ÀGORA DE SALUT VII ISSN: 2443-9827. DOI: http://dx.doi.org/10.6035/AgoraSalut.2020.7.29 - pp. 287-295

Big Data y *Big Five*
Análisis de los adjetivos de personalidad en el lenguaje escrito mediante Ngram Viewer

Andrei Valentin Ronai
andreyronai@gmail.com

Julio González-Álvarez
gonzalez@psb.uji.es

Resumen

El presente trabajo estudia la evolución del uso de adjetivos relacionados con la personalidad humana en el lenguaje escrito español entre los años 1950 y 2008. Los adjetivos corresponden al modelo teórico *The Big Five Structure* elaborado por Goldberg (1990) y Costa y McCrae (1992), analizados mediante la herramienta Google Ngram Viewer, una aplicación que trabaja con *Big Data* de carácter léxico, en el que alrededor del 5 % de los libros impresos desde 1500 ha sido volcado en una base de datos. Los adjetivos utilizados, inicialmente en inglés, han pasado por un proceso de retrotraducción para asegurar su concordancia semántica en castellano. Los resultados obtenidos muestran distintos cambios de tendencias en la población, especialmente en los factores 3 (responsabilidad) y 4 (neuroticismo), que mantienen una tendencia de evolución inversa. Los factores 1 (extraversión) y 5 (apertura a la experiencia) evolucionan de forma creciente, mientras que el factor 2 (amabilidad) experimenta un pronunciado descenso. Estos resultados pueden deberse a diversos cambios socioculturales y lingüísticos que modificaran las connotaciones y los matices de los adjetivos referentes a cada factor, así como a cambios en las necesidades y las percepciones de la población.

Palabras clave: *Big Five*, *Big Data*, Ngram, personalidad, lenguaje escrito.

Abstract

This paper studies the evolution of the use of adjectives related to the human personality in Spanish written language between 1950 and 2008. The adjectives correspond to the theoretical model "The Big Five Structure" elaborated by Goldberg (1990) and Costa and McCrae (1992), analyzed using the Google Ngram Viewer tool, an application that works with lexical Big Data, in which about 5% of printed books since 1500 have been dumped into a database. The adjectives used, initially in English, have gone through a retro-

translation process to ensure their semantic concordance in Spanish. The results obtained show different changes in population trends, especially in factors 3 (Responsibility) and 4 (Neuroticism) that maintain a trend of inverse evolution. The factors 1 (Extraversion) and 5 (Openness to experience) evolve increasingly, while factor 2 (Kindness) experiences a steep decline. These results may be due to various sociocultural and linguistic changes that modify the connotations and nuances of the adjectives referring to each factor, as well as changes in the needs and perceptions of the population.

Keywords: Big Five, Big Data, Ngram, personality, written language.

Introducción

La hipótesis léxica afirma que las diferencias humanas más importantes en las transacciones entre personas llegarán a codificarse como términos únicos en algunos o todos los idiomas del mundo (Saucier y Simonds 2006). Esta hipótesis ha tenido una gran relevancia en el desarrollo de teorías de la personalidad basadas en el léxico. *The Big Five Factor Structure* es una clasificación de los constructos de personalidad realizada por Goldberg (1990) y Costa y McCrae (1992) que se materializa en un modelo de cinco factores: 1, extraversión; 2, amabilidad; 3, responsabilidad; 4, neuroticismo y 5, apertura a la experiencia. Su desarrollo deriva de los estudios de Allport y Odbert (1936), que ordenaron más de 18 000 términos en cuatro listas de palabras clasificadas como rasgos estables. Cattell (1943) utilizó dicha clasificación para construir una serie de escalas formadas en su mayor parte por términos bipolares. Tras su análisis, y guiándose por las correlaciones obtenidas, agrupó los términos relacionados en 35 conjuntos bipolares, que se emplearon en varios estudios, midiendo a través de métodos de rotación oblicua la correlación entre las variables. Catell afirmó haber identificado 12 factores oblicuos, pero, a pesar de esto, análisis posteriores mediante procedimientos de rotación ortogonal resultaron en solo 5 factores replicables, publicados posteriormente por McCrae y Costa (1985; 1987). Tras diversas críticas relacionadas con la metodología usada durante el desarrollo del modelo de 5 factores, los investigadores han empleado paneles de jueces para estimar la relevancia social de diferentes rasgos. Saucier y Goldberg (1996) emplearon 25 jueces estudiantes y, primero, estimaron la familiaridad de los 492 adjetivos de personalidad de Goldberg (1990) en una escala de 0 a 9 y, después, eliminaron las 57 palabras menos populares, lo que resultó en una clasificación de 435 términos.

Posteriormente, en el 2010, Google tuvo la iniciativa de desarrollar una herramienta innovadora de gran utilidad para lingüistas y otros investigadores cuyos proyectos requieren el análisis de grandes cantidades de datos (*Big Data*) de carácter léxico. Se trata de Ngram Viewer, una aplicación en línea que permite realizar análisis cuantitativos de la frecuencia de uso de palabras y expresiones utilizando una base de datos de más de 5 195 769 libros digitalizados, aproximadamente un 5 % de los libros jamás publicados (Michel et al. 2010). Su funcionamiento se basa en el uso de *n-grams*, es decir, palabras o expresiones formadas por *n* partes. Ofrece la posibilidad de elegir entre distintos corpus lingüísticos, ajustar el rango temporal y revisar los libros de donde se recoge la información expuesta en la gráfica. Aplicaciones de esta herramienta se pueden observar en el estudio de Roivainen, quien afirmó que «Si la hipótesis léxica es correcta, podemos asumir que el lenguaje cotidiano refleja la estructura de la personalidad humana» (Roivainen 2013, 418) y que la frecuencia de uso es una buena medida de la importancia práctica y el poder descriptivo de los

términos, pero es el análisis factorial quien ofrece información sobre las relaciones entre conceptos.

Combinando los recursos descritos, este estudio propone indagar de manera exploratoria la hipotética evolución de los constructos de personalidad reflejada en los adjetivos contenidos en el corpus español de Google Ngram Viewer. Para ello, se ha partido de una selección de los adjetivos ingleses de Saucier y Goldberg (1996), se han traducido al castellano con asistencia experta y se ha analizado su evolución temporal a lo largo de la segunda mitad del siglo XX hasta el presente, concretamente desde 1950 hasta el 2008, último año del corpus español. Aunque se trata de un primer estudio exploratorio, nuestra hipótesis de partida es que los componentes de personalidad reflejados en el uso de los adjetivos en el idioma español han sufrido algún tipo de variación a lo largo de las décadas analizadas.

Método

Procedimiento

Los adjetivos son comúnmente las palabras más utilizadas a la hora de describir la personalidad (Saucier y Simonds 2006), por eso se ha utilizado la lista de adjetivos correspondientes a los 5 factores, elaborada por Saucier y Goldberg (1996), de los cuales se han elegido aquellos cuya correlación coincide con la dirección del factor, es decir, se han omitido aquellos que correlacionan negativamente. Esto ha derivado en una muestra de 250 adjetivos. Dado que los adjetivos originales pertenecen al idioma inglés, se ha llevado a cabo una retrotraducción, que consiste en traducir los adjetivos y que estos sean revisados por una persona bilingüe que haya vivido tanto en España como en el Reino Unido o los Estados Unidos de América y conozca las diferencias culturales que existen en su uso lingüístico. Algunos términos comparten la misma traducción, con lo cual la lista resultante se queda en 241 adjetivos. Los criterios de decisión han sido, por un lado, que su traducción existiera como término único en el lenguaje castellano y, por otro, que en el análisis mediante Google Ngram Viewer emitiera resultados en su forma de adjetivo. Los adjetivos han sido analizados en su forma masculina, femenina, singular y plural, con la etiqueta «_ADJ», para asegurar que los términos son seleccionados únicamente en su función de adjetivos y no en el desempeño de otra función gramatical, y el comando + para sumar las frecuencias relativas en las formas gramaticales descritas y obtener la frecuencia global de cada adjetivo.

Software

Ngram Viewer permite ajustar distintos parámetros en el análisis de los *ngrams* introducidos, como el corpus lingüístico, el rango temporal o las categorías gramaticales, entre otros. También permite revisar los libros fuente de la información, disponibles en Google Books, o descargar las bases de datos que usa la herramienta. El resultado del análisis se compone de gráficas con los porcentajes de aparición de uno o varios términos en función del número de libros publicados por año. Con la base de adjetivos resultantes, más de 800 términos incluyendo las formas de género y número, se ha procedido al análisis mediante Ngram Viewer en el corpus *Spanish*, con un suavizado de 3 años.

Análisis estadísticos

Los datos obtenidos, extraídos mediante el código fuente de la página, se han exportado a un documento Word para depurarlos y después a un documento Excel, donde se han agrupado en gráficas para su interpretación y se ha añadido el análisis de tendencias mediante modelos lineales y cuadráticos.

Resultados

Las siguientes gráficas muestran la evolución de los cinco factores desde 1950 hasta el 2008, así como las líneas de tendencia y su correspondiente ecuación. En el factor 5, *apertura a la experiencia*, se han omitido los adjetivos *profundo* e *intelectual* ya que creaban irregularidades en la gráfica debido a sus elevadas frecuencias. Algunos adjetivos no disponen de una traducción única al castellano, por lo que han sido omitidos. Algunas formas del plural no emitían resultados con la etiqueta «_ADJ» en Ngram Viewer, por lo que han sido omitidos también.

El factor 1, *extraversión*, mantiene un patrón constante hasta 1955, momento en el que su frecuencia relativa sube de forma muy rápida hasta 1986, donde tiende a estabilizarse. Esto implica un incremento en el uso de adjetivos que describen a una persona o un comportamiento extrovertido.

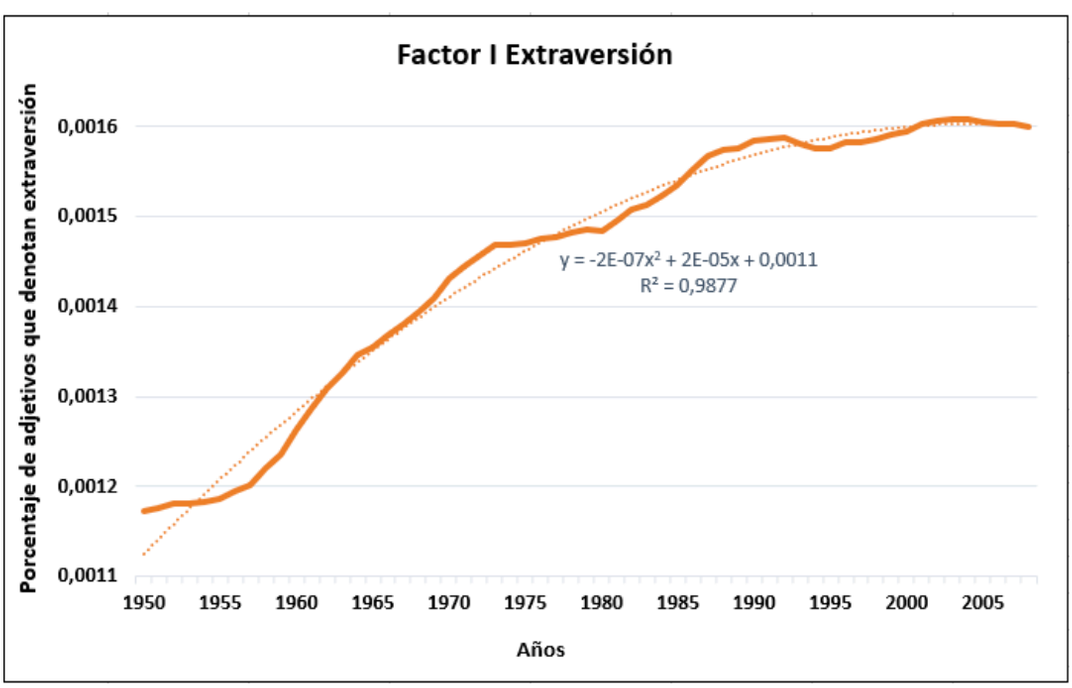

Figura 1. Evolución del factor e*xtraversión* y su tendencia.

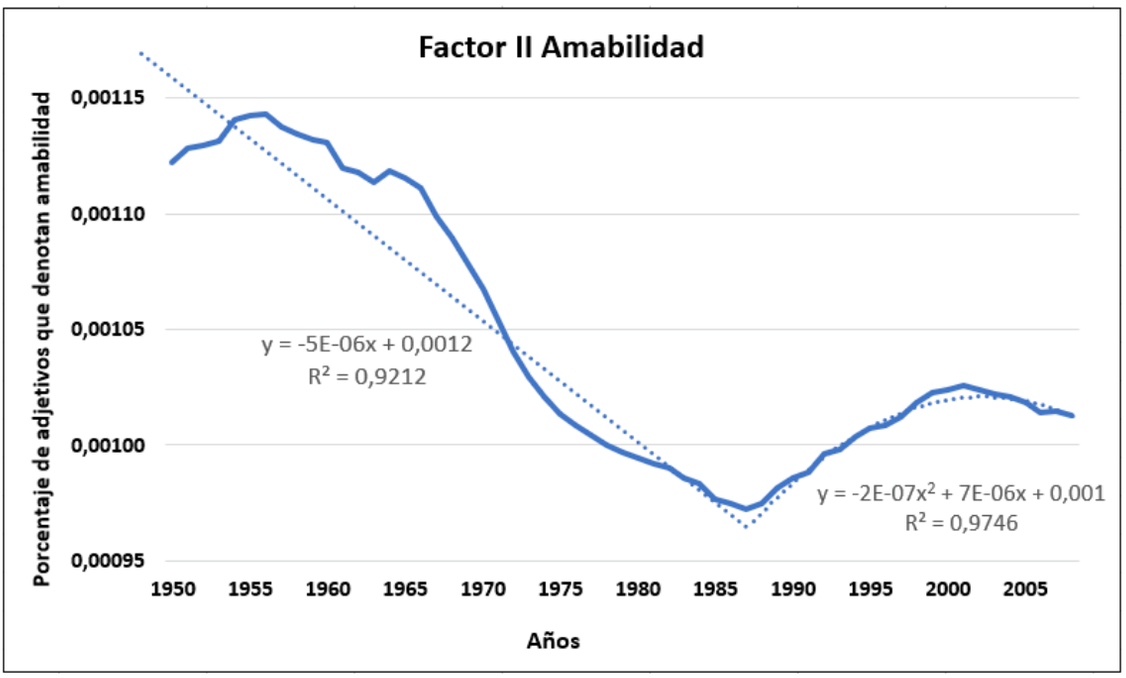

Figura 2. Evolución del factor *amabilidad* y su tendencia.

Por el contrario, el factor 2, *amabilidad*, va decreciendo de forma constante y notoria de 1956 hasta 1986 y, a partir de este año, comienza a subir de forma progresiva y suave hasta el 2001, donde adquiere una tendencia decreciente.

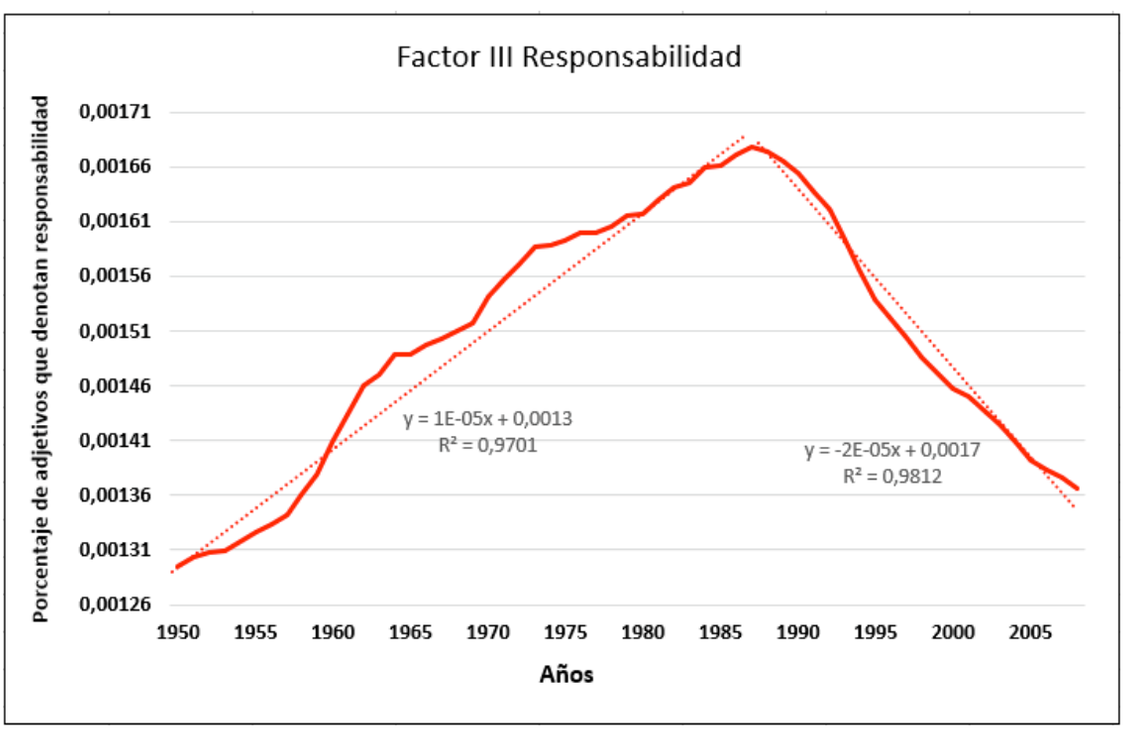

Figura 3. Evolución del factor *responsabilidad* y su tendencia.

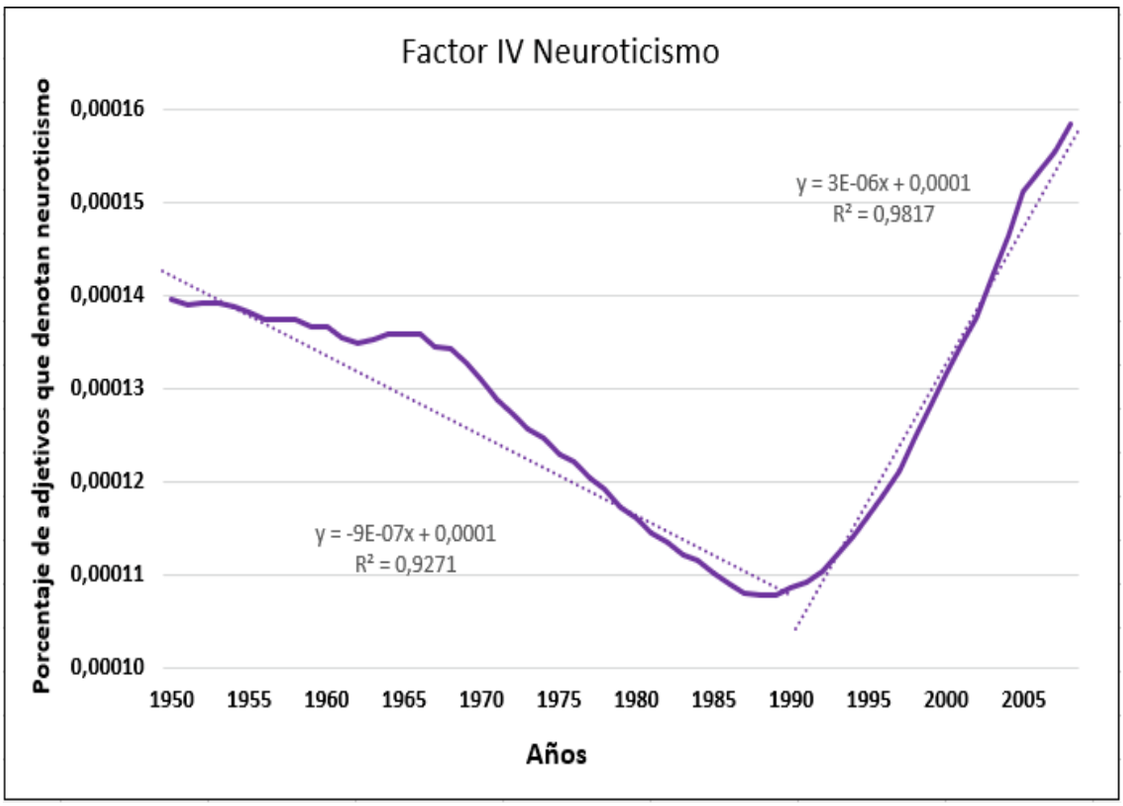

Figura 4. Evolución del factor *neuroticismo* y su tendencia.

El factor 3, *responsabilidad*, mantiene un crecimiento pronunciado hasta 1987, momento en el que comienza a descender. La forma resultante de la gráfica indica una disminución en el uso de adjetivos que denotan responsabilidad o un posible cambio cultural a la hora de reconocer y describir esta cualidad.

El factor 4, *neuroticismo*, presenta un cambio muy distintivo. Decrece suavemente hasta 1966, momento en el que experimenta un descenso muy pronunciado hasta 1988. A partir de ahí evoluciona de forma rápida y ascendente, superando el punto inicial. Esto, a diferencia del factor 3, indica un aumento en la expresión de términos referentes al componente emocional.

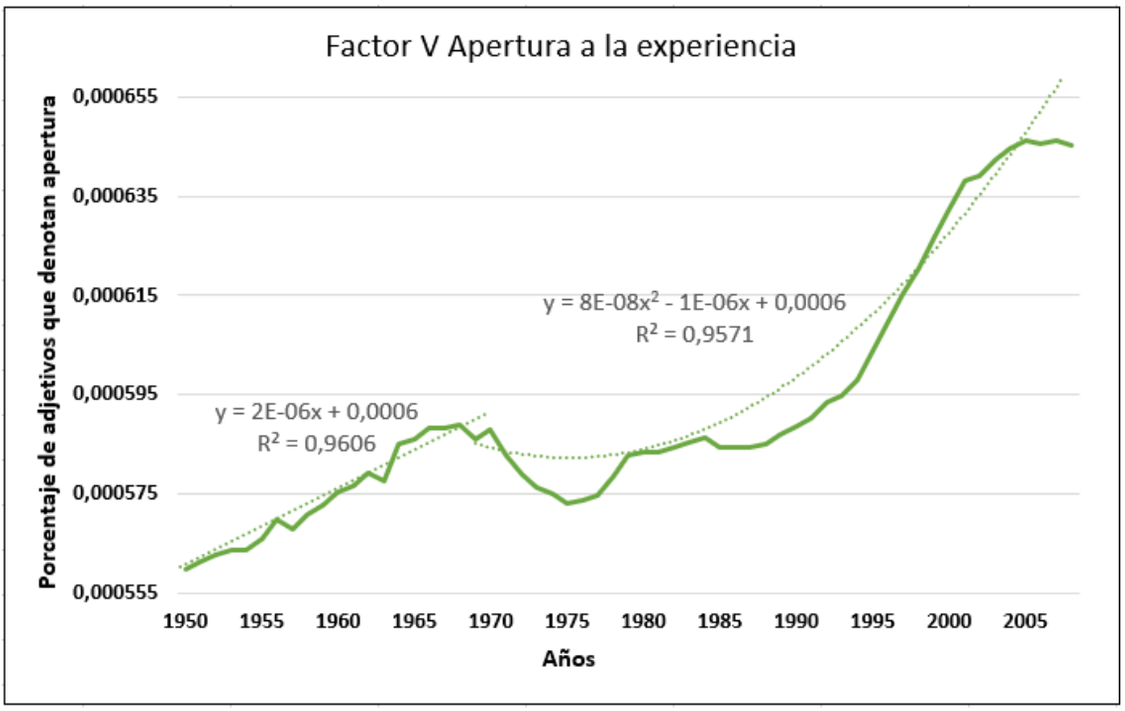

Figura 5. Evolución del factor *apertura a la experiencia* y su tendencia.

Finalmente, el factor 5, *apertura a la experiencia*, experimenta distintos cambios, siendo el factor con mayor variación a partir de 1950. Hasta 1970 muestra un incremento en su expresión con picos y altibajos. A partir de ese momento, decrece hasta 1975, donde vuelve a aumentar hasta que en el 2004 tiende a estabilizarse a un nivel mucho más alto que el inicial.

Discusión y conclusión

Los resultados obtenidos muestran cambios en los factores de personalidad descritos por Goldberg (1990) y Costa y McCrae (1992) en su expresión durante el último medio siglo en el lenguaje escrito español. Estas modificaciones pueden obedecer a diversas razones, como cambios socioculturales y sociodemográficos que hayan introducido nuevas formas de describir a una persona o que hayan ampliado la gama de etiquetas utilizadas con este fin. Por otro lado, acontecimientos históricos han podido influir en las connotaciones y matices de los adjetivos, propiciando cambios de uso en contextos distintos. Este estudio revela cómo, en el ámbito social, los constructos de personalidad evolucionan y median las tendencias personales de una comunidad. Conocer la personalidad de un individuo permite generar expectativas de la forma de sentir, pensar o comportarse de la persona en el futuro o en circunstancias nuevas (Roivainen 2013). Esto podría extrapolarse a todo un bloque social, permitiendo conocer mejor sus necesidades y fortalezas basándose en su uso del lenguaje, pues, tal como afirma Srivastava (2010), las dimensiones de los cinco factores reflejan las preocupaciones sociales de los perceptores. Un ejemplo aplicado de este tipo de análisis se puede apreciar en el estudio realizado con monjas de los conventos de las Escuelas de Hermanas de Notre Dame, en el cual se encontró que aquellas monjas que en su autobiografía usaban una mayor proporción de palabras que expresaban emociones positivas vivieron de media 9,4 años más que las del cuartil más bajo (Danner, Snowdon y Friesen 2001). De una forma semejante, el uso del lenguaje de

una región condiciona la actitud y la conducta de sus habitantes e influye en su bienestar.

Si observamos los factores 3 y 4, se puede apreciar un patrón de evolución inverso. A medida que el factor *responsabilidad* aumenta, el factor *neuroticismo* disminuye; y cuando, a partir del año 1987, el factor 3 comienza a disminuir de forma rápida, el factor 4 experimenta lo contrario, un incremento con tendencia exponencial. Esto puede implicar que una persona con alta responsabilidad tiende a un menor grado de neuroticismo o mayor estabilidad emocional, y un menor grado de responsabilidad se asocia a más neuroticismo. En el ámbito social, estos datos indican que, a medida que decrece la percepción de responsabilidad en la población, aumenta la manifestación de elementos asociados al neuroticismo, ya sean conductas, actitudes o estilos de afrontamiento. Este patrón se observa en el análisis de frecuencia de uso de los adjetivos referentes a cada factor en el lenguaje escrito. Futuros estudios podrán comprobar si esta relación existe más allá de la frecuencia de uso, ya sea a nivel factorial o mediante experimentos conductuales.

Finalmente, cabe mencionar que esta investigación cuenta con diversas limitaciones. Por un lado, los estudios de partida son del año 1996. Aunque tienen una base muy amplia, es posible que los *Big Five* hayan cambiado junto a la sociedad en la última década. Por otro lado, dicho estudios se realizaron en el idioma inglés y, aunque los adjetivos han pasado por un proceso de retrotraducción, los propios factores podrían estar compuestos por diferentes adjetivos en el idioma castellano. Por último, la base de datos de Google Ngram está compuesta únicamente por libros, no permitiendo así analizar la prensa, las revistas, las redes sociales ni otro tipo de lenguaje escrito de la sociedad actual.

Referencias bibliográficas

Allport, Gordon W. y Henry S. Odbert. 1936. Trait-names: A psycho-lexical study. *Psychological monographs*, 47(1): I.

Cattell, Raimon B. 1943. «The description of personality: Basic traits resolved into clusters». *The journal of abnormal and social psychology*, 38(4): 476.

Costa, Paul T. y Robert R. McCrae. 1992. «Four ways five factors are basic». *Personality and Individual Differences*, 13(6): 653-665.

Danner, Deborah D., David A. Snowdon y Wallace V. Friesen. 2001. «Positive emotions in early life and longevity: Findings from the nun study». *Journal of Personality and Social Psychology*, 80(5): 804-813.

Goldberg, Lewis R. 1990. «An alternative "description of personality": the big-five factor structure». *Journal of personality and social psychology*, 59(6): 1216.

McCrae, Robert R. y Paul T. Costa. 1985. Comparison of EPI and psychoticism scales with measures of the five-factor model of personality. *Personality and individual Differences*, 6(5): 587-597.

McCrae, Robert R. y Paul T. Costa. 1987. «Validation of the five-factor model of personality across instruments and observers». *Journal of personality and social psychology*, 52(1): 81.

Michel, Jean-Baptiste, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, The Google Books Team, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak y Erez Lieberman Aiden. 2010. «Quantitative analysis of culture using millions of digitized books». *Science,* 331(6014): 176-182.

Roivainen, Eka. 2013. «Frequency of the use of English personality adjectives: Implications for personality theory». *Journal of Research in Personality*, 47(4): 417-420.

Saucier, Gerard y Lewis R. Goldberg. 1996. «Evidence for the *Big Five* in analyses of familiar English personality adjectives». *European Journal of Personality*, 10(1): 61-77.

Saucier, Gerard y Jennifer Simonds. 2006. «The structure of personality and temperament». En *Handbook of Personality Development*, ed. Daniel K. Mroczek y Todd D. Little. New York / London: Lawrence Erlbaum Associates Publishers.

Srivastava, Siddhartha. 2010. «The five-factor model describes the structure of social perceptions». *Psychological Inquiry*, 21(1): 69-75.