrammées (analyse directe, analyses agrégées, mots et réponses caractéristiques) et où la partie manuelle du traitement statistique ferait l'objet d'interfaces très conviviales (élimination des mots outils par exemple).

Est-il possible d'aller plus loin dans l'analyse qu'ont faite les auteurs? Probablement, mais au prix d'une quantité de travail supplémentaire démesurée par rapport aux avantages que l'on peut en retirer.En fait, les travaux supplémentaires que l'on peut faire serviront plus à rassurer les sceptiques et à conforter les résultats qu'à apporter des éléments nouveaux. On peut par exemple valider les représentations graphiques (graphiques 1, 2, et 3) par des zones de confiance des points, calculées à partir de réplication bootstrap. On peut introduire plus d'information morpho-syntaxique, en travaillant sur un fichier lemmatisé, dans le même esprit, pour vérifier que la lemmatisation ne bouleverse pas les résultats. Mais l'expérience que nous avons de ce type de traitement nous fait penser que les conclusions des auteurs ne seront pas changées par ces perfectionnements. Je crois que ce travial donne de façon très honnête et claire, une excellente idée de l'intérêt, et aussi des difficultés de l'analyse textuelle des réponses aux questions ouvertes.

Eduardo García Jiménez Universidad de Sevilla

El trabajo que presentan Elena Abascal y María de los Ángeles Franco abre un frente de debate de impredecibles consecuencias, al menos esa es nuestra esperanza.

Hagamos una primera aproximación. A saber: está el tema de los “datos” en forma de “textos”, del “análisis” y, ligado a éste último, el de la “codificación” o la “post-codificación”. Pero hay más, están “la fundamentación teórica” del análisis y, como no podía ser menos, la propia “interpretación” de los resultados y la “extracción de conclusiones”. El artículo que comentamos permite, pues, abordar una serie de temas que a menudo se relegan dentro de la vorágine que supone la planificación y el desarrollo de una encuesta.

Comencemos por el concepto de dato, sin que ello implique ninguna prioridad conceptual. Si consideramos que un dato es una elaboración primaria, una forma verbal como “carrera”, “gusta” o “salidas” ¿es un dato? Dicho otro modo, en relación con la pregunta que se plantea en la entrevista del artículo que comentamos «*¿Cuáles son los motivos principales por los que has elegido esta titulación?*», las formas gráficas “dere-

cho", "trabajar"o "ser" ¿constuyen un dato? y los textos del segmento "salidas profesionales", "muchas salidas" o "más salidas" ¿son datos?

Cuando a alguno de los encuestados a los que se refiere el artículo responde a una pregunta cerrada, en la que se le pide que identifique los estudios que realiza, lo hace limitado por determinadas opciones de respuesta (p.e. Economía, Ingeniería Industrial, Enfermería, etc.). La respuesta "Economía", correspondiente a la pregunta «*¿Qué estudias»?* ¿es un dato de la misma naturaleza que las respuestas "carrera" o "muchas salidas"?, que se obtienen cuando se pregunta por los motivos del estudiante para elegir una titulación.

No son de la misma naturaleza. El dato en el caso de la pregunta «*¿Cuáles son los motivos principales por los que has elegido esta titulación?»* no es "tienen más salidas" o "relacionado", sino el conjunto de la respuesta ofrecida por el entrevistado a esa pregunta. Para que la respuesta a una pregunta abierta pueda considerarse un dato, deberíamos contemplarla en su conjunto. En términos de su análisis, deberíamos considerar como dato al menos un segmento de texto que permitiera que las palabras signifiquen su sentido; eso que Confucio llamó la rectificación de los nombres.

Las autoras del artículo lo dicen bien cuando afirman que el examen de las formas y su frecuencia se queda muy limitado, pues su significado depende mucho de la frase o contexto en que se encuentre ubicada y esto impide hacer afirmaciones sobre relaciones que se puedan establecer.

La idea de dato nos lleva a las de análisis, categorización y codificación. Tomemos una definición de análisis de datos textuales: conjunto de manipulaciones, transformaciones, operaciones, reflexiones y comprobaciones realizadas a partir de textos con el fin de extraer significado relevante para el problema de investigación. Desde ella volvámonos a situar la idea de dato antes debatida, qué debe ser objeto de manipulación, transformación, reflexión y comprobación: ¿las formas empleadas? ¿los segmentos de texto? ¿las unidades temáticas? Si consideramos que un dato textual debe reflejar lo que las formas empleadas (palabras) quieren decir, el análisis debería tomar como unidad segmentos de texto con sentido, es decir, frases o ideas que rectifiquen los conceptos a los que se refieren. De lo contrario, podemos acercar el análisis textual a la interpretación jeroglífica, ya que se nos obligaría a interpretar una serie de segmentos de texto desde la oscuridad contextual y, como ya comentaremos, sin el apoyo de unos fundamentos teóricos.

Cuando agrupamos conceptualmente las unidades que son cubiertas por un mismo tópico estamos dentro de un proceso de categorización, si además le asignamos un indicativo (código) propio de la categoría en la que se incluye, estamos codificando. Al separar entre sí los segmentos de texto con sentido, como proponíamos más arriba, estamos categorizando, si también le ponemos un código al segmento así obtenido lo habremos codificado. Hemos llegado justo a aquello que se quería evitar; recordemos el artículo: *La estadística textual (...) no parte de una reducción de la información a priori sino que utiliza toda la información disponible sobre el encuestado*.

Si no categorizamos ni codificamos a priori ¿no tendremos que hacerlo a posteriori? Es posible que la interpretación de los datos termine llevándonos a una post-categorización de los mismos: ¿puede el investigador interpretar los resultados del análisis textual sin contar con sus categorías mentales previas? ¿De qué otro modo pueden

interpretarse las representaciones sobre el plano? ¿Cómo se puede dar sentido a las clases? ¿Por qué se eligen variables como el sexo, la titulación y las notas en Secundaria y ESO? ¿Cuáles son las ideas previas sobre las respuestas de los estudiantes? Si no se alberga preconcepción alguna sobre el cariz de las respuestas de chicos y chicas ¿por qué se incluye la variable sexo o se omiten otras variables? En definitiva, la categorización e incluso la codificación son necesarias para reducir conceptualmente la complejidad de las respuestas, si categorización y codificación no se plantean antes de iniciar la recogida de datos, o a lo sumo antes de iniciar el análisis, ¿no tendremos que utilizarlas al final del estudio para poder interpretar con sentido (desde alguna finalidad, objetivo o problema) los resultados del análisis textual?

¿Qué aporta entonces la estadística textual? Entre otras cosas: facilita el proceso de categorización del investigador; le permite explorar desde la significación de un texto lo que las personas quieren realmente decir cuando responden a una pregunta abierta; le ayuda a obtener evidencias para sus propias categorías de análisis. Así, el análisis textual, cuando se utiliza en un sentido inductivo, está al principio y no al final de un proceso de investigación, de modo que sus conclusiones son sobre todo un punto de partida clave para profundizar en un problema.

Un elemento más a considerar es el de los fundamentos teóricos de la investigación. La categorización a priori de los datos textuales implica la existencia de una teoría o modelo de referencia, más o menos formalizado. Por oposición, de una categorización a posteriori se deduce una ausencia de una teoría o modelo de referencia. Pero, como se señalaba más arriba, resulta complicado construir una línea argumentativa *a forteriori* con el simple apoyo de los datos.

Si la interpretación de los datos de una forma coherente requiere de un referente teórico previo ¿qué puede aportarnos el análisis textual? El análisis textual puede ayudarnos a establecer hipótesis de trabajo iniciales que faciliten la identificación de constructos, la iluminación de relaciones entre conceptos o variables, en suma, a enriquecer la comprensión sobre un fenómeno objeto de estudio. Efectivamente, con el análisis textual podemos establecer hipótesis novedosas o rivales, explorar nuevas relaciones o perspectivas sobre un problema.

No obstante, también puede utilizar el análisis textual en la comprobación de hipótesis previas. Es decir, además de favorecer la formulación de hipótesis, el análisis textual puede confirmar hipótesis o supuestos de partida cuando existe una teoría previa. Desde esta perspectiva, la identificación de los segmentos de texto puede hacerse en función de un referente teórico, que también puede ser la referencia fundamental en la interpretación de los ejes factoriales y las clases.

Mónica Bécue Universitat Politècnica de Catalunya

Introducción

Me parece muy interesante que la revista Metodología de las Encuestas haya publicado el artículo "Análisis textual de encuestas: aplicación al estudio de las motivaciones

de los estudiantes en la elección de su titulación" por E. Abascal Fernández y M.A. Franco Manero. Dicho artículo ofrece una presentación de los métodos de análisis textual, y de su aplicación al análisis de respuestas a preguntas abiertas en las encuestas, que será de gran utilidad para los investigadores y profesionales que utilizan las encuestas por cuestionario.

Mediante el soporte de una encuesta a los estudiantes, tratada de forma detallada, se muestran claramente las aportaciones del análisis textual al tratamiento de encuestas. Esta metodología ofrece la ventaja de un tratamiento automático de las respuestas libres, relegando la interpretación a la fase final y permitiendo así una mayor objetividad.

Resultados ofrecidos por el análisis textual

El cuestionamiento abierto proporciona una información específica, distinta de la que podría aportar un cuestionamiento cerrado, como lo han mostrado varios estudios comparativos (ver por ejemplo, Lebart y col., 2000). Cuando se utiliza una pregunta abierta, se persiguen objetivos que sólo el cuestionamiento abierto permite alcanzar. En efecto, además de desear conocer la situación, actitud u opinión de los entrevistados, se desea recoger opiniones que no se pueden resumir en pocas palabras, evaluar el grado de interés del entrevistado (respuesta larga y argumentada o respuesta lacónica), tener en cuenta el nivel de lenguaje, o captar matices tal y como es la implicación personal. Así, las respuestas citadas en la tabla 4, p. 207, muestran que las respuestas de los hombres son, en media, más largas que las respuestas de las mujeres. Sería interesante estudiar si esta tendencia se verifica sobre el conjunto de todas las respuestas, lo que conduciría a emitir la hipótesis de que la elección de la carrera motiva más a los hombres que a las mujeres. O, para citar otro ejemplo, Abascal y Franco muestran que, en el marco de esta encuesta, no es lo mismo que una carrera guste y que guste *desde siempre*, etc.

El análisis textual comporta una serie de herramientas que se enmarcan en el análisis estadístico multidimensional descriptivo, frecuentemente llamado “Análisis de datos”. El enfoque de estas herramientas no lleva a emitir aserciones apoyadas en pruebas estadísticas sino a subrayar diferentes rasgos presentes en las observaciones que permiten orientar investigaciones posteriores y/o emitir nuevas hipótesis. Los resultados así obtenidos presentan una gran riqueza y diversidad. Así, el estudio muestra, a la vez, que las salidas profesionales son importantes a la hora de escoger una carrera pero que también es relevante el gusto por determinados estudios; o que, para una carrera dada, la motivación de los hombres y la de las mujeres son muy similares, aunque determinadas carreras atraen más a las mujeres que a los hombres y viceversa, etc..

Calidad de la información

No está de más insistir sobre la importancia de la calidad de la recogida de información, particularmente importante en el caso de las preguntas abiertas.

Las preguntas abiertas deben interesar y motivar, deben ser comprensibles y no restarse a diferentes interpretaciones. Además, deben plantear una sola pregunta a la vez. No son de la misma naturaleza que las preguntas de una entrevista en profundidad. La recogida de los datos textuales requiere una buena formación de los entrevistadores.

En caso de una encuesta cara a cara o por teléfono, se debe anotar la respuesta del entrevistado, integralmente, sin resumirla mediante palabras-claves y sin hacer hablar al entrevistado en tercera persona.

En el momento de la captura informática de las respuestas, se deben evitar los errores de trascripción, emplear una puntuación clásica y evitar las abreviaciones.

Normalización y lematización

Aunque las respuestas abiertas presenten menos problemas que otro tipo de texto, es conveniente "normalizar" el texto. Esta operación comprende una cuidadosa corrección ortográfica (facilitada por el empleo de un corrector automático que, desgraciadamente, puede revelarse insuficiente), emplear solamente caracteres en minúsculas (excepto para la inicial de los nombres propios), emplear siempre una misma notación para una misma palabra (por ejemplo en caso de siglas que pueden venir separadas por puntos o no), asegurarse que determinados signos juegan un papel unívoco, etc.

Si se dispone del recurso de un analizador morfosintáctico, se pueden lematizar las respuestas abiertas, es decir, trasformar las diversas formas verbales de un verbo en su infinitivo, y hacer el análisis dichas respuestas además del análisis efectuado a partir de las formas gráficas. La comparación de los resultados resulta siempre enriquecedora. Al respecto, Lebart y col., 2000 ofrecen los resultados obtenidos con una encuesta en castellano, sin y con lematización.

Conclusión

Plantear una pregunta abierta o cerrada es una elección que se hace en el momento de construir el cuestionario. Esta elección depende de métodos disponibles para tratar las respuestas abiertas. Esperamos que el artículo de E. Abascal y M.A. Franco, así como de la serie de comentarios publicados en este número, haya convencido a los lectores del interés de los métodos reagrupados bajo el nombre de "análisis textual".

Como se ha comentado, las preguntas abiertas no sustituyen a las cerradas, pero se insertan en el conjunto del cuestionario para complementar una batería de preguntas cerradas. Solamente así, las respuestas obtenidas son interpretables.

La gran disponibilidad de textos en soportes informatizados ha conducido a la emergencia de un nuevo campo de investigación y aplicación conocido bajo el nombre de minería de textos. En este campo, las respuestas abiertas ocupan todavía un lugar muy secundario. No obstante, se puede pensar que el tratamiento conjunto de datos abiertos y datos cerrados o, dicho de otra forma, de textos y de metainformación sobre los textos o bien de minería simultánea de textos y datos, se irá imponiendo. Las respuestas abiertas y cerradas de encuesta constituyen un claro ejemplo de la necesidad de explotar conjuntamente los dos tipos de información.

Referencias

Lebart L., Salem A., Bécue M. (2000). *Análisis estadístico de texto*s. Milenio: Lérida (España), con prólogo de Daniel Peña.

Elena Abascal Fernández y
María de los Ángeles Franco Manero Universidad Pública de Navarra

Los comentarios y reflexiones recogidos en este artículo junto con las exposiciones realizadas por R. Álvarez y K. Fernández en este mismo número de la revista, han ampliado y enriquecido considerablemente la visión del análisis textual de las respuestas de preguntas libres que proporcionamos en el artículo.

Agradecemos sinceramente los comentarios y contribuciones aportados, pues desarrollan distintos aspectos de la metodología que por razones obvias no pudimos contemplar en un único artículo, cuyo objetivo principal era presentar, por primera vez en la revista Metodología de Encuestas, una forma de enfrentarse al análisis de las preguntas de respuesta libre y favorecer la reflexión sobre esta forma de estudio.

También agradecemos que los investigadores más relevantes en este campo hayan aceptado participar en este proceso.

En general, estamos de acuerdo con todos los comentarios realizados. Entre todos muestran la complejidad y riqueza de la metodología. Durante más de treinta años de investigación científica en la estadística textual se han planteado y resuelto muchos problemas aunque aún quedan algunos aspectos mejorables.

Los autores hacen referencia a diferentes aspectos metodológicos que no han sido utilizados en el estudio publicado. Evidentemente, la profundización en el análisis mediante estas técnicas: estudio de la longitud de las respuestas, de la riqueza de las formas utilizadas, la lematización, validación de los resultados mediante *boot-strap*, etc. permiten ir más lejos en el análisis. En realidad, el artículo se basa en un trabajo más profundo, en el que se realizaron los análisis suplementarios citados y otros, como la utilización de diferentes distancias para obtener respuestas características. Sin embargo, como asegura L. Lebart que probablemente ocurriría, no aportan grandes ventajas y sí suponen alargar enormemente la presentación del trabajo.

Un aspecto que no contemplamos en el estudio ampliado fue la validación de las representaciones gráficas por intervalos de confianza de los puntos, calculados a partir de replicación *boot-strap*. Este procedimiento, como hemos podido comprobar posteriormente, es muy interesante; sin embargo, era un procedimiento costoso, dado que no estaba pre-programado en el paquete estadístico utilizado.

Como resumen de todo lo expuesto en estos artículos, podríamos decir que cuando se plantea elegir entre preguntas cerradas y abiertas hay que pensar, por una parte, los aspectos relativos a la calidad de la información recogida y, por otra, el método de análisis.

Los métodos de estadística textual proporcionan herramientas extraordinarias para poder extraer la información contenida en las respuestas. Es el procedimiento de análisis que más se aproxima a la realidad. Ahora bien, como siempre ocurre, esta metodología no está exenta de dificultades. Cuando se trata de comprimir miles de palabras en unos resultados concisos, siempre hay una simplificación que puede producir alguna deformación. Por otra parte, como manifiesta L. Lebart, cada análisis textual es una verdadera investigación. Aunque se ha expuesto una metodología de análisis, ésta no es totalmente automática, el investigador dispone de muchas opciones y tiene que tomar decisiones no

excluyentes o realizar el análisis de varias formas diferentes para comparar los resultados. Aquí el arte y la experiencia del investigador enriquecen el estudio.

Estamos convencidas de que la estadística textual constituye hoy en día una metodología idónea para el análisis de las preguntas de respuesta libre, así como otro tipo de textos, abierta a todas las aportaciones que la enriquezcan.