Big Data & Big Five

Analysis of personality adjectives in written language using Ngram Viewer

Author: Andrei Valentin Ronai (X-9084556-Q)
Tutor: Julio González Álvarez

Introduction

The lexical hypothesis states that the most distinctive, significant and widespread phenotypic attributes tend to be encoded as unique words in the conceptual repository of language throughout the world (Saucier and Simonds, 2006). This hypothesis had a great relevance in the development of theories of personality based on the lexicon. "The Big Five Factor Structure" is a classification of personality constructs made by Goldberg (1990), Costa and McCrae (1992), based on Cattell's studies on the description of personality. This model is grouped into five factors: (I) extraversion, (II) Agreeableness, (III) responsibility, (IV) neuroticism and (V) openness to experience. In 2010, Google released an innovative highly useful tool for linguists and other researchers interested in projects that require lexical Big Data analysis. This is Ngram Viewer, an online application that allows the performance of quantitative analysis of the frequency of written words and expressions using a database of 5,195,769 digitized books, about 4% of the books ever published (Jean-Baptiste, 2010), that are shown in line graphs. Its operation is based on the use of *n-grams*, which are words or expressions formed by *n* parts. It offers the possibility of choosing between different linguistic corpus, depending on the language in which the experiment is being carried out, adjusting the time range and It allows to review the books where the information is obtained. (Figure 1). Combining the described resources, this study proposes to investigate the personality constructs in the Spanish corpus, on the basis of the studies of Goldberg, Costa and McCrae and their classification of the five personality factors analyzed with Google Ngram Viewer.

Factor	Adjetives	Translation	Correlation Big 5	Correlation LEXESP
Extraversion	Aggressive	agresivo	0.61*	48+42
	Social	social	0.58*	1141
	Verbal	locuaz	0.53*	10
	Assertive	resuelto	0.53*	103+28
	Bold	atrevido	0.53*	56+16
		audaz		47
		valiente		63
	Enthusiastic	entusiasta	0.50*	28
	Spirited	enérgico	0.49*	32+25
	Confident	confiado	0.49*	36+19
	Daring	atrevido	0.46*	56+16
	Merry	alegre	0.44*	161
Agreeableness	Sympathetic	simpático	0.62*	70+47
	Kind	amable	0.60*	127
	Warm	cariñoso	0.56*	29+20
		acogedor	0.56*	13+14
	Courteous	cortés	0.53*	66
	Understanding	comprensivo	0.53*	15+15
	Polite	respetuoso	0.52*	31+20/43+24
		educado		43+24
	Considerate	considerado	0.51*	166+4
	Affectionate	afectuoso	0.51*	8+6
	Sincere	sincero	0.49*	55+30
	Sentimental	sentimental	0.48*	106
Conscientiousness	Organized	organizado	0.65*	95+42
	Precise	preciso	0.61*	370+132
	Responsible	responsable	0.59*	353
	Thorough	riguroso	0.58*	51+56
		exhaustivo		21+14
	Orderly	ordenado	0.57*	65+31
	Efficient	eficiente	0.57*	26
	Practical	práctico	0.54*	126+347
	Systematic	sistemático	0.54*	24+61
		metódico		15+10
	Dependable	confiable	0.51*	3
	Reliable	fiable	0.49*	22
Neuroticism	Moody	malhumorado	0.53*	19+5
	Touchy	delicado	0.51*	97+98
		sensible		118
	Temperamental	temperamental	0.51*	6
	Emotional	emocional	0.49*	65
	Jealous	celoso	0.47*	29+21
	Envious	envidioso	0.47*	9+2
	Possessive	posesivo	0.46*	7+5
	Fretful	fastidioso	0.45*	10+6/8
		irritable		8
	Impatient	impaciente	0.44*	49
	Nervous	nervioso	0.42*	202+88
Opennes/ Intellect	Intelligent	inteligente	0.55*	222
	Intellectual	intelectual	0.50*	251
	Smart	listo	0.49*	99
	Complex	complejo	0.48*	221+98
	Philosophical	filosófico	0.47*	55+73
	Bright	brillante	0.44*	217
	Innovative	innovador	0.44*	11+6
	Deep	profundo	0.43*	261+275
	Knowledgeable	experto	0.43*	95+23
	Ingenious	ingenioso	0.43*	35+17

Figure 2. List of adjectives used grouped according to Big Five factors and ordered according to their correlation.

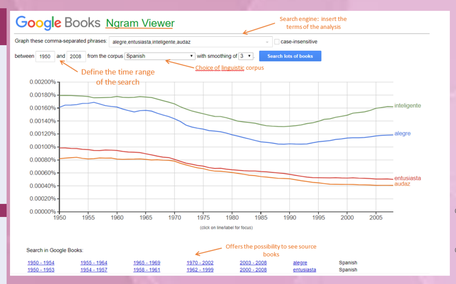

Figura 1. Ngram Viewer and its different elements

Abstract

The present study investigates the personality factors corresponding to "The Big Five Structure" in the written language of the Spanish corpus between 1950 and 2008, through the Ngram Viewer tool developed by Google. As main results, significant decreases have been found in Neuroticism and Kindness, a significant increase in Extraversion and linear trend in Responsibility and Openness to experience. These changes may be due to sociodemographic events that changed the connotations and nuances of the adjectives referring to each factor.

Keywords: Big Five, Big Data, Ngram, personality, written lenguaje, Spanish

Method

Adjectives are very advantageous words to describe personality, because they simultaneously include both desirable and undesirable dimensions. (Saucier & Simonds, 2006). For this reason, we have used the list of adjectives corresponding to the 5 factors elaborated by Saucier and Goldberg (1996), of which the most representative of each dimension have been chosen according to their degree of saturation. As a cut-off point, those whose correlation is higher than 0.40 were chosen. In addition, the resulting list went through a second screening based on the frequency of its use in Spanish lenguage according to the LEXESP corpus database developed by Sebastián Gallés (Gallés et al., 2000). The 10 most frequent, and with the highest correlation with the Big Five, were chosen for each factor (Figure 2). Some factors have more adjectives because its translation from English corresponds to several terms in Spanish. In this way, we captured the full meaning of the original expression.

Results

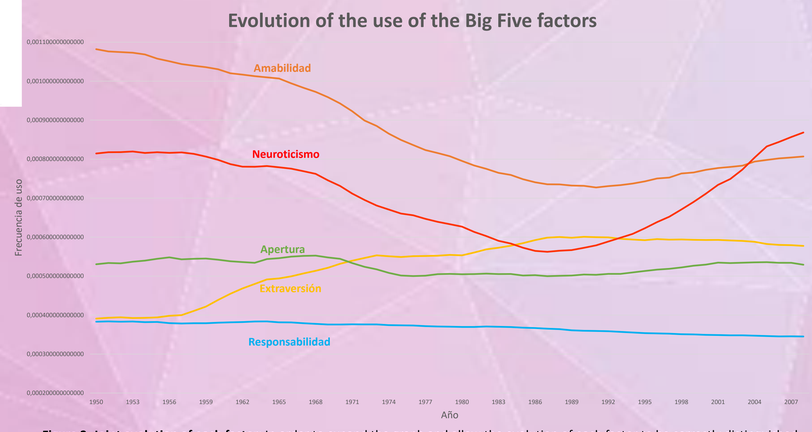

Figure 3. Joint evolution of each factor: In order to expand the graph and allow the evolution of each factor to be correctly distinguished, some of them have been multiplied: amabilidad x10; responsabilidad x1,5; neuroticismo x10 y apertura x2

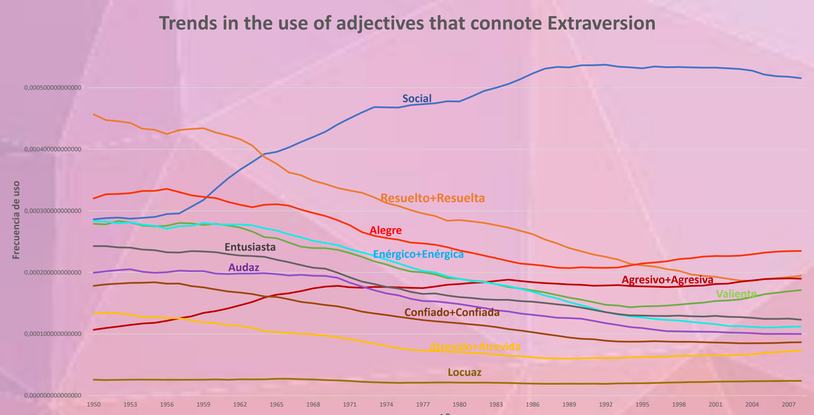

Figure 4. Extraversion: To expand the graph and allow the evolution of each term that refers to extraversion to be correctly distinguished, some of them have been multiplied: (resuelto+resuelta) x20; (agresivo+agresiva) x20; (atrevido+atrevida) x25; audaz x25; valiente x20; (enérgico+enérgica) x20; (confiado+confiada) x20; entusiasta x25; locuaz x40; alegre x20.

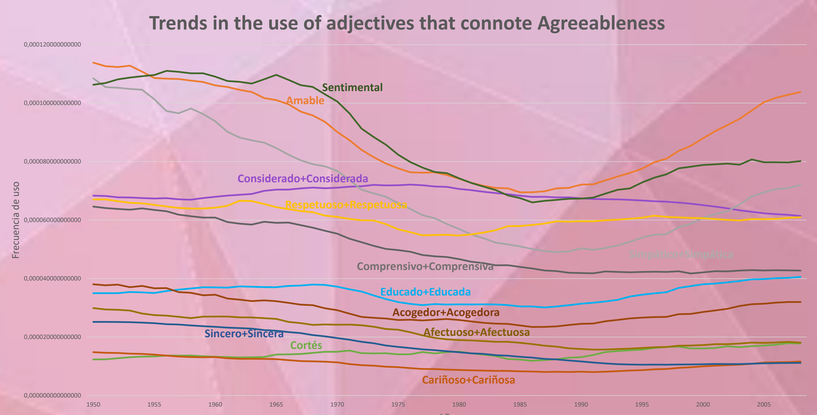

Figure 5. Agreeableness: To expand the graph and allow the evolution of each term that refers to agreeableness to be correctly distinguished, some of them have been multiplied: (considerado+considerada)x3; amable x10; (simpático+simpática)x10; (respetuoso+respetuosa)x10; (educado+educada)x10; cortés x10; (cariñoso+cariñosa)x3; (acogedor+acogedora)x15; (comprensivo+comprensiva)x10; (afectuoso+afectuosa)x15; sentimental x10;

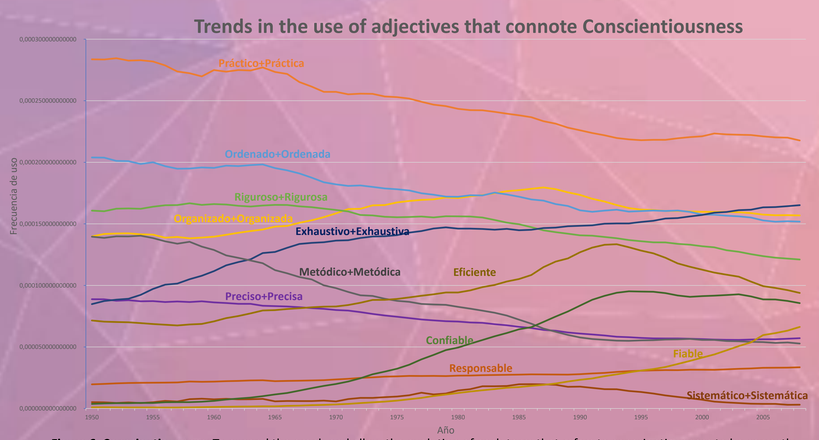

Figure 6. Conscientiousness: To expand the graph and allow the evolution of each term that refers to conscientiousness to be correctly distinguished, some of them have been multiplied: (práctico+práctica) x5; (organizado+organizada) x5; (ordenado+ordenada) x10; (riguroso+rigurosa) x10; (exhaustivo+exhaustiva) x20; (sistemático+sistemática) x1000; (metódico+metódica) x20; eficiente x5; fiable x20; confiable x20

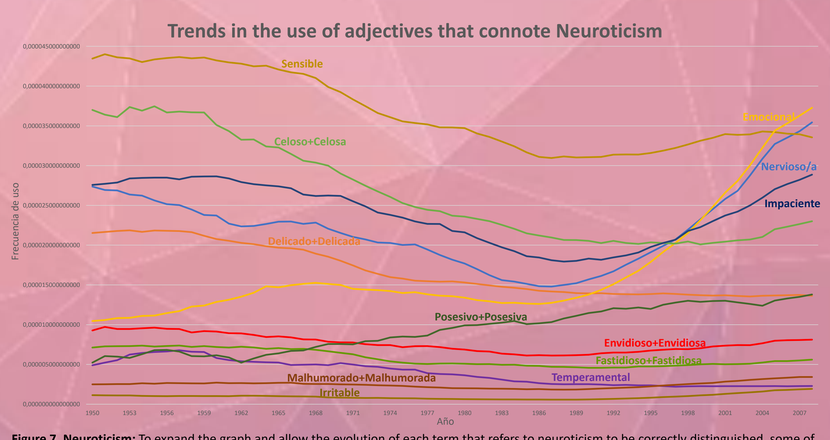

Figure 7. Neuroticism: To expand the graph and allow the evolution of each term that refers to neuroticism to be correctly distinguished, some of them have been multiplied: (nervioso+nerviosa) x1,7; sensible x2; emocional x1,7; temperamental x5; (celoso+celosa) x5; impaciente x10; (malhumorado+malhumorada) x4; (fastidioso+fastidiosa) x7; irritable x2; (envidioso+envidiosa) x10; (posesivo+posesiva) x10

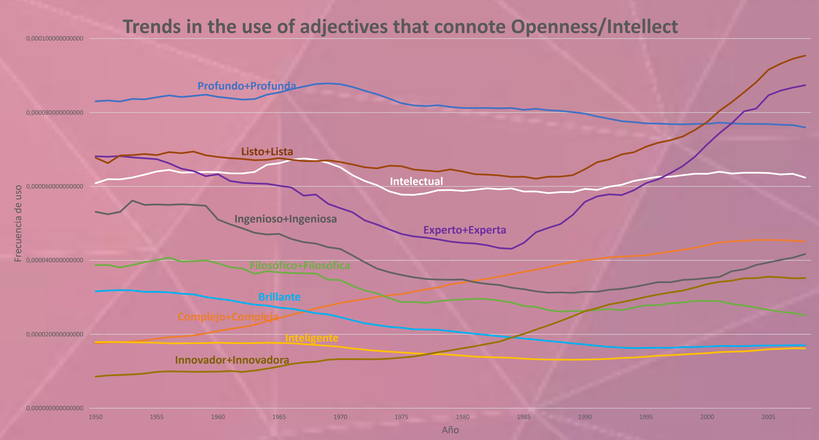

Figure 8. Openness/Intellect: To expand the graph and allow the evolution of each term that refers to openness/intellect to be correctly distinguished, some of them have been multiplied: (experto+experta) x40; (listo+lista) x10; (ingenioso+ingeniosa) x10; (innovador+innovadora) x5

Conclusions

Figure 3 shows the evolution of the five dimensions of the Big Five Structure in the Spanish written language from 1950 to 2008. The factor I, **Extraversion**, maintains a constant pattern until 1957, when its relative frequency rises very quickly until 1986, where it stabilizes. Looking closely at the terms corresponding to this factor in Figure 4, we observe that the pattern of evolution is very similar to that of the adjective “social”, while the others follow a decreasing curve. On the other hand, the factor II, **Agreeableness**, is decreasing constantly and notoriously until 1995 and starting this year, it begins to rise progressively and smoothly. Most of the adjectives used maintain a more or less stable beginning and ending, while “sentimental”, “amable” and “simpático/a” are those that experience more fluctuation.

The factor III **Conscientiousness** is intriguing, since its evolution over the last 58 years is minimal. As shown in Figure 3, it maintains a constant and linear pattern until approximately 1990, when it experiences a slight and minimal decrease. This implies that over the years, the connotations of a responsible personality have remained stable. Analyzing the terms separately, those who experience a greater increase are “fiable”, “confiable”, “eficiente” and “exhaustivo”and "exhaustive", while “preciso”, “ordenado” and “práctico” decrease steadily.

The most representative change is seen in factor IV, **Neuroticism**, whose associated adjectives have decreased the frequency of their use until 1986, moment from which their relative frequency increases very sharply until exceeding their initial use according to the time range of this study. Looking in detail, Figure 7 shows that these changes are marked mainly by the adjectives “emocional”, “nervioso/a” and “impaciente”, while the other terms analyzed follow a pattern with linear trend.

Finally, the fifth factor, **Openness/Intellect**, experienced a smooth fluctuation between 1962 and 1974, ending at approximately the same level as it began. The main adjectives that mark this process are “listo/a”, “experto/a”, “complejo/a” and “innovador/a”. These results indicate that 4 of the 5 personality factors described by Goldberg (1990), Costa and McCrae (1992) have experienced changes in their expression during the last half century in written language. These changes can be due to diverse reasons, such as socio-cultural and socio-demographic changes that include new ways of describing a person or expanding the range of labels used for this purpose. On the other hand, historical events could influence the connotations and nuances of adjectives, promoting changes of use in different contexts.

Shortcomings

This study has several limitations, because it is a novel exploratory study in the Spanish corpus. The first one is the reduced number of adjectives used, since only 10 have been used for each factor. As a suggestion for later experiments, it would be convenient to use a larger sample of 100 adjectives and increase the time interval of the analysis. In addition, it would be interesting to concretely relate the found changes with sociodemographic events in order to find a more concise explanation of them.

References

Jean-Baptiste Michel*, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, William Brockman, The Google Books Team, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak, and Erez Lieberman Aiden*. *Quantitative Analysis of Culture Using Millions of Digitized Books*. *Science* (Published online ahead of print: 12/16/2010)
Roivainen, E. (2013). Frequency of the use of English personality adjectives: Implications for personality theory. *Journal of Research in Personality*, *47*(4), 417-420.
Goldberg, L. R. (1990). An alternative" description of personality": the big-five factor structure. *Journal of personality and social psychology*, *59*(6), 1216.
Saucier, G., & Goldberg, L. R. (1996). Evidence for the Big Five in analyses of familiar English personality adjectives. *European journal of Personality*, *10*(1), 61-77.
Sebastián-Gallés, N., Martí, M. A., Carreiras, M., & Cuetos, F. (2000). LEXESP: Una base de datos informatizada del español. *Universitat de Barcelona, Barcelona*.
Allport, G. W., & Odbert, H. S. (1936). Trait names: A psycholexical study. Psychological Monographs, 47, 211