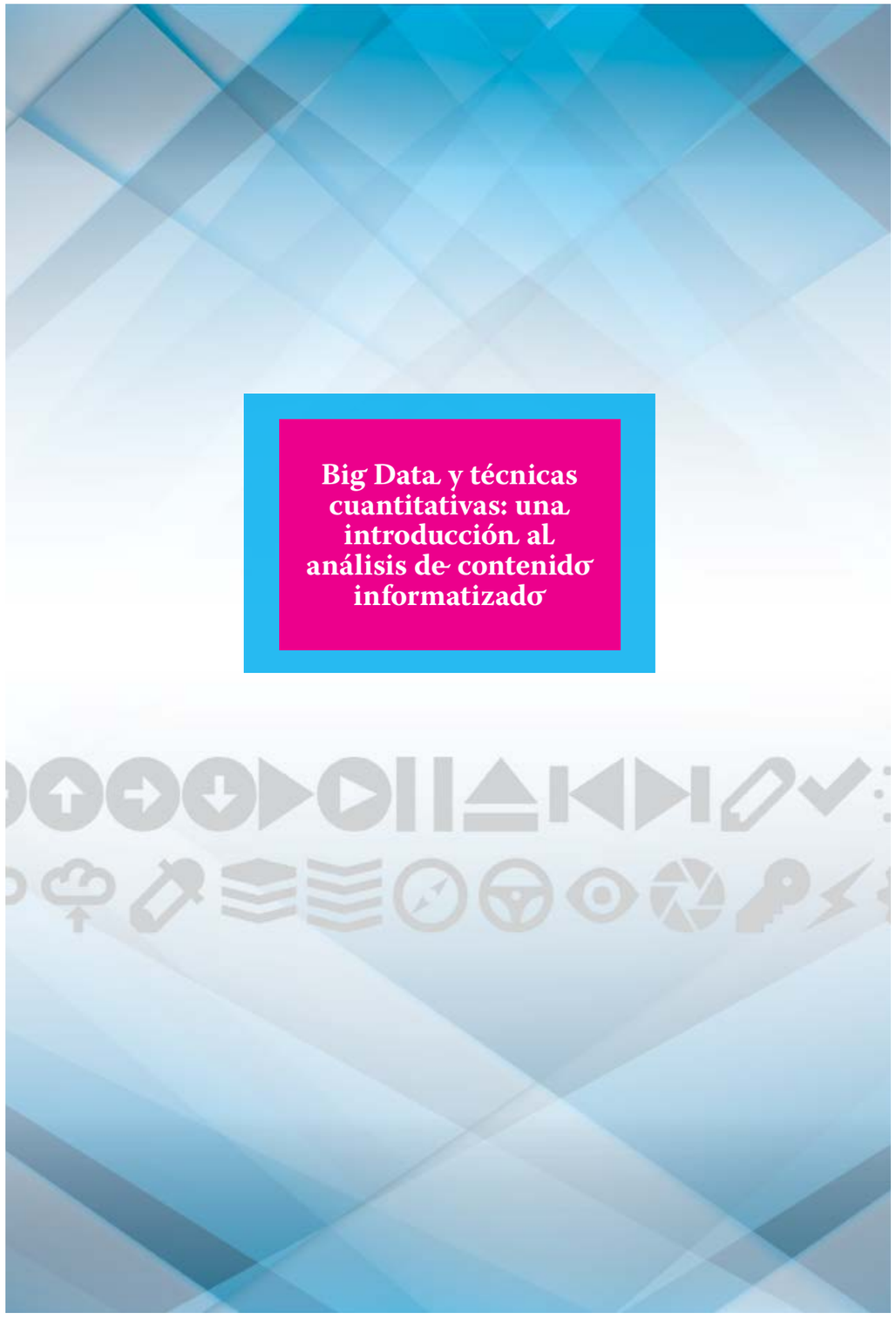

Big Data y técnicas cuantitativas: una introducción al análisis de contenido informatizado

RESUMEN

El siglo XXI viene marcado por la hegemonía de Internet como paradigma organizador de la convivencia: se han creado medios específicamente arraigados al ciberespacio, o redes sociales que estimulan el intercambio de usuario a usuario y de usuario a colectivo, entre otras formas expresivas. La ascensión de Internet tiene un efecto muy positivo en la investigación en Ciencias Sociales: la esfera digital, caracterizada por su elevado dinamismo, se distingue asimismo por la vocación de permanencia que poseen los contenidos. Internet se erige como una prótesis de la memoria colectiva que, de una forma relativamente sencilla, permite adentrarse en extensos depósitos de información. En las siguientes líneas, mostraremos una nueva perspectiva metodológica para diseccionar la memoria colectiva digital: el análisis de contenido informatizado. Esta técnica, de carácter cuantitativo, ha dado un extraordinario impulso en las tres últimas décadas gracias al desarrollo general de la informática. Tras describir sus características generales, expondremos un ejemplo práctico que sintetiza la fiabilidad y la validez de esta herramienta, particularidades que han llevado al análisis de contenido informatizado a conformar una emergente línea de estudios dentro de la Comunicación Social, pero también a erigirse como una útil herramienta de marketing –al servicio de las grandes corporaciones- para detectar las opiniones de las tendencias colectivas.

PALABRAS CLAVE: Análisis de contenido informatizado; Internet; Concordance; CatPac; AntConc

ABSTRACT

The XXI century is marked by the hegemony of the Internet as an organizing paradigm of coexistence: there have been created specifically rooted in cyberspace media or social networks that stimulate the exchange of user -to-user and user group, among other expressive forms. The rise of the Internet has a very positive effect on social science research: the digital sphere, characterized by high dynamism, is also distinguished by the intention of permanence of the contents. Internet stands as prosthesis of collective memory, in a relatively simple manner; it offers insights into vast repositories of information. In the following lines, we show a new methodological approach to dissect the digital collective memory: the computerized content analysis. This quantitative technique has an extraordinary momentum in the last three decades due to the general development of computing. After describing their general characteristics, we will present a practical example that synthesizes the reliability and validity of this tool, characteristics that have led to computerized content analysis to shape an emerging line of research in social communication, but also to establish itself as a useful -at marketing tool serving the large corporations to detect the opinions of collective trends.

KEY WORDS: Computerized content analysis; Internet; Concordance; CatPac ; AntConc

Daniel Barredo Ibáñez, Ph. D

Investigador de la Universidad de las Américas

daniel.barredo@udla.edu.ec

ARTÍCULO PRESENTADO PARA REVISIÓN: 17 DE AGOSTO DE 2014
ARTÍCULO ACEPTADO PARA PUBLICACIÓN: 26 DE SEPTIEMBRE DE 2014

INTRODUCCIÓN

En siglos anteriores, los lingüistas, los sociólogos, los analistas de los medios de comunicación de masas y, en general, los investigadores en Ciencias Sociales se enfrentaban a un problema de almacenamiento: las huellas colectivas tendían a desaparecer o se imprimían en soportes frágiles como el papel, las paredes, etcétera.

Sin embargo, en el siglo XXI asistimos a un cambio de paradigma: Internet se erige como uno de los principales organizadores de la convivencia. En Internet se vuelca y se archiva el conocimiento; podría describirse como una memoria colectiva digital o una exomemoria o memoria externa, definida como:

Una suerte de almacén de recuerdos situado en el exterior de la mente individual, que está formado por un gran número de pensamientos, más o menos relevantes, fijados de alguna manera en un soporte accesible, que puede ser compartido en diferentes momentos temporales por un número amplio de individuos de la especie humana (Pestano y Amézquita, 2006: 3).

La *exomemoria* opera como una "prótesis" (García Gutiérrez, 2003: 28), que acumula el conocimiento universal. Esa es una de las grandes ventajas de la galaxia Internet: la vasta acumulación de retazos cognitivos. Se está produciendo un trasvase progresivo desde las llamadas ciudades reales hacia las llamadas ciudades virtuales, en aclaración de Graham (1998) [1] . Los medios de comunicación se han adaptado a las nuevas rutinas del ciberespacio o se han creados medios nativos digitales muchos de ellos autogestionados o cogestionados por los usuarios (Bowman y Willis, 2003); y han emergido con fuerza las redes sociales, las cuales ocasionan un poderoso impacto simbólico en instituciones tan señeras como la Corona española; como ya avistamos en un trabajo anterior (Barredo, 2013), en 2012 el Rey D. Juan Carlos I pidió perdón a la sociedad española, tras la filtración vía redes de unas polémicas fotos en las que el monarca participaba en una cacería. Fue esta la primera vez en la historia de la democracia en que la monarquía española se disculpaba ante la nación; y fue, recordémoslo, gracias a la presión ciudadana canalizada a través del "espacio abstracto" (Oller y Barredo, 2012: 16) que es Internet.

El ciberespacio, con sus diversas formas de participación, configura una mayor interpretación de la esfera que origina la opinión pública (Ruiz, Domingo, Micó et al., 2011). Es, en cierto sentido, un reflejo que estructura y que complementa al colectivo: sus contenidos imprimen en bytes y en ficheros las huellas del presente cotidiano. Para entender -por continuar con el mismo ejemplo-, las disculpas vertidas por SS.MM. D. Juan Carlos I, ningún analista podría deshacerse de Internet. Las reacciones publicadas en los medios convencionales no explican por sí solas la humilde actitud adoptada por el anterior titular de la Corona española, sino que hay que remontarse a la actitud de los usuarios en red, o a novedosos efectos como el llamado "remolino informativo" (Oller y Barredo, 2012: 17), para poder interpretar los hechos desde un marco de objetividad.

Todavía en 2014, en la mayoría de los países, los medios convencionales abanderan jerárquicamente la hegemonía mediática; pero como ya vio Bucy (2003), el ciberespacio tiende a reemplazar a la televisión -reinante hasta ahora- en la escala de consumo e importancia. Y más allá; en un intento por recuperar la credibilidad perdida, los medios han comenzado a aceptar a los usuarios como redactores (Bowman y Willis, 2003). El llamado open journalism -periodismo abierto a las contribuciones de las audiencias- ocasiona que, en los cinco enfrentamientos entre Real Madrid y Barcelona celebrados en la temporada 2010-11, el 30% de las palabras totales de las coberturas de MARCA.com (principal cabecera deportiva de España y una de las líderes en América Latina) fueron gestionadas directamente por los usuarios, sin intervención directa de un periodista (Barredo y Oller, 2013).

I.- TÉCNICAS ANALÍTICAS MEDIANTE LA ASISTENCIA INFORMÁTICA

La ascensión de Internet tiene un efecto muy positivo en la investigación en Ciencias Sociales: la esfera digital, caracterizada por su elevado dinamismo, se distingue asimismo por la vocación de permanencia que poseen los contenidos. De una forma relativamente sencilla, el investigador puede adentrarse en extensos depósitos de información; no solo en los textos literarios o periodísticos -como en las hemerotecas tradicionales-, sino en una modalidad que Yus (2010: 177) ha definido como "texto escrito oralizado", y la cual refleja los intentos de los usuarios por volcar su espontaneidad en la ciberesfera. Hay muchas posibilidades: desde los chats o los foros -donde la información suele guardarse en los llamados logs u hojas de registro, como explica Crystal (2002)-, hasta las llamadas entrevistas corales (Barredo y Oller, 2012), que son unas entrevistas que estructuran los usuarios de los cibermedios alrededor de un argumento planeado por la propia cabecera.

Sea para verificar un conjunto de efectos como neologismos, modismos o xenismos; sea para averiguar las tendencias editoriales de los cibermedios; sea para cotejar la importancia asignada a una marca, tema o protagonista, lo cierto es que las extensas cantidades de materiales textuales digitalizados permiten diferentes usos dentro de la investigación en Ciencias Sociales. Son múltiples las posibilidades metodológicas, casi tantas como las necesidades del proyecto en cuestión. Pero en las siguientes páginas vamos a centrarnos en explicar una de esas técnicas analíticas mediante la asistencia informática: el análisis de contenido informatizado.

Esta técnica consiste, básicamente, en analizar un texto digitalizado mediante un software específico. El proceso que resulta permite acometer un examen cuantitativo y muy preciso (Igartua, 2006), sobre una cantidad ingente de material textual. Existen numerosos programas, los cuales esencialmente se dividen entre los que coleccionan las palabras jerárquicamente -como el LIWC [2]-, y que permiten agrupar los términos en categorías, una discriminación muy útil en, por ejemplo, los estudios de Psicología; y esos otros programas que presentan los recuentos no jerarquizados, en bloque -como Concordance, AntConc o CatPac[3] -, los cuales ofrecen una gran apertura y flexibilidad, ventajas que aprovechan disciplinas como la Lingüística o el Periodismo.

Un análisis de contenido informatizado, a diferencia de un análisis de contenido manual, cuenta entre sus ventajas con el hecho de que sus resultados pueden ser comprobados por cualquier otro investigador de una forma relativamente sencilla (Popping, 2000). Se reducen, además, muchos de los errores de los análisis de contenido manuales, por el mismo motivo en que es más probable cometer un error de cálculo en un conteo con lápiz y papel que en una calculadora. Pero es que son trabajos que ofrecen una enorme estabilidad, porque como asegura Diefenbach (2001: 15), los computadores no emiten juicios de valor, "a menos que se les dé comandos específicos explicando cómo hacerlo".

Aunque esta técnica, como explica el profesor Watt (2013), tiene sus precedentes en los estudiantes que realizaban anotaciones sobre la Biblia en los siglos VII–VII, fue a partir de los años sesenta cuando comenzaron a cosecharse asombrosos resultados en las Ciencias Sociales. La fecha no es azarosa: en los años sesenta, justamente, aparecieron dos programas tan importantes como el General Inquirer (Stone, Dunphy, Smith y Ogilvie, 1966) o el WORDS (Iker y Klein, 1974). Pero tenían algunos inconvenientes, entre los que sobresalen la lentitud o la carestía de los computadores. Unas décadas más tarde, esos problemas se han solucionado parcialmente: la informática de usuario alcanza potentes velocidades a precios relativamente bajos pero, en paralelo, existe software de análisis en código libre (como por ejemplo el mencionado AntConc), o con licencias a precios tan asequibles como la de CatPac . Y no olvidemos que ese mismo programa, sin ir más lejos, se utiliza en lugares tan diversos como las Universidades de Harvard o John Hopkins, o corporaciones como AT&T o la Ford Motor Company, por citar algunas instituciones o compañías . Pero es que cualquier persona u organización puede comprar una licencia o descargar libremente uno de estos programas, y competir con los productos originados en las Universidades más prestigiosas o en los departamentos de análisis de mercados de las grandes corporaciones. Y esa es otra ventaja del análisis de contenido informatizado: los instrumentos de investigación son, en estos tiempos, accesibles a la mayor parte de la comunidad científica. Los precios no son ya barreras discriminatorias, como tampoco aspectos como las interfaces de los programas, las cuales han ido adoptando el diseño tipo Windows, con lo cual se han transformado en sistemas más cómodos e intuitivos.

II.- ANÁLISIS PIONEROS SOBRE LA CORONA ESPAÑOLA: ZUGASTI VS BARREDO

Para mostrar las ventajas, de manera empírica, que supone la utilización de este procedimiento asistido, queremos mostrar los resultados de una investigación publicada recientemente.

En 2004, el profesor Ricardo Zugasti se doctoró con una tesis titulada Monarquía, prensa y democracia en la transición española: una relación de complicidad (1975-1978). Era este un trabajo pionero en el cual Zugasti afirmaba que durante el periodo histórico de la Transición política española, los medios habían adoptado un consenso periodístico alrededor de determinados temas, como por ejemplo la monarquía. El consenso, según ese estudio, se había mantenido estable durante la mayor parte de la democracia:

1. Dentro de las ciudades virtuales anotamos, por ejemplo, las tiendas que permiten adquirir bienes y servicios; las comunidades de intercambio de opinión, llamados a propósito foros, etcétera.

2. Para más información se aconseja visitar su página web oficial: http://www.liwc.net/.
3. Para más información, se aconseja visitar sus páginas web: http://www.concordancesoftware.co.uk/, http://www.galileoco.com/N_catpac.asp y http://www.antlab.sci.waseda.ac.jp/antconc_index.html .

Entre 1975 y 1978, periodo en el que institucionalmente se completó la transición, la prensa forjó la imagen de Juan Carlos I que se ha mantenido hasta nuestros días, caracterizada fundamentalmente por el énfasis puesto en su papel como actor democratizador. Una representación regia que, aunque influida decisivamente por la complicidad periodística con la Corona cuyo origen fue la situación creada tras la muerte de Franco, ha proseguido en líneas generales hasta la actualidad (Zugasti, 2007: 350).

El profesor Zugasti había centrado su análisis en procedimientos cualitativos como la interpretación de las noticias aparecidas en los diarios de la época, o en las entrevistas a periodistas activos durante la Transición española.

Ocho años después, Barredo (2012) se doctoró en la Universidad de Málaga (España) con una tesis titulada El tabú de la expresividad real. Análisis del tratamiento informativo del rey Juan Carlos I en ABC.es y ELPAÍS.com (2009 - 2011). En uno de los apartados de la tesis, el autor coleccionó todas las noticias publicadas sobre el rey Juan Carlos en dos cibermedios representantes de dos tendencias informativas: ABC.es (conservador) y ELPAÍS.com (progresista). Para conseguir la máxima homogeneidad temática posible, ciñó la fecha de búsqueda a los años 2009 y 2011. En uno de esos apartados, publicado en Barredo (2013) se cotejaron mediante el software Concordance casi cuatro mil contenidos, con un volumen de cerca de dos millones de palabras.

En el estudio de Barredo, se obtuvieron las siguientes estadísticas:

1) Macroestructuralmente, el 79,70% de las palabras – tema sobre el rey Juan Carlos de una tendencia y otra, eran las mismas.

2) Microestructuralmente, eran idénticas (en las informaciones sobre el monarca español):

2.1.) El 80% de las 20 siglas más citadas.

2.2.) El 100% de los 20 lugares más repetidos.

2.3.) El 90% de los 20 políticos más frecuentes.

En definitiva, en su estudio de carácter cuantitativo Barredo sostenía que el consenso de la Transición (avistado por Zugasti mediante técnicas cualitativas), permanecía todavía vigente en dos de las cabeceras más representativas del paradigma español. Macro y microestructuralmente, y sin realizar más labores que las del filtrado y gestión, Barredo complementó el estudio doctoral de Zugasti: ambos alcanzaron las mismas conclusiones, a través de procedimientos complementarios.

CONCLUSIONES

Internet, según hemos explicado en las páginas anteriores, ofrece nuevas preguntas de investigación, en tanto que se está produciendo una readaptación de las esferas gestoras de la opinión pública. Se han multiplicado las posibilidades analíticas: y esa es una de las ventajas de la exomemoria digital, porque como subrayan Pestano y Amézquita (2006) el ser humano, durante siglos, ha empleado distintos soportes para volcar un reflejo de su identidad. La ciberesfera posee a priori unas dimensiones infinitas, por lo que genera asimismo un mayor volumen de contenidos; pero es que los contenidos no son solo los resultados de los productores profesionales -como sucedía en siglos anteriores- sino también el fruto de la colaboración ciudadana. Anteriormente no existían huellas de la memoria colectiva como los chats, los foros, las redes sociales, los comentarios de las noticias, etcétera. Todos esos aportes, que acumulan vastas cantidades de material humano digitalizado, exigen la incorporación de novedosos instrumentos de investigación, la búsqueda de nuevas técnicas que, hasta fechas cercanas, eran solo difícilmente manejables. De entre ellas sobresale el análisis de contenido informatizado, el cual presenta algunas propiedades que lo convierten en un atractivo instrumento metodológico, como son:

a) La precisión y la estabilidad instrumental.

b) La objetividad o posibilidad de alejar las injerencias personales del investigador o la discriminación de los errores humanos.

c) La velocidad de los computadores y el coste cero o muy asequible del software específico.

d) La facilidad de gestionar un proyecto de investigación con un reducido número de investigadores.

e) Los bajos costes generales para llevar a cabo un proyecto que lo emplee como metodología principal. Una ventaja a tener en cuenta especialmente en, por ejemplo, los proyectos comparativos internacionales.

f) La explotación de los enormes yacimientos de material textual digitalizado.

g) La apertura instrumental, la cual facilita enormemente la adaptación de la investigación a las nuevas interrogantes suscitadas desde el ciberespacio.

h) Los cada vez más intuitivos diseños de las interfaces del software.

La interrelación de estas ventajas convierte al análisis de contenido informatizado en una de las metodologías más apetecibles y con mayor proyección dentro de las Ciencias Sociales contemporáneas. ◆

REFERENCIAS BIBLIOGRÁFICAS

◆ *BARREDO, Daniel (2013). El Tabú Real. La imagen de una monarquía en crisis. Córdoba: Berenice.*

◆ *BARREDO, Daniel y OLLER, Martín (2013). Análisis de las tendencias macroestructurales de la lengua española en los cibermedios: las entrevistas corales de MARCA. com protagonizadas por los Clásicos entre Real Madrid CF y el Barcelona FC (2010–2011). Signo y Pensamiento, 63, 130 - 150.*

◆ *BARREDO, Daniel y OLLER, Martín (2012). Las entrevistas corales de MARCA.com: un ejemplo de periodismo ideante. Actas del IV Congreso Internacional Latina de Comunicación Social, Universidad de La Laguna, Tenerife, España. Consultado el 01/04/2013 de: http://www.revistalatinacs.org/12SLCS/2012_actas/118_Barredo.pdf.*

◆ *BOWMAN, Shayne y WILLIS, Chris (2003): "We Media: How Audiences Are Shaping the Future of News and Information". The Media Center at the American Press Institute. Consultado el 20/12/2012 de: http://www.hypergene.net/wemedia/download/we_media.pdf.*

◆ *BUCY, Erik P. (2003). Media Credibility Reconsidered: Synergy Effects between On-Air and Online News. Journalism & Mass Communication Quarterly, 80(2), 247 - 264.*

◆ *CRYSTAL, David (2002). El lenguaje e Internet. Madrid: Cambridge University Press.*

◆ *DIEFENBACH, Donald L. (2001). "Historical foundations of computer-assisted content analysis". En: WEST, Mark D. (Ed.). Theory, method, and practice in computer content analysis <pp. 13 - 41>. Westport, CT: Ablex Publishing.*

◆ *GARCÍA GUTIÉRREZ, Antonio (2003). Redes digitales y exomemoria. I/C: Revista Científica de Información y Comunicación, 1, 21 - 39. Consultado el 24/04/2013 de: http://icjournal.files.wordpress.com/2013/01/completo_1.pdf.*

◆ *GRAHAM, Stephen (1998). The end of geography or the explosion of place? Conceptualizing space, place and information technology. Progress in Human Geography, 22, 165 - 185.*

◆ *IGARTUA, Juan José (2006). Métodos cuantitativos de investigación en comunicación. Barcelona: Bosch.*

◆ *IKER, Howard P. y KLEIN, Robert H. (1974). WORDS: A computer system for the analysis of content. Behavior Research Methods and Instrumentation, 6, 430 - 438.*

◆ *OLLER, Martín y BARREDO, Daniel (2012). La Sociedad de los Ideantes: Repensando los conceptos de opinión y esfera pública y las teorías democráticas relacionadas con el fenómeno comunicativo ciudadano. Tenerife: Sociedad Latina de Comunicación Social. Consultado el 01/04/2013 de: http://www.revistalatinacs.org/067/cuadernos/29_Oller.pdf.*

◆ *PESTANO RODRÍGUEZ, José Manuel y AMÉZQUITA CASTAÑEDA, Irma (2006). Exomemoria audiovisual e Internet. Un proyecto necesario. Razón y Palabra, 49. Consultado el 23/04/2013 de: http://www.razonypalabra.org.mx/anteriores/n49/bienal/Mesa%201/Pestano%20y%20Am%E9zquita.pdf.*

◆ *POPPING, Roel (2000). Computer - assisted text analysis. Londres: Sage.*

◆ *RUIZ, Carlos, Domingo, David, Micó, Josep Lluís, et al. (2011). Public Sphere 2.0? The Democratic Qualities of Citizen Debates in Online Newspapers. The International Journal of Press/Politics, 16(4), 463 - 487.*

◆ *STONE, Philip J., DUNPHY, Dexter C., SMITH, Marshall S. y OGILVIE, Daniel M. (1966). The General Inquirer: a computer approach to content analysis. Cambridge, MA: MIT Press.*

◆ *ZUGASTI, Ricardo (2004). Monarquía, prensa y democracia en la transición española: una relación de complicidad (1975-1978) <tesis doctoral>. Pamplona: Universidad de Navarra.*

◆ *ZUGASTI, Ricardo (2007). La forja de una complicidad. Monarquía y Prensa en la Transición española (1975 – 1978). Madrid: Fragua.*

◆ *YUS, Francisco (2010). Ciberpragmática 2.0. Nuevos usos del lenguaje en Internet. Barcelona: Ariel.*

◆ *WATT, Rob J. C. (2013). "Epilogue". En: BARREDO, Daniel (Autor). Monarquía, consenso y democracia. Análisis de contenido informatizado de las coberturas sobre el rey Juan Carlos I en ABC.es y ELPAÍS.com (2009 - 2011). Quito: CIESPAL.*