HECHOS Y PALABRAS: LA REALIDAD COLOMBIANA VISTA A TRAVÉS DE LA PRENSA ESCRITA

*Juan Manuel Caicedo**
Alejandro Gaviria **
*Javier Moreno****

Este artículo presenta la primera aplicación a la realidad colombiana de *culturomics*, un nuevo método de investigación en ciencias sociales que describe tendencias culturales, sociales y lingüísticas con base en el análisis cuantitativo de textos digitalizados. En principio, la mayor o menor aparición de ciertas palabras o expresiones en millones de textos digitalizados revela cambios relevantes en la cultura, la sociedad o el lenguaje. Dicho de otra manera, la dimensión estadística de los textos escritos puede proporcionar información útil sobre ciertos aspectos de la realidad.

El artículo utiliza la totalidad de las noticias y comentarios publicados durante los últimos veinte años en tres medios escritos de circulación nacional: *El Tiempo*, *Semana* y *Dinero*. En números redondos, analiza más de dos millones de artículos que contienen unos seiscientos millones de palabras. Los cambios en la aparición de ciertas palabras, *desempleo*, *recesión*, *corrupción*, *magistrados*, entre otras, ayudan a entender algunos aspectos de la realidad contemporánea. Contar palabras permite contar historias.

Este trabajo es una de las primeras aplicaciones de *culturomics* basada enteramente en publicaciones periódicas. Las aplicaciones más conocidas y difundidas usan libros publicados en el transcurso de varias décadas e incluso de siglos. Este trabajo, en cambio, utiliza

* Magíster en Sistemas. Cursa estudios de posgrado en la Universidad de Carnegie Mellon, Pensilvania, Estados Unidos, [juan@cavorite.com].

** Doctor en Economía. Profesor asociado y decano de la Facultad de Economía de la Universidad de los Andes en Bogotá, Colombia, [agaviria@uniandes.edu.co].

*** Doctor en Matemáticas. Postdoctoral fellow en la Universidad de Waterloo, Ontario, Canadá, [bluelephant@gmail.com]. Agradecemos los comentarios de los asistentes al seminario de investigación del CEDE de la Universidad de los Andes. Fecha de recepción: 9 de febrero de 2012, fecha de modificación: 10 de abril de 2012, fecha de aceptación: 11 de abril de 2012.

artículos publicados en periódicos durante un periodo más breve. Por tanto, hace hincapié no en los cambios culturales de larga duración, sino en cambios institucionales y sociales de corto y mediano plazo. Además, es quizá la primera aplicación de *culturomics* a una realidad local, a un periodo específico en un país particular. Las aplicaciones anteriores son transnacionales, abarcan una realidad más amplia, al menos geográfica y socialmente.

El artículo muestra que algunos fenómenos económicos –como el desempleo y el crecimiento económico– son descritos o seguidos adecuadamente por los cambios en la mención de las palabras correspondientes: *desempleo* y *recesión* en este caso. Muestra que la frecuencia de aparición de *verano* e *invierno* sigue de cerca las fluctuaciones de la temperatura del océano Pacífico. Y revela que, desde una perspectiva de mediano plazo, la aparición de la palabra *corrupción* no ha crecido, la sigla *farc* suele ir acompañada del vocablo *secuestros* y la palabra *magistrados* aparece un mayor número de veces que *congresistas*.

Pero más que examinar problemas de fondo se describen los aspectos metodológicos. El artículo describe una base de datos, presenta un método de análisis y muestra su potencial mediante una serie de ejemplos. La sección 1 revisa los antecedentes y repasa la literatura relevante. La sección 2 describe los datos. La sección 3 compara, para algunos fenómenos socioeconómicos, el comportamiento de los indicadores con el de la frecuencia de aparición de las palabras correspondientes. La sección 4 utiliza frecuencias de palabras para estudiar varios fenómenos de difícil medición. El propósito de esta sección, la más polémica, es mostrar que este método permite desarrollar un nuevo tipo de indicadores en las ciencias sociales. La última sección presenta algunas ideas para futuras investigaciones.

MOTIVACIÓN Y ANTECEDENTES

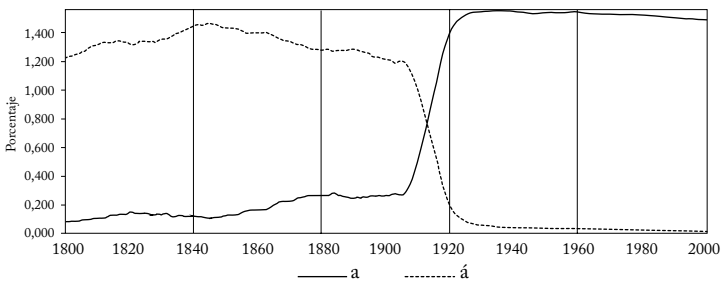
Culturomics es el análisis cuantitativo de tendencias culturales, sociales y lingüísticas con base en libros, periódicos y otros textos digitales disponibles en Internet o en medios similares. Este tipo de análisis usa millones de páginas de texto para estudiar la evolución de patrones culturales y para identificar cambios significativos en la opinión pública. En opinión de Michel et al. (2011), el análisis cuantitativo de textos es un nuevo método de análisis en las ciencias sociales, cuya virtud radica no en el estudio minucioso de algunos textos seminales –la estrategia tradicional de las ciencias sociales– sino en la lectura automatizada de millones de textos de diversa calidad y trascendencia.

Culturomics compensa con volumen su falta de discernimiento; es un método de fuerza bruta.

Michel et al. usan un corpus de más de cinco millones de libros en inglés (el 4% de los libros publicados en ese idioma en todos los tiempos) para analizar, entre otras cosas, la evolución de la gramática inglesa, el auge y la caída de la reputación política, científica y artística, y algunos casos de censura contra artistas judíos. Muchas otras aplicaciones son posibles. Este tipo de análisis permitiría estudiar, por ejemplo, la cambiante popularidad de algunas teorías científicas (la teoría de la evolución), de ciertas ideologías (el marxismo) e incluso de varias formas de pensamiento (los sesgos étnicos o raciales).

En general, las fluctuaciones en el uso de ciertas palabras dan información relevante sobre el mundo del lenguaje y las ideas, sobre la realidad exterior y sobre lo que ha ocurrido (y está ocurriendo) en la mente humana. Algunos ejemplos bastan para ilustrar este tipo de análisis. La gráfica 1 muestra la frecuencia de la preposición *a* con dos grafías distintas: una con acento y otra sin acento[1]. La preposición acentuada (á) era la más utilizada en el siglo XIX, pero la preposición sin acento (*a*) se convirtió en la norma de uso general en la primera mitad del siglo XX. La transición fue rápida, tomó aparentemente menos de una década. En teoría, la existencia de las academias de la lengua hace que algunos cambios ortográficas sean más rápidos en el español que en el inglés (ibíd.).

Gráfica 1
Frecuencia de las preposiciones *a* y á



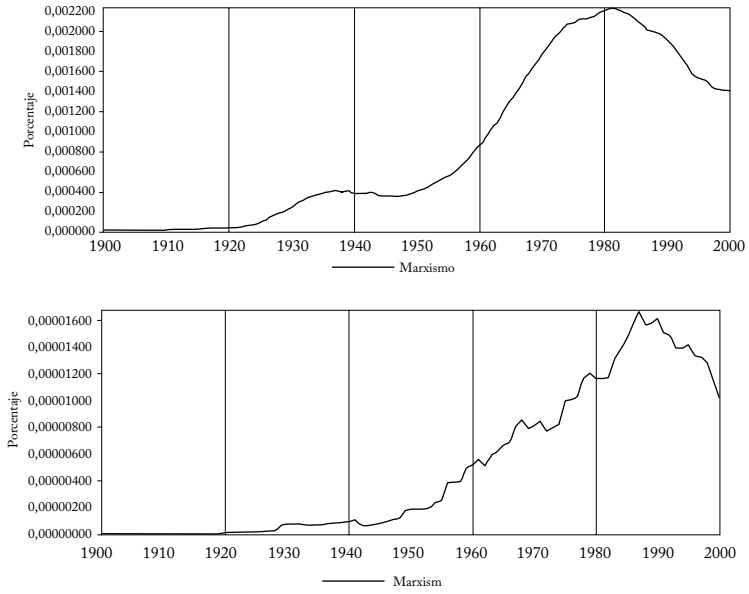
Fuente: books.google.com/ngrams.

La gráfica 2 muestra, para todo el siglo XX, la frecuencia de la palabra *marxismo* en el corpus de libros en español y de la expresión equiva-

[1] Como se explica más adelante, la frecuencia se calcula como el número de apariciones de la palabra en cuestión (*a*) en el universo de textos analizados en un año dado dividido por el número total de palabras en esos textos en el mismo año. La gráfica puede reproducirse fácilmente en http://ngrams.googlelabs.com/.

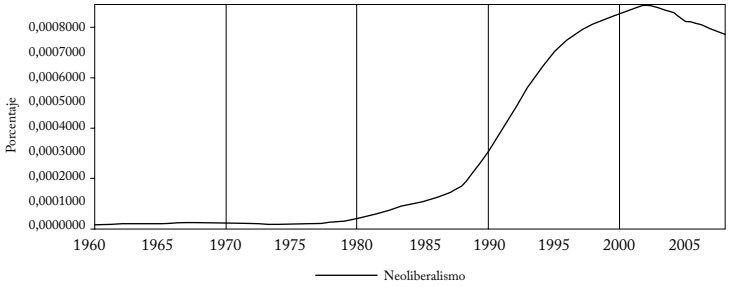
lente, *marxism*, en el corpus de libros en inglés. En ambos idiomas, la frecuencia aumenta casi de manera continua entre 1920 y 1980. En español empieza a disminuir en 1980; en inglés, unos pocos años más tarde. La evolución es similar en ambos casos, pero la frecuencia es mucho mayor en los textos publicados en español. La gráfica 3 repite el análisis para la palabra *neoliberalismo*. El auge comienza lentamente en los años ochenta, toma fuerza en los noventa y empieza a revertirse en el año 2002, coincidiendo paradójicamente con la recuperación de la economía mundial y de las economías latinoamericanas que aplicaron, años atrás, las recetas neoliberales. Finalmente, la gráfica 4 muestra los cambios en la influencia de Francia y Estados Unidos en las letras hispanas. Francia dominó hasta finales del siglo XIX y Estados Unidos empezó a consolidar su dominio en la segunda mitad del siglo XX. Francia es el pasado, Estados Unidos el presente; pero no necesariamente el futuro.

Gráfica 2
Marxismo en el siglo XX: inglés y español



Fuente: books.google.com/ngrams.

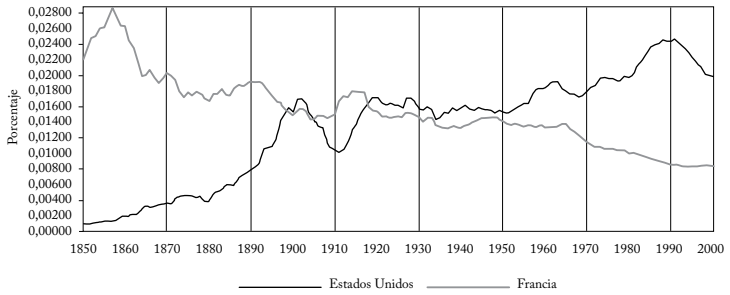
Gráfica 3
Auge del uso del término *neoliberalismo* en los años noventa



Fuente: books.google.com/ngrams.

Culturomics se ha usado recientemente para estudiar el origen y el desarrollo de algunas ideas fundamentales en economía, sociología y ciencia política. Por ejemplo, Ravallion (2011) muestra que a comienzos de los años sesenta las expresiones *pobreza*, *desigualdad* y *crecimiento económico* comenzaron a ganar popularidad. Este cambio refleja un cambio intelectual de fondo, una reconceptualización de la idea de desarrollo económico: el desarrollo empezó a ser visto como un problema tratable y no solo como un reto intelectual sino también como una responsabilidad inaplazable de la comunidad internacional.

Gráfica 4
Influencias foráneas en el mundo hispano: Francia y Estados Unidos



Fuente: books.google.com/ngrams.

Michel et al. (2011) y Ravallion (2011) usan las mismas herramientas, el conteo de palabras en un corpus de cinco millones de libros y 360 mil millones de palabras, para estudiar la importancia atribuida a algunas enfermedades infecciosas. Con las notables excepciones del sida y la hepatitis, las enfermedades infecciosas perdieron figuración desde la segunda mitad del siglo XX. Ravallion anota, además, que

las menciones al sida superan, en todo momento, su impacto sobre la morbilidad y la mortalidad; resultado que ilustra un hecho esencial: el análisis de textos revela aspectos relevantes de la realidad, pero tiene, más que otro tipo de análisis, un sesgo cultural, está sesgado por las creencias, teorías, opiniones y modas de cada momento. *Culturomics* estudia la realidad a través del filtro de la cultura.

Aquí cabe una aclaración: Michel et al. (2011) excluyen intencionalmente cualquier tipo de publicación periódica. Esta decisión no se explica en el artículo, pero no es arbitraria. Los libros ofrecen una perspectiva decantada, más de largo plazo, desconectada de las fluctuaciones bruscas de la opinión pública. Al limitar el análisis a los libros, se reduce el ruido coyuntural pero, por ello mismo, se pierde especificidad: los libros no dan cuenta de la manera como la sociedad responde e interpreta el flujo de información, imperfecto y a veces contradictorio, que se produce a diario. Para explorar fenómenos sociales sensibles a una información que cambia y se adapta en tiempo real conviene utilizar más bien archivos de noticias.

Antes del lanzamiento de la base de datos de libros digitalizados de Google que popularizó el tipo de análisis descrito, Glaeser y Goldin (2002) examinaron la frecuencia de aparición de las palabras *corrupción* y *fraude* en *The New York Times* y un conjunto de diarios regionales para construir un indicador de la trayectoria de la corrupción en Estados Unidos. Su análisis muestra que la frecuencia de esas palabras aumentó durante la primera parte del siglo xix y disminuyó súbitamente después de 1870. En opinión de los autores, esta trayectoria replica la evolución de la corrupción en Estados Unidos a pesar de los sesgos mediáticos ya mencionados[2]. En síntesis, la palabra escrita puede dar cuenta de la trayectoria de algunos fenómenos sociales de difícil medición.

Más recientemente, Leetaru (2011) utilizó un archivo de treinta años de noticias recopiladas por servicios de inteligencia de Estados Unidos e Inglaterra para medir, mediante un análisis automatizado del tono de los artículos (positivo o negativo y en qué grado), la opinión global sobre eventos como la *Primavera Árabe* o la *Guerra en los Balcanes*. El análisis muestra que las crisis políticas suelen ser precedidas por una caída significativa en el tono de los artículos. Con métodos automatizados de geoposición de textos, Leetaru hizo una estima-

[2] Como reconocen Glaeser y Goldin (2002), en este caso el indicador propuesto, basado en las menciones de prensa, tiene un problema adicional: la aparición en la prensa, esto es, el reporte escrito de la corrupción, puede afectar directamente el fenómeno que se trata de medir. La prensa no solo refleja, también puede influir en el fenómeno analizado.

ción aproximada de la localización de Osama Bin Laden antes de su muerte. El autor propone el uso de estas metodologías para predecir eventos de importancia global de manera similar a como Bollen et al. (2011) anticipan movimientos del mercado de valores mediante un análisis de frecuencias del caudal de Twitter. Estos resultados evidencian el potencial descriptivo de *culturomics*.

La dificultad de este tipo de análisis radica entonces en la interpretación; en la necesidad de discernir, para el fenómeno en cuestión, qué tanto corresponde el cambio de frecuencia de las palabras a una faceta real y qué tanto a una distorsión mediática. La distorsión depende, en general, del fenómeno estudiado, del momento histórico y de las publicaciones. Cuando existen cifras objetivas, como ocurre para algunos fenómenos económicos, la comparación de la aparición de las palabras y la realidad del fenómeno da pistas sobre los sesgos culturales y de opinión. Cuando no existen cifras objetivas, ambos aspectos son difíciles de separar; los gráficas dicen tanto de los vaivenes de la realidad como de los ciclos de la cultura y la opinión.

Este artículo utiliza un archivo de noticias para estudiar la realidad, la opinión, la cultura y la economía del país durante los últimos veinte años. El análisis es más sugestivo que definitivo. Plantea muchos interrogantes, revela algunos sesgos y sugiere algunos temas de investigación.

BASE DE DATOS Y CÁLCULO DE FRECUENCIAS

El corpus de noticias utilizado incluye todos los artículos publicados en las versiones electrónicas del periódico *El Tiempo* y de las revistas *Semana* y *Dinero*. Los archivos de noticias tienen una cobertura temporal distinta. El archivo de *El Tiempo* comienza en 1991, el de *Semana* en 1980 y el de *Dinero* en 1993. Los tres archivos se extienden hasta el 31 de julio del 2011, fecha de corte del análisis. Las tres publicaciones mencionadas tienen los archivos electrónicos de noticias más antiguos y completos de los medios impresos del país. Al menos en términos de contenido, estas publicaciones pueden considerarse representativas de la prensa escrita de circulación nacional.

El método para construir los archivos de noticias es simple. Primero, un programa recorre los sitios web de las publicaciones seleccionadas y descarga la totalidad de los artículos. Luego elimina los elementos adicionales (barras de navegación, enlaces, imágenes, anuncios publicitarios, etc.) y almacena una versión simplificada de cada artículo. Finalmente, descarta los artículos (más de 100.000)

con el mismo título y el mismo contenido. El análisis final se basa en un archivo depurado que contiene una sola copia de los artículos.

El archivo analizado contiene casi dos millones de artículos. En números redondos, el 90% proviene de *El Tiempo*, el 6,5% de *Semana* y el 3,5% de *Dinero*. La gráfica 5 muestra que el número de artículos varía de manera sustancial de un año al siguiente. El número de artículos de *Semana* y *Dinero* aumentó de manera considerable después de 2005, como consecuencia de la introducción de blogs y de artículos informativos que no hacen parte de las ediciones impresas. El número de artículos de *El Tiempo* no cambió grandemente entre 1993 y 2010, con la excepción de un bache (inexplicado) en los años 2004 y 2005. En los años ochenta, la muestra sólo contiene unos pocos artículos de *Semana*: menos de 2.000 anuales en promedio.

Una vez depurado el archivo de noticias, se identificaron el título, la fecha de publicación y el texto completo de cada uno de los artículos y se almacenaron en registros separados. Las letras mayúsculas se convirtieron a minúsculas y la totalidad del texto se dividió en *n-gramas*. Un *n-grama* es una secuencia de *n* palabras consecutivas dentro de un texto determinado. Así, por ejemplo, la división de un texto en *1-gramas* arroja un listado de todas las cadenas de caracteres separadas por espacios o signos de puntuación, incluidas las palabras (*partido* o *Colombia*), los números (1984 y 8.000) y otras expresiones (como *M-19* o *F1*). La división del mismo texto en *2-gramas* arroja secuencias tales como *derechos humanos* o *nueva constitución*. La división en *3-gramas* muestra secuencias tales como *5 a 0* o *Valle del Cauca*.

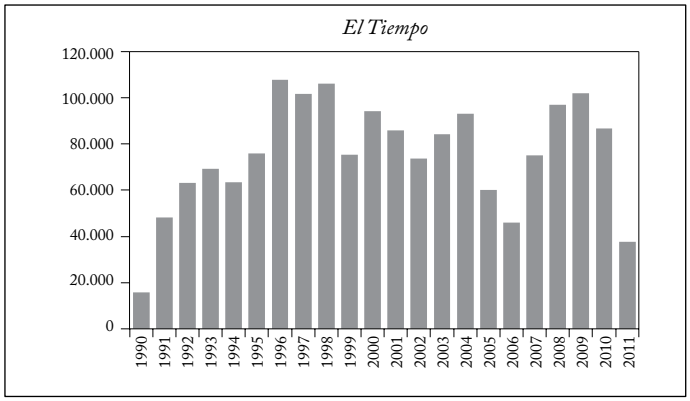
La frecuencia de aparición se calcula como el cociente entre el número de apariciones de un *n-grama* en todos los artículos publicados durante un mes dado en una de las tres publicaciones analizadas y el número total de *1-gramas* publicados durante el mismo mes en la misma publicación. Los *n-gramas* que aparecen menos de diez veces en un mes se excluyeron del análisis.

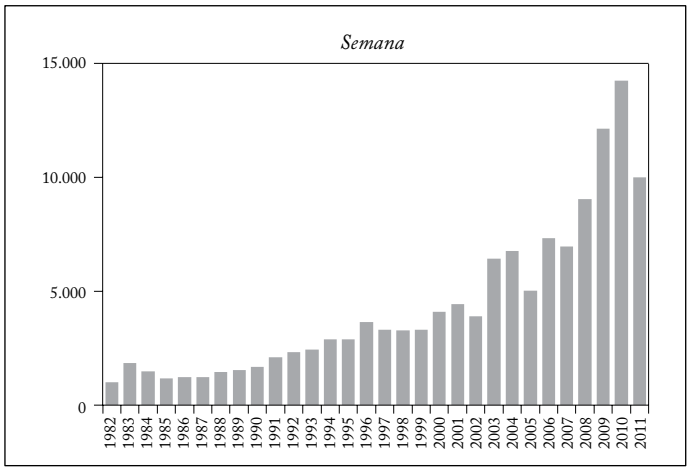
El corpus contiene más de 600 millones de 1-*gramas*. La distribución por publicación de los *1-gramas* es similar, pero no idéntica, a la distribución de los artículos: el 86,5% corresponde a *El Tiempo*, el 9,5 a *Semana* y el 4,0% restante a *Dinero*. La participación de *Semana* y *Dinero* es mayor en la distribución de *1-gramas* que en la de artículos, habida cuenta de la mayor longitud de los artículos publicados en estos medios de circulación semanal o quincenal con respecto a los publicados en *El Tiempo*, de circulación diaria.

El análisis de las secciones siguientes se limita al periodo 1992-2011. Antes de 1992, el corpus incluye apenas unos pocos artículos,

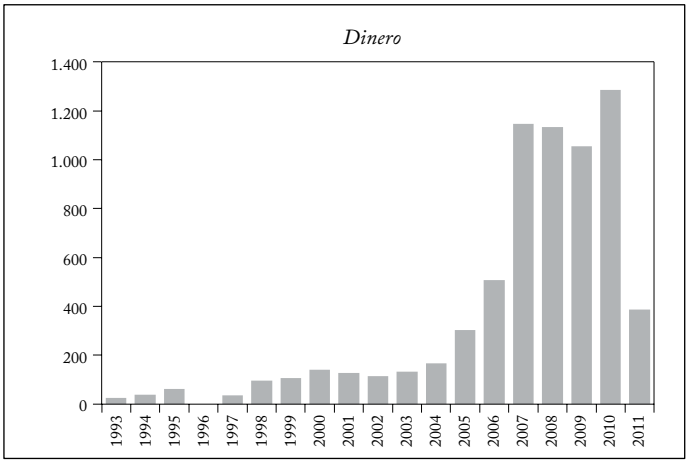
provenientes en su gran mayoría de *Semana*. En cambio, en el periodo estudiado el número de artículos permite analizar los cambios en las frecuencias de palabras de escasa aparición (*bonanza*, *desempleo*, *sequía*, etc.). El periodo de análisis coincide con los primeros veinte años de la Constitución Política de Colombia. Aunque el periodo se escogió por razones pragmáticas, asociadas a la disponibilidad de información, tiene también un sentido o significado histórico.

Gráfica 5
Cantidad de artículos por publicación





Gráfica 5
Cantidad de artículos por publicación (continuación)



Como se dijo antes, las menciones a las palabras de interés (*corrupción*, p. ej.) están siempre normalizadas por el número total de palabras o *1-gramas* en la totalidad del archivo de noticias. En principio, el aumento del número de artículos, como resultado, por ejemplo, de la inauguración de contenidos virtuales, no es un problema: la frecuencia de la palabra *corrupción*, para usar el mismo ejemplo, no tiene por qué aumentar si aumenta el número de artículos o el volumen de información.

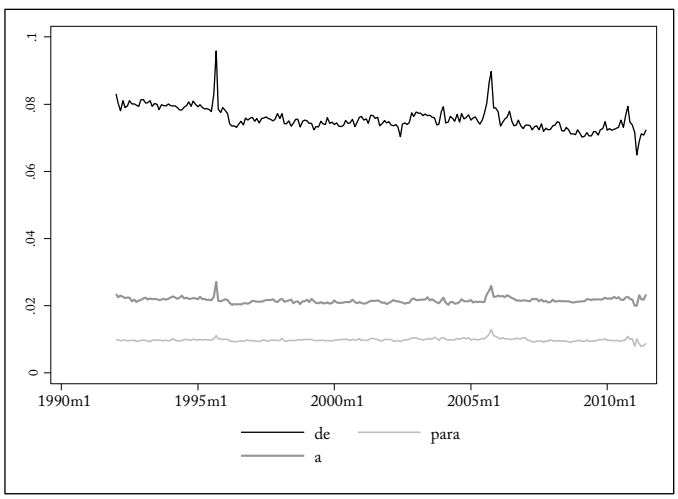
La gráfica 6 muestra la frecuencia de las preposiciones más comunes del idioma español: *a*, *de* y *para*. Esta frecuencia no debería cambiar de un año a otro a pesar del aumento del número de artículos publicados; por tanto, la existencia de cambios abruptos o de tendencias bien definidas podría indicar la presencia de sesgos o problemas en los archivos. La gráfica muestra que los cambios son marginales, tal vez asociados a distorsiones aleatorias y a algunos baches en el archivo de noticias (en septiembre de 1995, septiembre de 2005 y octubre de 2010). Este resultado descarta, en principio, la presencia de grandes errores de construcción o programación.

Si el contenido de una publicación cambia en forma sustancial, por ejemplo, si se concentra en la información internacional o deportiva, los cambios en la frecuencia darían, en teoría, una idea equivocada de la trayectoria de ciertos fenómenos: una reducción de la frecuencia de la palabra *corrupción* obedecería no tanto a una disminución del fenómeno, como a un descenso de la cobertura mediática, asociado,

a su vez, a los cambios en el contenido de la publicación. En general, el sesgo derivado de los cambios en el contenido de las publicaciones periódicas se puede atenuar cambiando la normalización de las series. La mención de las palabras de interés puede dividirse ya no por el número total de *1-gramas*, sino por el número total de apariciones de algunas palabras genéricas que captan indirectamente las posibles variaciones de importancia o contenido. La aparición de *corrupción* podría dividirse por las apariciones de *política*, *gobierno* o *presidente*, palabras que reflejan, en términos generales, la importancia del cubrimiento local dentro del contenido general del periódico en cada momento. En general, los cambios en la normalización no afectaron los resultados de manera significativa.

Por último, el análisis siguiente depende en buena medida de la comparación de series de tiempo. La comparación se basa en la inspección visual y en el simple cálculo de correlaciones. Una comparación más sofisticada podría usar, por ejemplo, los análisis de supervivencia comunes en biología (Jones y Crowley, 1989) o los indicadores de bondad de ajuste basados en conceptos de entropía (Cowell et al., 2011). La sofisticación metodológica, sin embargo, no necesariamente resulta más informativa.

Gráfica 6
Frecuencia de algunas preposiciones



ALGUNOS EJEMPLOS: REALIDADES Y PALABRAS

Esta sección presenta cinco ejemplos de fenómenos distintos que comparten una misma característica: todos se miden de manera sistemática mediante indicadores conocidos y probados. El análisis propuesto compara, en todos los casos, la evolución de dos series: el indicador del fenómeno (desempleo, p. ej.) y la frecuencia de la palabra correspondiente (*desempleo*). La comparación entre indicadores y frecuencias revela, por una parte, la pertinencia del método y, por otra, la magnitud y dirección de algunos sesgos mediáticos. En síntesis, el análisis permite entender de qué manera la prensa escrita refleja (y al mismo tiempo distorsiona) la realidad[3].

Desempleo

La gráfica 7 muestra la tasa trimestral de desempleo de las siete principales ciudades del país y la frecuencia de la palabra *desempleo* en el archivo de noticias de *El Tiempo* del mismo trimestre. Las series abarcan el periodo comprendido entre el primer trimestre de 1993 y el segundo de 2011. Ambas series se filtraron con base en un promedio móvil de un año (cuatro trimestres). La tasa de desempleo se limita a las siete principales ciudades para asegurar la comparabilidad de los datos a lo largo del periodo. El análisis no cambia si el archivo de *El Tiempo* se complementa con archivos de *Semana* y *Dinero*: el grueso de las noticias sobre el tema en cuestión proviene de *El Tiempo*.

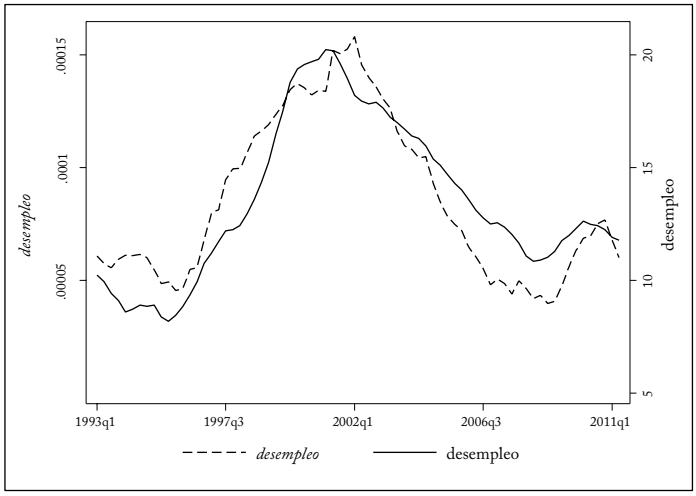
La correlación entre ambas series es evidente. El coeficiente de correlación es de 0,90 en todo el periodo. La tasa de desempleo y la frecuencia de la palabra *desempleo* crecieron a un ritmo similar durante la crisis de finales de los noventa, pero el descenso de ambas series fue distinto. La frecuencia (una medida del interés mediático) descendió más rápidamente que la tasa de desempleo (una medida objetiva de la desocupación). La inercia de la realidad fue aparentemente mayor que la inercia del interés mediático. Dicho de otra manera, el fenómeno del desempleo fue más duradero que las noticias y comentarios al respecto.

En 2008, coincidiendo con la crisis internacional y el aumento del desempleo interno, el interés mediático revivió. La frecuencia de la palabra *desempleo* aumentó de manera desproporcionada entre 2009 y 2010. Podría decirse que la prensa reaccionó de más ante el repunte del desempleo. En términos más generales, la reducción injustificada del interés mediático después de la crisis de los noventa y el aumento

[3] Una versión del buscador está disponible en http://ngrams.cavorite.com.

desproporcionado después de la crisis internacional de 2008 sugieren que los medios escritos son más sensibles al agravamiento de un problema social que a su persistencia. La prensa escrita perdió interés en un problema duradero y acuciante. Los periodistas y comentaristas volvieron a ocuparse del tema cuando la situación empeoró.

Gráfica 7
Desempleo y *desempleo* en *El Tiempo*



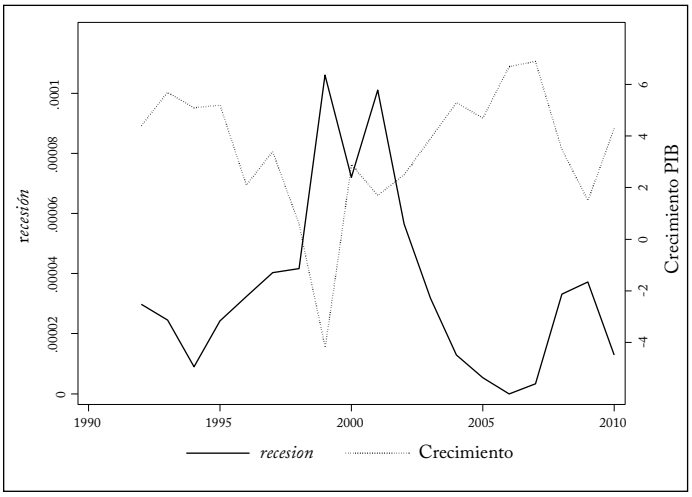
En suma, se destacan dos hechos: la alta correlación entre las dos series y la respuesta asimétrica de la prensa escrita: el olvido relativo de un problema persistente y la reacción abrupta ante su empeoramiento.

Recesión

La gráfica 8 muestra la tasa anual de crecimiento económico y la frecuencia de la palabra *recesión* en *El Tiempo*. Las series cubren el periodo 1992-2010. La frecuencia corresponde al promedio móvil de doce meses. La serie de crecimiento corresponde, por su parte, a la tasa anual de crecimiento del PIB.

Las conclusiones de este ejemplo son similares a las del ejemplo anterior. Como en el caso del desempleo, la correlación entre las dos series es evidente. La frecuencia de *recesión* aumentó cuando cayó la tasa de crecimiento y viceversa. La correlación es de -0,82. Solo hay una discordancia notable: la leve desaceleración económica de 2002 estuvo acompañada de un aumento desproporcionado de la frecuencia. Pero, en general, el comportamiento de ambas series es similar.

Gráfica 8
Crecimiento del PIB y *recesión* en *El Tiempo*



El anexo muestra una gráfica complementaria que relaciona la frecuencia de la palabra *recesión*, ya no con la tasa de crecimiento económico, sino con una variable idéntica pero inversa: *10 – tasa de crecimiento anual*. Esta gráfica permite observar más claramente el movimiento conjunto de ambas series. La caída de la frecuencia coincide con la recuperación de la economía. A diferencia del ejemplo anterior, en el cual el desempleo cayó más lentamente que la frecuencia de la palabra que lo designa, en este ejemplo la frecuencia de *recesión* cayó a un ritmo similar al de la recuperación económica. Mientras que el conteo de palabras no captó plenamente la trayectoria asimétrica de la tasa de desempleo (empeoramiento súbito y mejoramiento lento), sí parece captar la trayectoria más simétrica de la tasa de crecimiento económico (empeoramiento y mejoramiento de duraciones semejantes).

Revaluación

La gráfica 9 presenta el índice de tasa de cambio real calculado por el banco JP Morgan[4]. La gráfica permite apreciar las fluctuaciones de la tasa de cambio real: el peso se depreció durante la crisis de los noventa, se apreció años durante la recuperación económica, se volvió a depreciar durante la crisis financiera de 2008 y se apreció en los últimos dos años de crecimiento acelerado. Los ciclos fueron

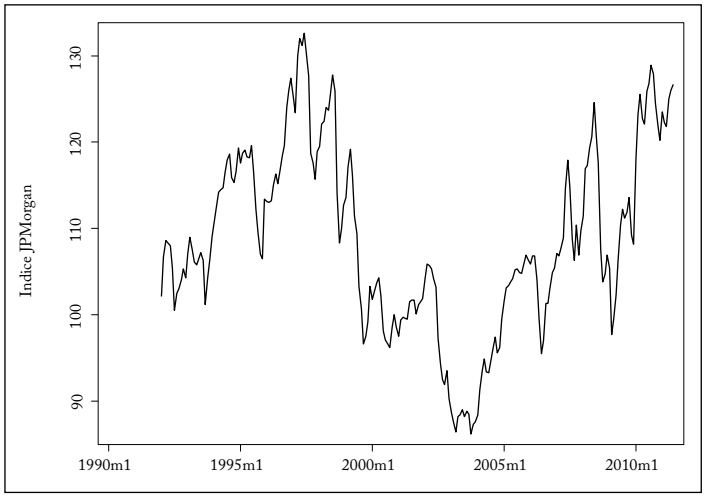
4 A diferencia del índice calculado por el Banco de la República (ITCR), este índice aumenta cuando el peso se valoriza (o el dólar se desvaloriza) y disminuye en caso contrario. Las conclusiones del análisis no cambian si se usa el ITCR.

pronunciados, no muy distintos a los experimentados por otros países latinoamericanos. El comportamiento del real brasileño, por ejemplo, fue muy similar.

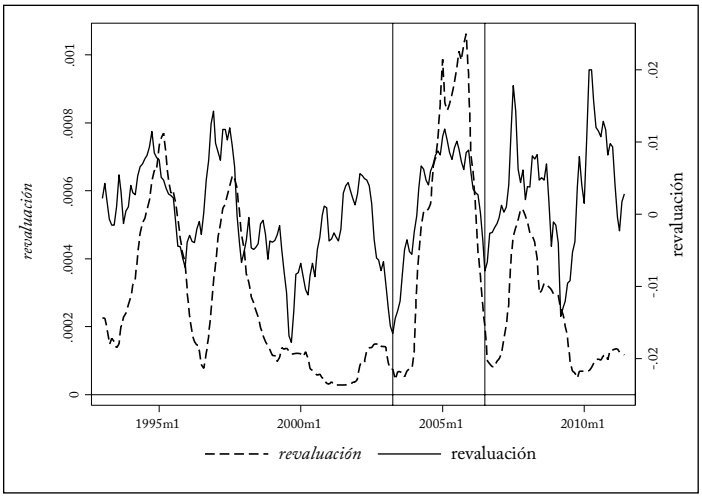
La gráfica 10 muestra el cambio porcentual mes a mes del ITCR y la frecuencia de la palabra *revaluación* en el archivo de *El Tiempo*. El periodo de análisis va de 1992 hasta 2011. Los datos corresponden a los promedios móviles anuales (12 meses). La gráfica distingue tres periodos: 1993(1)-2003(4), 2003(5)-2006(7) y 2006(7)-2011(6). El co-movimiento de las series analizadas, la revaluación real y la frecuencia de la palabra *revaluación* fueron diferentes en cada uno de los periodos señalados.

El coeficiente de correlación entre la revaluación y *revaluación* es alto, cercano a 0,5 entre 1993 y 2011. Pero el valor del coeficiente cambió de manera sustancial a lo largo del periodo: 0,50 entre 1993 y 2003, 0,90 entre 2003 y 2006, y 0,11 entre 2001 y 2007. En el primer periodo (1993-2003) hubo tres eventos de revaluación. Los dos primeros, ambos anteriores a la crisis de finales de los noventa, estuvieron acompañados de un aumento moderado de la frecuencia. Por el contrario, el último evento posterior a la crisis y de menor magnitud no suscitó un cambio sustancial en la frecuencia; no mereció mayor atención de la prensa.

Gráfica 9
Índice de tasa de cambio real (ITCR) de JP Morgan



Gráfica 10
Revaluación y *revaluación* en *El Tiempo*



En el segundo periodo (2003-2007), la revaluación generó mucho mayor interés mediático: la frecuencia aumentó de manera sustancial, mucho más rápidamente que la revaluación. El mayor interés mediático pudo haber sido impulsado por la respuesta oficial a la presión de los exportadores e industriales. Entre 2003 y 2007, el Gobierno del entonces presidente Uribe estableció una serie de subsidios a los exportadores y trató fallidamente de fijar un piso a la tasa de cambio en diciembre de 2004[5]. Aparentemente el activismo oficial se tradujo en más noticias y comentarios sobre la revaluación. Sea como fuere, el interés mediático por la revaluación creció notablemente en este periodo.

Pero el interés mediático en la revaluación parece haberse desvanecido en los últimos años. En el tercer periodo (2007-2011) ha pasado casi desapercibida. La frecuencia apenas aumentó a pesar del fuerte aumento del ITCR. Como conjetura, se podría decir que hay menos noticias porque el Gobierno ha hecho menos, pero también que el Gobierno ha hecho menos porque hay menos noticias. La prensa alimenta las preocupaciones del Gobierno y viceversa. En teoría, este

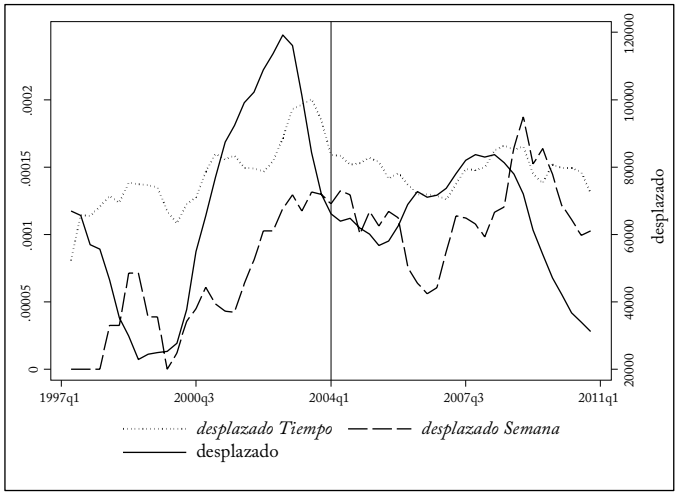
[5] La presión del Ejecutivo sobre el Banco de la República fue entonces un secreto a voces. Ver, por ejemplo: "Los ex codirectores del Banco, Carlos Caballero Argáez, Salomón Kalmanovitz y Sergio Clavijo, así como el decano de la Facultad de Economía de la Universidad de Los Andes y ex director de Planeación Nacional, Juan Carlos Echeverry cuestionaron públicamente las presiones que ejerció el Ejecutivo ese 20 de diciembre de 2004, insinuando que decretaría la emergencia económica e impondría un control de cambios en Colombia" [http://www.primerapagina.com.co/MostrarDocumentoPublico.aspx?id=1113575].

tipo de retroalimentación positiva abre la posibilidad de equilibrios múltiples: unos de obsesión mediática y otros de desatención o indiferencia. El hecho cierto es que el mismo fenómeno primero fue cubierto obsesivamente y después casi olvidado por completo.

Desplazados

La gráfica 11 estudia el cubrimiento del desplazamiento forzado. La gráfica muestra el flujo mensual de desplazados según los registros oficiales de Acción Social y la frecuencia de la palabra *desplazado* (y sus variantes[6]) tanto en *El Tiempo* como en *Semana*. Las series fueron suavizadas con base en promedios móviles de doce meses. El periodo de análisis va de enero de 1997 a junio de 2011. Antes de 1997, el número de desplazados era insignificante según los registros oficiales.

Gráfica 11
Desplazados y *desplazados* en *El Tiempo* y *Semana*



El análisis se divide en dos periodos: antes y después de enero de 2004, esto es, antes y después de la sentencia de la Corte Constitucional que ordenó al Gobierno, entre otras cosas, dar prioridad a la atención de emergencia a los desplazados y garantizar su acceso a los servicios sociales básicos. Entre 1997 y 2004, el flujo de desplazados aumentó sustancialmente, de 20.000 a comienzos del periodo a 120.000 en los años intermedios y a 60.000 a finales de 2003. Este aumento estuvo acompañado de un incremento de la aparición de la palabra

[6] Las palabras afines fueron *desplazado*, *desplazada*, *desplazados* y *desplazamiento*. El análisis suma la aparición de cada una de esas palabras.

desplazado y sus variantes. Aparentemente la prensa escrita reaccionó al aumento del flujo de personas desplazadas por la violencia. La reacción fue menor, si se quiere, que en los ejemplos anteriores, pero fue notable en todo caso.

Después de 2004, el co-movimiento de las series es menos evidente. Las fluctuaciones de la frecuencia poco tuvieron que ver con las fluctuaciones del flujo de desplazados. Además, la caída del flujo posterior a 2007 no estuvo acompañada de una caída consustancial de la frecuencia. Quizá porque la cantidad de desplazados siguió creciendo a pesar de la reducción del flujo o porque la sentencia de la Corte Constitucional fue una fuente ocasional de noticias, en parte por las polémicas constantes entre el Gobierno y la Corte.

Antes de 2004, el coeficiente de correlación entre el flujo de desplazados y la frecuencia de la palabra *desplazado* (y sus variantes) fue de 0,60 en *El Tiempo* y en *Semana*. Después de 2004, el coeficiente de correlación cayó a 0,15 en *El Tiempo* y a -0.08 en *Semana*. El coeficiente de correlación de las dos series de frecuencias de *desplazado* fue de 0,69 en todo el periodo, de 0,88 en el periodo inicial (antes de 2004) y de 0,64 en el periodo final (después de 2004). En general, *El Tiempo* y *Semana* le dieron un cubrimiento similar: el promedio de las frecuencias fue semejante y los patrones temporales fueron también parecidos.

En suma, el mayor flujo de desplazados sí suscitó el interés de la prensa. Creció rápidamente con el aumento de los flujos y no cayó con su disminución. Aparentemente, el cubrimiento respondió tanto a los flujos como a la cantidad total de desplazados.

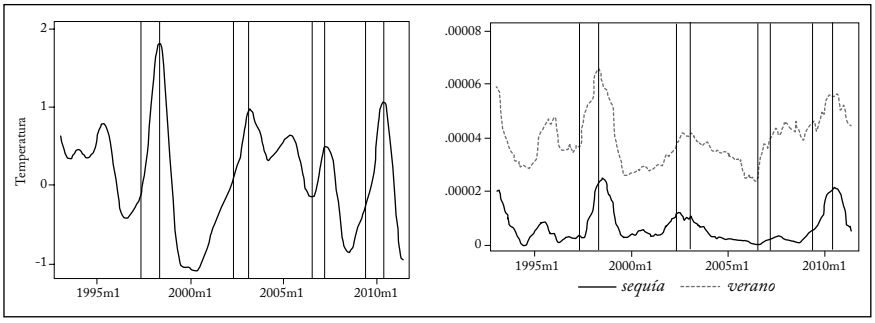
Clima: *"El Niño"* y *"La Niña"*

Las noticias de prensa captan la dinámica del clima y se pueden usar como una medida indirecta del impacto de algunos fenómenos climáticos globales. La gráfica 12 presenta, a la izquierda, la temperatura del Pacífico ecuatorial en la llamada zona 3.4[7] y, a la derecha, la frecuencia de las palabras *sequía* y *verano* en los archivos de *El Tiempo*. Los datos corresponden a los promedios móviles de doce meses. Ambas figuras muestran, mediante líneas verticales, las últimas cuatro apariciones del fenómeno de El Niño, un aumento atípico de la temperatura del océano Pacífico[8].

[7] La Zona 3.4 está ubicada entre la latitud 5º N y 5º S y entre los meridianos 170 y 120. La temperatura de esta zona se usa para monitorear la presencia de "El Niño" y "La Niña".

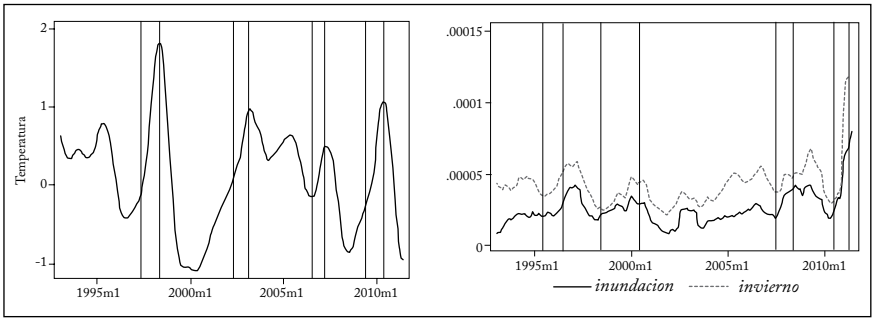
[8] Ver, por ejemplo, http://es.wikipedia.org/wiki/El_Niño.

Gráfica 12
El "Niño", *sequía* y *verano* en *El Tiempo*



El aumento de la temperatura del Pacífico afecta los patrones de lluvia y genera periodos prolongados de sequía que afectan las cosechas, el volumen de los embalses, etc. La gráfica muestra que la frecuencia de aparición de las palabras *sequía* y *verano* aumentó de manera notable durante los periodos en que ocurrió este fenómeno climático. El aumento fue especialmente notorio en dos momentos: a finales de los noventa y al final del periodo de análisis, en el año 2010.

Gráfica 13
La Niña, *inundación* e *invierno* en *El Tiempo*



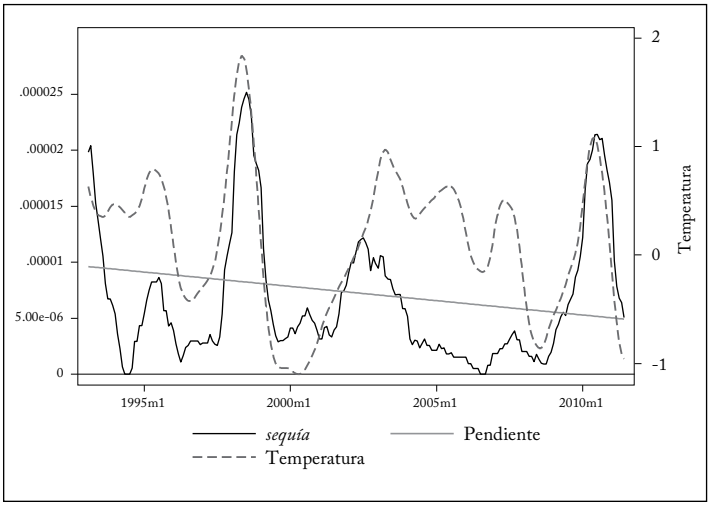
La gráfica 13 repite el análisis para la fase de enfriamiento de la temperatura, conocida como el fenómeno de "La Niña"[9]. Los resultados son similares. La figura de la izquierda ilustra los periodos de caída de la temperatura y la de la derecha, la frecuencia de las palabras *inundación* e *invierno*. De nuevo, los aumentos de la frecuencia de las palabras en cuestión coincidieron con la llegada de "La Niña". El aumento al final del periodo es particularmente notable, refleja el mucho mayor

[9] Ver http://es.wikipedia.org/wiki/La_Niña_(clima).

impacto del último evento de "La Niña", en la primera mitad del año 2011. La gráfica sugiere la existencia de un evento extremo en comparación con los eventos precedentes.

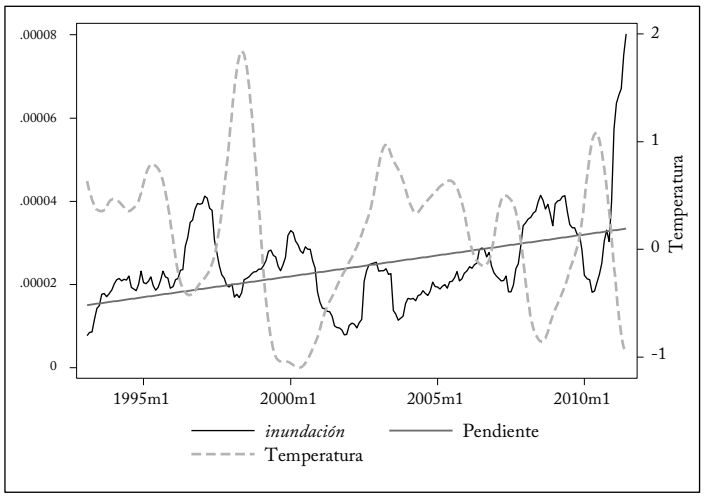
La gráfica 14 muestra los co-movimientos de la temperatura del Pacífico ecuatorial y la frecuencia de la palabra *sequía*. La relación es evidente, sorprendente incluso. El coeficiente de correlación es de 0,61 para todo el periodo. Si la frecuencia de *sequía* da una idea indirecta del impacto de las distorsiones climáticas –a mayor impacto, más noticias–, la gráfica 6 sugiere que el impacto de "El Niño" no ha aumentado en las últimas dos décadas. Todo lo contrario. La pendiente de la línea es negativa y significativamente diferente de cero. El evento de 2010 tuvo un impacto considerable, pero no implica, por sí solo, un agravamiento de los efectos económicos, sociales y ambientales de las sequías.

Gráfica 14
Temperatura y *sequía* en *El Tiempo*



La gráfica 15 repite el análisis para la frecuencia de la palabra *inundación*. El coeficiente de correlación es de nuevo alto: -0,54 para todo el periodo. Las conclusiones son en este caso opuestas a las del caso anterior. La pendiente de la línea de regresión es positiva y estadísticamente significativa. Los datos parecen consistentes con la idea de un agravamiento gradual, no espectacular pero sí notable. Esta conclusión depende, sin embargo, del evento extremo de 2011 y no debería considerase definitiva.

Gráfica 15
Temperatura e *inundación* en *El Tiempo*



En suma, la frecuencia de las palabras mencionadas da una idea del impacto general de los dos eventos climáticos en las dos últimas décadas. El análisis no es concluyente. La evidencia no sugiere un empeoramiento de las sequías, pero sí de las inundaciones. Sin embargo, los resultados dependen de los eventos extremos de finales del periodo. Sea como sea, la frecuencia de las noticias es una forma de medir, al menos preliminarmente, el impacto de los eventos climáticos.

20 AÑOS, CINCO HISTORIAS: REALIDADES COMO PALABRAS

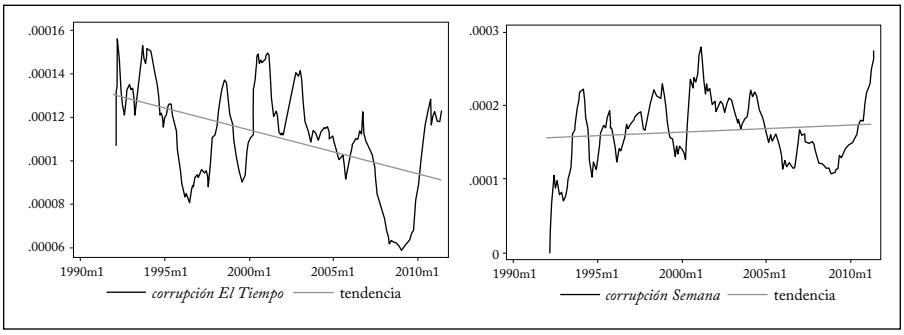
Esta sección presenta cinco ejemplos, cinco estudios de caso que ilustran el uso de *culturomics* como método de cuantificación de fenómenos que, por su misma naturaleza, son difíciles de cuantificar. Los ejemplos se refieren a temas centrales de la realidad colombiana de los últimos veinte años: la corrupción, la guerra, el optimismo económico y el equilibrio de poderes (los congresistas frente a los jueces y el presidente frente a los mandatarios locales). Dadas las dificultades obvias de medición, estos temas no han sido cuantificados de manera sistemática. Ninguno de ellos cuenta con indicadores conocidos y respetados. Las comparaciones entre indicadores y frecuencias no son por tanto posibles. Las frecuencias son, en este caso, los indicadores: la forma imperfecta de cuantificar los cambios y las tendencias de la corrupción, el conflicto, el entusiasmo y la distribución del poder.

Corrupción

La gráfica 16 muestra, para el periodo enero de 1992-julio de 2011, la evolución de la frecuencia de la palabra *corrupción* y sus accidentes[10]. Las series corresponden al promedio móvil de doce meses: inicialmente se calcularon las frecuencias mensuales y luego los promedios móviles anuales. La gráfica presenta por separado las frecuencias correspondientes a *El Tiempo* y *Semana*. En *Dinero*, el número de noticias es relativamente menor y la aparición de la palabra *corrupción* es muy escasa.

Ambas gráficas cuentan una historia similar. Revelan, por ejemplo, grandes fluctuaciones alrededor de una tendencia más o menos horizontal. En *El Tiempo* (gráfica izquierda), la tendencia es negativa; en *Semana* (gráfica derecha), es positiva. Pero más allá de estas diferencias, la gráfica sugiere, en esencia, una considerable inercia de la corrupción: los escándalos ocurren cada cierto tiempo pero no parece existir una tendencia clara. En suma, la corrupción fluctúa en el corto plazo pero es constante desde una perspectiva de más largo plazo[11]. Todo cambia y todo sigue igual.

Gráfica 16
Corrupción en *El Tiempo* y *Semana*



[10] El análisis muestra la frecuencia conjunta de las palabras *corrupción*, *corrupta*, *corruptas*, *corrupto* y *corruptos*. Las conclusiones no cambian si se incluyen otras palabras relacionadas como *desfalco*, *peculado*, *robo al erario*, etc.

[11] En cada momento, abrumados por los eventos de la coyuntura, los comentaristas políticos tienden a percibir la corrupción actual como la peor en mucho tiempo, en otras palabras, tienden a confundir las fluctuaciones con la pendiente. En diciembre de 1997, un reconocido periodista escribió: "Nunca antes el país había presenciado tan impresionante sucesión de hechos escandalosos. Trátese de peculados o desfalcos en entidades del Estado, de narcomicos en el Congreso, de testaferratos o de simple venalidad administrativa, el panorama de la corrupción en Colombia es francamente desolador". En octubre de 2011, otro periodista escribió: "Lo que se robó en Colombia en los últimos años no tiene antecedentes y no es que fuéramos el paraíso anti-corrupción". El presentismo domina las opiniones sobre la corrupción.

El conteo de noticias, opiniones y comentarios no es un indicador perfecto de la corrupción. Este indicador está sesgado por los eventos más costosos o por algunos casos que concentran, por razones muchas veces fortuitas, la atención de la opinión pública. En coyunturas específicas, el indicador recoge los sesgos ideológicos o los intereses políticos de los directores y editores de los medios de comunicación. En fin, los cuestionamientos abundan. Pero este tipo de análisis no debería descartarse fácilmente. En cierta medida, equivale a un ejercicio memorístico –contar para recordar–, a una forma de contrarrestar los juicios impresionistas del presente con los juicios del pasado, de comparar la indignación de hoy con la de ayer.

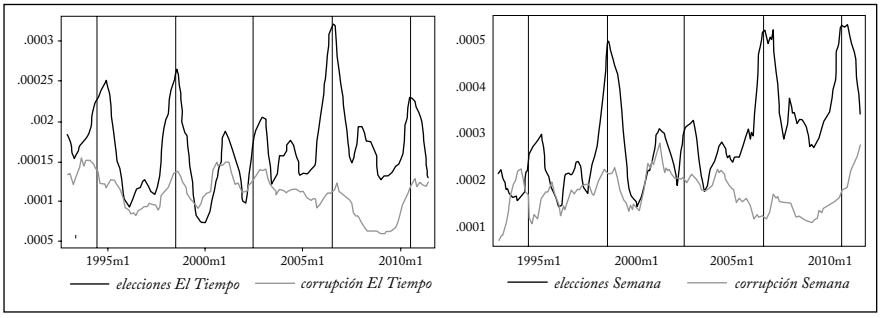
Como se dijo en la sección 1, Goldin y Glaeser (2001) usaron un indicador similar para estudiar la evolución de la corrupción en Estados Unidos en un horizonte de largo plazo. Más recientemente, Goel, Nelson y Naretta (2011) usaron la frecuencia de búsqueda de la palabra *corrupción* en Internet para hacer comparaciones entre países. Los indicadores tradicionales de corrupción se basan en opiniones, en las cuales influye –en la mayoría de los casos– el cubrimiento de la prensa. Los indicadores que aquí se proponen se basan en la intensidad del cubrimiento, en la idea de que la cambiante realidad de un fenómeno complejo puede cuantificarse, en cierta medida al menos, con base en su cubrimiento mediático.

Volviendo a la gráfica 16, hay un hecho peculiar que merece un comentario aparte. En ambas figuras, tanto en la de *El Tiempo* como en la de *Semana*, la frecuencia de la palabra *corrupción* cayó en forma notable entre finales de 2005 e inicios de 2010, y luego aumentó también de manera notable. Al parecer, los medios analizados se desentendieron de la corrupción durante buena parte del segundo periodo del ex presidente Uribe (2006-2010) y luego, como si tuvieran que ponerse al día, volvieron a preocuparse por el tema con intensidad renovada. Después de una calma de varios años, vino la tempestad mediática de los meses recientes.

Las razones de este comportamiento no son fáciles de precisar. Pero la gráfica 17 da algunas pistas. La gráfica muestra conjuntamente la frecuencia de las palabras *corrupción* y *elecciones,* y señala las fechas de elecciones presidenciales durante el periodo. La frecuencia de *corrupción* aumentó cíclicamente en los meses anteriores y posteriores a las elecciones presidenciales: subió y cayó coordinadamente con la frecuencia de *elecciones*. Esta regularidad mediática tuvo una excep-

ción notable: las elecciones de 2006, las únicas elecciones de todo el periodo en las que el presidente en ejercicio fue candidato[12].

Gráfica 17
C*orrupción* y *elecciones en El Tiempo* y *Semana*



Durante los meses que precedieron y sucedieron a las elecciones de 2006, las noticias, comentarios y opiniones sobre la corrupción (una medida indirecta de la intensificación de las denuncias y los debates al respecto) no aumentaron de manera sustancial como lo habían hecho en el pasado durante periodos similares. Pero en las elecciones de 2010, ya con el presidente en ejercicio por fuera de la contienda, todo pareció volver a la normalidad: la "corrupción" creció sustancialmente antes y después de las elecciones. En apariencia, las denuncias y debates que se habían postergado salieron a flote súbitamente. En suma, más que un aumento permanente de la corrupción, el crecimiento súbito de la frecuencia noticiosa al final del periodo de análisis podría indicar una suerte de actualización, de desfogue.

Más allá de los ciclos y las fluctuaciones temporales, los datos sugieren que la corrupción permaneció más o menos constante durante los últimos veinte años. Al menos, la frecuencia de *corrupción* no muestra una tendencia clara, ni positiva ni negativa. Las variaciones fueron muchas, pero la tendencia no cambió notablemente.

Conflicto

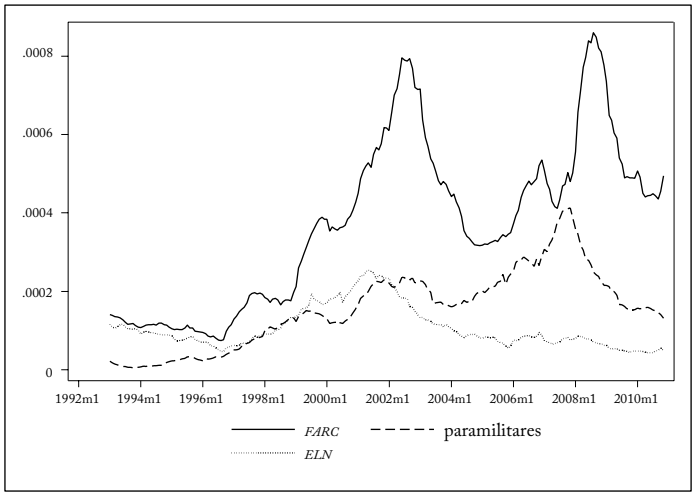
El conflicto colombiano concentró la atención de los medios de comunicación durante los últimos veinte años. El fortalecimiento de los grupos armados durante la primera mitad de los años noventa, las posteriores negociaciones con las FARC, la subsecuente ofensiva mili-

[12] En *El Tiempo* el coeficiente de correlación entre *corrupción* y *elecciones* fue de 0,60 antes de 2002 y de 0,35 después. En *Semana* fue de 0,36 y de -0.07 respetivamente.

tar, los acuerdos con los paramilitares y los rescates de los secuestrados produjeron muchas noticias, comentarios y editoriales de prensa. En teoría, al menos, la frecuencia de aparición de las palabras *FARC*, *ELN* y *paramilitares* ilustra la manera como los medios de comunicación dieron cuenta de la cambiante realidad del conflicto colombiano. Las palabras permiten, en suma, tomarle el pulso a la obsesión mediática con el conflicto.

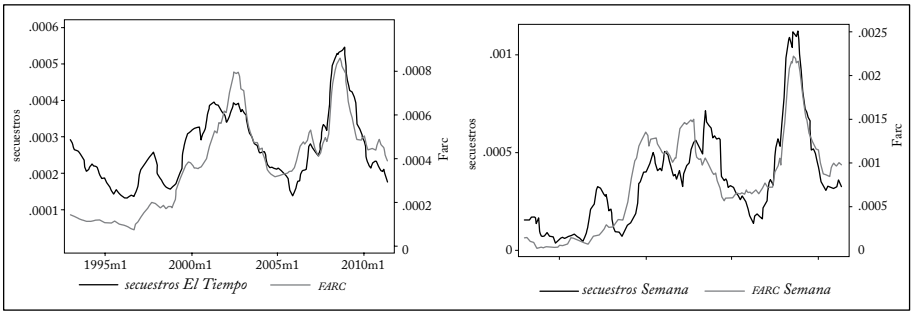
La gráfica 18 muestra la frecuencia mensual de las palabras *FARC*, *ELN* y *paramilitares* en el periodo 1992-2011[13]. El análisis corresponde en este caso a los archivos del diario *El Tiempo* (el análisis conjunto de los tres medios disponibles es casi idéntico). La frecuencia de la sigla *FARC* supera ampliamente, en más de diez veces, la de palabras como *desempleo* y *corrupción*. Supera incluso la de expresiones genéricas como *Congreso* y *elecciones*. La frecuencia de las palabras *ELN* y *paramilitares* es relativamente menor, pero no insignificante. En general, la importancia mediática del conflicto fue enorme. Las FARC tuvieron dos momentos de ebullición mediática (en 2002 y 2009), los paramilitares uno (en 2008) y el ELN otro (en 2001). La desaparición de la frecuencia de aparición de *ELN* fue gradual y continua; la de *paramilitares*, mucho más abrupta.

Gráfica 18
farc, *eln* y *paramilitares* en *El Tiempo*



[13] La frecuencia de *paralimitares* corresponde a la frecuencia de sus accidentes y del mismo término en inglés: *paramilitar*, *paramilitares*, *paramilitaries*, *paramilitarism*, *paramilitarismo*, *paramilitarizado* y *paramilitary*. Hay problemas en inglés pues el archivo de *Semana* incluye algunos documentos académicos escritos en este idioma.

Gráfica 19
farc y *secuestros* en *El Tiempo* y *Semana*



La gráfica 19 da algunas pistas sobre las causas de los grandes altibajos en el cubrimiento mediático de la guerrilla de las FARC. La gráfica muestra, tanto para *El Tiempo* como para *Semana*, el cambio mensual de la frecuencia de las palabras *farc* y *secuestros*. Las coincidencias son enormes. Ambas series se mueven de manera casi sincrónica. El coeficiente de correlación es de 0,78 en *El Tiempo* y de 0,87 en *Semana*. La evidencia indica que la visibilidad de las FARC estuvo asociada esencialmente al tema del secuestro. Los rescates y las liberaciones, en particular, parecen haber generado todo tipo de noticias, reacciones y comentarios que, en conjunto, aumentaron de manera sustancial la visibilidad mediática de este grupo. De nuevo, La gráfica sugiere que los secuestros (y los secuestrados) garantizaron a las FARC una gran visibilidad a pesar de su debilitamiento militar. La fórmula *farc* = *secuestros* resume bien esta historia mediática.

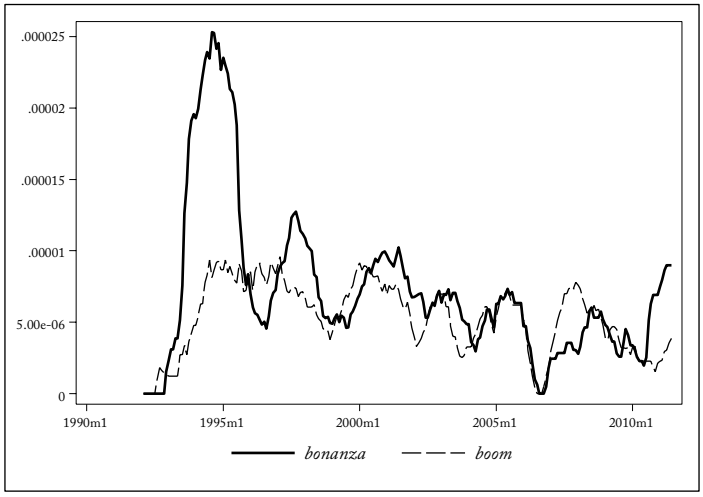
En síntesis, los secuestros de las FARC fueron el tema predominante en el cubrimiento del conflicto colombiano. El cubrimiento tuvo dos o tres momentos de ebullición pero, en general, el interés fue sostenido durante al menos una década.

Bonanza

La gráfica 20 muestra la frecuencia de las palabras *bonanza* y *boom* en el archivo de *El Tiempo*: las conclusiones no cambian si se incluyen los otros dos medios. En principio, esta serie mide, de manera indirecta, el entusiasmo colectivo ante las buenas noticias económicas, originadas, por ejemplo, en un descubrimiento petrolero o minero o en un aumento sustancial de los precios de los principales productos de exportación. Los datos sugieren que el mayor entusiasmo colectivo de las últimas dos décadas ocurrió entre 1993 y 1995 como consecuencia de los hallazgos petroleros de Cusiana y Cupiaga. La prensa

reaccionó mucho más fuertemente ante el descubrimiento de un nuevo yacimiento que ante los altos precios del petróleo y del carbón de los últimos años. Este resultado indica, en últimas, la existencia de una realidad sociológica relevante (un sentimiento colectivo de abundancia, en este caso) que pudo haber incidido en las decisiones públicas y privadas.

Gráfica 20
Bonanza y *boom* en *El Tiempo*



En teoría, los descubrimientos petroleros crearon una sensación de abundancia de recursos y ausencia de restricciones, impulsaron un auge en el consumo público y privado y pudieron, incluso, haber sembrado la semilla de la crisis de finales de los años noventa, la peor en la historia moderna del país (Echeverry, 1996). Más allá de las consecuencias, la evidencia sugiere que en la primera mitad de los años noventa, más que en cualquier otro momento de la últimas dos décadas, la idea de una bonanza o de un *boom* económico captó la imaginación de mucha gente. La prensa a veces sirve de termómetro del entusiasmo colectivo.

División de poderes

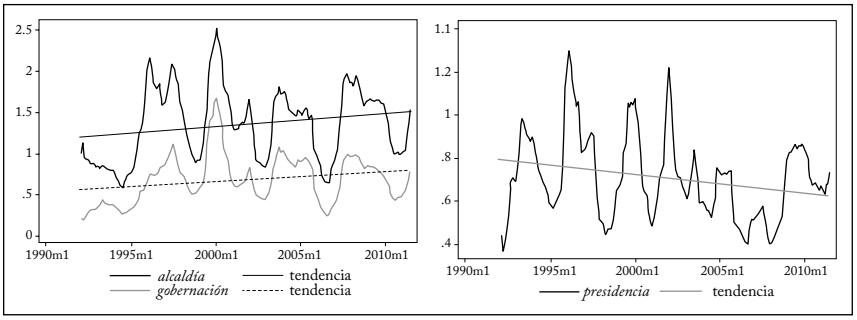
La Constitución de 1991 redefinió la estructura de poder. Formalmente, la descentralización dio mayor poder a los departamentos y municipios. De la misma manera, la independencia del Banco de la República y la creación de la Corte Constitucional le restaron poder a la rama ejecutiva. Pero los cambios institucionales no siempre tienen

consecuencias reales. La estructura de poder no solo depende de la Constitución o de las instituciones formales. Otras factores, económicos y sociológicos, pueden ser determinantes.

La frecuencia de aparición de algunas palabras puede dar alguna idea de los cambios reales (no formales) en la estructura de poder. Por ejemplo, si la frecuencia de las palabras *alcaldía* y *gobernación* aumenta con respecto a la de *presidencia*, podría hablarse de un mayor protagonismo político de los poderes territoriales o de una mayor visibilidad de los mandatarios locales y, por lo tanto, de un profundización efectiva de la descentralización que trasciende los meros cambios institucionales. Asimismo, si la frecuencia de la palabra *magistrado* (y sus accidentes) aumenta con relación a la de la palabra *congresista* (y sus accidentes), podría hablarse de una transferencia de poder hacia el poder judicial.

La gráfica 21 muestra, para el periodo 1992-2011, la frecuencia de las palabras *alcaldía*, *gobernación* y *presidencia*. Los datos corresponden al diario *El Tiempo*. Las series se normalizaron con base en la mención de la palabra *elecciones* para corregir por los ciclos electorales: la frecuencia de las palabras en cuestión tiende a aumentar, por razones obvias, en los periodos de elecciones. Los resultados muestran, por una parte, un aumento tendencial en la frecuencia de *alcaldía* y *gobernación* y, por otra, una disminución en la frecuencia de *presidencia*. Las pendientes son estadísticamente significativas en cada una de las gráficas.

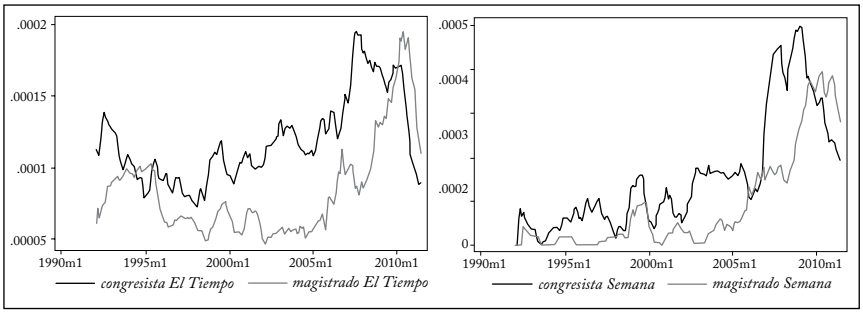
Gráfica 21
Alcaldía, *gobernación* y *presidencia* en *El Tiempo*



Este resultado sugiere que la descentralización sí vino acompañada de una mayor visibilidad mediática de los centros regionales de poder. La mayor visibilidad puede indicar, a su vez, una transferencia real de poder de la nación a las regiones o, simplemente, mostrar un

mayor interés de la prensa nacional por la suerte de los municipios y departamentos. Sea como fuere, la gráfica da información, en principio relevante, sobre una faceta no estudiada de la descentralización.

Gráfica 22
Magistrados y *congresistas* en *El Tiempo* y *Semana*



La gráfica 22 muestra la frecuencia de las palabras *magistrados* y *congresistas*. La serie de la izquierda corresponde a *El Tiempo* y la de la derecha a *Semana*. Ambas series cuentan una historia similar: un aumento sostenido de la frecuencia de aparición de *magistrados* y sus accidentes desde mediados de la década anterior. En 2010, por primera vez en las dos últimas décadas, la frecuencia de *magistrados* superó a la de *congresistas*. El mayor protagonismo mediático de los magistrados quizá tuvo mucho que ver con los escándalos de la parapolítica y de las interceptaciones telefónicas. Pero también puede reflejar un cambio estructural, no asociado a una coyuntura específica: la mayor injerencia de los magistrados en las decisiones públicas.

En los últimos meses, ambas series cayeron abruptamente, pero, al mismo tiempo, se mantuvo la prominencia mediática de los magistrados. En general, el resultado sugiere un cambio significativo en la estructura de poder[14].

CONCLUSIONES

Este artículo presenta un análisis preliminar de algunos aspectos de la realidad colombiana basado en el conteo de palabras en tres medios

[14] En una entrevista publicada en *El Espectador* (18 de diciembre de 2010) el abogado y columnista Yesid Reyes hizo una interesante observación sobre la vida pública de su padre, el presidente de la Corte Suprema, Alfonso Reyes Echandía, inmolado en la toma y retoma del Palacio de Justicia: "la exposición de mi padre a la prensa en el año 1985, cuando era el presidente de la corporación, fue mínima. No tengo idea de cuántas veces saldría en la prensa, pero en todo caso no fueron más de tres o cuatro: dos de ellas antes de morir, durante la toma del Palacio".

escritos de circulación nacional. El análisis tiene una dificultad obvia: las noticias no son neutrales; incorporan necesariamente los sesgos de los editores y comentaristas de los periódicos bajo escrutinio. No obstante, la sección 3 muestra que describen adecuadamente la cambiante realidad de algunos fenómenos socioeconómicos. La sección 4 muestra, de otro lado, que el análisis permite captar la dinámica de otros fenómenos que, por su misma naturaleza, son difíciles de medir o cuantificar. Las descripciones no son definitivas, pero plantean preguntas interesantes, sugieren hipótesis no triviales y pueden servir de punto de partida para investigaciones posteriores.

En esencia, este artículo describe una base de datos mediante una serie de ejemplos que, en conjunto, dan algunas luces sobre las transformaciones económicas y sociales ocurridas en Colombia durante los últimos veinte años. Pero el objetivo es más ilustrativo que descriptivo, más de forma que de fondo. Más que medir o explicar algunos fenómenos socioeconómicos, el artículo quiere mostrar la utilidad de un método novedoso, de una nueva herramienta de investigación en ciencias sociales.

Este es el primer artículo de *culturomics* sobre Colombia. No será el último. Algunas ideas sobre posibles investigaciones o análisis posteriores son obvias. Los trabajos futuros podrían retomar algunos de los temas aquí planteados: la corrupción, el cubrimiento periodístico de las políticas económicas, los determinantes del cubrimiento del conflicto, etc. Podrían también explorar otros temas: las relaciones internacionales, la percepción de inseguridad, el cubrimiento relativo de las regiones, el papel del Banco de la República, etc.

Metodológicamente, las posibilidades de investigación son variadas. Valdría la pena, por ejemplo, estudiar la coexistencia de dos o más palabras en los artículos y comentarios de prensa. Este tipo de enfoque permitiría ir más allá del simple análisis de series de tiempo y brindaría información relevante sobre relaciones causales entre las variables de interés. Por ejemplo, valdría la pena conocer la medida en que las palabras *regalías* y *corrupción* (o *crisis* y *pobreza*, o *salario mínimo* e *inflación*) vienen juntas en la prensa. En otros términos, se podría pasar del análisis univariado al multivariado.

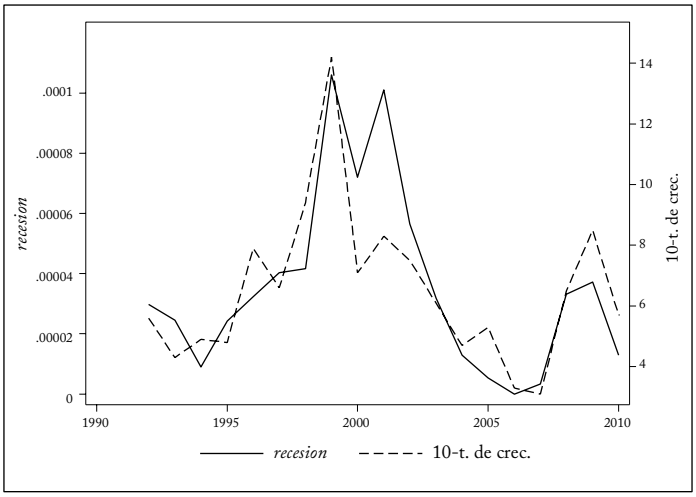
También sería útil estudiar el tono de la información. El conteo no discrimina entre cobertura positiva o negativa, mucho menos entre las posibles variaciones en el tono de las noticias y comentarios. Convendría, por ejemplo, analizar el tono de la cobertura mediática de una institución determinada (el Banco de la República), de una figura política (el presidente Uribe) o de un país (Venezuela). Convendría,

en últimas, complementar el análisis de frecuencias con información sobre el sentido y el tono de la cobertura.

También valdría la pena estudiar las diferencias entre noticias y opinión. Las noticias y las opiniones pueden reflejar la realidad de manera diferente y podrían estudiarse por separado. Este tipo de análisis daría algunas luces sobre los sesgos de los medios y los cambios en la opinión publicada. Por último, este método se podría combinar con encuestas de opinión para analizar las interacciones, no siempre obvias, entre opinión pública y publicada.

ANEXO

Gráfica A1
10 – tasa de crecimiento del pib y *recesión* en *El Tiempo*



REFERENCIAS BIBLIOGRÁFICAS

1. Cowell, F. A., E. Flachaire y S. Bandyopadhyay. "Inequality, entropy and goodness of fit", Document de Travail n°2011-23, Groupement de Recherche en Economie Quantitative d'Aix-Marseille, UMR-CNRS 6579, École des Hautes Etudes en Sciences Sociales, 2011.
2. Echeverry, J. C. "The fall in Colombian savings during the 1990s. Theory and evidence", *Borradores de economía* 3593, Banco de la República, 1996.
3. Glaeser, E. L. y C. Goldin. "Corruption and reform: Introduction", en *Corruption and reform: Lessons from America's economic history*, NBER, 2006, pp. 2-22.
4. Jones, M. P. y J. Crowley. "A General class of nonparametric tests for survival analysis", *Biometrics* 45, 1, 1989, pp. 157-170.

5. Michel, J. B., Y. K. Shen, A. P. Aiden et al. "Quantitative analysis of culture using millions of digitized books", *Science* 331, 6014, 2011, pp. 176-182.
6. Goel, R., M. Nelson y M. Naretta. "The internet as an indicator of corruption awareness", *European Journal of Political Economy* 28, 1, 2012, pp. 64-75.