University of Brasília (UnB)
Institute of Psychology
Department of Social and Work Psychology
Social, Work and Organizations Psychology Graduate Program (PSTO)

THE PERSONALITY LEXICON IN BRAZILIAN PORTUGUESE:

STUDIES WITH NATURAL LANGUAGE

Alexandre José de Souza Peres

Supervisor: Prof. PhD. Jacob Arie Laros

Brasília, Distrito Federal

January 2018

University of Brasília
Institute of Psychology
Department of Social and Work Psychology
Social, Work and Organizations Psychology Graduate Program (PSTO)

THE PERSONALITY LEXICON IN BRAZILIAN PORTUGUESE: STUDIES WITH NATURAL LANGUAGE

Alexandre José de Souza Peres

Doctoral dissertation elaborated under the supervision of Prof. PhD. Jacob Arie Laros, and presented to the Social, Work and Organizations Psychology Graduate Program of the University of Brasília, as partial requirement for the degree of Doctor in Social, Work and Organizations Psychology.

Brasília, Distrito Federal

January 2018

University of Brasília

Institute of Psychology, Department of Social and Work Psychology

Social, Work and Organizations Psychology Graduate Program

Dissertation approved by:

_____

Prof. PhD. Jacob Arie Laros (Supervisor)
Social, Work and Organizations Psychology Graduate Program, University of Brasilia

_____

Prof. Dr. Carlos Henrique Sancineto da Silva Nunes
Programa de Pós-Graduação em Psicologia, Universidade Federal de Santa Catarina

_____

Prof. PhD. Guilherme Veiga Rios
Programa de Pós-Graduação em Linguística, Universidade de Brasília

_____

Profa. Dra. Cristiane Faiad de Moura
Department of Social and Work Psychology, University of Brasilia

_____

Prof. Dr. Josemberg Moura de Andrade
Department of Social and Work Psychology, University of Brasilia

Brasília, Distrito Federal

January 2018

Procura da poesia

(...)

Chega mais perto e contempla as palavras.

Cada uma

tem mil faces secretas sob a face neutra

e te pergunta, sem interesse pela resposta,

pobre ou terrível, que lhe deres:

Trouxeste a chave?

(...)

Carlos Drummond de Andrade in A Rosa do Povo (1945)

From March 1979

Tired of all who come with words, words but no language

I went to the snow-covered island.

The wild does not have words.

The unwritten pages spread out on all sides!

I come upon the tracks of roe deer in the snow.

Language but no words.

Tomas Tranströmer in The Wild Market Square (1983, trad. Robin Fulton)

Dedicatory

This dissertation is dedicated to Professor Jacob Laros and to Professor Luiz Pasquali, who advised me during my postgraduate studies with inspiring wisdom, not only embracing my projects, but also kindly respecting my ideas and decisions, and patiently guiding me through my doubts and after my stumbles. It is an honor to be an apprentice and a friend of such masters.

Agradecimentos (Acknowledgements)

Um doutoramento não é resultado apenas da dedicação individual do estudante. Nesse sentido, faço os seguintes agradecimentos. À Renata e ao Bartholomeu, pelo amor e companheirismo. Aos meus pais, que sempre se esforçaram para oferecer aos filhos e a outros familiares oportunidades para o desenvolvimento educacional e profissional. Aos meus irmãos, cunhados e ao restante da minha grande e querida família mineira. À Thaís, Dudu e Raphael, a quem espero inspirar em seguir a carreira acadêmica. Aos meus professores da Educação Básica, que me ensinaram a aprender e inspiraram minha predileção pela ciência, artes e filosofia. Gostaria de lembrar especialmente das professoras Laudiene, Eliana, Magda e Marta de Carmo do Paranaíba, e dos professores Luiz Antônio, Ana Eugênia, Marta, Fábio, Deize, Rosana e outros de Uberlândia. Aos meus amigos dessa época, especialmente Fred e Estevão. Aos meus professores da Universidade Federal de Uberlândia, em especial Luiz Avelino, Dárcio, Sueli, Maria do Carmo, Renata, Marília, Nilton, Miltom, Joaquim e Ederaldo. Aos meus colegas e amigos da graduação e do movimento estudantil, em especial a *party* Prole da Cabra, Getúlio, Frank, Charlie e Giuliano, além da Lúcia e do Jimmy. Aos meus professores da Universidade de Brasília, em especial aqueles que abriram espaço para a psicologia da personalidade em suas disciplinas ou se envolveram de alguma forma com meu doutoramento, como os Professores Cláudio, Fábio e Kátia. Aos professores que gentilmente aceitaram participar das bancas de qualificação e defesa, Cris, Josemberg, Guilherme, Carlos e Primi. Aos amigos Ricardo Primi, Daniel Kinpara, Eloisa Vidal e ao Prof. Dalton Andrade que me auxiliaram em diferentes momentos e incentivaram meu projeto. Aos colegas de curso e laboratório, Talita, Renata, Gina, Luís, Camila, Carlos, entre outros. Ao Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep), que me concedeu afastamento para conclusão do curso. Aos amigos do Inep, em especial a Cintia, o pessoal do Café Chicago, Carlos, Geraldo, Luciano, Maria Fernanda, Matthias, Marcelinho, Marcelo, Gabi e Mariângela, além dos que chegaram mais tarde, Priscila, Elenita, Robson, Maria Regina, Bolivar e tantos outros. Aos meus estagiários, Paula, Rafaela e Leonardo. Aos professores com quem trabalhei no Inep, Elaine, Maria Luiza, Chico e Quintana. Aos tantos amigos pesquisadores na área de políticas e avaliação educacional, avaliação psicológica e psicometria. Aos pesquisadores envolvidos no desenvolvimento do R e de seus pacotes, que possibilitaram a operacionalização da pesquisa descrita nesta tese. À Morcegão FM pela trilha sonora. Meu agradecimento a todos.

Summary

List of figures

List of tables

General Abstract

This dissertation consists of three studies concerning the lexical approach of research in the field of personality, with a focus on Brazilian culture and natural language. The first study is of a theoretical nature and explores some of the criticisms regarding the lexical approach to personality research with its origin in the psychological study of natural language and cross-cultural psychology, as well as methodological propositions coming from these fields. A historical review of the development of the lexical approach is also presented. The second manuscript reports a study that aimed to generate a set of Brazilian Portuguese personality descriptors using the social network Twitter as a trait source. As a result, we obtained a list of potentially relevant descriptors for the construction of a Brazilian personality taxonomy, with 1,454 adjectives, six names, 10 pronouns, and 383 nouns. The third manuscript reports dimensional analyses of a corpus recovered from Twitter regarding 172 adjectives and 86,899 subjects. The results suggest two suitable candidate models for future research, one with seven and another with 14 dimensions. Methodological and theoretical issues and the potential contributions from these studies for future research in the field of personality are also discussed.

Keywords: personality; personality structure; personality traits; lexical hypothesis; text mining; Brazilian Portuguese.

Resumo Geral

Esta tese é composta por três estudos relacionados à abordagem léxica na pesquisa em personalidade, com foco na cultura brasileira e no estudo da linguagem natural. No primeiro estudo, de caráter teórico, exploramos algumas das críticas relacionadas à hipótese léxica a partir das perspectivas do estudo psicológico da linguagem natural e da psicologia transcultural, bem como propostas metodológicas oriundas desses dois campos. Uma revisão histórica do desenvolvimento da hipótese léxica também é apresentada nesse manuscrito. Já no segundo manuscrito, relatamos um estudo que objetivou gerar uma lista de descritores da personalidade para o português brasileiro utilizando a rede social Twitter como fonte. Como resultado, obtivemos uma lista com 1.454 adjetivos, seis nomes, 10 pronomes e 383 substantivos, potenciais descritores para a construção de uma taxonomia brasileira da personalidade. No terceiro manuscrito relatamos um estudo relacionado à análise da dimensionalidade de um corpus também obtido no Twitter, com 172 adjetivos e 86.899 sujeitos. Os resultados sugeriram dois promissores modelos a serem utilizados em futuras pesquisas, um com sete e outro com 14 dimensões. Também são discutidas questões metodológicas e teóricas, além das potenciais contribuições desses estudos para a pesquisa futura em personalidade.

Palavras-chave: personalidade; estrutura da personalidade; traços da personalidade; hipótese léxica; mineração de texto; português brasileiro.

Presentation

The central theme of this dissertation is the investigation of the lexical approach in the context of Brazilian culture using natural language registers as the primary source of personality trait-descriptive terms and data. Three studies are presented in this dissertation as independent manuscripts that are followed by a final considerations section, in which we seek to synthesize and contextualize the main findings.

Manuscript 1, "Lexical approach of personality research, cross-cultural psychology, and natural language", has a theoretical character and consists of a literature review. This study presents a historical introduction to the development of the lexical approach of personality research and the theoretical models constructed from it, like the five-factor model of personality, for example. This manuscript focuses on some of the major criticisms of the lexical approach in the study of the personality, and pursues to indicate methodological and theoretical directions for future research in the area from perspectives related to the psychological study of natural language and cross-cultural psychology. With this study, we aimed to contextualize the theoretical and methodological aspects of the dissertation, as well as the results reported in the second and third manuscripts.

The study reported in Manuscript 2, "The lexicon of personality in Brazilian Portuguese: Searching for descriptive terms in natural language", is empirical and exploratory. The primary objective of this study was to prospect personality trait-descriptive terms employing text-mining techniques in public messages written by users of social networks, specifically Twitter. The data collection procedures aimed to find terms that people spontaneously use to describe themselves and others. The search was restricted to users located in Brazil, for a certain period, and to texts in Portuguese.

After text cleaning procedures, we obtained a list of potential personality descriptors organized by word classes (e.g., adjectives, nouns and adverbs).

Finally, in the study reported in Manuscript 3, "Developing dimensional models for a Brazilian personality lexicon based on text mining of Twitter: Adjectives", we investigated the dimensionality of a term-document matrix with 172 adjectives and 86,899 subjects using the topic modeling technique Latent Dirichlet Allocation. The data collection procedures aimed to find terms that people use to describe themselves, and the search was restricted to users located in Brazil, and to posts in Portuguese.

Cross-validation analyses suggested that models with seven and 14 latent dimensions (i.e., topics) were the most suitable for the data. We compared the semantic content of these models with the formulations of factors from prominent models (e.g., the Big Five model of personality). The results indicated that these two models are promising candidates for future research, with a preference for the model with 14 topics that showed more internal semantic coherence.

We also examined models with latent structures similar to prominent theoretical models found in the literature (e.g., the three-factor model, the Big Five model, the six-factor model, and Cattell's model of 16 primary factors). Corroborating the cross-validation analyses, the qualitative interpretation of the results indicated that the semantic content of the investigated theoretical models lacked interpretability and was not congruent with the formulations of these models of reference.

Manuscript 1

Lexical approach, cross-cultural psychology, and natural language

Abstract

The lexical approach, or lexical hypothesis, is a theoretical perspective on personality psychology on basis of which some of the main theoretical models of the area were developed, such as Cattell's model of 16 personality factors and the five-factor model or Big Five. The lexical approach is based on the idea that personality-descriptive terms can be retrieved from the lexicons of idioms, since most socially relevant and salient personality traits are supposed to have been encoded in the natural languages of different cultures in the course of their history. This manuscript presents a historical review of the development of the lexical approach and highlights potential contemporary methods for the investigation of the lexical hypothesis that have as origin the psychological study of natural language and cross-cultural psychology.

Keywords: personality; lexical hypothesis; personality taxonomy; big five; natural language; cross-cultural psychology.

Manuscrito 1

Hipótese léxica, psicologia transcultural e linguagem natural

Resumo

A abordagem léxica, ou hipótese léxica, é uma perspectiva teórica na psicologia da personalidade a partir da qual alguns dos principais modelos teóricos da área foram desenvolvidos, como o modelo de Cattell dos 16 fatores primários e o modelo dos cinco fatores ou Big Five. Essa abordagem fundamenta-se na ideia de que a maioria das características da personalidade socialmente relevantes e salientes teria sido codificada na linguagem natural das diferentes culturas ao longo de sua história, ou seja, que termos descritores de traços da personalidade podem ser retirados dos léxicos dos idiomas. Este manuscrito apresenta uma revisão histórica do desenvolvimento da abordagem léxica, bem como destaca contribuições à investigação da hipótese léxica oriundas do estudo psicológico da linguagem natural e da psicologia transcultural.

Palavras-chave: personalidade; hipótese léxica; taxonomia da personalidade; Big Five; linguagem natural; psicologia transcultural.

The lexical approach, or lexical hypothesis, is a theoretical perspective in personality psychology. Some of the most prominent theoretical models of the area were developed following this approach, such as the Cattell's model of the 16 primary personality factors, and the five-factor model of personality, also known as *Big five* (John, Angleitner, & Ostendorf, 1988). The lexical hypothesis originated primarily from the idea that personality traits can be identified in language lexicons since most of the socially relevant and salient personality traits would have been encoded into natural language (De Raad & Mlacic, 2015; Goldberg, 1981; John et al. 1988). A personality trait can be conceptualized as "an enduring personality characteristic that describes or determines an individual's behavior across a range of situations" ("Trait", 2015).

According to the lexical hypothesis, the vocabulary (i.e., the lexicon) of personality contained in natural language offers an extensive, albeit finite, a set of attributes (i.e., traits) that people throughout generations have found to be relevant in their daily interactions. Therefore, it is considered possible to elaborate a personality taxonomy from the words used by people in their daily lives in different cultures to refer to themselves, others and the world. John et al. (1988) defined the lexical hypothesis as follows:

> Those individual differences that are most salient and socially relevant in people's lives will eventually become encoded into their language; the more important such a difference, the more likely is it to become expressed as a single word. (p. 174)

Traditionally, psycholexical models of personality such as the *Big Five* were developed according to the following approach. First, a thesaurus is examined and a list of words is developed representing personality traits. Second, psychological instruments

are developed with items using the words from the list. Third, psychometric analyses (e.g., Exploratory Factor Analysis) are performed to assess the latent dimensionality of the instruments and the relevance of each trait (De Raad & Mlacic, 2015; Digman, 1990; John et al., 1988). The construction of robust psychometric instruments (Smith, Fischer, Vignoles, & Bond, 2013) enabled cross-cultural comparative research. As a consequence, the hypothesis regarding the cross-cultural universality of these models emerged (Allik, Realo, & McCrae, 2013; Costa & McCrae, 2014; De Raad et al., 2010; Gurven, von Rueden, Massekoff, Kaplan, & Vie, 2013; Schmitt, Allik, McCrae, & Benet-Martínez, 2007).

From this information, it is possible to apprehend two limitations in the more traditional research strategies adopted under the lexical approach in the study of personality. The first is that the taxonomic models of personality traits were substantially derived from the lexicon of the English language and the replicability of the models in other cultures was verified mostly by translation, adaptation and collection of evidence of validity and reliability in other languages and cultures. This perspective, named *etic* approach, denotes an universalist perspective concerned with the replicability of hypothetically universal personality models in different cultural contexts (Allik, Massoudi, Realo, & Rossier, 2012; Cheung, Van de Vijver, & Leong, 2011; Church, 2008; Valchev et al., 2012).

The second limitation is that the research work occurred mostly in non-naturalistic scenarios, where the use of words for personality description was measured in researcher-controlled test settings, starting from a pre-selected set of words withdrawn from dictionaries, especially adjectives (De Raad & Mlacic, 2015). The review of the literature conducted by Passos and Laros (2014), for example, confirms this limitation

as it reveals that 78.6% of the Big Five studies between 2008 and 2013 used surveys for data collection.

In this sense, this article aimed to explore these two sets of limitations, since they are central to the theoretical and empirical development of research with the lexical approach. In the next section, we present a brief history of the lexical approach with the objective of describing its underlying theoretical and methodological aspects. Subsequently, we discuss some of the leading criticisms towards the lexical approach and, finally, we present potential contributions from cross-cultural psychology and from the psychological study of natural language to address the methodological issues discussed.

Historical aspects of the development of the lexical approach

The pioneers: Galton, Rümelin, Klages, Partridge and Perkins.

Galton, in England, is credited as the first scientist to have had the idea that would later be known as the lexical hypothesis (John et al., 1988). In the paper *Measurement of Character,* Galton (1884) presented the idea of consulting a dictionary to obtain a notion of how many words could express the most obvious aspects of a person's character. He estimated that at least a thousand words in the English language would serve for this purpose, each with different shades of meaning and with senses shared with some other words. Galton (1884) argued that the simplest and most accurate measure of character should be based on the statistics of the behavior of individuals in routine activities carefully recorded, verified, evaluated, and re-evaluated. Nevertheless, Galton also warned that caution is needed regarding the use of different words to distinguish character aspects. Shortly afterwards, in Germany, Rümelin (1890, as cited in De Raad & Mlacic, 2015) made a similar suggestion.

After Galton's work, other efforts were made to develop lists of words on basis of lexicons that would represent personality traits. Examples are the work of Partridge (1910) and Perkins (1926) in the United States. Partridge argued that the study of mental differences could begin with the dictionary, with the analysis of the exact meaning, etymology, and the frequency of words by people to describe mental traits. To that end, Partridge identified 750 words. Similarly, Perkins identified about 3,000 words in the dictionary that presumably expressed traits and ideals, and planned to perform analyses with adjectives that describe the traits, and adverbs that describe modifying actions. Perkins planned to organize them according to their current use, considering its obsolescence, its social desirability, and the groups of meaning. Perkins' idea was to make use of the words that people developed to express the traits to study personality and character.

However, it is the German philosopher and psychologist Klages the author of the work which is considered the first formal articulation of the lexical approach in the study of personality, published in 1926 (English version of 1929). In his book on personality theory, Klages (1929) argued that the instrument of psychological discovery resides in the natural language developed throughout the history of humankind, from generation to generation. According to him, the natural language offers words that denote from the simplest to the most complex processes, conditions, and properties of what he called the inner life. For him, it would be an act of arrogance to attempt to invent a psycholinguistic terminology, for it would lead only to poor interpretations and distortions of the meanings of existing words. Klages considered that in the German language there would be approximately 4,000 words to describe the traces of the so-called internal states. It is interesting to note that Klages referred in his book to "traits of character which have been fixed in language" (p. 47).

Baumgarten (1933) and Allport and Odbert (1936).

Later, the German Franziska Baumgarten sought to test systematically the proposition made by Klages (1929) regarding the use of words to describe personality. According to John et al. (1988), she published in 1933 an extensive list of terms describing personality traits derived from both German-language dictionaries and German publications related to the study of character. Baumgarten selected the terms that she judged to be the most frequently used to describe personality, but did not propose any classification criteria or classified them in any way. Her list consisted of 941 adjectives and 688 nouns.

The work of Klages and Baumgarten in Germany had a significant influence on the subsequent development of the trait approach to the study of personality in the United States (John et al., 1988). Allport and Odbert (1936) cited directly the affirmation of Klages (1929) that the examination of words and phrases would allow greater knowledge than could be provided by observation, apparatus, and experiments. In referring to the "problem of trait-names" (p. V), Allport and Odbert argued for the necessity of knowing whether the terms adopted to describe personality are in fact referring to qualities or attributes denoting dispositions or psychological traits or whether they are just hypotheses and verbal ambushes. They also argued that the solution to seeking, identifying, and naming mental structures and substructures was the use of verbal symbols, even if they are ambiguous and problematic.

Allport and Odbert (1936) elaborated a list with terms relevant to the study of personality from a dictionary of the English language. The criterion to include a term in the list was its ability to discriminate the behavior of one human being from another (e.g., *affectionate, weak, irascible*). The list of Allport and Odbert (1936) contained 17,953 terms organized - often arbitrarily - into four categories or columns. Column I

lists 4,504 “neutral terms designating possible personal traits” (p. 38), that is, terms that “symbolize the most clearly ‘real’ traits of the personality” (p. 26). These terms, as well as the definition of stable traits as internal and causal trends, served as a guide for most subsequent taxonomic research (John et al., 1988). Column II listed 4,541 terms that would describe temporary moods or activities. Column III is composed of 5,226 weighted terms regarding social or character judgments of individual behavior or designating influence on others. For last, Column IV contained 3,682 metaphorical or ambiguous terms that did not meet the criteria for inclusion in the first three columns.

Raymond B. Cattell and the 16 primary personality factors.

In 1945, Cattell began efforts to apply factor analysis techniques to measure personality, with the goal of deriving the psychological equivalent of the periodic table (Revelle, 2009). The initial work of Cattell was to reduce the number of terms listed by Allport and Odbert (1936). He started with Column I, with the primary objective of discovering major dimensions of personality in the English language (Cattell, 1943; De Raad & Mlacic, 2015; John et al., 1988). Cattell (1943) added more than a hundred terms regarding temporary states and removed several words considered rare or archaic from the list of Allport and Odbert.

Then, Cattell conducted a series of studies of semantic reduction and factor analysis (Cattell, 1943, 1945, 1947). After over a decade of research, Cattell (1957, cited by John et al., 1988) developed the famous instrument *Sixteen Personality Factors Questionnaire* or 16PF (De Raad & Mlacic, 2015), composed of 12 personality factors, added to four dimensions specific to the domain of the questionnaire. John et al. (1988) provide a detailed review of Cattell’s efforts. These authors warned that Cattell adopted

several arbitrary procedures and did not present in his numerous publications more detailed information allowing the replication of the method of his studies.

Throughout his work, Cattell (H. E. P. Cattell and Schuerger, 2003) defended that psychology should develop measurement procedures for three distinct domains: personality, ability, and motivation or dynamical drives. Cattell also identified three primary data sources to explore these domains. The first source is L-data, which is the observation and recording of information about the behavior of subjects in natural or real-life settings. The second source is Q-data or questionnaire data, which consists of self-description information obtained in response to multiple-choice or open-ended questions. The third source is T-data or data from objective tests, which involves objective measurement of behavior such as standardized tests and experiments that do not require any self-examination by the subject. Cattell defended that the three sources are complementary and used them to identify personality traits in his studies.

Fiske (1949), Tupes & Cristal (1961), and Norman (1963).

The work of Cattell motivated several further studies that sought to replicate the model with 16 primary personality dimensions. However, these attempts have repeatedly failed. Waller (1999) cites three replication studies of Cattell's model that became influential in the later research with the lexical approach: Fiske (1949), Tupes and Cristal (1961), and Norman (1963). These three studies have in common the fact that they have reached solutions with five orthogonal factors to explain the sources of variance in Cattell's scales, and that the interpretation of the solutions showed a high degree of internal consistency.

Fiske (1949) named the five dimensions he found (i) Confident Self-Expression, (ii) Social Adaptability, (iii) Conformity, (iv) Emotional Control, and (v) Inquiring

Intellect. Tupes and Crystal (1961) gave other names to the five factors they found: (i) Surgency, (ii) Agreeableness, (iii) Dependability, (iv) Emotional Stability, and (v) Culture. Finally, Norman (1963) adopted the nomenclature of Tupes and Cristal (1961), modifying only the name of the third factor for Conscientiousness. Goldberg (1993) refers to Fiske as the "accidental discoverer" and to Tupes and Cristal as "the true fathers" of the five-factor personality model. According to De Raad and Mlacic (2015), since the study of Norman (1963) the five factors found in these studies were referred to as "Norman five" and later were called "Big Five" by Goldberg (1981).

As these studies started from Cattell's list of 35 clusters of terms, Tupes and Cristal (1961) and Norman (1963) cogitated the possibility of finding additional or different dimensions beyond the five mentioned factors (Waller, 1999). According to Waller (1999), this problem led Norman (1963) to suggest that it was necessary to make a return to the complete set of traits in natural language with the objective of finding new personality indicators that would not be present in the five factors.

Warren T. Norman.

As Waller (1999) described, Norman (1967) made a return to the dictionary with the aim of developing a new taxonomy "sufficiently exhaustive, precise, and well-structured to be useful for purposes of scientific communication and assessment" (Norman, 1967, p. 2). Norman intended to reach an exhaustive taxonomy in the sense that it had as source of data the "set of all perceptible variations in performance and appearance between persons or within individuals over time and of varying situations that are of social significance, of sufficiently widespread occurrence and retained as a subset of descriptive predicates in the natural language" (Norman, 1967, p. 2). Norman defended a precise taxonomy in two ways: (i) to exclude vague or ambiguous terms, and

(ii) to organize the terms in well-defined subsets based on criteria such as the evaluation of similarities of meanings, desirability, endorsement probabilities, and level of difficulty, among others. Finally, the taxonomy pursued by Norman should be well structured in the sense that the relationship between the groups of terms could be determined.

Initially, Norman (1967) complemented the list of Allport and Odbert (1936) with a search in a dictionary that resulted in the identification of 9,046 additional terms. Many of these new terms already had variations present in the existing list. This way, Norman's initial list was composed of 18,125 terms. This list was then submitted to a series of dimension reduction analyses. Norman delimited 12 categories of terms, organized into four sets: (i) stable biophysical traits; (ii) temporary status and activities; (iii) social roles, relationships and effects; and (iv) other excluded categories. An agreement from three of the four judges participating in the research was required to classify a term into a category.

The category of excluded terms was composed of evaluative, physical or anatomical descriptors, ambiguous, vague, obscure, rare, and difficult terms. In this round of reduction, 60% of the terms from the initial list were excluded. As a result, Norman (1967) came up with a list of 8,081 terms divided into the first three categories. Each category consisted of primary trait terms, difficult terms, and trait terms that were slang or weird. The first category of stable biophysical traits with 2,797 terms was used for data collection. The subsequent analyses led Norman (1967) to exclude another 1,200 terms, whose degree of knowledge, validity (i.e., whether the term was commonly used), and content specificity were considered uncertain or dubious. Finally, 1,566 terms remained in Norman's taxonomy.

Lewis Goldberg.

After the work of Norman (1963, 1967), Goldberg (1981, 1982, 1990, 1992) would formalize some of the foundations of the lexical approach to the study of personality. Goldberg's studies would also shape the methodological aspects that are contemporaneously adopted in the field (De Raad & Mlacic, 2015). The elaboration of the axiom of the lexical hypothesis is attributed to Goldberg (John et al., 1988):

> Those individual differences that are of the most significance in the daily transactions of persons with each other will eventually become encoded into their language. The more important is such a difference, the more will people notice it and wish to talk about it, with the result that eventually they will invent a word for it. (Goldberg, 1981, p. 142)

For Goldberg (1981), this axiom has a corollary:

> The more important an individual difference is in human transactions, the greater the number of languages that will have a term for it. In the strongest form of this corollary, we should find a universal order of emergence of the individual differences encoded into the set of all the world's languages. (p. 142)

In addition to formalizing the hypothesis that personality traits are encoded in languages and that regularity exists between different languages in that sense (i.e., universal personality traits), Goldberg (1981) also discussed some methodological issues central to the study of structural personality models. Goldberg argued that a dimensional or ordered description (e.g., adjectives) is preferable to a discrete or categorical one (e.g., nouns), once it would allow a description in terms of a continuum and the use of categories or types as special cases. Goldberg argued that the dimensions of individual differences should be analyzed both from the perspective of the unipolarity

of terms (i.e., two antonyms would be considered two separate dimensions), and their bipolarity (i.e., two antonyms would be considered as two poles of the same dimension). Regarding the association between personality dimensions, Goldberg argued that orthogonal models (i.e., with uncorrelated dimensions) would be the best for predictive purposes, but oblique models (i.e., with correlated dimensions) would be more realistic. As for the level of description, Goldberg argued that it should be neither very concrete (i.e., specific) nor very abstract (i.e., global). Finally, Goldberg reasoned that the search for the universal lexicon of personality must start from a previously defined structure.

Starting from a list of 2,797 terms that comprised the category of stable biophysical traits developed by Norman (1967), Goldberg (1982, 1990) elaborated a list of 1,710 adjectives. Subsequently, Goldberg (1982) re-filtered his list, eliminating terms that did not appear as entries in an English dictionary, and excluded ambiguous, slang, unisex, over evaluation, metaphorical, difficult, and redundant terms. Goldberg (1982) also added 61 other terms to the list, reaching 566 terms.

Latterly, Goldberg (1990) conducted three studies with the aim of demonstrating the generality of the five-factor personality model. In the first study, he used a list of 1,431 terms grouped into 75 *clusters* proposed by Norman (1967). In the second Goldberg adopted a list of 479 terms considered most common, and arranged them in 133 *clusters* of synonyms. Lastly, in the third study, he used another list of 100 *clusters* grouping 339 terms. Goldberg concluded that there was sufficient evidence for a general structure of the Big Five.

In 1992, Goldberg organized a scale with 100 unipolar markers and one with 50 Big Five bipolar markers. The objective was to replace the markers proposed by Norman (1963) and offer an alternative to the NEO (Costa & McCrae, 2014) and Hogan (cited by Goldberg, 1992) scales. Besides that, Goldberg (1992) also aimed to construct

standardized markers for the Big Five. Goldberg continues to contribute to the development of the lexical approach.

Paul Costa & Robert McCrae.

The work with the lexical hypothesis of the Costa & McCrae duo had as a starting point a *cluster* analysis of the 16PF of Cattell with the objective of determining possible differences in the structure of personality related to age (Costa & McCrae, 1976). Costa and McCrae came to two recurring *clusters*, Extraversion and Neuroticism, and a third dimension that they interpreted as Openness to Experience. In 1985, Costa and McCrae (cited by De Raad & Mlacic, 2015) published the NEO Personality Inventory (NEO-PI), dedicated to measuring these dimensions. Later, in 1989, they published a new version of the instrument adding the Agreeableness and Conscientiousness dimensions, called NEO Five-Factor Inventory (NEO-FFI), which was followed by at least three versions of NEO (Costa & McCrae, 2014).

Goldberg (1993) affirmed that the efforts of Costa and McCrae during the 1980s allowed both to become the most influential and prolific proponents of the five-factor model. In fact, in 2016, the duo was among the 300 scientists most cited in the world in all fields (Webometrics, 2016). For Goldberg, this success was due to the large number of articles the pair had published and especially to the strategy they adopted to use the NEO scales as a reference to integrate several other systems and instruments of personality.

Besides the psychometric work with NEO scales, Costa and McCrae also endeavored to formulate a theory of personality with the Big Five (McCrae, 2011). The duo participated in more than a hundred articles and book chapters seeking to present and defend this theory. This production can be understood from the argument that Costa

and McCrae have developed at least five lines of evidence that support the model (Costa & McCrae, 1992; McCrae, 2011). First, the Big Five would represent long-lasting dispositions through behavioral patterns in longitudinal and cross-observer studies. Second, the five factors would be related to several personality systems, such as the Minnesota Multiphasic Personality Inventory and the Myers-Briggs Type Indicator, for example (De Raad & Mlacic, 2015). Third, the Big Five would be universal, found in different groups of sex, race/color, age, and language, even if there is some variation between cultures. Fourth, evidence of heredity suggests a biological basis for the five factors. Fifth, evidence of studies that the *Big Five* are important influences in various aspects of people's lives, such as vocational interests, religiosity, drug use, etc.

The return to Europe: Dutch and German taxonomic projects.

The lexical approach in the study of personality would return to Europe in the mid-1970s, especially in The Netherlands and later in Germany, at about the same time as the work of Goldberg and Costa and McCrae took shape. De Raad and Mlacic (2015) call the experiences in these two countries as Dutch and German taxonomic projects. In The Netherlands, Hofstee and Brokken were the pioneers in the development of what John et al. (1988) identified, at the time, as the only taxonomy that was not based on the English language. According to John et al., the work of the Dutch group sought to avoid subjective decisions such as those that marked, for example, the work of Cattell. Therefore, some methodological strategies were developed. John et al. (1988) highlighted that the Dutch team developed procedures to secure the objectivity of the identification process that specify the domains; to ensure that the structures found can be generalized among judges and data sources; and to improve the interpretation of factors and other structural categories through consensus data obtained independently.

Inspired by the work of Norman (1967), the Dutch research group began its work with the inspection of a dictionary by two independent researchers, which resulted in a combined list of 8,690 words, according to John et al. (1988). The group excluded difficult-to-understand adjectives, jargon, metaphorical or purely evaluative terms, medical, physical or anatomical terms, and terms representing moods or temporary states. In the end, 2,635 terms were consensually excluded. The next stage consisted of the classification of the terms, a procedure performed through operational definitions expressed in sentences that would represent significant and heuristic criteria to retain only relevant terms for personality description (John et al., 1988).

Two criteria were adopted to classify the terms (Brokken, 1978, as cited in De Raad & Mlacic, 2015). The first was the Nature criterion, which states that an adjective should fit meaningfully in the sentence to be considered a useful descriptor for personality: "He (She) is ... by nature" (De Raad & Mlacic, 2015, p. 8). The second was the Person criterion, which states that an adjective relevant for personality description should answer the question, "Mr./Ms. X., what kind of person is he/she?" (De Raad & Mlacic, 2015, p. 8). Later, the same type of identification sentence was used for the development of a taxonomy of verbs (De Raad, Mulder, Kloosterman, & Hofstee, 1988), while taxonomy of nouns followed criteria relating to the noun's ability to describe, typify, or characterize a person (De Raad & Hoskens, 1990).

De Raad and Mlacic (2015) described that further Dutch studies have shown evidence that a five-factor structure is clearer for the list of adjectives, while verbs and nouns can be interpreted by some of the Big Five or by a mixture of them. De Raad and Mlacic also argued that the Dutch taxonomy influenced lexical studies in other languages, such as the Italian and Hungarian.

The German project (Angleitner, Ostendorf, & John, 1990) initiated its efforts from the analysis of a dictionary of the German language, searching for adjectives and nouns that would represent personality types and attributes. The operational definition of Allport and Odbert (1936) that a term would be relevant to the personality if it distinguished the behavior of two individuals, as well as those of the Dutch group, were adopted by Angleitner et al. (1990). To these operational definitions, the German researchers also adopted specific sentences for the nouns. Terms applicable to all human beings (e.g., born), terms referring to the geographical origin, nationality, and occupation of individuals, as well as metaphorical terms or that described only one part of the person's body were excluded (Angleitner et al., 1990).

The German list was divided into six categories: (i) stable traits; (ii) states and moods; (iii) activities; (iv) social aspects of personality; (v) talents and abilities, and (vi) appearance. De Raad and Mlacic (2015) report that the findings of later German studies corroborated the five-factor structure and that the German method influenced most of the taxonomies developed later in languages such as Italian, Czech, Polish, Filipino, Croatian, Slovak, and Spanish.

The lexical approach and cross-cultural psychology

The acronym WEIRD is often used to refer to the problem of bias in the selection of research samples in studies published in the world's most influential journals in the behavioral sciences. According to Henrich, Heine, and Norenzayan (2010a, 2010b), in the leading periodicals of the field, 96% of the participants were WEIRD, that is, from Western, Educated, Industrialized, Rich, and Democratic countries. Henrich et al. (2010a, 2010b) argue that these subjects have particularly unusual characteristics if compared to the rest of the human species, often being outliers. Similarly, in the

journals with the greatest impact in the area of personality psychology, according to Allik (2013), most authors are linked to institutions from the United States of America, as are the journals themselves. Besides that, there is also a growing tendency for self-citation in these publications. This information corroborates the perception that in the history of the development of the lexical approach, the majority of researchers, theoretical models, and psychometric instruments most influential internationally are WEIRD, mainly from the United States of America.

In this context, cross-cultural psychology seeks to promote advancements in psychological science by pondering the weight of culture. The development of this field can be divided into three phases, which are also major objectives of the area (Cheung et al., 2011; Smith et al., 2013). The first is the *etic* imposition and concerns the investigation of the generality and validity of models and theories developed in WEIRD countries in other cultural contexts. The second is the *emic* approach or indigenous (i.e., autochthonous or native) psychology, which is focused on the study of phenomena specific to cultures and on the investigation of the validity of theories that are intended to be universal (Cheung et al., 2011). The third is the emic-etic approach, which seeks to approximate and integrate the two first objectives (Cheung et al., 2011). Daouk-Ory, Zeinoun, Choueiri, and Van de Vijver (2016) also call the *emic* approach as *local*, the *etic* approach as *global*, and the *emic-etic* as GloCal. Relatedly, the study of personality in the cultural context, including research with the lexical hypothesis, can be understood in terms of these three approaches (Cheung et al., 2011; Church, 2008).

The lexical approach can be considered as fundamentally *emic*, once it aims to derive local personality latent dimensions (Cheung et al., 2011; Daouk-Ory et al., 2016). Usually, the research protocol involves the analysis of the dictionary of a given language, followed by the elaboration of lists of personality descriptors terms, and

dimensionality reduction analyzes. The data is also local since participants rate themselves and others using the terms previously identified.

However, after its initial development up to the 1990s in WEIRD countries (e.g., the United States, The Netherlands, and Germany), psycholexical research has been conducted from a predominantly *etic* perspective. The focus was extended to outside the culture where the models were developed. The usual research strategy involves the translation and adaptation of terms and scales developed in WEIRD countries (Daouk-Öyry et al., 2016; Gurven et al., 2013). In Brazil, for example, there are numerous examples of the adaption or construction of scales starting from foreign models (e.g., Andrade, 2008; Passos & Laros, 2015; Hauck et al., 2012; Hutz et al, 1998).

This approach is also exemplified by large-scale cross-cultural studies, which have WEIRD models, markers and instruments as a reference (e.g., Allik & McCrae, 2004; Bartram, De Fruyt, Bolle, McCrae, Terracciano, & Costa, 2009; De Raad et al., 2010; McCrae & Terracciano, 2005a, 2005b; Schmitt et al., 2007; Zecca et al., 2012). These cross-cultural *etic* studies are based on tests of invariance or equivalence of measures that assess the multi-group comparability of constructs or scores (Byrne & Van de Vijver, 2014; Church et al., 2011). Cheung et al. (2011) emphasize that in addition to its integrator potential, the great advantage of these cross-cultural *etic* approaches is the large databases.

Despite its potential, there are different criticisms regarding the *etic* perspective (Cheung et al., 2011, Church, 2008, Church et al., 2011). Let us take the Big Five case as an example. As described by Daouk-Ory et al. (2016), on the one hand there is a set of evidence collected about the universality of this model, especially in Germanic and Romantic languages such as English, Dutch, German, French, Italian, and Spanish (Allik et al., 2013; Cheung et al., 2011; Fetvadjiev & Van de Vijver, 2015). On the

other hand, there are several studies that have failed to find evidence to prove the stability of the five factors and their facets between different cultures. For example, there are studies that found models with three (De Raad et al., 2010), six (Nel et al., 2012), seven (Almagor, Tellegen, & Waller, 1995), and nine factors (Nel et al., 2012, Daouk-Ory et al., 2016). In relation to this problem of cross-cultural replicability, De Raad et al. (2010) concluded that they do not believe "that a final canonical response can be given, considering the incompatibility of specific language structures, each with different trait variables and different participants" (p. 171). Church et al. (2011), whose study focused on the question of the invariance of the measure, concluded that the issue of the validity of cross-cultural comparisons has not yet been resolved.

Cheung et al. (2011) consider that in addition to the question of measurement invariance, there are other methodological limitations of cross-cultural nature in the *etic* approach. These restrictions are related to the constructs (e.g., differences in constitutive and operational definitions of constructs), method (e.g., differences in response styles), and the items (e.g., differential item functioning). Additionally, there is a gap between the theoretical development regarding cross-cultural differences and the explanations of the results of equivalence tests (Cheung et al., 2011). That is, it still too early to theorize about cross-cultural differences and similarities regarding the constructs, differential item functioning, relations between factors, and error variances. Daouk-Ory et al. (2016) go further and argue that even the axiom and corollary of the lexical hypothesis formalized by Goldberg (1981) are challenged, since there is evidence that single words are not sufficient to represent all relevant terms of personality, for example. Daouk-Ory et al. conclude that this paradigm produced “results that are neither culturally specific, nor adequately comparable across cultures” (p. 6).

The lexical hypothesis also inspired intranational investigations of local languages that aimed to identify autochthonous constructs and structures in non-WEIRD countries (Smith et al., 2013). Saucier and Goldberg (2001), Church et al. (2011), and Daouk-Öyry et al. (2016) present reviews that mention some of *emic* studies, citing productions from various countries. Examples are: Cheung et al. (2001) and Cheung, Van de Vijver, and Leong (2011) in China; Isaka (1990) in Japan; Katigbak et al. (2002) in the Philippines; Nel et al. (2012), Valchev et al. (2012), Valchev, Van de Vijver, Nel, Rothman and Meiring (2013) in South Africa; and Ortiz et al. (2007) in Mexico.

While there is evidence of culturally universal components of the personality (Allik & McCrae, 2004; McCrae & Costa, 1997), there is also evidence of components specific to cultures (Cheung et al., 2001; Saucier & Goldberg, 2001). In the perspective of cross-cultural psychology, *etic* and *emic* approaches can be integrated. Such integration may result in the development of a personality theory that incorporates universal (i.e., common) and unique (i.e., culturally-specific) aspects of languages, as Cheung et al. (2011) and Daouk-Öyry et al. (2016) defend. To achieve this integration, Cheung et al. (2011) advocate combining culturally specific components to the models developed in WEIRD countries. Daouk-Öyry et al. (2016) amplify this perspective by defending that there is not an exclusively *etic* or *emic* focus, but that the knowledge of each language and culture should devote the methods to be employed, thereby allowing unique and universal components to emerge that would be otherwise restrained. Integrative proposals of these two perspectives have been explored, as described in studies such as Allik et al. (2011); Arzu, Lee, Ashton, and Somer (2008); Cheung et al. (2008); Benet-Martinez and John (2000), De Raad, Blas, and Perugini (1998); De Raad, Perugini, Hrebícková, and Szarota (1998); and Di Blas and Forzi (1998).

The lexical approach and the psychological study of natural language

The hypothesis that the personality taxonomy is found in the natural expressions of language (Goldberg, 1981; Klages, 1929) may potentially fail to be genuinely investigated with questionnaires with restricted sets of traits (i.e., items) pre-selected by scientists and administered in controlled scenarios. As Cheung et al. (2011) highlighted, the psycholexical studies typically made use of the dictionaries as the primary source of personality traits. This perception is corroborated by Daouk-Öyry et al. (2016), who did a systematic review of studies published between 1970 and 2012 and classified them regarding the sources used for personality descriptors selection. Daouk-Öyry et al. affirmed that 80% of the studies with an *etic* approach and 84% of those with an *emic* approach used dictionaries as the source of trait-descriptive terms. This strategy, however, is not necessarily suitable for the investigation of all languages and cultures, as shown by the study by Nel et al. (2012), for example. To investigate 11 South African languages they had to use interviews as a primary source since there were no consolidated dictionaries for those languages.

Uher (2013, 2015a, 2015b, 2015c) makes a series of criticisms and propositions for the psycholexical approach from epistemological, meta-theoretical, and methodological nature perspectives. One of the criticisms is that the use of standardized instruments in the lexical approach can have as a consequence a departure from the physical representations and daily beliefs of lay people regarding personality traits. It is interesting to note that the theorists of the lexical approach have already argued against the criticism made precisely by the frequent use of lay observers in psychological research (Ashton & Lee, 2005). For Uher (2015c), "Research on 'personality' is intimately connected to people's everyday beliefs, not only because beliefs form part of the set of phenomena commonly conceived of as 'personality'" (p. 643). Thus, Uher

(2013) advocates adopting an approach that brings together the behavioral repertoire and the environmental situations. This defense is very close to that made by natural language scholars, as we shall see below.

The research field related to the study of natural language and its relations with psychological variables (e.g., personality) has been growing (Pennebaker, Mehl, & Niederhoffer, 2003). The approaches developed in this field are promising solutions to tackle issues such as those raised by Daouk-Öyry et al. (2001) e Uher (2013, 2015a, 2015b, 2015c). According to Park et al. (2015) and Tausczik and Pennebaker (2010), more than 100 studies were published exploring the link between language use and a number of psychological correlates. Pennebaker et al. (2003) in their review of the psychological aspects of the use of natural language, state that by *natural* they refer to "relatively open-ended responses to questions, natural interactions, and written or spoken texts" (p. 549). That is, open situations and registers that can capture a free linguistic expression.

In the research with the lexical hypothesis, it is possible to find several examples of sources of personality descriptors that made use of natural language, not having been restricted to the dictionaries or lists of adjectives, for example. Daouk-Öyry et al. (2016) elaborated the most comprehensive list available with these types of sources, which include oral records, print media, literary texts, etc. Allik et al. (2011) used literary and academic texts; Cheung et al. (1996) used literary texts, proverbs and the spoken language; and Nel et al. (2012) and Valchev et al. (2013) recorded and transcribed the audio of interviews. Polzehl (2015), on the other hand, published a book entirely devoted to the automated evaluation of the personality through voice recording and speech analysis, including acoustic measures.

The most common methods for natural language analysis can be divided into three broad sets (Pennebaker et al., 2003). The first is the judgment based on the thematic analysis of content, which involves the identification of thematic references in text samples based on empirically defined coding systems. The second set consists of word pattern analysis, which emerged in the context of artificial intelligence. Park et al. (2015) named this strategy an open approach. These methods exploit text from patterns identified by covariance between large text samples. That is, without previously defining categories of words or psychological dimensions. Pennebaker et al. (2003) highlighted the method of latent semantic analysis (LSA), which would be similar to a factorial analysis of individual words.

The third set of methods identified by Pennebaker et al. (2003) includes word counting strategies. These are based on the assumption that words carry information that is beyond its literal meaning and that are independent of its semantic context, thus involving both the content (i.e., what is being said) and the style (i.e., how it is being said, for example, passive or active voice, use of metaphors, etc.). Park et al. (2015) named this third set as closed-vocabulary approach. One of the most widely adopted methods of this set is the Linguistic Inquiry and Word Count - LIWC (Tausczik & Pennebaker, 2010).

Specifically, in the field of personality, Yarkoni (2010) identified three challenges related to research involving natural language. The first concerns the access of naturalistic textual samples, written or spoken (i.e., texts produced in natural circumstances). According to Yarkoni, most studies are based on non-naturalistic samples, that is, researchers require participants to speak or write about a specific subject (e.g., describing themselves and others, recounting their life history, etc.).

The second challenge identified by Yarkoni (2010) is related to the size and scope of the textual samples. The author points out that most studies in the field were restricted to only a few thousand words per participants, which would prevent a more reliable estimation of the frequency only for aggregate categories, not regarding the use of individual words. Another limitation of most studies is that they usually adopt textual samples from one or a few occasions, which does not allow the analysis of the time stability between the use of words and personality. Likewise, the results may be susceptible to the influence of more transient factors, such as humor (Mohammad & Kiritchenko, 2015).

The third challenge identified by Yarkoni (2010) concerns the modeling of the relationship between language and personality in a more detailed fashion since most studies carry out modeling broad semantic categories such as the Big Five factors. Usually, the researcher adds large sets of words in these categories, rather than analyzing them individually. For Yarkoni, this approach may hide the specificity of the relationship between the use of words or specific categories of words and personality, limiting the discovery of new relationships. That is, by using broad categories defined *a priori*, it is possible that the probability of finding such relationships is being limited.

Lexical approach, natural language, and online social media.

A new research front has been developed to overcome the mentioned challenges, seeking to take advantage of the current computational resources and the large volume of data (i.e., big data) with naturalistic records of human behavior in virtual environments, such as social media. For instance, only in one of the available social networks, Twitter, users make approximately 6,000 posts per second, 350,000 per minute, 500 million per day, and 200 billion per year (Internet Live Stats, nd).

Park et al. (2015) point out the potential benefits of exploiting this gigantic volume of data from social media. Like Yarkoni (2010), Park et al. argue that social media is a natural social setting that is part of the daily lives of many people. Second, this data can be easily accessed, even retroactively, which avoids the costs of large sample studies. Third, social media users offer a great deal of information about themselves - we also highlight that users are also evaluating other people. Fourth, citing Back et al. (2010), Park et al. (2010) argue that people usually present their true self in networks, not just their idealized versions.

Yarkoni (2010) investigated the relationship between language and personality in blogs. The author invited by e-mail about 5,000 users of a blog hosting system, resulting in a sample of 694 blogs written in English. Participants completed a demographic questionnaire and two questionnaires measuring the dimensions and facets of the Big Five (Goldberg et al., 2006). Two analytical procedures were conducted, one at the level of categories of words and another at the level of words.

In the first procedure, Yarkoni (2010) analyzed the correlations between the Big Five factors and 66 categories of words (e.g., negative emotions and affections, such as sadness, anxiety, etc.) defined in the dictionary of the LIWC software (Tausczik & Pennebaker, 2010). In the second procedure, Yarkoni investigated the correlations between the Big Five measures and 5,068 words, which were selected based on two criteria: words from blogs with more than 50,000 words; and words whose frequency was greater than 5,000, considering all blogs. Yarkoni pointed out that most of the words, (i.e., about 10,000), had a frequency lower than two, that is, they appeared only once in the analyzed texts. Other examples of studies with blogs are Li and Chignell (2010) and Iacobelli, Gill, Nowson, and Oberlander (2011).

Qiu et al. (2012) investigated how personality is manifested and perceived on Twitter. They analyzed 142 participants recruited through a "snowball" sampling procedure involving college campus participants in exchange for credit, and through a virtual workspace called Amazon Mechanical Turk, in exchange for US$ 0.50. They expected participants with between 20 and 1,000 English-language posts in a given period of one month. The participants of the study of Qiu et al. (2012) answered the Big Five Inventory and a demographic questionnaire, and also informed their identity as Twitter users. To assemble the database, the researchers manually copied and pasted into text files 28,978 posts, an average of 204.07 posts and 2,362.72 words per participant. In addition to these data, a hetero-report was realized by eight assistant research students using the Big Five Inventory containing the evaluation of the participants' posts. As in the study conducted by Yarkoni (2010), Qiu et al. (2012) processed the data using the LIWC software (Tausczik & Pennebaker, 2010), adopting the categories of words available in it. Mohammad and Kiritchenko (2015) also conducted studies with Twitter investigating the lexicon of emotions expressed through keywords (i.e., #hashtags) in user posts and associating them with personality measures.

Park et al. (2015) analyzed the written language of 66,732 users of the social network Facebook. The data were collected over a period of approximately two years and consisted of all status messages written by participants. In total, more than 15 million messages and 4,107 words on average per user were collected. Participants also answered questionnaires related to the Big Five factors. The analyses involved three steps: data extraction, data dimensionality reduction, and regression modeling and machine learning. The data were organized into 24,530 words and phrases and 2,000 topics. The results of the dimensionality reduction analyses were combined and used as predictors of the factors and facets of the Big Five. The data found in the natural

language on Facebook were therefore used to predict the Big Five. Other studies related to personality evaluation using Facebook were reviewed by Limas, Primi, and Carvalho (2014), and Carvalho and Pianowski (2017).

Poddar, Kattagoni, and Singh (2015) adopted as a data source the biographical texts from a website about 574 famous personages of history (e.g., Einstein, Goethe). Through a technique they termed the "adjectival marker", they extracted adjectives that appeared in lists related to the Big Five model and Jung's personality typology. The personages (i.e., the texts referring to them) were separated into two groups, one training and one testing. Using regression models and machine learning, Poddar et al. evaluated the predictive power of adjectives concerning four Big Five factors (i.e., Agreeableness, Conscientiousness, Extraversion, and Imaginative), considering the personality classification of these characters made by http://www.celebritytypes.com/ as an independent variable. The model was evaluated using the correlation between the Big Five four-factor classification and the Jung typology. The results indicated accuracy above 80%.

Despite notable advances regarding the use of alternative data sources and statistical analysis presented in the studies cited above, some methodological challenges remain. Yarkoni (2010), for example, points out limitations to his study: selection bias (only part of the bloggers provided their electronic addresses, and, of these, only a portion agreed to participate in the study); the low magnitude of the identified correlations (the highest correlation identified between a Big Five factor and a word category was .23); and the method was based only on counting the frequency of the words, disregarding contextual and semantic factors. These same limitations are present in the studies of Qiu et al. (2012), Poddar et al. (2015), and, in part, of Park et al. (2015), for example.

Another common limitation is that Yarkoni (2010), Qiu et al. (2012), and Park et al. (2015) worked with a design in which they search for a previously defined personality model, the Big Five, instead of exploring the possibility of deriving a new model from the collected data. The question that remains is whether exploratory dimensionality analysis of these data would reveal a structure or constructs other than the Big Five.

Final Considerations

This work aimed to demonstrate that, although originating from the idea that most of the relevant personality traits would be encoded in natural language in different languages (Goldberg, 1981), the lexical approach historically has been devoted to deriving the vocabulary from the personality mainly from the examination of dictionaries (Daouk-Öyry et al., 2016). Thus, closed lists of words, especially adjectives, became the primary source of items in the form of sentences or markers for the construction of psychometric instruments. This practice left behind the rich source of information that the study of the use of natural language would be able to offer (Uher, 2013, 2015a, 2015b, 2015c). In addition, the concern about the development of a universal personality taxonomy for the whole human species has led researchers to adopt an *etic* methodological approach, conducted through surveys that seek to cross-culturally replicate models derived from WEIRD countries (Cheung, 2011).

To deal with these issues, we discussed methodological potentialities from two areas, the cross-cultural psychology and the psychological study of natural language. Cross-cultural psychology brings important contributions to the study of personality insofar as it recognizes the importance of both the investigation of common or universal aspects of personality, and of unique or culture-specific aspects as well. In this

integrative *emic-etic* perspective, it is advocated that the sources for obtaining personality trait-descriptive terms and the research methods should be selected considering the specificities of the language and culture under analysis.

In a complementary way, the study of natural language in psychology broadens the perspectives of analysis in personality research, diversifying the sources to obtain descriptive terms. Audios, videos, literary, academic and biographical texts, and recordings of human behavior in online social media have become alternatives for the lexical study of personality. With the advancing of methodological and analytical tools, the study of natural language in an integrative cross-cultural perspective seems to be the main developing path for the theoretical and empirical construction of the lexical hypothesis in the future.

References

Allik, J. (2013). Personality psychology in the first decade of the new millennium: A bibliometric portrait. *European Journal of Personality, 27*(1), 5-14.

Allik, J., & McCrae, R. (2004). Toward a geography of personality traits: Patterns of profiles across 36 cultures. *Journal of Cross-Cultural Psychology, 35*, 13-28.

Allik, J., Massoudi, K., Realo, A., & Rossier, J. (2012). Personality and culture. Cross-cultural psychology at the next crossroads. *Swiss Journal of Psychology, 71*(1), 5-12.

Allik, J., Realo, A., & McCrae, R. (2013). Universality of the five-factor model of personality. In P. T. Costa & T. Widiger (Eds.), *Personality disorders and the Five Factor Model of Personality* (pp. 61-74). Washington: American Psychological Association.

Allik, J., Realo, a., Mottus, R., Pullmann, H., Trifonova, a., McCrae, R. R., … Korneeva, E. E. (2011). Personality profiles and the "Russian Soul": Literary and scholarly views evaluated. *Journal of Cross-Cultural Psychology*, *42*(3), 372–389. doi:10.1177/0022022110362751.

Allport, G., & Odbert, H. (1936). Trait names: a psycho-lexical study. *Psychological Monographs, 47*(211), 1-38.

Almagor, M., Tellegen, A., & Waller, N. G. (1995). The Big Seven model: A cross-cultural replication and further exploration of the basic dimensions of natural language trait descriptors. *Journal of Personality and Social Psychology, 69*(2), 300-307. doi: 10.1037/0022-3514.69.2.300.

Andrade, J. M. (2008). *Evidências de validade do Inventário dos Cinco Fatores de Personalidade para o Brasil.* Tese de Doutorado, Universidade de Brasília, Brasília.

Angleitner, A., Ostendorf, F., & John, O. P. (1990). Towards a taxonomy of personality descriptors in German. A psycho-lexical study. *European Journal of Personality, 4*, 89–118.

Ashton, M. C., Lee, K., de Vries, R. E., Perugini, M., Gnisci, A., & Sergi, I. (2006). The HEXACO model of personality structure and indigenous lexical personality dimensions in Italian, Dutch, and English. *Journal of Research in Personality, 40*, 851–875. doi:10.1016/j.jrp. 2005.06.003.

Asthon, M. C., & Lee, K. (2005). A defence of the lexical approach to the study of personality structure. *European Journal of Personality, 19*, 5-24.

Bartram, D. (2013). Scalar equivalence of OPQ32: Big Five profiles of 31 countries. *Journal of Cross-Cultural Psychology, 44*(1), 61–83. doi:10.1177/0022022111430258

Benet-Martinez, V., & John, O. P. (2000). Toward the development of quasi-indigenous personality constructs: Measuring los cinco grandes in Spain with

indigenous Castilian markes. *American Behavioral Scientist, 44*(1), 141-157. doi: 10.1177/00027640021956035.

Boies, K., Lee, K., Ashton, M. C., Pascal, S., & Nicol, A. A. M. (2001). The structure of the French personality lexicon. *European Journal of Personality, 15*(4), 277-295. doi: 10.1002/per.411.

Byrne, B. M., & van de Vijver, F. J. R. (2014). Factorial structure of the Family Values Scale from a multilevel-multicultural perspective. *International Journal of Testing, 14*(2), 168-192. doi: 0.1080/15305058.2013.870903.

Carvalho, L. F., & Pianowski, G. (2017). Pathological personality traits assessment using Facebook: Systematic review and meta-analyses. *Computers in Human Behavior, 71*, 307-317. doi: 10.1016/j.chb.2017.01.061

Cattell, H. E. P. , & Schuerger, J. M. (2003). Essentials of 16PF assessment. Hoboken, NJ, USA: John Wiley & Sons, Inc.

Cattell, R. B. (1943). The description of personality: Basic traits resolved into clusters. *Journal of Abnormal and Social Psychology*, *38*, 476–507.

Cattell, R. B. (1945). The description of personality: Principles and findings in a factor analysis. *American Journal of Psychology*, *58*(1), 69-90.

Cattell, R. B. (1947). Confirmation and clarification of primary personality factors. *Psychometrika*, *12*, 197–220.

Cheung, F. M., Cheung, S. F., Zhang, J, Leung, K., Leong, F., & Huiyeh, K (2008). Relevance of Openness as a personality dimension in Chinese culture: Aspects of its cultural relevance. *Journal of Cross-Cultural Psychology, 39*(1), 81–108. doi:10.1177/0022022107311968.

Cheung, F. M., Leung, K., Fan, R. M., Song, W.-z., Zhang, J.-x., & Zhang, J.-p. (1996). Development of the Chinese Personality Assessment Inventory. *Journal of Cross Cultural Psychology, 27*(2), 181-199. doi: 10.1177/0022022196272003

Cheung, F., Leung, K., Zhang, J., Sun, H., Gan, Y., Song, W., & Xie, D. (2001). Indigenous Chinese personality constructs: Is the Five-Factor model complete? *Journal of Cross-Cultural Psychology, 32*, 407-433.

Cheung, F., van de Vijver, F. J. R., & Leong, F. (2011). Toward a new approach to the assessment of personality in culture. *American Psychologist, 66*, 593-603.

Church, A. (2008). Current controversies in the study of personality across cultures. *Social and Personality Psychology Compass, 2*, 1930-1951.

Church, A. T., Alvarez, J. M., Mai, N. T. Q., French, B. F., Katigbak, M. S., & Ortiz, F. A. (2011). Are cross-cultural comparisons of personality profiles meaningful? Differential item and facet functioning in the Revised NEO Personality Inventory. *Journal of Personality and Social Psychology, 101*(5), 1068-1089. doi:10.1037/a0025290

Church, A. T., Reyes, J. A. S., Katigbak, M. S., & Grimm, S. D. (1997). Filipino personality structure and the Big Five model: A lexical approach. *Journal of Personality, 65*(3), 477- 528. doi: 10.1111/j.1467-6494.1997.tb00325.x.

Costa, P. T., Jr., & McCrae, R. R. (1976). Age differences in personality structure: A cluster analytic approach. *Journal of Gerontology*, *21*, 564–570.

Costa, P. T., Jr., & McCrae, R. R. (1992). Four ways five factors are basic. *Personality and Individual Differences, 13,* 653-665.

Costa, P. T., Jr., & McCrae, R. R. (2009). The Five-Factor Model and the NEO Inventories. In: Butcher, J. N. (Ed.). *Oxford handbook of personality assessment* (pp. 299-322). New York: Oxford University Press.

Costa, P. T., Jr., & McCrae, R. R. (2014). The NEO inventories. In R. P. Archer, & S. R. Smith, (Eds.), *Personality Assessment* (2nd ed., pp. 229-260). New York: Routledge.

Daouk-Öyry, L., Zeinoun, P. , Choueiri, L., & van de Vijver, F. J. R. (2016). Integrating global and local perspectives in psycholexical studies: A GloCal approach. *Journal of Research in Personality*. Publicação *online* adiantada. doi: http://dx.doi.org/10.1016/j.jrp. 2016.02.008

De Fruyt, F. Bolle, M. D., McCrae, R. R., Terracciano, A., & Costa, P. T., Jr. (2009). Assessing the universal structure of personality in early adolescence: The NEO-PI-R and NEO-PI-3 in 24 cultures. *Assessment, 16*(3), 301-311. doi: 10.1177/1073191109333760

De Raad, B., & Hoskens, M. (1990). Personality-descriptive nouns. *European Journal of Personality, 4*, 131–146.

De Raad, B., & Mlacic, B. (2015). The lexical foundation of the Big Five – Factor Model. In T. Widiger (Ed.), *The Oxford handbook of the Five Factor Model.* Retrieved from www.oxfordhandbooks.com. doi: 10.1093/oxfordhb/9780199352487.013.12.

De Raad, B., Barelds, D., Levert, E., … Katigbak, M. S. (2010). Only three factors of personality description are fully replicable across languages: A comparison of 14 trait taxonomies. *Journal of Personality and Social Psychology, 91*(1), 160-173.

De Raad, B., Di Blas, L., & Perugini, M. (1998). Two independently constructed Italian trait taxonomies: Comparisons among Italian and between Italian and Germanic languages. *European Journal of Personality, 12*(1), 19-41. doi: 10.1002/(sici)1099-0984(199801/02)12:1<19::aid-per290>3.3.co;2-y.

De Raad, B., Mulder, E., Kloosterman, K., & Hofstee, W. K. B. (1988). Personality-descriptive verbs. *European Journal of Personality, 2*, 81–96.

De Raad, B., Perugini, M., Hrebícková, M., & Szarota, P. (1998). Lingua franca of personality: Taxonomies and structures based on the psycholexical approach. *Journal of Cross Cultural Psychology, 29*(1), 212-232. doi: 10.1177/0022022198291011

Di Blas, L., & Forzi, M. (1998). An alternative taxonomic study of personality-descriptive adjectives in the Italian language. *European Journal of Personality, 12*(2), 75-101. doi: 10.1002/(sici)1099-0984(199803/04)12:2<75::aid-per288>3.0.co;2-h.

Digman, J. (1990). Personality structure: emergence of the five-factor model. *Annual Review of Psychology*, *41*, 417-440.

Fetvadjiev, V., & Van de Vijver, F. J. R. (2015). Universality of the five factor model of personality. In J. D. Wright (Ed.), *International Encyclopedia of Social and Behavioral Sciences* (2ª ed., vol 9, pp. 249-253). Oxford, United Kingdom: Elsevier.

Fiske, D. W. (1949). Consistency of the factorial structures of personality ratings from different sources. *Journal of Abnormal and Social Psychology, 44,* 329-344.

Galton, F. (1884). Measurement of character. *Fortnightly Review, 36,* 179-185.

Goldberg, L. R. (1981). Language and individual differences. The search for universals in personality lexicons. In L. Wheeler (Ed.), *Review of personality and social psychology* (Vol. 2, pp. 141–165). Beverly Hills, CA: Sage.

Goldberg, L. R. (1982). From ace to zombie. Some explorations in the language of personality. In C. D. Spielberger & J. N. Butcher (Eds.), *Advances in personality assessment* (Vol. 1, pp. 203–234). Hillsdale, NJ: Lawrence Erlbaum.

Goldberg, L. R. (1990). An alternative "description of personality." The Big Five factor structure. *Journal of Personality and Social Psychology, 59*, 1216–1229.

Goldberg, L. R. (1992). The development of markers for the Big Five factor structure. *Psychological Assessment, 4*, 26–42.

Goldberg, L. R. (1993). The strcture of phenotypic personality traits. *American Psychologist, 48*(1), 26-34.

Gurven, M., von Rueden, C., Massenkoff, M., Kaplan, H., & Vie, M. (2013). How universal is the big five? Testing the five-factor model of personality variation among forager-farmers in the Bolivian Amazon. *Journal of Personality and Social Psychology, 104*(2), 354-370.

Hauck, N., F., Machado, W., Teixeira, M., & Bandeira, D. (2012). Evidências de validade de marcadores reduzidos para a avaliação da personalidade no modelo dos cinco grandes fatores. *Psicologia: Teoria e Pesquisa, 28*(4), 417-423.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010a). The weirdest people in the world? *Behavioral and Brain Sciences, 33*(2-3), 1-75.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010b). Beyond WEIRD: Towards a broad-based behavioral science. *Behavioral and Brain Sciences, 33*(2-3), 111-135.

Hutz, C., Nunes, C. H. Silveira, C., Serra, A., Anton, M., & Wieczorek, L. (1998). O desenvolvimento de marcadores para a avaliação da personalidade no modelo dos cinco grandes fatores. *Psicologia: Reflexão e Crítica, 11,* 395-409.

Iacobelli, F., Gill, A. J., Nowson, S., & Oberlander, J. (2011). In S. D'Mello, A. Graesser, B. Schuller, & J. Martin (Eds.), Affective computing and intelligent interaction (pp. 568-577). doi: 10.1007/978-3-642-24571-8.

Internet Live Stats. (s.d.). *Twitter Usage Statistics.* Retrieved from http://www.Internetlive stats.com/twitter-statistics/

Isaka, H. (1990). Factor analysis of trait terms in everyday Japanese language. *Personality and Individual Differences, 11*(2), 115-124.

John, O. P. , Angleitner, A., & Ostendorf, F. (1988). The lexical approach to personality: A historical review of trait taxonomic research. *European Journal of Personality, 2*, 171-203.

Katigbak, M. S., Church, A. T., Guanzon-Lapeña, M. A., Carlota, A. J., & Del Pilar, G. H. (2002). Are indigenous dimensions culture-specific? Philipine inventories and the Five-Factor Model. *Journal of Personality and Social Psychology, 82*, 89-101.

Klages, L. (1929). *The Science of Character (6ª ed.)*. (W. H. Johnston, Trad.) London: Unwin Brothers Ltd.

Li, J., & Chignell, M. (2010). Birds of a feather: How personality influences blog writing and reading. *International Journal of Human-Computer Studies*, *68*(9), pp. 589-602. doi:10.1016/j.ijhcs.2010.04.001.

Limas, A. F., Primi, R., & Carvalho, L. F. (2014). Avaliação da personalidade por redes sociais online: uso do facebook na área. *Revista Sul Americana de Psicologia, 2*(1), 1-25.

McCrae, R. R. (2011). Personality theories for the 21st century. *Teaching of Psychology, 38*(3), 209-214. doi: 10.1177/0098628311411785

McCrae, R. R., & Costa, P. T., Jr. (1997). Personality trait structure as a human universal. *American Psychologist, 52*, 509-516.

McCrae, R. R., & Terracciano, A. (2005a). Universal features of personality traits from the observer's perspective: Data from 50 cultures. *Journal of Personality and Social Psychology, 88*(3), 547-561.

McCrae, R. R., & Terracciano, A. (2005b). Personality profiles of cultures: Aggregate personality traits. *Journal of Personality and Social Psychology, 89*(3), 407-425.

McCrae, R. R., Costa, P. T., Jr., & Martin, T. A. (2005). The NEO–PI–3: A more readable revised NEO personality inventory. *Journal of Personality Assessment, 84*(3), 261-270. doi: 10.1207/s15327752jpa8403_05.

Mohammad, S., & Kiritchenko, S. (2015). Using hashtags to capture fine emotion categories from tweets. *Computational Intelligence, 31*(2), 301-326.

Nel, J. A., Valchev, V. H., Rothmann, S., van de Vijver, F. J. R., Meiring, D., & de Bruin, G. P. (2012). Exploring the personality structure in the 11 languages of South Africa. *Journal of Personality, 80*(4), 915-948. doi: 10.1111/j.1467-6494.2011.00751.x

Norman, W. T. (1963). Toward an adequate taxonomy of personality attributes. Replicated factor structure in peer nomination personality ratings. *Journal of Abnormal and Social Psychology, 66*(6), 574–583.

Norman, W. T. (1967). *2800 personality trait descriptors: Normative operating characteristics for a university population.* Unpublished manuscript: Department of Psychology, University of Michigan, Ann Arbor, MI.

Ortiz, F., Church, A., Vargas-Flores, J., Ibáñez-Reyes, J., Flores-Galaz, M., Iuit-Briceño, J., & Escamilla, J. (2007). Are indigenous personality dimensions culture-specific? Mexican inventories and the Five-Factor Model. *Journal of Research in Personality*, 618-649.

Park, G., Schwartz, A., Eichstaedt, J., Kern, M., Kosinski, M., Stillwell, D., et al. (2015). Automatic personality assessment through social media language. *Journal of Personality and Social Psychology, 108*(6), 934-952.

Partridge, G. E. (1910). *An outline of individual study.* NewYork: Sturgis & Walton.

Passos, M. F. D., & Laros, J. A. (2014). O modelo dos cinco grandes fatores de personalidade: Revisão de literatura. *Peritia, 21*, 13-21.

Passos, M. F. D., & Laros, J. A. (2015). Construção de uma escala reduzida de cinco grandes fatores de personalidade. *Avaliação Psicológica, 14*(1), 115-123.

Pennebaker, J., Mehl, M., & Niederhoffer, K. (2003). Psychological aspects of natural language use: Our words, our selves. *Annual Review of Psychology*, *54*(1), 547-577.

Perkins, M. L. (1926). The teaching of ideals and the development of the traits of character and personality. *Proceedings of the Oklahoma Academy of Science, 3*(2), 344-347.

Poddar, S., Kattagoni, V., & Singh, N. (2015). *Personality mining from biographical data with the "Adjectival Marker" technique*. Manuscrito não-publicado, Center for Exact Humanities, International Institute of Information Technology, Hyderabad, India.

Polzehl, T. (2015). *Personality in speech. Assessment and automatic classification*. Suíça: Springer.

Qiu, L., Lin, H., Ramsay, J., & Yang, F. (2012). You are what you tweet: Personality expression and perception on Twitter. *Journal of Research in Personality, 46*, 710-718.

Revelle, W. (2009). Personality structure and measurement: The contributions of Raymond Cattell. *British Journal of Psychology, 100*, 253-257.

Saucier, G., & Goldberg, L. R. (2001). Lexical studies of indigenous personality factors: Premises, products, and prospects. *Journal of Personality, 69*(6), 847-879. doi: 10.1111/1467-6494.696167.

Schmitt, D., Allik, J., McCrae, R., & Benet-Martínez, V. (2007). The geographic distribution of big five personality traits: Patterns and profiles of human self-description across 56 nations. *Journal of Cross-Cultural Psychology, 38*, 173-212.

Smith, P. B., Fischer, R., Vignoles, V., & Bond, M. (2013). *Understanding social psychology across cultures: Engaging with others in a changing world* (2nd Edition ed.). London: Sage.

Tausczik, Y., & Pennebaker, J. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology, 29*(1), 24-54.

Trait. (2015). *APA dictionary of psychology*. Washington, DC: American Psychological Association.

Tupes, E. C., & Christal, R. E. (1961). Recurrent personality factors based on trait ratings. *USAF ASD Technical Report No. 61-97*. U.S. Air Force: Lackland Air Force Base, San Antonio, TX.

Uher, J. (2013). Personality psychology: Lexical approaches, assessment methods, and trait concepts reveal only half of the story — Why it is time for a paradigm shift. *Integrative Psychological and Behavioral Science, 47*, 1-55. doi: 10.1007/s12124-013-9230-6.

Uher, J. (2015a). Developing "personality" taxonomies: Metatheoretical and methodological rationales underlying selection approaches, methods of data generation and reduction principles. *Integrative Psychological and Behavioral Science, 49*, 531-589. doi: 10.1007/s12124-014-9280-4.

Uher, J. (2015b). Interpreting "personality" taxonomies: Why previous models cannot capture individual-specific experiencing, behaviour, functioning and development. Major taxonomic tasks still lay ahead. *Integrative Psychological and Behavioral Science, 49*, 600-655. doi: 10.1007/s12124-014-9281-3.

Valchev, V. H., van de Vijver, F. J., Nel, J. A., Rothmann, S., & Meiring, D. (2013). The use of traits and contextual information in free personality descriptions across ethnocultural groups in South Africa. *Journal of Personality and Social Psychology, 104*(6), 1077-1091.

Valchev, V., Nel, J., van de Vijver, F., Meiring, D., Bruin, G., & Rothman, S. (2012). Similiraties and differences in implicit personality concepts across ethnocultural groups in South Africa. *Journal of Cross-Cultural Psychology, 44*(3), 365-388.

Vasconcelos, S., & Hutz, C. (2008). Construção e validação de uma escala de abertura à experiência. *Avaliação Psicológica, 7*(2), 135-141.

Waller, N. G. (1999). Evaluating the structure of personality. In C. R. Cloninger, (Ed.), *Personality and psychopathology* (pp. 155-200). Washington, Estados Unidos: American Psychiatric Press, Inc.

Wasti, S. A., Lee, K., Ashton, M. C., & Somer, O. (2008). Six Turkish personality factors and the HEXACO model of personality structure. *Journal of Cross-Cultural Psychology, 39*(6), 665–684. doi:10.1177/0022022108323783.

Webometrics (2016). *1040 Highly cited researchers (h>100) according to their google scholar citations public profiles*. Retrieved from http://www.webometrics. info /en/node/58.

Yarkoni, T. (2010). Personality in 100,000 words: A large-scale analysis of personality and word use among bloggers. *Journal of Research in Personality, 4*(3), 363-373.

Zecca, G., Verardi, S., Antonietti, J., Dahourou, D., Adjahouisso, M., Ah-Kion, J., ..., Rossier, J. (2012). African cultures and the five-factor model of personality: Evidence for a specific pan-African structure and profile? Journal of Cross-Cultural Psychology 44(5), 684-700. doi: 10.1177/0022022112468943.

Manuscript 2

The lexicon of personality in Brazilian Portuguese: Searching for descriptive terms in natural language

Abstract

This study aimed to generate a set of personality trait-descriptive terms in Brazilian Portuguese, exploring the natural language through mining public texts written by users of the online social network Twitter. To perform the search for descriptors, we employed as a keyword the Brazilian Portuguese equivalent of *I am* (i.e., *Eu sou*). The search was configured to retrieve posts made in Portuguese, from users located in Brazil, and made between March 19 and March 25 in 2016. Data collection resulted in the recovery of 6,303 messages made by 5,493 unique users. The data were submitted to text cleaning procedures and converted to a term-document matrix. Then, the terms were organized according to grammatical classes of words. A total of 1,454 adjectives, six names, 10 pronouns, and 383 nouns were collected. The adjectives and nouns are potentially relevant descriptors for the construction of a Brazilian taxonomy of personality. The main result of this study is a list of descriptors organized by word class with descriptive statistics of the frequency with which each term was found, according to the search criteria employed.

Keywords: personality; lexical hypothesis; text mining; big data, Brazilian Portuguese; Twitter.

Manuscrito 2

O léxico da personalidade no português brasileiro: buscando termos descritores na linguagem natural

Resumo

Este estudo objetivou gerar uma lista de termos descritores da personalidade no português brasileiro, explorando a linguagem natural a partir da mineração de textos públicos escritos por usuários da rede social *online Twitter*. Para realizar a busca dos termos, utilizou-se como palavra-chave a expressão *sou*. A busca foi configurada para recuperar postagens feitas em português brasileiro no período de 19 a 25 de março de 2016 por usuários localizados no Brasil. A coleta de dados resultou na recuperação de 6.303 postagens realizadas por 5.493 diferentes usuários. Os dados foram submetidos a procedimentos de limpeza e convertidos em uma matriz documento-termo. Então, os termos foram organizados em classes gramaticais de palavras. Ao final, foram encontrados 1.454 adjetivos, seis nomes, 10 pronomes e 383 substantivos, potencialmente descritores relevantes para a construção de uma taxonomia brasileira da personalidade. Apresenta-se, como resultado, uma lista dos descritores organizada por classe de palavras e que informa as estatísticas descritivas com a frequência com que cada termo foi encontrado, de acordo com os critérios de busca empregados.

Palavras-chave: personalidade; hipótese léxica; mineração de texto; big data; português brasileiro; *Twitter*.

The lexical hypothesis, or psycholexical approach, is based on the idea that the personality traits were codified in the natural language in the different cultures in the course of its history. It is presumed, therefore, that a personality taxonomy can be drawn from natural language, that is, from the words that people use to describe the personality characteristics of their own and of others (De Raad & Mlacic, 2015; Goldberg, 1981). From the lexical hypothesis, historically important theoretical models in the personality psychology domain were developed, such as Cattell's model of 16 primary factors and the five-factor model, also known as Big Five (De Raad & Mlacic, 2015; John, Angleitner, & Ostendorf, 1988).

Daouk-Öyry, Zeinoun, Choueiri e Van de Vijver (2016) describe that research in the lexical approach occurs in two major phases. The first stage is the identification of personality descriptors and consists of five steps. The first step is to select the source of the descriptors (e.g., dictionaries, literary or academic texts, and online social media). The second step is the definition of the ciriteria of inclusion and exclusion of the descriptors, for example, removal of obscure, unfamiliar or difficult to understand terms. In this step, the word classes that will be part of the taxonomy are also defined (e.g., adjectives, nouns, and verbs). The third step is to select the descriptors in the data source and involves, for example, the analyses of judges or some method of word sampling. The fourth step is categorization, in which the descriptors are organized according to grammatical criteria such as word classes or other criteria relevant for personality description. The last step is the semantic reduction, using criteria such as synonymy and antonym.

The second phase described by Oyry-Daouk et al. (2016) is the identification of the factorial structure, composed of two steps. The first is data collection, in which participants classify themselves or others using the descriptors selected from the

procedures of the first phase. In the second step, data analysis is conducted, basically involving analyses of dimensionality (e.g., factor analysis) of the collected data to obtain a parsimonious model.

Some of the seminal studies in the history of lexical hypothesis followed all or some of the steps of the two phases described by Daouk-Öyry et al. (2016), as Cattell (1943), Goldberg (1981, 1982), Norman (1967), Brokken (1978, as cited in De Raad & Mlacic, 2015), and Angleitner, Ostendorf, and John (1990). The research with the lexical hypothesis, however, does not occur without criticism of its assumptions, methods, and procedures (Ashton & Lee, 2005; Uher, 2013). In Manuscript 1 of this dissertation, Peres (2018) highlighted two sets of limitations related to the methodological strategies most traditionally adopted in psychology studies.

The first set of limitations pointed out by Peres (2018) is related to cultural and cross-cultural aspects that marked the history of the development of this approach. The axiom of lexical hypothesis conceptually represents an *emic* research perspective (Cheung, Van de Vijver & Leong, 2011), that is, focused on the study of phenomena specific to the cultures and on the validity of supposedly universal theories in specific cultures. Psycholexical research has as primary objective the investigation of the natural language in the different cultures in search of personality descriptors specific to the language and culture in analysis, but with the aspiration to elaborate a universal taxonomy. Despite the *emic* character of the lexical hypothesis axiom, psycholexical research has been conducted from an *etic* imposed perspective since after its initial development between the 1940s and 1980s in countries like United States, The Netherlands, and Germany - all speakers of Germanic languages (Cheung et al., 2011;. Daouk-Öyry et al., 2016; Peres & Laros, 2018) predominantly models and theories developed in these countries were investigated.

The major objective of the *etic* approach is to investigate whether models and theories developed in a given culture can be validly generalized to other cultures, assuming the postulation that these models and theories would be universal (Cheung et al., 2011). The term *etic imposed* is related to the fact that, generally, theories and research samples in psychology comes from Western, educated, industrialized, rich and democratic countries, and are generalized to poor or developing countries. In personality research, this perspective is represented, for example, by the practice of translation, adaptation, and collection of evidence of validity and reliability of psychometric instruments elaborated in other countries, very often the United States. Henrich, Heine, and Norenzayan (2010a, 2010b) describe the possible impacts of this research perspective on psychological science as a whole and Allik (2013) and Cheung et al. (2011) for the area of personality, specifically.

It has been argued, therefore, that the construction of a universal personality taxonomy (i.e., generalizable to all cultures) is impaired by the domination of a strictly *etic* perspective (Cheung et al., 2011; Daouk-Öyry et al., 2016; Peres & Laros, 2018). Such a practice limits or even neglects the emergence of specific aspects of languages and cultures. This issue debated and studied in the area of personality, including concerning the five-factor personality model (Allik, Realo, & McCrae, 2013; Fetvadjiev & Van de Vijver, 2015). Some researchers argue that the *emic* and *etic* approaches should be integrated to address this problem (Cheung et al., 2011; Daouk-Öyry et al., 2016; Valchev et al., 2012). Such integration will occur to the extent that common or universal aspects of personality are integrated with unique or culturally specific aspects. In *emic-etic* perspective, it is also recognized that the idiosyncrasies of a given language and culture should guide the methodological choices, like the selection of sources from which to withdraw personality descriptors, for instance.

The second set of limitations pointed out by Peres (2018) is related to the sources that researchers traditionally use to obtain personality descriptors. As highlighted, the lexical hypothesis points to natural language as the source for these terms (Goldberg, 1981). However, according to Cheung et al. (2011), typically this source is the dictionary of the language under analysis, an observation reinforced by the study of Daouk-Öyry et al. (2016). These authors carried out a systematic review of the literature focusing on papers that report on the method used to compile lists of personality descriptors. Daouk-Öyry et al. reviewed 25 articles published between 1970 and 2015, available on crawlers PsycINFO and Social Science Citation Index. In 80% (20) of these studies dictionaries were used exclusively as personality descriptors sources.

The use of dictionaries as sources was the strategy adopted in classical studies of the lexical hypothesis, as by Allport and Odbert (1936) and Norman (1967) in the United States, Brokken (1978, as cited in De Raad & Mlacic, 2015) in the Netherlands, and Angleitner et al. (1990) in Germany, among others (Peres & Laros, 2018). More recently, studies have been carried out in languages such as Lithuanian (Livaniene & De Raad, 2016), Polish (Szarota, Ashton, & Lee, 2007), Canadian French (Boies, Ashton, Pascal, Nicol, 2001), Croatian (Mlacic & Ostendorf, 2005), Spanish Castilian (Benet-Martinez & John, 2000), and Turkish (Somer & Goldberg, 1999).

On the one hand, dictionaries offer a quite extensive organized compilations of the lexical units (e.g., words) of a language, including the definition of thousands of entries and even offering synonyms and antonyms. However, on the other hand, they may not be synchronized with the current use of these units, that is, with natural language. For example, a dictionary does not report on the social context of the use of a word or on the difficulty of individuals of the language-speaking population in understanding the senses of a word.

The restricted use of the dictionary is one of the major criticisms concerning the lexical approach. From an epistemological perspective, Uher (2015), for example, criticizes the widespread use of dictionaries, describing it as a decontextualized lexical approach. Uher criticizes the practice of examining the dictionary, making a selection of certain descriptors, and finally translating them into questionnaire items. For Uher, "the construction of meanings for the items studied and for the results obtained largely relies on the researchers" (p. 557).

From the perspective of the study of natural language in social psychology (Pennebaker, Mehl, & Niederhoffer, 2003), Chung and Pennebaker (2008) presented criticisms similar to those elaborated by Uher (2015). For them, the practice of using judges to define what and which would be the most appropriate and frequently used by people to describe personality traits occurs without information as to how much the judgment of these experts approaches the actual use of the terms. Consequently, according to these researchers, factor analyzes were not yet undertaken from data that represented the use of everyday language at a high frequency. Another critique described by Chung and Pennebaker is related to the relevance of the traits when using psychometric instruments with closed lists of items, which restricts the variables of interest. That is, one assembles a set of items to form a measure, but as a consequence, this limited set of variables will be able to predict only a few behaviors at a given moment (Chung & Pennebaker, 2008).

Other studies, seeking to circumvent this obstacle and to capture the spontaneous expression of natural language better, have used mixed sources to obtain terms describing the personality. Some scholars combined the examination of dictionaries with the analysis of literary and scholarly texts of languages such as Persian (Farahani, De Raad, Farzad, & Fotoohie, 2016), Hindi (Singh, Misra, & De Raad, 2013), and

Russian (Allik et al., 2011), among others (Daouk-Öyry et al., 2016). Another set of studies adopted as sources free descriptions made by laypersons, such as semi-structured interviews, as did researchers in South Africa (Nel et al., 2012; Valchev, Van de Vijver, Nel, Rothman, & Meiring, 2007). In Japan, Isaka (1990) used texts in which participants freely described targets, for example, an ideal man and woman, five known persons, five famous people, and a pair of unpleasant man and woman. In other studies, researchers combined dictionaries with other textual sources, such as literary and journalistic texts, and free descriptions made by lay people. Examples are Hahn (1992, as cited in Hahn, Lee, & Ashton, 1999) in South Korea, Cheung et al. (1996) in China, and Katigbak, Church, Guanzon-Lapeña, Carlota, and Pilar (2002) in the Philippines.

With the popularity of computers and the web, possible sources for obtaining personality descriptors in different means of expression of natural language have multiplied exponentially. Newspapers, literary texts, academic texts, wikis, blogs, social networks that make use of text, image, audio, videos and symbols (e.g., emoticons), and various other types of digital media are available for interested researchers. Several studies are being carried out with alternative sources, such as written essays by students describing themselves (Chung & Pennebaker, 2008), written records of self-narratives (Hirsh & Peterson, 2009; Pennebaker & King, 1999), flow of consciousness reports (Lee et al., 2007), individual conversations records (Mehl, Gosling, & Pennebaker, 2006; Laserna, Seih, & Pennebaker, 2014; Polzehl, 2015), blogs (Iacobelli, Gill, Nowon, & Oberlander, 2011; Li & Chignell, 2010; Yarkoni, 2010), biographical texts (Poddar, Kattagoni, & Singh, 2015), and social networks such as Twitter (Mohammad & Kiritchenko, 2015;. Qiu et al., 2012), Facebook (Limas, Primi, & Carvalho, 2014; Park et al., 2015), and YouTube (Yeo, 2010; Biel, Aran, & Gatica-Perez, 2011).

The research with the lexical hypothesis in Brazil

In Brazil, research with the lexical approach seems to occur mainly through the translation, adaptation, and elaboration of items from lists of terms descriptors, instruments, models and theories elaborated in other countries, especially the United States. There are studies whose general objective was to translate, adapt, and collect evidence of the validity of foreign instruments for Brazil. Other studies used a set of procedures to construct or adapt instruments. For example, some studies have drawn from the international literature and lists of foreign adjectives, items and tests (e.g., Goldberg 1992; International Personality Item Pool, 2016; Peabody & De Raad, 2002). The researchers of these studies complemented the analyzes with Brazilian dictionaries, national studies, analysis of judges, and analysis of semantic validation with community participants (e.g., students) or experts (e.g., linguists). Examples of both cases are Andrade (2008); Nunes and Hutz (2007); Hutz, Silveira, Serra, Anton, and Wieczorek (1998); Passos and Laros (2015); Primi, Ferreira-Rodrigues, and Carvalho (2014); and Vasconcelos and Hutz (2008). There are also several studies that made use of instruments already developed or adapted (e.g., Fujita, Nakano, & Rondina, 2015; Noronha, Martins, Ferraz, & Mansão, 2015).

Only one Brazilian initiative was identified that followed the procedure of initiating the construction of a personality taxonomy starting from the examination of a dictionary. The research of the Brazilian group (Guzzo, Pinho, & Carvalho, 2002; Pinho, 2005; Pinho & Guzzo, 2003) was carried out in five subsequent stages, basically involving analyses made by judges: (i) selection of all 35,834 adjectives from one Brazilian Portuguese dictionary of 1996 (Guzzo et al., 2002); (ii) selection of personality descriptive adjectives by two judges graduating in psychology, according to 11 exclusion criteria inspired in Angleitner et al. (1990), resulting in 5,774 words

(Guzzo et al., 2002; Pinho & Guzzo, 2003); (iii) the classification and selection of the remaining adjectives from the second stage, following the criteria of the usefulness of the adjective for personality description, frequency of use in professional practice, and clarity, resulting in a list of 938 words; and (iv) classification of the adjectives of this list in the categories of tendencies, social aspects, physical characteristics or appearance, temporary states or conditions, and terms of limited utility (Pinho, 2005). No other study that has sequenced this project was identified in a search in the Brazilian Virtual Health Library of Psychology (2016).

A final set of studies was identified in the area of information techonology (Barros, Nunes, & Matos, 2012; Cardoso, Carvalho, & Nunes, 2014; Cardoso & Nunes, 2014; Nunes, Bezerra, & Oliveira, 2012; Nunes, Teles, & Souza, 2013). These studies aimed to develop computational tools that perform an automated evaluation of the personality of users of electronic systems focusing on recommender systems. These systems aim to customize the user experience according to their habits and preferences.

In one of these studies, Nunes et al. (2012) proposed a markup language to "standardize and help disseminate and share the use of information concerning the personality of users among applications that take into account psychological aspects in computational decision-making processes" (p. 267). Two other studies have explored the association of scores on Big Five inventories with the pace of typing (Porto et al., 2012) and with posts on the social network Twitter (Nunes et al., 2013). However, these two studies used small samples with less than 100 participants. Porto et al. (2012) did not report the statistical fit of the cluster analysis employed. Nunes et al. (2013) reported low correlation coefficients, less than 0.14, among the scores of the 28 subjects in the five-factors inventory, based on the analysis of the posts made by the participants on Twitter. Although they are innovative and contribute to the development of the area

in Brazil, these studies lack adequate psychometric analysis and evidence of validity and reliability.

The limitations regarding the usual lexical hypothesis investigation strategies (Cheung et al., 2011; Chung & Pennebaker, 2008; Peres & Laros, 2018; Uher, 2015) can be identified in the research with this approach in Brazil. That is, it can be affirmed that many Brazilian studies follow a predominantly *etic* perspective since the development of instruments for collecting data are based on foreign personality descriptors, models and, theories. It is also relevant to observe that the data collection procedures usually occur in testing environments, with predetermined sets of items, which means that natural language does not seem to be much exploited in these works.

The present study

The objective of the present study was to develop a list of personality descriptors obtained from the mining of a natural language source, specfically public texts written by users of the online social network Twitter. We planned to present, as a final result, the descriptors organized by class of words (e.g. adjectives, nouns and adverbs), accompanied by descriptive statistics with the frequency with which each term was found in this social network, according to the search criteria employed.

Twitter is a social network on the internet, created in 2006 and characterized by allowing users to post short messages of up to 140 characters (until 2017), called tweets ("Twitter", 2016). Users can read and send posts through different interfaces, such as website, mobile instant messaging, and applications installed on mobile devices, such as tablets and smartphones. The posts (i.e., messages) from Twitter users can be public or protected. When protected, the user specifies that only certain people can read the messages. However, posts are by default public. The public posts can be consulted by anyone using the social network's own search tool, or, more systematically, by a

specialized software. For a software performing such a systematic search on the social network, Twitter requires requesting permission to access – the author of this manuscript received such consent.

This study aims is to explore a source of personality descriptors characterized by a spontaneous expression of natural language, empirically verifying the frequency and predilection for the use of descriptors by people. This approach also aims to reduce the weight of the researcher's decisions concerning the usefulness or relevance of the terms for the construction of a personality taxonomy, avoiding subjective judgments about the relevance of the traits, and unfounded choices of variables of interest. Also, we understand that this study fits in an eminently *emic* perspective, not starting from any theoretical model (e.g., Big Five), but of the assumptions of the own lexical hypothesis.

Method

Data collection procedures and research corpus

Data collection took place through the extraction (i.e., web scraping) of public posts from Twitter users. The search adopted as a criterion the expression "I am" (i.e., “Eu sou”) to find texts in which, in describing themselves, people have adopted terms that may represent personality traits. The search was performed between March 19 and March 25 of 2016, in Portuguese, considering the universe of users located in Brazil and merging the most recent and popular posts of the period. The software has been configured to get up to 15,000 posts. These posts constituted the corpus to be analyzed. For these procedures, the "twitteR" package of the statistical software "R" (Gentry, 2015) was used.

Although we did not analyze the demographic profile of our sample, a recent study reveals some general characteristics of Brazilian users of Twitter (Global Web Index, 2015). According to this study, 58% of Brazilian users are male. Considering the age, 15% of the users are between 16 and 20 years old, 21% between 21 and 24, and 22% between 25 and 34. Regarding relationship status, 38% are single, 30% are married, 17% are in another kind of relationship, and 6% are divorced or widower. The Brazilian users are interested in subjects such as movies (79%), cars (57%), music (75%), travels (71%), food and restaurants (65%), electronic games (55%), personal finances (57%), and fashion (39%). According to the study, the Brazilian users express opinions regarding movies (50%), alcohol consumption (40%), music (40%), travels (40%), food and restaurants (25%), shopping (40%), electronics (50%), and fashion (50%).

Data analysis

In the first stage of the analysis, the corpus was submitted to text clean procedures. All capitalized and accented letters were converted to lowercase letters without accents. Numbers, symbols (e.g., emoticons), scores and URLs (i.e., links to internet addresses) were removed. Also, stop words such as conjunctions and articles have been removed (see Appendix 1). For this work, the "tm" package (Feinerer & Hornik, 2015) of the software "R" was used. After initial cleaning, the corpus was converted into a term-document matrix, that is, an array in which each term (i.e., word) is counted for each document (i.e., post) in which it appears. The frequency with which each term appeared in the searches was calculated, that is, the frequency of each term in the corpus was summed up.

In the second stage, the terms resulting from the previous step were classified by the researcher in different classes of words (i.e., adjectives, nouns, adverbs, pronouns,

names, and contractions) and organized into lists. In this stage, only the words that complemented the sentence “I am ...” were maintained, such as *muggle*, *stylish*, *spoiled*, *goddess*, *Einstein*, *crazy*, *blind*, etc. As a result, the first list of terms was generated. In the third step, the list was rearranged to identify unique terms, grouping the frequency of forms in masculine and feminine (e.g., *perfeito* and *perfeita*), and spellings (e.g., *trouxa*, *trouxaaa*, *troxa*) of the same word. Thus, the second list was generated without repeating different forms of the same terms. The methodological steps of the present study are synthesized in Figure 1. The results are presented next, accompanied by a more detailed description of the organization of the two lists.

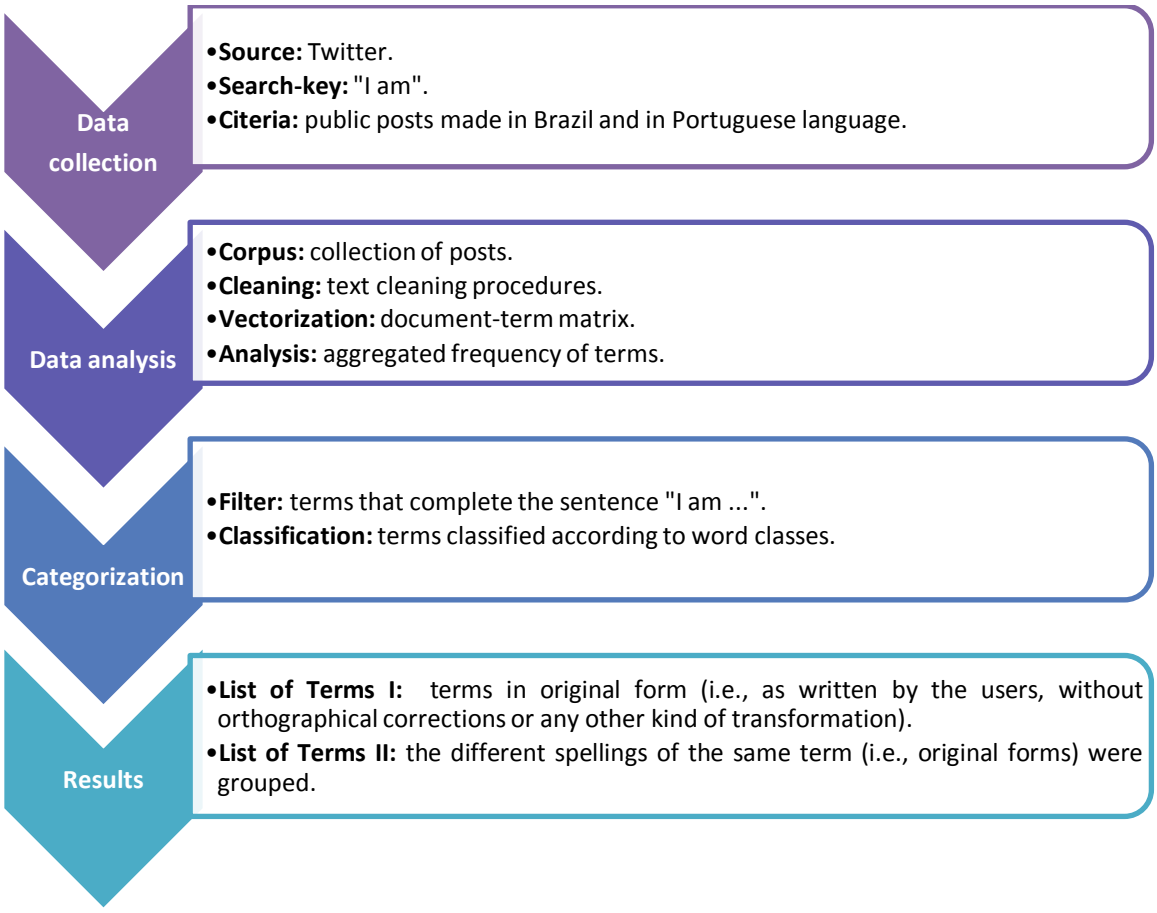

Figure 1. Methological steps of the study.

Results

The search retrieved 6,303 messages posted by 5,493 unique users. These messages concerned a sampling of the period from March 19 to 25, 2016, according to Twitter's search policy. The vast majority (97.5%) of users whose posts were retrieved contributed with one or two tweets of up to 140 characters. These data are presented in Table 1. From these messages, 13,653 terms were extracted, already excluded numbers, symbols, scores, URLs and stop words (see Appendix 1). Posts and terms have been converted into a term-document matrix.

Table 1
Number of posts recovered and users

Posts	Users		
	N	%	Accumulated
1	4.936	89,86	89,86
2	421	7,66	97,52
3	84	1,53	99,05
4	29	0,53	99,58
5	7	0,13	99,71
6	9	0,16	99,87
7	1	0,02	99,89
8	2	0,04	99,93
9	3	0,05	99,98
10	1	0,02	100,00
Total	5.493	100,00	

After classifying the terms according to word classes and keeping only descriptors that completed the sentence “I am ...”, an initial list was reached with 1,454 adjectives, seven adverbs, six names, 10 pronouns, and 383 nouns. Table 2 presents the number and frequency of adjectives and nouns as well as examples. Figure 2 graphically illustrates the 200 most frequently found descriptors (in Portuguese) in the search corpus. The complete list is presented in Appendix 2, which also includes the contractions, names, pronouns and adverbs found. In this list, the frequency with which each term was found in the search is displayed.

In this second step, the terms were kept as written by the users, except for the initial text cleaning performed in the previous step. In order to capture the real use of words, we did not drop any term that denoted personality description, even though the term belonged to one of the categories of words often eliminated in other studies (John et al., 1988; Peres & Laros, 2018). For example: gentilics (e.g., *Eskimo*); professions or occupations (e.g., *artist*); parts of the body (e.g., *ear*); references to the physical constitution (e.g., *beautiful*); references to ideologies (e.g., *Marxist*); slang and vulgar terms (e.g., *scrotum*); references to animals (e.g., *chameleon*); terms from foreign languages (e.g., *Jedi*); and neologisms (e.g., *falsiane*, *dboa*), etc.

Figure 2. Wordcloud with the 200 most frequent adjectives (*adjetivos*) and nouns (*substantivos*) found in the search. The words are the originals in Portuguese.

Table 2
Examples of terms (in Portuguese) with duplicity and without orthographical correction

Frequency	Adjectives		Nouns	
	Quantity	Examples	Quantity	Examples
> 100	5	*mimada, coisado, antigo, trouxaaaaaaaaa, apaixonadaaaaaaa*	2	*porquinha, deusa*
51 to 100	8	*falador, amigão, lindaa, mala, loucaaaa, chateada, favorita, precoce*	2	*burrão, mulherão*
26 to 50	18	*sonsa, falsa, loucona, apaixonante, normal, lerdaaa, viva, certinha, doente*	4	*filho, amorzinho, peixes, burrona, menino*
11 to 25	63	*amarga, causador, esperta, fracaa, velhaca, odiada, viciado, atrasada*	7	*bocado, viaaado, ninja, caiacara, cantor, menininha, cachorrona*
< 10	1,360	*estilosa, chorão, dramático, esperto, extremista, inferior, sortudo, confuso*	368	*palhaço, nadaaaa, desastre, musa, passatempo, grude, asno, espetáculo*
Total	1,454		383	

In the third stage, different spellings of the same word were grouped and their frequency added together. For example, the adjective of two genders *aborrecido* (i.e., *bored*) appeared with the masculine (*aborrecido*) and feminine (*aborrecida*) spellings once each. Thus, the two spellings were grouped in *aborrecido(a),* with a frequency equal to two. The same was done with words found with different spellings, such as the adjective *louco* (i.e., *crazy*), which besides appearing with the spelling in both genders, was also written as *loko*, *loka*, *loco,* and *loca*, for example. However, the diminutives (e.g., *bonzinho*, diminutive of *good*) and augmentatives (e.g., *loucão*, augmentative of *louco*) of a word have not been agglutinated, since they may take different meanings or magnitudes. There was no deletion of terms at this stage. The frequencies and some examples of terms resulting from this reorganization are presented in Table 3. It is

possible to notice that this procedure resulted in 1,118 adjectives (reduction of 336 terms) and 332 nouns (reduction of 51 terms). The complete list of words resulting from this stage of analysis is not included in this manuscript, due to space constraints.

Table 3
Examples of unique descriptors (in Portuguese) after orthographical correction

Frequency	Adjectives		Nouns	
	Quantity	Examples	Quantity	Examples
> 100	5	*mimado, coisado, antigo, trouxa*	2	*porquinha, deusa*
51 to 100	10	*falador, amigão, lindo, louco, chateado*	2	*burrão, mulherão*
26 to 50	25	*orgulhoso, apaixonante, fraco, falso*	3	*filho, menino, burrona*
11 to 25	47	*causador, velhaco, desgraçado*	10	*cachorrona, amor, ninja*
< 10	1.031	*dramático, medroso, chorão, bravo*	315	*satanás, rato, florzinha*
Total	1.118		332	

Discussion

The objective of this study was to develop a list of terms describing personality from the Portuguese language lexicon spoken in Brazil using a natural language source, specifically the online social network Twitter. From the mining of public messages posted by Brazilian users in this social network, a list was obtained with 1,118 adjectives and 332 nouns. Considering the spontaneous expression of the authors of the messages, that is, without ortographical corrections and not changing the gender of the words, we obtained a list of 1,454 adjectives and 383 nouns. Even though they were not

the target of the search, adverbs, contractions, names, and pronouns that could potentially serve as personality descriptors were also identified.

As we could see, the analysis of the use of the language in social media presents peculiar challenges. For example, besides unintentional orthographical errors and typos, we also identified alternative spellings of many words made intentionally, like the repetition of letters to give a certain emphasis to the sentence (e.g., *felizzzz* [*happyyyy*], *apaixonaaaada* [*in looove*], etc.), or the use of abbreviations (eg, *vc* [*u*] instead of *você* [*you*]). Also, the use of foreignisms, neologisms, and vulgar terms were frequent observed.

Notwithstanding these challenges, the analysis of natural language in an environment in which individuals express themselves spontaneously makes it possible to identify personality descriptors that people actually employ in their daily lives. Several of the terms listed in Appendix 2 of this manuscript, for example, do not appear in lists of descriptors, such as those carefully elaborated by the team responsible for the first psycholexical study which used as the data source a Brazilian dictionary (Guzzo & Carvalho, 2002; Pinho & Guzzo, 2003). In this study we found a great number of adjectives such as *grudento* (*sticky*), *hiperativo* (*hyperactive*), *enganador* (*deceitful*), *desnaturado* [*denatured*], etc. that were not included in the list of the study that used the Brazilian dictionary as a data source. Also many vulgar terms, foreignisms and neologisms were found.

As pointed out in the introduction to this manuscript, one of the main set of critiques of the lexical hypothesis and the development of personality taxonomies concerns the potential violation of the axiom of this hypothesis, regarding the analysis of the use of natural language for the identification of descriptors. The use of dictionaries as a source of descriptors and psychometric instruments with limited sets of

items for data collection is one of the major criticisms regarding the lexical approach. According to the critics, this approach potentially limits psychological research, since decisions about the relevance of traits fall far too much on researchers' decisions, whose instruments may be unable to measure the most salient personality traits in a particular culture.

An open approach such as the one adopted in this study has the potential to circumvent these questions since the research corpus is formed by spontaneous records of natural language. That is, this approach offers the possibility of analyzing the natural use of language, capturing people's daily use of language to describe and express their personality and that of other people. Thus, such approach is hoped to explore what is the fundamental idea of the lexical hypothesis and to reduce the weight of the judgment of the researcher or the judges employed in the research.

Another set of criticisms is related to the predominantly *etic* cross-cultural paradigm of lexical studies. This perspective may restrict or even prevent the emergence of specific aspects of languages and cultures under analysis. An *emic-etic* perspective has been proposed (Cheung et al., 2011; Daouk-Öyry et al., 2016; Valchev et al., 2012) that aims to integrate personality aspects considered universal to aspects specific to cultures. In this approach, the specificities of languages and cultures must guide methodological decisions. An open strategy for data collection, as was used in this study, has a promising potential not only for the autochthonous development of personality models but also for the verification of the universality of personality components, such as the Big Five factors.

Regarding the identification of personality descriptors, subsequent studies should be concerned with new categorizations of words recovered from social media, beyond their grammatical classes. For this purpose, criteria concerning the relevance of a given

word for the description of personality (Angleitner et al., 1990) and for the study of natural language (Pennebaker & King, 1999; Pennebaker et al., 2003) should also be adopted.

We also highlight that there is still an extensive research agenda regarding the semantic reduction of the descriptors and the identification of uderlying latent structure. New data collections should be planned and performed in such a way as to obtain a data matrix suitable for multivariate analyzes, such as cluster or factor analysis. As can be seen from the analysis of the data presented, there is a very low proportion of words per user: there were only one or two posts for 97.52% of the 5,493 users examined. In this way, a very sparse therm-document matrix was obtained. A possible solution to this problem is to design a data collection procedure able to retrieve more messages per subject.

Another set of important limitations of this study that should be considered relates to the search configuration. First, the search is not exhaustive or census-based, having been restricted to the sentence "I am ..." when we could have used the expressions "you are", "he/she is", "we are", and "they are". In addition to increasing the diversity and volume of the research corpus, this strategy would allow comparing data from hetero and self-report. The format of the search also did not allow to capture the use of verbs, for example. Second, the search must be performed at different times, avoiding waves of use of certain expressions occasioned by cultural and social events (e.g., religious holidays, political events, etc.). Third, although it is already a sampling procedure, the Twitter search can be improved by configuring it to retrieve messages from users in different geographical locations in Brazil (e.g., Municipalities or States).

We hope that the list elaborated in this study will serve as a consultation guide for future studies. Guzzo et al. (2003), for example, concluded that this type of research is

rare in Brazil and this remains true 15 years later, since no new initiatives of this nature were identified in the Brazilian literature. In the scope of this doctoral dissertation, the present study has the role of guiding the collections and analyses that will be carried out in the next study, whose results will complement the list presented here. As defended by Guzzo et al. (2003), even incomplete, lists such as these can assist researchers in the selection procedures of items to compose psychometric instruments.

References

Allik, J. (2013). Personality psychology in the first decade of the new millennium: A bibliometric portrait. *European Journal of Personality, 27*(1), 5-14. doi: 10.1002/per.1843.

Allik, J., Realo, A., & McCrae, R. (2013). Universality of the five-factor model of personality. In P. T. Costa & T. A. Widiger (Eds.), *Personality disorders and the five factor model of personality* (pp. 61-74). Washington: American Psychological Association.

Allik, J., Realo, A., Mottus, R., Pullmann, H., Trifonova, a., McCrae, R. R., … Korneeva, E. E. (2011). Personality profiles and the "Russian Soul": Literary and scholarly views evaluated. *Journal of Cross-Cultural Psychology*, *42*(3), 372–389. doi:10.1177/0022022110362751.

Allport, G., & Odbert, H. (1936). Trait names: a psycho-lexical study. *Psychological Monographs*, *47*(211), 1-38.

Andrade, J. M. (2008). *Evidências de validade do inventário dos cinco grandes fatores de personalidade para o Brasil.* Tese de Doutorado, Universidade de Brasília, Brasília.

Angleitner, A., Ostendorf, F., & John, O. P. (1990). Towards a taxonomy of personality descriptors in German. A psycho-lexical study. *European Journal of Personality*, *4*(2), 89–118. doi: 10.1002/per.2410040204.

Asthon, M. C., & Lee, K. (2005). A defence of the lexical approach to the study of personality structure. *European Journal of Personality*, *19*(1), 5-24. doi: 10.1002/per.541.

Benet-Martinez, V., & John, O. P. (2000). Toward the development of quasi-indigenous personality constructs: Measuring los cinco grandes in Spain with indigenous Castilian markes. *American Behavioral Scientist, 44*(1), 141-157. doi: 10.1177/00027640021956035.

Biblioteca Virtual em Saúde Psicologia (Brasil). (2016, March 10). *BVS Psicologia Brasil*. Retrieved from: www.bvs-psi.org.br.

Biel, J. I., Aran, O., & Gatica-Perez, D. (2011). You are known by how you vlog: Personality impressions and nonverbal behavior in YouTube. In *Fifth International AAAI Conference on Web and Social Media*. Retrieved from https://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/viewFile/2796/3220.

Boies, K., Lee, K., Ashton, M. C., Pascal, S., & Nicol, A. A. M. (2001). The structure of the French personality lexicon. *European Journal of Personality, 15*(4), 277-295. doi: 10.1002/per.411.

Cardoso, G. G., & Nunes, M. A. S. N. (2014). Inferindo personalidade por meio de histórias. *Revista Brasileira de Computação Aplicada*, *6*(2), 113-122. doi: 10.5335/rbca.2014.3782.

Cardoso, G. G., Carvalho, I. C., & Nunes, M. A. S. N. (2014). Prospecção sobre a extração da personalidade do usuário e seu uso em sistemas computacionais. *Scientia Plena*, *10*(7), 1-8.

Cattell, R. B. (1943). The description of personality: Basic traits resolved into clusters. *Journal of Abnormal and Social Psychology*, 38, 476–507.

Cheung, F. M., Leung, K., Fan, R. M., Song, W. Z., Zhang, J.X., & Zhang, J.P. (1996). Development of the Chinese Personality Assessment Inventory. *Journal of Cross Cultural Psychology, 27*(2), 181-199. doi: 10.1177/0022022196272003.

Cheung, F. M., Van de Vijver, F., & Leong, F. (2011). Toward a new approach to the assessment of personality in culture. *American Psychologist, 66*(7), 593-603. doi: 10.1037/a0022389.

Chung, C. K., & Pennebaker, J,. W. (2008). Revealing dimensions of thinking in open-ended self-descriptions: An automated meaning extraction method for natural language. *Journal of Research in Personality*, *42*(1), 96-132. doi:10.1016/j.jrp. 2007.04.006.

Daouk-Öyry, L., Zeinoun, P. , Choueiri, L., & Van de Vijver, F. J. R. (2016). Integrating global and local perspectives in psycholexical studies: A GloCal approach. *Journal of Research in Personality*. Advance online publication. doi: http://dx.doi.org/10.1016/ j.jrp. 2016.02.008

De Raad, B., & Mlacic, B. (2015). The lexical foundation of the big five factor model. In T. Widiger (Ed.), *The Oxford handbook of the five factor model.* Retrieved from www.oxfordhandbooks.com. doi: 10.1093/oxfordhb/9780199352487.013.12.

Farahani, M. N., De Raad, B., Farzad, V., & Fotoohie, M. (2016). Taxonomy and structure of Persian personality-descriptive trait terms. *International Journal of Psychology, 51*(2), 139-149. doi:10.1002/ijop. 12133.

Feinerer, I., & Hornik, K. (2015). tm: Text Mining Package (Version 0.6-2) [Software]. Available from *The Comprehensive R Archive Network*: http://cran.r-project.org/ web/packages/tm/index.html.

Fetvadjiev, V., & Van de Vijver, F. J. R. (2015). Universality of the five factor model of personality. In J. D. Wright (Ed.), *International Encyclopedia of Social and Behavioral Sciences* (2nd ed., Vol 9, pp. 249-253). Oxford: Elsevier.

Fujita, A. T. L., Nakano, T. C., & Rondina, R. C. (2015). Características de personalidade e dependência nicotínica em universitários. *Avaliação Psicológica, 14*(1), 73-81. doi: 10.15689/ap. 2015.1401.08.

Gentry, J. (2015). twitteR: R Based Twitter Client (Version 1.1.9) [Software]. Available from *The Comprehensive R Archive Network*: http://cran.r-project.org/web/packages/twitteR/ index.html.

Goldberg, L. R. (1981). Language and individual differences. The search for universals in personality lexicons. In L. Wheeler (Ed.), *Review of personality and social psychology* (Vol. 2, pp. 141–165). Beverly Hills, CA: Sage.

Goldberg, L. R. (1982). From ace to zombie. Some explorations in the language of personality. In C. D. Spielberger & J. N. Butcher (Eds.), *Advances in personality assessment* (Vol. 1, pp. 203–234). Hillsdale: Lawrence Erlbaum.

Goldberg, L. R. (1992). The development of markers for the big five factor structure. *Psychological Assessment, 4*(1), 26–42. doi: 10.1037/1040-3590.4.1.26.

Gurven, M., von Rueden, C., Massenkoff, M., Kaplan, H., & Vie, M. (2013). How universal is the big five? Testing the five-factor model of personality variation among forager-farmers in the Bolivian Amazon. *Journal of Personality and Social Psychology, 104*(2), pp. 354-370. doi: 10.1037/a0030841.

Guzzo, R. S. L., Pinho, C. C. M., & Carvalho, C. F. C. (2002). Construção da taxonomia brasileira para descritores da personalidade. *Psicologia: Reflexão e Crítica, 15*(1), 71-75. doi: 10.1590/S0102-79722002000100009.

Han, D. W., Lee, K., & Ashton, M. C. (1999). A factor analysis of the most frequently used Korean personality trait adjectives. *European Journal of Personality, 13*(4), 261-282. doi: 10.1002/(SICI)1099-0984(199907/08)13:4<261::AID-PER340>3.0.CO;2-B.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010a). The weirdest people in the world? *Behavioral and Brain Sciences, 33*(2-3), 61-83. doi: 10.1017/S0140525X0999152X.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010b). Beyond WEIRD: Towards a broad-based behavioral science. *Behavioral and Brain Sciences, 33*(2-3), 111-135. doi: 10.1017/S0140525X10000725.

Hirsh, J., & Peterson, J. (2009). Personality and language use in self-narratives. *Journal of Research in Personality, 43*(3), 524-527. doi: 1016/j.jrp. 2009.01.006.

Hutz, C. S., Nunes, C. H, Silveira, A. D., Serra, J., Anton, M., & Wieczorek, L. S. (1998). O desenvolvimento de marcadores para a avaliação da personalidade no modelo dos cinco grandes fatores. *Psicologia: Reflexão e Crítica, 11*(2), 395-411. doi: 10.1590/S0102-79721998000200015.

Iacobelli, F., Gill, A. J., Nowson, S., & Oberlander, J. (2011). Large scale personality classification of bloggers. In S. D'Mello, A. Graesser, B. Schuller, & J. Martin (Eds.), *Affective computing and intelligent interaction* (pp. 568-577). doi: 10.1007/978-3-642-24571-8.

International Personality Item Pool (2016, March, 23). *International Personality Item Pool: A scientific collaboratory for the development of advanced measures of personality traits and other individual differences*. Retrieved from http://ipip.ori.org/.

Global Web Index (2015). *#QuemUsaOTwitter?*. Retrieved from https://globalwebindex.net

Isaka, H. (1990). Factor analysis of trait terms in everyday Japanese language. *Personality and Individual Differences, 11*(2), 115-124. doi: 10.1016/0191-8869(90)90003-A.

John, O. P. , Angleitner, A., & Ostendorf, F. (1988). The lexical approach to personality: a historical review of trait taxonomic research. *European Journal of Personality, 2*(3), 171-203. doi: 10.1002/per.2410020302.

Katigbak, M. S., Church, A. T., Guanzon-Lapeña, M. A., Carlota, A. J., & Del Pilar, G. H. (2002). Are indigenous dimensions culture-specific? Philipine inventories and the five-factor model. *Journal of Personality and Social Psychology, 82*(1), 89-101.

Laserna, C. M., Seih, Y. T., & Pennebaker, J. W. (2014). Um... who like says you know: Filler word use as a function of age, gender and personality. *Journal of Language and Social Psychology, 33*(3), 328-338. doi: 10.1177/0261927X14526993.

Lee, C., Kim, K., Seo, Y., & Chung, C. (2007). The relations between personality and language use. *The Journal of General Psychology, 134*(4), 405-413. doi: 10.3200/GENP. 134.4.405-414.

Li, J., & Chignell, M. (2010). Birds of a feather: How personality influences blog writing and reading. *International Journal of Human-Computer Studies, 68*(9), 589-602. doi:10.1016/j.ijhcs.2010.04.001.

Limas, A. F., Primi, R., & Carvalho, L. F. (2014). Avaliação da personalidade por redes sociais online: uso do *facebook* na área. *Revista Sul Americana de Psicologia, 2*(1), 1-25.

Livaniene, V., & De Raad, B (2016). The factor structure of Lithuanian personality-descriptive adjectives of the highest frequency of use. *International Journal of Psychology*. Advance online publication. doi: 10.1002/ijop. 12247.

Mehl, M., Gosling, S., & Pennebaker, J. (2006). Personality in its natural habitat: Manifestations and implicit folk theories of personality in daily life. *Journal of Personality and Social Psychology, 90*(5), 862-877.

Mlacic, B., & Ostendorf. F. (2005). Taxonomy and structure of Croatian personality-descriptive adjectives. *European Journal of Personality, 19*(2), 117-152. doi: 10.1002/per.539.

Mohammad, S., & Kiritchenko, S. (2015). Using hashtags to capture fine emotion categories from tweets. *Computational Intelligence, 31*(2), 301-326. doi: 10.1111/coin.12024.

Nel, J. A., Valchev, V. H., Rothmann, S., Van de Vijver, F. J. R., Meiring, D., & de Bruin, G. P. (2012). Exploring the personality structure in the 11 languages of South Africa. *Journal of Personality, 80*(4), 915-948. doi: 10.1111/j.1467-6494.2011.00751.x.

Norman, W. T. (1967). *2800 personality trait descriptors: Normative operating characteristics for a university population.* Unpublished manuscript: Department of Psychology, University of Michigan, Ann Arbor, MI.

Noronha, A. P. P. , Martins, D. F., Campos, R. R. F., & Mansão, C. S. M. (2015). Relações entre afetos positivos e negativos e os cinco fatores de personalidade. *Estudos de Psicologia (Natal)*, *20*(2), 92-101. doi: 10.5935/1678-4669.20150011.

Nunes, C. H. S. S., & Hutz, C. S. (2007). Construção e validação da escala fatorial de socialização no modelo dos cinco grandes fatores de personalidade. *Psicologia: Reflexão e Crítica*, *20*(1), 20-25. doi: 10.1590/S0102-79722007000100004.

Nunes, M. A. S. N., Bezerra, J. S., & Oliveira, A. A. (2012). Personalityml: A markup language to standardize the user personality in recommender systems. *GEINTEC*, *2*(3), 255-273. doi: 10.7198/S2237-0722201200030006.

Nunes, M. A. S. N., Teles, F. R., & Souza, J. G. (2013). Inferindo personalidade via tweets. *GEINTEC*, *3*(3), 45-57. doi: 10.7198/S2237-0722201300030004.

Park, G., Schwartz, A., Eichstaedt, J., Kern, M., Kosinski, M., Stillwell, D., et al. (2015). Automatic personality assessment through social media language. *Journal of Personality and Social Psychology, 108*(6), 934-952. doi: 10.1037/pspp0000020.

Passos, M. F. D., & Laros, J. A. (2015). Construção de uma escala reduzida de cinco grandes fatores de personalidade. *Avaliação Psicológica*, *14*(1), 115-123.

Peabody, D., & De Raad, B. (2002). The substantive nature of psycholexical personality factors: A comparison across languages. *Journal of Personality and Social Psychology*, *83*(4), 983-997. doi: 10.1037//0022-3514.83.4.983.

Pennebaker, J., & King, L. (1999). Linguistic styles: Language use as an individual difference. *Journal of Personality and Social Psychology*, *77*(6), 1296-1312.

Pennebaker, J., Mehl, M., & Niederhoffer, K. (2003). Psychological aspects of natural language use: Our words, our selves. *Annual Review of Psychology*, *54*(1), 547-577.

Peres, A. J. S. (2018). *Hipótese léxica, psicologia transcultural e linguagem natural.* Unpublished manuscript, Institute of Psychology, University of Brasilia, Brasília, Distrito Federal, Brazil.*

Pinho, C. C. M. (2005). *Taxonomia brasileira da personalidade: um estudo dos adjetivos da língua portuguesa*. Tese de Doutorado, Pontifícia Universidade Católica de Campinas, Campinas.

Pinho, C. C. M., & Guzzo, R. S. L. (2003). Taxonomia de adjetivos descritores da personalidade. *Avaliação Psicológica*, *2*(2), 81-97.

Poddar, S., Kattagoni, V., & Singh, N. (2015). *Personality mining from biographical data with the "Adjectival Marker" technique*. Manuscrito não-publicado, Center for Exact Humanities, International Institute of Information Technology, Hyderabad, India.

Polzehl, T. (2015). *Personality in speech. Assessment and automatic classification*. Bern: Springer.

Porto, S. M., Costa, W. S., Silva, E. P. , Barros, S. L. A., Nunes, M. A. S. N., & Matos, L. N. (2012). Personalikey – um software para extração de traços de personalidade através do ritmo de digitação. *GEINTEC*, *3*(1), 76-01. doi: 10.7198/S2237-0722201300010007.

Primi, R., Ferreira-Rodrigues, C. F., & Carvalho, L. F. (2014). Cattell's Personality Factor Questionnaire (CPFQ): Development and preliminary study. *Paidéia*, *24*(57), 29-37. doi: 10.1590/1982-43272457201405.

Qiu, L., Lin, H., Ramsay, J., & Yang, F. (2012). You are what you tweet: Personality expression and perception on Twitter. *Journal of Research in Personality, 46*(6), 710-718. doi: 10.1016/j.jrp. 2012.08.008.

Singh, J. K., Misra, G., & De Raad, B. (2013). Personality structure in the trait lexicon of Hindi, a major language spoken in India. *European Journal of Personality*, *27*(6), 605-620. doi: 10.1002/per.1940.

Somer, O., & Goldberg, L. R. (1999). The structure of Turkish trait-descriptive adjectives. *Journal of Personality and Social Psychology, 76*(3), 431-450.

Szarota, P. , Ashton, M. C., & Lee, K. (2007). Taxonomy and structure of Polish personality lexicon. *European Journal of Personality*, *21*(6), 823-852. doi: 10.1002/per.635.

Twitter. (n.d.). In *Wikipedia.* Retrieved March 23, 2016, from https://en.wikipedia.org/wiki/ Twitter.

Uher, J. (2013). Personality psychology: lexical approaches, assessment methods, and trait concepts reveal only half of the story — Why it is time for a paradigm shift. *Integrative Psychological and Behavioral Science*, *47*(1), 1-55. doi: 10.1007/s12124-013-9230-6.

Uher, J. (2015). Developing "personality" taxonomies: Metatheoretical and methodological rationales underlying selection approaches, methods of data generation and reduction principles. *Integrative Psychological and Behavioral Science*, *49*(4), 531-589. doi: 10.1007/s12124-014-9280-4.

Valchev, V. H., Van de Vijver, F. J., Nel, J. A., Rothmann, S., & Meiring, D. (2013). The use of traits and contextual information in free personality descriptions across ethnocultural groups in South Africa. *Journal of Personality and Social Psychology, 104*(6), 1077-1791. doi: 10.1037/a0032276.

Valchev, V., Nel, J., Van de Vijver, F., Meiring, D., Bruin, G., & Rothman, S. (2012). Similiraties and differences in implicit personality concepts across ethnocultural groups in South Africa. *Journal of Cross-Cultural Psychology, 44*(3), 365-388. doi: 10.1177/0022022112443856.

Vasconcelos, S. J. L., & Hutz, C. S. (2008). Construção e validação de uma escala de abertura à experiência. *Avaliação Psicológica, 7*(2), 135-141.

Yarkoni, T. (2010). Personality in 100,000 words: A large-scale analysis of personality and word use among bloggers. *Journal of Research in Personality, 4*(3), 363-373. doi: 10.1016/j.jrp. 2010.04.001.

Yeo, T. P. E. (2010). Modeling personality influences on YouTube usage. In *Fourth International AAAI Conference on Web and Social Media*. Retrieved from http://www.aaai.org/ocs/index.php/ICWSM/ICWSM10/paper/view/1493.

* The manuscript is part of this dissertation as Manuscript 1.

Appendix 1

List of "stop words" that were automatically removed from the posts using the "tm" package (Feinerer & Hornik, 2015).

a	estão	hão	nossas	têm
à	estão	havemos	nosso	temos
ao	estas	havia	nossos	tenha
aos	estava	hei	num	tenham
aquela	estavam	houve	numa	tenhamos
aquelas	estávamos	houvemos	o	tenho
aquele	este	houver	os	tenho
aqueles	esteja	houvera	ou	ter
aquilo	estejam	houverá	para	terá
as	estejamos	houveram	pela	terão
às	estes	houvéramos	pelas	terei
até	esteve	houverão	pelo	teremos
com	estive	houverei	pelos	teria
como	estivemos	houverem	por	teriam
da	estiver	houveremos	qual	teríamos
das	estivera	houveria	quando	teu
de	estiveram	houveriam	que	teus
dela	estivéramos	houveríamos	quem	teve
delas	estiverem	houvermos	são	tinha
dele	estivermos	houvesse	se	tinha
deles	estivesse	houvessem	seja	tinham
depois	estivessem	houvéssemos	seja	tínhamos
do	estivéssemos	isso	sejam	tive
dos	estou	isto	sejamos	tivemos
e	eu	já	sem	tiver
é	foi	lhe	será	tivera
é	fomos	lhes	será	tiveram
ela	for	mais	serão	tivéramos
elas	fora	mas	serei	tiverem
ele	foram	me	seremos	tivermos
eles	foram	mesmo	seria	tivesse
em	fôramos	meu	seriam	tivessem
entre	forem	meus	seríamos	tivéssemos
era	formos	minha	seu	tu
era	fosse	minhas	seus	tua
eram	fosse	muito	só	tuas
éramos	fossem	na	somos	um
essa	fôssemos	não	sou	uma
essas	fui	nas	sua	verbo ser
esse	há	nem	suas	você
esses	há	no	também	vocês
esta	haja	nos	te	vós
está	hajam	nós	tem	
estamos	hajamos	nossa	tém	

Reference

A Portuguese stop word list. (2016, March 10). *Snowball*. Retrieved from: http://snowball.tartarus.org/algorithms/portuguese/stop. txt.

Appendix 2

List of terms in Portuguese in alphabetical order, with their respective frequency between parenthesis.

Adjectives

abandonada (1)
aberta (1)
aberto (1)
abestado (1)
abobada (1)
abobadinha (1)
abobado (1)
aborrecida (1)
aborrecido (1)
abrigada (2)
absoluta (2)
abusadao (1)
abusivo (1)
aceito (1)
acesa (1)
acomodado (1)
acompanhante (1)
acostada (6)
acostumada (1)
acostumado (4)
adepta (1)
adepto (1)
adestrador (1)
adimirado (2)
adivinha (1)
admirador (3)
adolescente (1)
adoravel (1)
adormecida (1)
adotada (1)
adulta (1)
adulto (3)
advinha (1)
afastada (1)
afetiva (5)
afiada (1)
afobado (1)
afrontada (2)
agil (2)
agitada (1)
agitado (1)
agoniada (1)
agoniado (2)
agradecida (3)
agradecido (1)
agressiva (1)
agressivo (1)
alcoalatra (1)
aleatorio (2)
alegre (1)
alegria (2)
alegrona (2)
aleijado (1)
alergico (1)
alheia (7)
alheiaa (2)
alone (1)
alta (1)
alto (1)
alucinada (1)
alucinado (1)
amada (3)
amado (5)
amante (1)
amarga (25)
amargo (2)
amarrado (1)
amavel (1)
ambiciosa (1)
ambulante (1)
amg - amigo(a) (1)
amiga (2)
amigao (84)
amigo (14)
amiguinha (1)
amiguinho (1)
amontoado (1)
amorosa (1)
amoroso (1)
amparada (4)
analfa (1)
analfabeto (1)
anao (1)
anarquista (1)
ancestral (1)
andante (2)
animada (4)
animal (3)
anormal (2)
ansiosa (1)
ansiosao (8)
ansioso (1)
anta (2)
antigo (130)
antinha (2)
antipatia (1)
antipatica (1)
antipatico (7)
antissocial (2)
apagada (2)
apaixonada (5)
apaixonadaaaaaaa a (105)
apaixonadaaaaaaa aa (1)
apaixonado (1)
apaixonante (39)
aparecida (6)
apegada (2)
apegado (5)
apelona (1)
apertado (3)
apoiada (1)
apologista (3)
apropriado (3)
aquariana (1)
armado (3)
arrepiada (2)
arretado (1)
arriada (1)
arrogante (2)
arrombada (10)
arrumada (1)
artilheiro (4)
aspirantes (1)
assanhadas (1)
assediada (1)
assertivo (1)
assexuada (1)
assistido (2)
assombrada (1)
assumidona (4)
assustada (1)
assustados (1)
astronauta (1)
ateia (1)
atendido (2)
atentado (1)
atentico (1)
atraente (1)
atrasada (23)
atrasadao (3)
atrasado (1)
atrativo (2)
atrevida (2)
atrevido (1)
atual (1)
ausente (1)
autista (2)
aventureira (2)
avoada (1)
azarada (2)
azarado (5)
aziada (1)
baao (1)
babaca (2)
babacona (12)
babada (2)
babado (1)
babona (1)
bacana (5)
bacano (2)
badalada (18)
baita (2)
baixa (2)
baixinha (2)
baixo (3)
bala (6)
baladeira (1)
banguela (1)
banido (9)
barata (1)
barato (2)
barraqueira (2)
barraqueiro (2)
barrigudinha (2)
barrigudo (2)
bbzinho - bebezinho (1)
beata (1)
bebada (5)
bebado (4)
bebezao (6)
bebezin (1)
bela (1)
beleza (1)
bella (2)
belo (1)
best (1)
big (1)
bilionario (1)
binitinho (2)
bipolar (2)
birrento (1)
bissexual (2)
bitch (3)
bitchcrazy (4)
bizarro (1)
blindado (1)
boa (1)
boazinha (1)
boba (1)
bobaaaa (13)
bobalhao (1)
bobao (1)
bobo (3)
bobona (9)
bolada (1)
bolado (2)
bom (1)
bonita (1)
bonitinha (5)
bonitinho (1)
bonito (5)
bonzin (1)
bonzinho (1)
booa (5)
borrachudo (2)
bossa (1)
bosta (1)
bostinha (19)
braba (2)
brabo (5)
brava (1)
bravo (7)
brigona (1)
briguenta (1)
brincalhao (7)
brincalhona (2)
brisado (2)
britada (1)
bruta (2)
brutas (1)
brutis (1)
bruto (1)
bugado (1)
bundona (7)
bunitin (3)
burra (1)
cabrao (1)
cachaceira (1)
cachaceiro (1)
cagada (2)
cagado (1)
cagao (4)
cagona (1)
caidinho (1)
caipira (1)
calaada (6)
calada (1)
caladao (1)
calado (1)
calma (2)

calmo (1)
canalha (4)
cansada (1)
cansado (5)
capaz (3)
capitalista (5)
caralhudo (1)
carente (1)
carentona (28)
careta (1)
carinhosa (1)
carinhoso (6)
caro (19)
casada (3)
casado (2)
casadona (6)
caseira (1)
caseiro (4)
castigada (1)
castrado (1)
casual (1)
cativo (1)
causador (25)
cauteloso (1)
cavalheiro (2)
cega (1)
cego (1)
ceguinho (1)
certinha (29)
certinho (2)
certo (3)
chapada (1)
chapadaaa (2)
chapado (1)
charmosa (1)
charmoso (1)
chata (2)
chateada (63)
chateado (5)
chatinha (1)
chato (1)
chatona (29)
chaveirinho (6)
chegada (33)
chegado (6)
cheia (15)
cheio (11)
cheirosa (1)
cheiroso (6)
chique (1)
chocada (1)
chorao (10)
chorona (14)
cismada (2)
ciumenta (2)
ciumento (1)
classica (4)
coisado (135)
coitada (1)
coitadinho (2)
coitado (1)
colorida (1)
colorido (1)
comandado (1)
competente (1)
competitiva (1)
competitivo (1)
completa (2)
completo (1)
complexa (2)
complexada (1)
complexado (1)
complicada (1)
complicado (2)
compreensiva (3)
compreensivo (1)
comprometida (2)
comprometido (2)
compromissada (1)
comum (1)
comunista (1)
conceitual (3)
confidente (2)
confusa (2)
confuso (7)
conhecida (1)
conhecido (1)
conquistada (1)
conquistador (1)
consciente (3)
considerado (1)
constrangedora (1)
construtiva (1)
contestador (1)
contra (3)
contrario (1)
controlada (1)
convencida (3)
convencido (1)
convincente (1)
corajosa (7)
corajoso (1)
corna (1)
corno (1)
corretissimo (1)
correto (1)
covarde (1)
cracudo (1)
crazy (1)
crente (1)
crianca (2)
criancinha (1)
criancona (1)
criativa (6)
criativo (2)
criatura (1)
criminosa (1)
cristalizada (2)
cristao (1)
critica (2)
criticado (3)
critico (2)
cuidadora (6)
culpa (2)
culpaaaaado (14)
culpada (1)
culpado (5)
cult (2)
culta (1)
cupido (1)
curiosa (1)
curioso (4)
cusao (12)
cusona (1)
custoso (1)
cuzao (1)
cuzona (4)
daantiga (1)
dado (2)
danadinho (2)
danado (1)
daora (3)
dark (4)
davida (3)
dboa (1)
dboazinha (1)
debochada (1)
debochado (5)
decidido (2)
dedicada (2)
dedicado (2)
defeito (1)
defeitos (5)
defensora (1)
delicada (1)
delinquente (1)
demente (1)
demonio (3)
dengosa (1)
depre (1)
depressiva (3)
deprimida (2)
deprimido (1)
derrotada (5)
derrotado (1)
desafinada (1)
desagradavel (1)
desajeitado (2)
desapegada (1)
desastrada (1)
desastrado (4)
desatualizado (1)
desbocada (1)
desbocado (1)
descansada (1)
descartavel (4)
descendente (3)
desconfiada (1)
desconfiado (2)
desconfortavel (1)
desconhecida (1)
descrente (1)
desencanada (1)
desesperada (2)
desesperado (1)
desgracada (8)
desgracado (15)
desgrassada (1)
desimpedida (1)
desinteressada (1)
desinteressante (1)
desleixado (5)
desligada (1)
desnaturada (1)
desnecessaria (1)
desocupada (1)
desocupado (2)
desorganizada (1)
desprezavel (1)
desprovida (1)
desprovido (1)
desrespeitada (1)
destemido (2)
destruidora (1)
detalhista (2)
determinada (1)
devota (1)
devoto (1)
diferentao (1)
diferente (3)
diferentona (5)
diferentonaaoo (12)
difernte (1)
dificil (1)
digna (2)
digno (2)
direita (1)
direitinho (3)
direito (1)
direta (17)
direto (1)
discreta (5)
discreto (1)
diva (1)
dividido (1)
divorciada (2)
doce (2)
docinho (1)
doente (33)
doentinha (1)
doida (1)
doidao (26)
doidinha (1)
doido (1)
doidona (16)
dorminhoca (13)
dozamigo (1)
dramatica (1)
dramatico (10)
drogadao (12)
drunk (1)
duente (1)
dura (1)
duro (19)
durona (3)
educada (3)
educado (6)
eficiente (6)
egoista (1)
elastica (1)
emburrado (1)
emocionada (3)
emocional (3)
emociono (3)
empregado (1)
encalhada (2)
encantado (1)
enganado (2)
enganador (1)
enjoada (1)
enjoadinha (4)
enjoado (1)
enorme (1)
enrolada (1)
enrolao (2)
enroscada (1)
entediado (1)
entendido (11)
entusiasta (1)
envergonhada (1)
envergonhadinho (3)
envergonhado (1)
envolvido (1)
equilibrada (1)
errada (3)
erradissima (4)
errado (1)
erradooo (22)
erradooooooo (1)
erro (3)
escandalosa (1)
esclarecido (1)
escolhido (2)
escondida (1)
escondido (1)
escorada (3)
escorpiana (1)
escrava (1)
escrota (8)
escroto (1)
escura (1)
escuro (2)
esganiaada (1)
eskimo (2)
esnobe (1)
especial (3)
especialista (6)
esperta (25)
esperto (10)
espirita (1)
esquecida (16)
esquecido (4)
esquerda (2)
esquezita (2)
esquisita (1)
esquisito (4)
estagiaria (8)
estilosa (8)
estorvo (1)
estragada (1)
estrago (1)
estranha (1)
estranho (1)
estranhona (17)
estressada (1)
estressado (9)
estrondosa (1)
estudada (1)
estudado (2)
estupida (1)

estupido (2)
eterna (1)
eterno (1)
exagerada (2)
exagerado (1)
exato (1)
exatooooo (4)
exausto (1)
excelente (2)
excluido (2)
exclusiva (2)
exclusivo (1)
executivo (1)
exemplar (6)
existencial (14)
exorcista (2)
expert (1)
extraordinaria (1)
extravagante (1)
extremista (10)
extremo (3)
extrovertida (3)
extrovertido (3)
facil (1)
fadada (4)
falador (98)
falida (1)
falsa (29)
falsiane (2)
famosa (1)
famosinho (8)
famoso (2)
fanatica (9)
fantasma (1)
fantastico (2)
farreiro (3)
farsa (1)
farta (1)
fascinada (9)
fascista (1)
fatal (1)
favelada (6)
favelado (1)
favorita (53)
favorito (1)
fechada (2)
fechado (5)
fei (1)
feia (1)
feio (2)
feiosa (3)
feioso (1)
feliiiiiiiz (1)
feliiiiiiz (1)
feliiz (1)
felix (1)
felixz (1)
feliz (1)
feminina (1)
feminista (5)
fera (1)
feraaaaaa (3)
ferida (1)
ferido (2)
ferrada (1)
festeira (5)
fiel (1)
fina (1)
fingida (4)
fino (1)
fluente (1)
foda (1)
fodaaaa (49)
fodao (1)
fodida (1)
fodido (1)
fodona (3)
fofa (1)
fofao (1)
fofinha (2)
fofinho (5)
fofo (2)
fofoqueira (2)
fofoqueiro (1)
fofuraaaaaaaa (1)
folgado (2)
forte (1)
fraca (17)
fracaa (19)
fracaaaa (1)
fracassada (1)
fracassado (4)
fracasso (1)
fraco (7)
fragil (1)
fraquinha (4)
fraquinho (1)
fresca (14)
fresco (6)
frescurinha (3)
fria (3)
friend (1)
frieza (2)
frio (1)
frouuxa (1)
frouxa (1)
frustada (1)
frustrada (1)
fudida (1)
fudido (1)
funqueira (1)
futuro (3)
galante (1)
gamadao (1)
gamadinho (1)
gamado (1)
gamer (1)
garantida (1)
garantido (1)
garota (1)
garotinha (1)
garotinho (4)
garoto (1)
gata (1)
gataa (18)
gatao (1)
gatinho (1)
gato (2)
gatos (9)
geminiana (1)
geminiano (3)
genial (1)
genio (1)
girl (1)
gladiadora (1)
glamourosa (1)
global (1)
golpista (1)
good (2)
gorda (1)
gordao (18)
gordinha (1)
gordinho (2)
gordo (3)
gostosa (1)
gostoso (17)
gostosooooo (12)
gotica (1)
gotico (2)
gracinha (20)
gracious (2)
grandao (1)
grande (1)
grandinha (3)
grandinho (1)
grata (1)
grato (19)
grosa (1)
grossa (2)
grosso (1)
grudenta (1)
guerreira (1)
guerreiro (2)
gulosa (1)
guri (1)
guria (2)
habilidoso (1)
hard (1)
hardcore (2)
hater (1)
hetera (1)
hetero (1)
heterofobica (19)
heterossexual (1)
hilario (1)
hiperativo (1)
hipocondriaca (1)
hippie (1)
homenzinho (9)
homicida (3)
homo (1)
homossequissual (1)
homossexual (1)
honesta (1)
honrada (1)
honrado (2)
horrivel (1)
horrorosa (1)
horroroso (2)
human (2)
humana (1)
humano (9)
humilde (1)
humorada (6)
humorado (4)
idiota (1)
idiotona (2)
idosa (1)
ignorada (2)
ignorante (1)
igual (4)
igualzinha (31)
iludida (1)
iludido (7)
iluminado (1)
imbecil (2)
imensa (1)
imigrante (1)
imortal (1)
impaciente (1)
imparcial (1)
implicante (1)
importante (1)
impossible (2)
impossivel (1)
impressionada (1)
impressionante (1)
imprestavel (1)
inaceitavel (2)
inativa (1)
incapaz (1)
incerto (1)
incoerente (7)
incomodada (1)
inconformado (1)
inconquistavel (2)
inconstante (1)
inconveniente (1)
incoviniente (4)
incrivel (12)
indecisa (1)
indeciso (1)
indefeso (5)
indelicada (1)
independente (1)
indiferente (1)
indignada (4)
infeliz (1)
infelizes (1)
inferior (10)
infiel (5)
informada (3)
ingenua (1)
ingrato (1)
inimiga (1)
inimigo (1)
inimigos (7)
inocente (1)
inofensiva (1)
inovador (1)
inquietaaao (1)
insane (1)
insegura (4)
inseguro (2)
insensavel (2)
insistente (2)
insone (4)
instavel (1)
insuportavel (1)
inteira (1)
inteiro (4)
intelectual (12)
inteligente (1)
intensa (1)
intenso (1)
interessante (1)
interesseira (3)
intima (1)
intimo (2)
intolerante (1)
introvertido (2)
ironica (2)
ironico (1)
irrelevante (1)
irresponsavel (1)
irritante (3)
irritantee (2)
jogador (3)
joia (1)
jovem (1)
jovenzinha (3)
julgada (4)
julgado (1)
jurassica (1)
justa (1)
justo (3)
ladra (1)
ladrao (3)
ladrona (1)
lagarto (1)
lamento (2)
larga (1)
largo (1)
leal (2)
legal (1)
legalll (5)
legalzao (1)
legalzinha (1)
legalzinho (1)
lenda (1)
lenta (8)
lento (2)
leoazinha (1)
leonino (1)
lerda (23)
lerdaaa (37)
lerdo (1)
lerdona (8)
lesa (1)
lesada (1)
lesadaaaa (4)
lesbica (1)
levada (6)
liberal (2)
libriana (5)
libriano (3)
ligada (1)
ligado (4)
ligadona (10)
limpinha (1)
linda (1)
lindaa (76)
lindao (1)
lindinha (3)
lindo (2)

lindona (1)
lindooo (1)
lindooooooooo (1)
linear (1)
liso (1)
livre (1)
loca (1)
local (17)
loco (1)
locona (7)
loka (1)
lokao (1)
loko (1)
lokona (3)
lord (1)
lorde (2)
louca (3)
loucaaaa (67)
louco (1)
loucona (48)
loucos (1)/
loved (1)
lutador (4)
macabra (1)
machine (1)
machista (1)
macho (6)
machucado (1)
maconheiro (2)
madura (1)
maduro (1)
magic (1)
magico (1)
magoada (1)
magoado (1)
maior (1)
mala (71)
malandro (2)
maldisposto (1)
maleavel (1)
maloqueiro (1)
maloquera (7)
maluca (2)
maluco (22)
maluko (1)
maluquinha (1)
maluquinho (1)
malvado (1)
malz (1)
manca (1)
maneira (1)
maneirinho (6)
maneiro (1)
manera (1)
manhosa (1)
maravilhosa (3)
maravilhoso (1)
maravilhousa (1)
marginal (1)
marrenta (1)
marrento (3)
marxista (1)
masculina (1)
masoq (1)
masoquista (1)
matador (1)
mau (2)
maximo (1)
medrosa (1)
medroso (9)
meiga (1)
meigo (3)
melhor (1)
melhorr (8)
melhorzinho (1)
melior (2)
melosa (1)
meloso (1)
menor (1)
mentirosa (1)
mentiroso (2)
merdinha (6)
merecedor (12)
metaleiro (1)
metida (2)
metido (6)
miga (1)
migo (1)
miguxa (1)
milagre (1)
milionaria (1)
militar (1)
mimada (214)
mimado (1)
mina (11)
miseravel (1)
misterio (1)
mistica (1)
mistura (1)
mita (1)
mito (1)
miudaaaa (1)
mlk - moleque (1)
moca (3)
mocinhaa (1)
moderna (1)
mole (1)
moleque (1)
monstra (2)
motivado (1)
mucha (1)
mudada (3)
mudo (9)
mula (1)
mutante (1)
nato (1)
natural (2)
necessario (1)
necessitada (2)
nerd (1)
nerdd (1)
nervosa (1)
nervoso (1)
neurada (1)
neuratica (1)
neuratico (1)
ninfomanaaca (1)
noia (1)
nojenta (1)
nojento (5)
nojinho (3)
nojo (1)
normal (2)
normali (42)
normalzinha (5)
normauu (1)
nostalgica (1)
nostlgica (1)
noturna (2)
noturno (1)
nova (3)
novato (1)
novinha (1)
novinho (2)
novo (1)
obcecada (1)
obcecado (2)
obesa (1)
obeso (1)
objetivo (1)
obrigada (2)
obrigadaaaa (3)
obrigadaaaaa (1)
obrigadaaaaaaaa (1)
obrigado (1)
obscuro (2)
obvio (2)
ocupada (1)
ocupado (2)
odiada (24)
older (1)
oposto (3)
opressorrrrrrrrr (1)
oprimidissimo (1)
orfa (3)
organizada (1)
organizado (2)
orgulho (2)
orgulhosa (16)
orgulhoso (27)
original (1)
otaria (1)
otariane (1)
otariazona (1)
otario (1)
otima (1)
otimo (1)
ousada (6)
ousado (1)
paciente (3)
palha (1)
palhaaa (2)
palhaaada (1)
palida (3)
panaca (1)
pancada (1)
panda (4)
panqueca (1)
parada (1)
parado (1)
paradona (2)
paranoica (1)
paranoico (1)
parceira (1)
parceiro (2)
parecida (2)
pasma (1)
paspalha (1)
passada (21)
passado (3)
passional (1)
passivo (1)
pecadora (7)
pegador (1)
pegadora (3)
pequena (1)
pequenininho (8)
pequenino (1)
pequeno (2)
perdedora (5)
perdida (1)
perdido (1)
perdoada (1)
perfeccionista (1)
perfect (1)
perfeita (1)
perfeito (6)
perigosa (5)
perigoso (1)
perseguido (1)
perturbada (1)
perturbado (2)
perversa (1)
pervertida (1)
pesada (1)
pesado (1)
pessima (1)
pessimista (8)
piadista (1)
pior (1)
pirada (3)
pirado (1)
pisciana (1)
pisciano (1)
pisico (5)
playboy (1)
plena (3)
pleno (1)
poderosa (18)
poderoso (1)
podre (1)
polemico (1)
pontual (1)
poor (1)
pop (1)
popozuda (2)
popular (1)
porreiro (1)
poser (1)
positiva (3)
positivo (3)
possessiva (1)
potranca (1)
praieiro (10)
pratico (7)
precioso (1)
precipitada (1)
precisa (1)
preciso (1)
precoce (52)
preconceituosa (1)
preconceituoso (1)
preguicosa (12)
preguicoso (16)
prendada (1)
preocupada (1)
preocupado (5)
preparada (1)
preparado (1)
prepotente (1)
presa (1)
prestativo (1)
primeira (1)
primeiro (3)
primor (1)
prisoner (8)
procuradaa (4)
profissional (2)
profundo (1)
proibidao (1)
proibido (1)
pronta (1)
prontinho (2)
pronto (1)
protagonista (3)
protegido (1)
protetora (2)
proxima (1)
psico (1)
psicopata (1)
psycho (1)
punheteiro (1)
pura (1)
purinho (2)
puro (1)
purple (6)
purpura (1)
purpurina (1)
quebrada (2)
quebradaaaa (5)
quebrado (1)
quebradora (1)
queer (2)
quente (1)
querida (1)
queridinha (3)
queridinho (1)
querido (1)
queridoooooo (1)
questionador (7)
quieta (1)
quietinho (1)
quieto (3)
racional (1)
racista (1)
radical (1)
rancorosa (2)
rancoroso (2)
rapidao (1)
rapido (1)
rara (5)
raro (1)
rascista (1)
rasgada (1)

real (4)
realidade (2)
realista (9)
realistona (1)
realizada (2)
realizado (1)
recebida (2)
redonda (1)
refugiada (2)
regular (1)
relaxada (3)
religiosa (4)
religioso (1)
resolvida (1)
respeitada (11)
responsavel (1)
retardad (1)
retardada (1)
retardadinha (1)
retardado (1)
revoltado (2)
revolucionaria (1)
rica (1)
ricaa (25)
ricaaaa (1)
rico (1)
ridicula (5)
ridiculo (1)
romantica (2)
romantico (7)
romantiquinha (8)
ruim (2)
ruinzao (2)
ruinzona (1)
sad (44)
sadica (2)
safada (1)
safadao (2)
safadinha (2)
safado (2)
safo (9)
sagiotariana (2)
sagita (2)
sagitariana (1)
sagitariano (1)
sagrado (1)
salgadinha (2)
saliente (1)
salvador (3)
samaritano (1)
santa (1)
santinho (1)
santo (1)
sapequinha (3)
sapiens (3)
saradinho (1)
sarcastica (1)
sarcastico (2)
satisfeita (1)
saudavel (8)
scientist (1)
sebosa (1)
seca (1)
seco (1)
secona (3)
secreto (2)
sedentaria (1)
seguida (2)
seguido (1)
seguidor (1)
segura (1)
seguro (2)
seletiva (1)
seletivo (3)
semelhante (3)
sensitiva (1)
sensivel (1)
sensual (1)
sentida (1)
sentido (1)
sentimental (1)
sentinela (12)
sequelado (1)
serena (2)
sereno (1)
serio (2)
serioooooo (21)
sexy (1)
silencioso (1)
simpatica (16)
simpatico (6)
simpatizante (2)
simples (1)
sincera (1)
sincero (1)
sinistro (1)
sobrenatural (1)
sobrevivente (4)
socialista (7)
sociavel (1)
sofredora (4)
sofrida (1)
solidario (4)
solitaria (1)
solitario (1)
solteira (1)
solteiro (7)
solto (1)
sonambula (1)
sonambulo (2)
sonhador (1)
sonho (4)
sonsa (50)
sortuda (1)
sortudo (8)
sozinha (1)
sozinho (23)
sozinhooooooo (23)
stressada (1)
suavao (1)
suave (1)
suavona (7)
suficiente (1)
suflair (1)
sufocada (1)
suicida (2)
suja (1)
sujeira (1)
sujo (2)
sumida (3)
super (1)
superdotada (6)
superior (1)
surdo (1)
surreal (2)
tadinha (1)
tadinho (1)
tagarela (3)
tarada (2)
taurina (1)
teimosa (1)
teimoso (13)
tensa (2)
tenso (3)
ternurinha (2)
terrivel (5)
terrorista (1)
tesao (2)
testemunha (1)
todinha (1)
todinho (6)
tola (1)
tolinha (1)
tomba (2)
torta (1)
torto (1)
tradicional (1)
tradicionalista (2)
traidor (1)
tranquila (1)
tranquilao (1)
tranquilo (1)
tranquilona (4)
traste (1)
tratada (8)
traumatizada (1)
treinada (1)
treinado (2)
triste (2)
tristeeeee (28)
trouxa (1)
trouxaaaaaaaaa (112)
trouxae (1)
trouxao (1)
trouxiane (2)
trouxxxxxxxaaaa (2)
troxa (1)
ultima (1)
ultimo (2)
ungida (1)
unica (1)
unicaaa (23)
unico (4)
urbana (5)
usada (14)
usado (1)
usurpador (1)
vacilao (2)
vacilona (9)
vacinado (3)
vagaba (3)
vagabunda (1)
vagabundo (6)
valente (1)
vazio (3)
veeeeei (1)
veeey (1)
vegetariana (2)
velha (2)
velhaca (25)
velho (1)
velhoi (12)
vencedor (2)
venenoso (3)
vesga (1)
viajada (3)
viajadona (2)
viajante (3)
viavel (1)
viciada (1)
viciado (24)
vidente (1)
vigiada (2)
vingativa (1)
vingativo (3)
violao (4)
violenta (5)
violento (1)
virado (7)
viralata (4)
virginiana (1)
virginiano (1)
viva (32)
vivaz (7)
vivida (1)
vossa (1)
vosso (1)
vulgar (1)
vulneravel (2)
winner (1)
xonado (2)
zangada (1)
zikada (1)
zoada (4)
zoado (1)

Nouns

abacaxi (1)
abelha (1)
advogada (3)
afronta (1)
agito (2)
alce (1)
alimento (1)
alma (5)
almaaaa (9)
amooooor (1)
amoor (1)
amor (3)
amore (1)
amori (1)
amorinha (2)
amorrr (1)
amorrrr (4)
amorrrrrr (1)
amorzinho (1)
angeell (1)
angel (1)
anjaaaaaa (6)
anjinho (1)
anjo (4)
aprendiz (5)
aranha (1)
aries (1)
arquiteta (1)
arroz (1)
artista (1)
asno (2)
assaltante (1)
assassino (1)
ateu (1)
atleta (1)
aviao (1)
baabyy (2)
baby (1)
bagulho (1)
baixaria (3)
balao (1)
balaozinho (1)
baleia (1)
banana (1)
bandido (2)
barra (1)
batata (4)
besta (1)
bicha (4)
bichao (9)
bichinha (1)
bichinho (2)
bicho (1)
bixa (1)
bixinha (2)
bixo (1)
black (2)
bocado (20)
bode (1)
bofe (1)
boi (1)
bola (1)
bolacha (7)
boleiro (3)
bolinho (1)

bombeiro (1)
bombom (1)
boneco (2)
borboleta (2)
borracha (1)
branca (1)
branco (3)
branquinha (2)
brilho (1)
brincadeira (3)
brinquedo (1)
bro - brother (1)
bronca (4)
brutalidade (1)
budista (1)
bunda (1)
burguesia (4)
burrao (81)
burro (1)
burrona (26)
burrrrooooooooo
ooooooo (1)
cabelo (1)
cabra (1)
cachorra (1)
cachorrao (3)
cachorrinho (1)
cachorro (2)
cachorrona (11)
cadela (1)
cadelaaa (4)
cafetao (1)
cafetina (1)
caiacara (12)
calango (2)
calculadora (2)
camaleao (1)
camarao (1)
camelo (1)
caminho (1)
campea (1)
campeao (1)
cancer (2)
canceriana (1)
canela (1)
cantor (12)
cantora (2)
capoeirista (3)
capricornio (1)
carne (1)
chef (1)
chefe (2)
chiclete (1)
chinelo (2)
chinesa (1)
chocolate (1)
chumbinho (2)
cigana (1)
cigano (1)
cinderela (3)
cobaia (1)
cocota (1)
codorna (1)
coelha (1)
coelhinho (1)
coelinha (1)
coisa (1)
coisinha (1)
colega (2)
coleguinha (1)
compositor (1)
conhaque (4)
coracao (2)
coroa (1)
coroinha (1)
coruja (2)
corujao (4)
costelinha (1)
coxa (1)
cozinheira (1)
cria (3)
criado (1)
cuuuuuuu (1)
dama (1)
deeeus (2)
derrota (1)
desastre (3)
descarte (2)
desenhista (1)
deusa (166)
deussss (1)
deusssss (1)
devil (2)
diabo (1)
dog (1)
dona (4)
dono (1)
doutor (1)
droga (1)
duende (1)
duque (2)
elfo (1)
emo (1)
encanador (2)
entregador (2)
escorpiao (2)
escoteiro (1)
escudo (6)
espada (4)
espetaculo (4)
esponja (1)
esposo (1)
estrela (1)
excecao (1)
exemplo (1)
falha (1)
familia (1)
feiticeira (1)
filha (1)
filho (32)
florzinha (1)
fogo (1)
fogueira (2)
frango (1)
fruta (1)
fruto (1)
gajo (10)
galinha (1)
gangsta (1)
geladeira (1)
gelo (2)
gent (1)
gente (3)
genteeee (1)
genteeeee (1)
gloria (1)
gospel (1)
grude (1)
guia (5)
hero (2)
heroi (1)
homem (1)
honey (1)
ilha (1)
imperator (1)
interface (1)
jacare (2)
jacarezinho (1)
javali (1)
jedi (1)
joguete (1)
judas (1)
juiz (1)
leader (1)
leao (1)
leitao (2)
libra (3)
lobo (1)
loira (1)
loirinha (1)
loiro (1)
macaco (1)
madrinha (1)
mae (1)
mainha (4)
marido (1)
mel (1)
melancia (1)
melao (2)
menina (1)
menininha (12)
menino (26)
mensageiro (1)
merda (1)
mestra (1)
mestre (1)
metralhadora (2)
mico (1)
monstro (1)
morango (1)
morcega (1)
morcego (8)
morena (1)
moreno (3)
morta (6)
morto (7)
mosquita (1)
mozao (1)
muleque (2)
mulheeeeeer (2)
mulher (1)
mulherao (80)
mulherrrr (1)
mulherzinha (2)
musa (1)
naada (1)
nada (1)
nadaaaa (3)
nadaaaaa (1)
nadaaaaaa (2)
nadaaaaaaaaaa
(2)
naja (1)
namorada (9)
namoradinha (1)
namoradinho (2)
namorado (1)
natureza (1)
ninja (13)
novidade (3)
objeto (1)
odalisca (1)
ogra (1)
ogrinha (2)
ogro (1)
orelha (1)
osso (1)
ouro (1)
ovelha (1)
pai (1)
painho (1)
paizinho (1)
palhaco (4)
pao (1)
passageira (9)
passageiro (1)
passarinha (1)
passarinho (1)
passaro (1)
passatempo (4)
pastora (7)
patrao (1)
patricinha (2)
patroa (1)
pavoa (1)
peao (1)
pedreiro (6)
peguete (5)
peixes (29)
pelicano (3)
pereba (1)
perigo (1)
perola (3)
personificacao
(1)
pesadelo (1)
piada (1)
pimenta (1)
pimentinha (3)
piranha (1)
pirata (1)
piriquito (1)
pitanguinha (1)
poeta (1)
poetisa (1)
policial (1)
porca (1)
porco (2)
porquinha (243)
porta (1)
poste (2)
prancipe (2)
princesa (1)
problema (1)
prostituta (1)
puta (1)
putinha (1)
putinhas (1)
puto (1)
puuuuta (2)
puuuutaaaaaaaaa
aa (1)
queen (1)
rapariga (3)
rastafari (1)
rato (2)
rei (1)
rosa (1)
rosinha (1)
ruela (1)
ruiva (1)
ruivo (2)
saco (1)
sapatao (1)
sapo (2)
sardinha (1)
sata (1)
satanas (1)
sereia (1)
servo (1)
socialite (2)
sogra (7)
spider (2)
sucesso (1)
terror (1)
tigrao (2)
titia (1)
titio (1)
tormenta (1)
toupeira (1)
touro (6)
trabalhador (3)
trabalhadora (2)
traficante (1)
urubu (1)
vaca (1)
vadia (6)
vadio (5)
vampira (1)
vampiro (3)
veneno (1)
viaaado (17)
viada (1)
viadao (1)
viado (1)
virgem (1)
vulcao (5)
wicca (1)
zoeira (1)
zombie (1)
zuera (1)
zumbi (2)

Adverbs

altamente (5)
brutalmente (1)
demais (2)
demaisss (1)
demaissss (1)
demasiado (1)
relativamente (1)

Contractions

daquela (1)
daquelas (4)
daquele (11)
daqueles (3)
dessas (36)
desses (30)
destas (3)

Names

bambam (3)
barbie (2)
einstein (7)
hulk (1)
mickey (1)
princesaphiona (10)

Pronoums

algo (2)
alguem (2)
algum (1)
alguma (15)
qualquer (3)
tudo (1)
tudooh (1)
tudoooo (1)
tuudoa (2)
tuuudo (1)

Manuscript 3

Developing dimensional models for a Brazilian personality lexicon based on text mining of Twitter: Adjectives

Abstract

We investigated a Brazilian personality lexicon using the social network Twitter as a source of descriptors and data. The dimensionality of a term-document matrix with 172 adjectives and 86,899 subjects was explored with Latent Dirichlet Allocation topic modeling. Cross-validation analyses suggested models with 7 and 14 topics as the most suitable for the data. We examined these two models and also five prominent theoretical models such as the Big Five, the three-factor, the six-factor and the 16PF model and compared the semantic content of the topics in our models with the content of factors from these prominent models. The results suggested that prominent models such as the Big Five did not emerge from our data. Furthermore, the interpretation of the models with seven and 14 topics indicated that these are promising candidate models for future research, with an inclination for the last model, whose dimensions showed more internal semantic coherence.

Keywords: personality; lexical hypothesis; text mining; machine learning; Brazilian Portuguese; Big Five.

Manuscrito 3

Desenvolvendo modelos dimensionais para o léxico brasileiro da personalidade com base na mineração de textos do Twitter: Adjetivos.

Resumo

Nós investigamos o léxico da personalidade brasileira usando a rede social Twitter como fonte de descritores e de dados. A dimensionalidade de uma matriz de documento-termo com 172 adjetivos e 86.899 indivíduos foi explorada com modelagem de tópicos Latent Dirichlet Allocation. As análises de validação cruzada sugeriram modelos com 7 e 14 tópicos como os mais adequados para os dados. Examinamos estes e outros cinco modelos teóricos proeminentes (e.g., três, cinco, seis e 15PF) e comparamos o conteúdo semântico dos tópicos em nossos modelos com o conteúdo de fatores desses modelos. Os resultados sugeriram que modelos proeminentes como o Big Five não emergiram dos dados. A interpretação dos modelos com sete e 14 tópicos indicou que estes são promissores modelos para pesquisas futuras, com uma inclinação para o último, cujas dimensões apresentaram maior coerência semântica interna.

Palavras-chave: personalidade; hipótese léxica; mineração de texto; aprendizagem de máquina; português brasileiro; *Big Five*.

The fundamental postulate of the lexical hypothesis is that the most salient and relevant personality traits become encoded in the natural language in different cultures throughout its history. Some of the most prominent personality models were developed following this postulate, such as the Cattell's 16 primary personality factors (16PF) and the five-factor model, or Big Five personality model, also known simply as Big Five (De Raad & Mlacic, 2015; Goldberg, 1981; John, Angleitner, & Ostendorf, 1988; Peres & Laros, 2018a). Although the substantial influence of these models, the psycholexical approach has been criticized regarding some of its epistemological and methodological aspects.

Peres (2018a; 2018b) identified two crucial sets of critiques regarding the research strategies traditionally adopted in the psycholexical approach. The first comes from the perspective of cross-cultural psychology (Cheung, Van de Vijver, & Leong, 2011; Daouk-Öyry, Zeinoun, Choueiri, & Van de Vijver, 2016). Although the lexical hypothesis epistemologically represents an *emic* perspective, the mainstream research in this field is conducted from an *etic* imposed perspective (Cheung et al., 2011). In other words, the *emic* perspective is an indigenous (i.e., autochthonous or native) approach dedicated to the study of culture-specific phenomena and the validity of potentially universal theories. In contrast, the *etic* imposed perspective is an approach concerned with the investigation of the generality and validity of models and theories developed in specific cultures to other cultural contexts.

The first and more influent psycholexical personality models were initially constructed following an *emic* perspective. It was in countries like the United States, The Netherlands, and Germany that models like the 16PF and the Big Five were developed (De Raad & Mlacic, 2015; John et al., 1988). After consistent models were found in these cultures, the next research endeavor was to investigate whether these

models were universal and composed by fundamental dimensions of personality. The main strategy to answer these questions was to assess the generalizability of the models to other cultures through the translation and adaptation of psychometric instruments. Notwithstanding the merit and the many fundamental contributions to personality research made with this strategy, the identification of indigenous models might be impaired (Peres, 2018a, 2018b).

An emic-etic integration has been proposed to tackle this issue (Cheung et al., 2011; Daouk-Öyry et al., 2016; Valchev et al., 2012). Such integration occurs by means of a combination of universal and culturally specific aspects of personality. The advocates of this perspective also argue that methodological choices in psycholexical research (e.g., selection of sources of personality descriptors, and data collection strategies) should be guided by the features of the investigated language and culture.

The second group of critiques comes from psychological studies of natural language. While the traditional psycholexical strategy has historically focused on dictionaries as the primary source for the identification of personality descriptors (Cheung et al., 2011; Daouk-Öyry et al., 2016; Uher, 2015), there are alternative sources that were not explored as intensely. Alternative sources can include journalistic, literary, and scholarly texts (e.g., Allik et al., 2011; Farahani, De Raad, Farzad, & Fotoohie, 2016), semi-structured interviews (Nel et al., 2012), free-descriptions of the self or of other people (Isaka, 1990), flow of consciousness reports (Lee, Kim, Seo, & Chung, 2007), conversations records (Polzehl, 2015), blogs (Yarkoni, 2010), and social networks such as Twitter and Facebook (Park et al., 2015; Peres & Laros, 2018b; Qiu et al, 2012).

The second critique in this group is related to the use of psychometric instruments with a limited set of items and administered in testing scenarios as the principal

procedure of data collection (Chung & Pennebaker, 2008). The limited number of items can circumscribe the emergence of latent traits to the content of these items. Also, the test setting can be restrictive in capturing the free expression of personality traits by the subjects.

This study seeks to contribute in overcoming some of the issues described above by exploring an online social network as a natural language source for gathering personality trait descriptors and by assessing its dimensionality. In a previous effort (Peres, 2018b), we elaborated a list of personality descriptors retrieved on Twitter ("Twitter", 2017). As a result, we obtained a list of 1,118 adjectives and 332 nouns, as well as adverbs, contractions, names, and pronouns. Although this effort was successful in the process of extracting data from the social network and in subsequent organization of the extracted terms according to classes of words and frequencies, we concluded that some improvements were necessary. For instance, it was required to expand the diversity and the volume of the data to allow the use of dimensionality reduction techniques. In our previous study (Peres & Laros, 2018b), we retrieved data from 5,435 Twitter users, from which 97.5% employed just one or two terms in their posts.

This way, from an *emic* perspective, the overall objective of this study was to investigate the lexical structure of personality in Brazilian Portuguese through text mining and dimensionality reduction analyses of a larger sample of Twitter users and posts. From an *emic-etic* perspective, this study also has the objective of comparing the structure found with prominent psycholexical models developed in other cultures. Next, before reporting the method and procedures of this study, we present a brief description of the most prominent psycholexical models in order to contextualize the analysis and the discussion of the results. The text mining techniques employed in this study are

described in the Method section (Blei, Ng, & Jordan, 2003; Leskovec, Rajaraman, & Ullman, 2014; Kosinski, Wang, Lakkaraju, & Leskovec, 2016; Silge & Robinson).

The psycholexical personality models

Cattell's 16 primary factors and five global factor model.

Cattell's 16PF model is composed of 16 primary personality factors. Each primary factor is named after the high pole of its scale, and it is also identified by a letter that indicates the alphabetical order in which it was empirically detected, as well its importance as a personality trait. The missing letters refer to factors that were dropped by 16PF authors. Fifteen of the primary factors are incorporated into five global or higher order factors. The exception is the Reasoning factor, which is not a personality factor, but an ability measure. Therefore, it is not nested in any global factor.

The five global factors are described in terms of the content of the primary factors (Figure 1), and the 16 primary factors can be described as follows, with examples of descriptors (H. E. P. Cattell and Schuerger, 2003):

1. Warmth (A): reserved, unemotional vs. warm, sympathetic.
2. Reasoning (B): low abstract reasoning vs. high abstract reasoning.
3. Emotional Stability (C): reactive, temperamental vs. calm, even-tempered.
4. Dominance (E): deferential, cooperative, docile vs. dominant, assertive, bossy.
5. Liveliness (F): serious, quiet, cautious vs. enthusiastic, animated, spontaneous.
6. Rule-Consciousness (G): careless of rules vs. dutiful, moralistic.
7. Social Boldness (H): timid, threat-sensitive vs. socially bold, fearless.
8. Sensitivity (I): unemotional, hard, cynical vs. empathic, sentimental, aesthetic.
9. Vigilance (L): trusting, tolerant, gullible vs. suspicious, skeptical, competitive.
10. Abstractedness (M): pragmatic, realistic vs. imaginative, contemplative.
11. Privateness (N): open, unguarded, genuine vs. private, guarded, calculating.

12. Apprehension (O): self-assured, placid vs. apprehensive, self-depreciating.

13. Openness to Change (Q1): prefers status quo vs. freethinking, experimenting.

14. Self-reliance (Q2): group-oriented, affiliative vs. individualistic, self-reliant.

15. Perfectionism (Q3): undisciplined, careless vs. organized, self-disciplined.

16. Tension (Q4): patient, relaxed, tranquil vs. impatient, tense, restless.

Figure 1. Cattell's 16PF five global factors (in the first level) and their primary factors (in the second level). Adapted from H. E. P. Cattell and Schuerger (2003). The letters between parentheses are the alphabetic designations of the 16PF primary scales, the lacking letters refers to Reasoning (B) and factors that were dropped by 16PF authors. The symbols represent the low (-) or the high (+) pole of each scale range.

H. E. P. Cattell and Schuerger (2003), Primi, Ferreira-Rodrigues, and Carvalho (2014), and John et al. (1988) offer an historical overview of the development of this model, also introduced in the first manuscript of this dissertation (Peres & Laros, 2018a).

The five-factor model or Big Five.

The five-factor model was developed on basis of Cattell's work, by scientists such as Fiske (1949), Tupes and Cristal (1961), Norman (1967), Goldberg (1981), and Costa and McCrae (1976). For a while, the model was known as the *Norman five*, later earning the alias *Big Five* (Goldberg, 1981). The history of the development of the five-factor model is reviewed in great detail by De Raad and Milacic (2015), Digman (1990), and John et al. (1988), and it was also subject of the first manuscript of this dissertation (Peres & Laros, 2018a). The Big Five is formed by the factors Surgency, also known as Extraversion or Extraversion-Introversion, Agreeableness, Conscientiousness, and Intellect, also called Openness to Experience, and Culture (Figure 2).

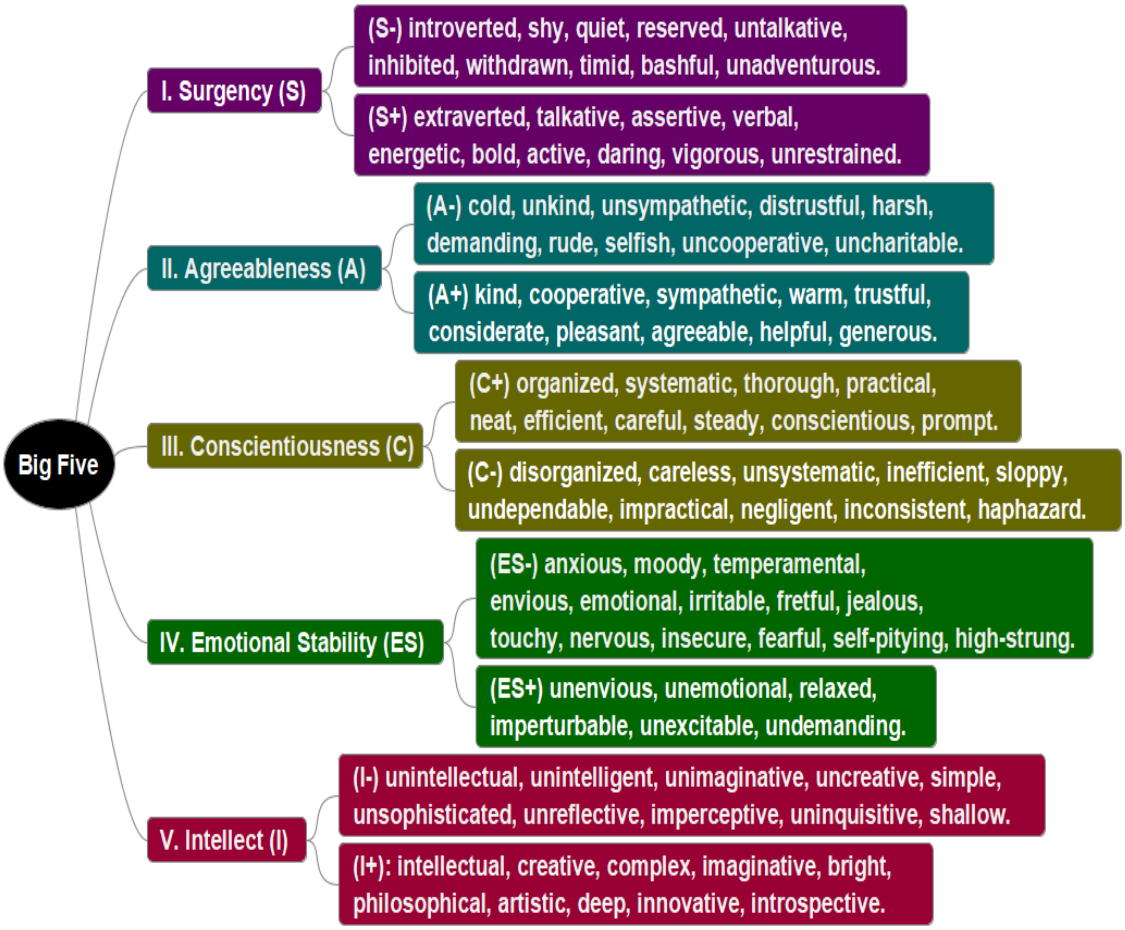

Figure 2. The Big Five model (Goldberg, 1992).

Eysenck's three-factor model.

Known as Eysenck´s PEN System, this personality model is formed by the factors Psychoticism, Extraversion, and Neuroticism (Eysenck, 1991; S. B. G. Eysenck, Eysenck, & Barret, 1985). The PEN model was not developed on basis of the psycholexical approach the same way as Cattell's 16PF or the Big Five (Revelle, 2016). Rather, Eysenck's research was strongly influenced by experimental psychology and focused on the biological basis of personality beyond the problem of the taxonomy (Eysenck, 1997). In fact, Eysenck was a critic of many aspects of the psycholexical approach (Eysenck, 1991). Nevertheless, the PEN model is often compared with the Big Five and the 16PF model due to the many empirical and theoretical similarities they share (Eysenck, 1991; Eysenck, 1992; Goldberg & Rosolack, 1994). Eysenck (1991) considered the Psychoticism dimension as a higher order factor of which Agreeableness and Conscientiousness are facets. Other authors see this dimension not as a higher order one, but as a blend of these two factors (Goldberg & Rosolack, 1994). The PEN factors can be described as follows (S. B. G. Eysenck et al., 1985):

- Psychoticism: aggressive, impulsive, antisocial, masculine, egocentric, etc.
- Extraversion: sociable, sensation-seeking, risk-taking, lack of reflection, etc.
- Neuroticism: anxious, depressed, low self-esteem, timid, moody, tense, etc.

Alternative models.

There are many controversies regarding the criteria to be used in determining the primary and universal personality dimensions (Costa & McCrae, 1992; De Raad et al., 2010; De Raad & Milacic, 2015; Eysenck, 1991, 1992; Goldberg & Rosolack, 1994). Albeit the importance of the 16PF, the PEN system, and the Big Five, there are other competing systems (De Raad and Milacic, 2015). Next, we introduce some of these models and the similarities between them.

One-factor models.

Some authors have argued favorably to a general factor of personality. Nevertheless, there are not yet concordance regarding the content of such factor. For instance, Musek (2007) suggested a Big One in which all the Big Five factor loaded positively in one factor, while De Raad et al. (2010) suggested that the general factor is characterized predominantly by Agreeableness and Conscientiousness, and by Emotional Stability to a smaller extent. The controversies regarding a general factor are also related to the method employed to identify this factor. Revelle and Wilt (2013) argued that the most common methods can generate confusing results, by considering the general factor as the first factor of a correlation or covariance matrix, or as the first factor resulting from a bifactor rotation, or as a forced bifactor model in confirmatory factor analysis.

Two-factor models.

Digman (1997) proposed a two-factor solution, with the higher-order factors α, related to Agreeableness, Conscientiousness, and Emotional Stability, and β, linked to Extraversion and Intellect. De Raad and Milacic (2015) mention two other two-factor structures, one in modern Greek (Saucier, Georgiades, Tsaousis, & Goldberg, 2005), Morality/Social Propriety and Dynamism, and one in Chinese (Zhou, Saucier, Gao, & Liu, 2009), Social Propriety and Dynamism.

Three-factor models.

Besides Eysenck's PEN system, there is another three-factor model, the Big Three, which were derived from the Big Five (De Raad & Milacic, 2015). This model is formed by Extraversion, Agreeableness, and Conscientiousness. De Raad and Milacic (2015) also mention another three-factor model, the Indian Triguna in Hindi (Singh,

Misra, & De Raad, 2013). This model has the indigenous factors *Sattvic* (e.g., well-behaved, virtuous, harmonious, etc.), *Rajasic* (hypocrite, insensitive, quarrelsome, etc. vs. friendly, smart, sociable, etc.), and *Tamasic* (restless, arrogant, egoist, frustrated, etc.).

Six-factor models.

There are two similar alternative propositions with six dimensions, the HEXACO (Ashton & Lee, 2007; Ashton, Lee, & Vries, 2014), and the Big Six (Saucier, 2009; Thalmayer & Saucier, 2014). Additionally to the Big Five factors, the two models have a sixth dimension. In the HEXACO, this factor is named Honesty-Humility and consists of traits such as sincerity, fairness, greed avoidance, and modesty (Ashton et al., 2014). In the Big Six, the additional factor is Negative Valence (Saucier, 2009) or Honesty/Propriety (Thalmayer & Saucier, 2014), and is composed by trait markers such as *cruel, corrupt, disgusting, wicked,* etc. According to De Raad and Milacic (2015), this factor can be viewed as derived from the Big Five factor Agreeableness.

Seven-factor models.

Saucier (2003) synthesized an alternative model with seven dimensions, named by him as "Multi-Language seven" or ML7. According to De Raad and Milacic (2015), this model was identified after the inclusion of evaluative and mood state terms in personality taxonomies in Hebrew (Almagor, Tellegen, & Waller, 1995), Spanish (Benet-Martinez & Waller, 1997), and Filipino (Church, Katigbak, & Reyes, 1998). The model is formed by the following factors, with examples of descriptors (Saucier, 2003):

- Gregariousness: talkative, sociable, noisy vs. quiet, seclusive, serious, etc.
- Self-Assurance: fearful, cowardly, weak vs. confident, brave, secure, etc.
- Temperamentalness: short-tempered, irritable, impatient, etc.

– Concern for Others: compassionate, helpful, generous, soft-hearted, etc.

– Conscientiousness: neat, orderly, meticulous vs. sloppy, forgetful, etc.

– Originality/Virtuosity: talented, imaginative, knowledgeable, artistic, etc.

– Negative Valence or Social Unacceptability: insane, weird vs. normal, etc.

Correspondence between the models.

Most comparisons between the psycholexical models use as reference the Big Five model. Digman's two-factor model reassembled the Big Five factors in two higher-order dimensions (De Raad & Milacic, 2015). The three-factor models, Big Three and PEN, can also be interpreted in the Big Five framework (De Raad & Milacic, 2015; Goldberg & Rosolack, 1994). Neither of the models has the factor Intellect, and the Big Three also dropped the Neuroticism dimension. The PEN model blended Agreeableness and Conscientiousness into a broad factor, Psychoticism. The six-factor models included one dimension to the Big Five factors, while the ML7 seems to have included two new ones (i.e., Negative Valence and Self-Assurance). Finally, the five global factors of the 16PF are comparable with the Big Five (H. E. P. Cattell and Schuerger, 2003). In Figure 3 we synthesized the comparisons between the two-factor, the PEN, the Big Five, the six-factor, and the16PF models.

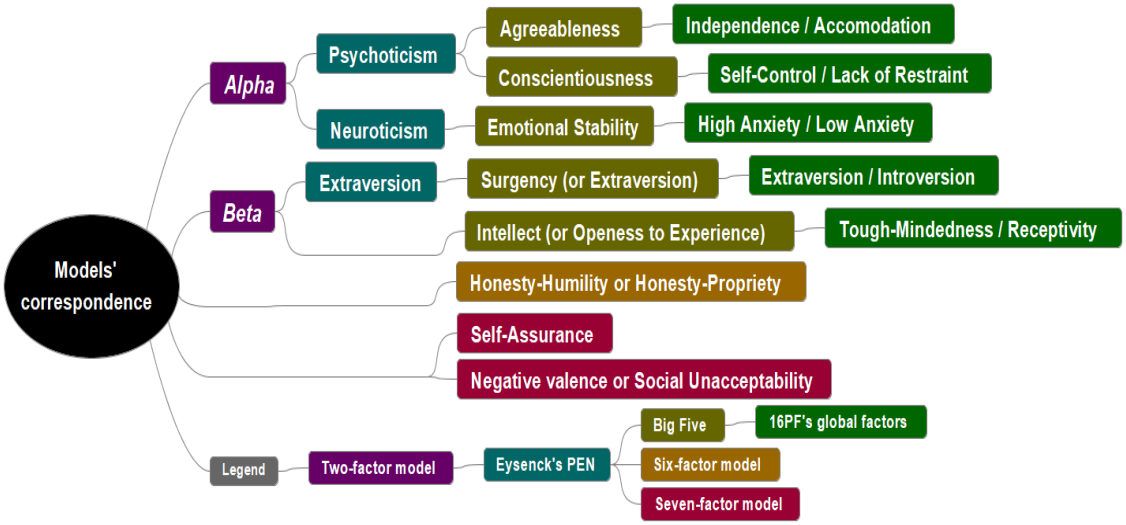

Figure 3. Presumed correspondence between psycholexical models of personality.

Method

In this section, we describe the procedures employed in this study and introduce the text mining techniques employed. The methodological steps of the present study are synthesized in Figure 4.

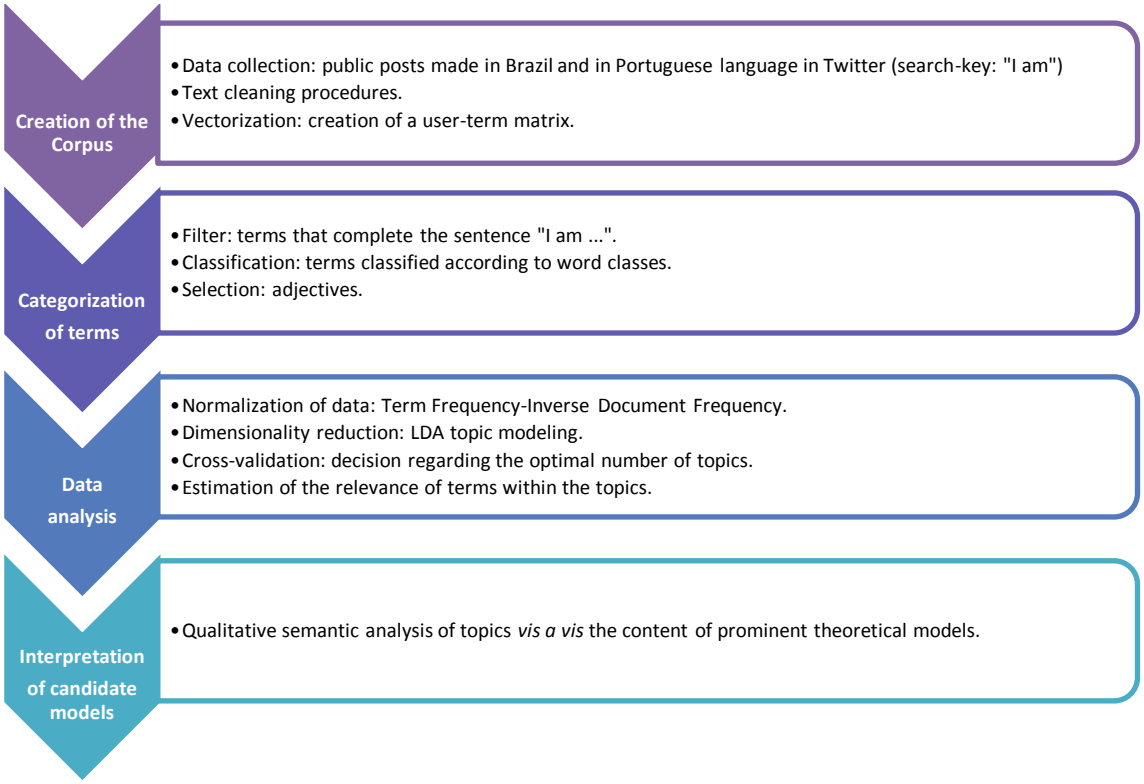

Figure 4. Methological steps of this study.

Creating the corpus: Data collection procedures

The data were collected on Twitter ("Twitter", 2017) during September 2016 in two stages. By the time of data collection, each post (i.e., tweet) were limited to 140 characters or less. First, we compiled tweets containing the search key *sou* (i.e., *I am*) aiming to identify potential Twitter users. In the second phase, with a list of the previously identified users, we collected all tweets written by each user containing the same search key. We created a single text document with the recovered tweets for each user. All the searches were set to find users in Brazil and tweets in Portuguese. Finally,

we converted the text documents into a single corpus (Feinerer, Hornik, & Meyer, 2008).

Text cleaning procedures

We listed all the terms presented in the corpus and then conducted a series of text cleaning procedures. First, we analyzed each term in the list, creating a correspondence table with the correct orthography and the classification according to a word class (i.e., adjectives, verbs, nouns, adverbs, pronouns, conjunctions, and interjections). This procedure was manual since we did not identify any available software that could handle Portuguese colloquial language such as we found on Twitter (e.g., words with many typos, orthographical errors, and other alternative orthographies that deviate from the formally written pattern).

In the second step, we removed punctuation, diacritics, white spaces, symbols (e.g., emoticons), names, numbers, URLs, and unintelligible terms. We also dropped hashtags and retweets ("Twitter", 2017). The third step included a series of text transformations. Terms with any orthographical error were corrected. We transformed verbs in all tenses into the corresponding present infinitive form. We substituted each adjective and noun in the feminine form by the masculine correspondent form. We also removed adjectives and nouns in the plural form, since we were only interested in self descriptions.

The third step was to create some specific n-grams ("*n*-gram", 2017; Silge & Robinson, 2017). We added a tag (i.e., "_not") to terms following the adverb *não* (i.e., *no* or *not*) or the conjunction *nem* (i.e., *nor* or *neither*). For example, if someone wrote *I am not friendly*, a new term was created with the tag *friendly_not*. We also created terms representing locutions formed by two or three words (e.g., *mal-educado* [*ill-mannered*] etc.).

Vectorization of the corpus

We converted the corpus into a term-document matrix and removed terms with the overall frequency lower than 100. Then we subset the matrix, selecting only adjectives as variables. This matrix was reduced, maintaining only terms with frequency greater than 100 and users that employed at least two terms in their tweets. The criteria adopted to define these thresholds are somewhat arbitrary. For a matter of comparison, in other studies with data from social networks, Kosinski et al. (2016) suggested thresholds of 50 Facebook Likes per user and a minimum of 150 users per Like, while Kosinski et al. (2013) used thresholds of a minimum of two Likes per user and a maximum of 20 users per Like. With the adopted thresholds, with aimed to balance the retaining of information and the reduction of the sparsity of the matrix.

Data Analyses

Normalization: Term frequency - inverse document frequency (TF-IDF).

Working in the context of text mining of big data (Leskovec et al., 2014), we had to deal with sparse term-document matrices in which most of the elements are zeroes and with an unbalanced distribution of terms over the documents. One consequence is that very frequent or very rare terms assume a disproportionate weight in the analysis, although the literature suggests that the best indicators of topics in a corpus often are relatively rare words (Leskovec et al., 2014). An advisable strategy to deal with this issue is to apply normalization techniques to the data, such as the TF-IDF, one of the most popular transformations in the text mining field. This way, we submitted the complete, the test and the training term-document matrices to TF-IDF normalization before conducting the dimensionality reduction analyzes.

The TF-IDF normalization is considered a measure of term importance in a corpus (Leskovec et al., 2014). In the TF-IDF scheme, the number of occurrences of each word

in each document in the corpus is counted. This count is then compared to an inverse document frequency count of occurrences of a word in the entire corpus (Blei et al., 2003). The TF-IDF is computed as follows (Leskovec et al., 2014). The term frequency (TF_{ij}) of term *i* in document *j* is f_{ij},, which is normalized by dividing it by the maximum occurrences of any term in the same document: $TF_{ij} = \frac{f_{ij}}{max_k f_{kj}}$. The term frequency TF will assume the value of 1 for the most frequent term in the document *j*, and fraction values between 0 and 1 will be assigned to the other terms in the same document. The IDF_i (i.e., inverse document frequency of a term) is given by $IDF_i = log_2(N/n_i)$, where n_i is the occurrence of term *i* in the *N* documents in the collection. Thus, the TF-IDF score for term *i* in document *j* is simply given by $TF\text{-}IDF = TF_{ij}\ x\ IDF_i$. The Table 7 in Appendix 1 present the results of TF-IDF normalization.

Topic modeling: Latent Dirichlet Allocation (LDA).

We used the LDA topic model to identify underlying dimensions in the *corpus*. The LDA was introduced by Blei et al. (2003) and became a popular model for uncovering latent topics in large text corpora and other kinds of discrete data (Griffiths, Steyvers, & Tenenbaum, 2007; Grün & Hornik, 2011; Kosinski et al., 2016; Liu, Tang, Dong, Yao, & Zhou, 2016; Poldrack et al., 2012). According to Blei and Lafferty (n.d.):

The idea behind LDA is to model documents as arising from multiple topics, where a topic is defined to be a distribution over a fixed vocabulary of terms. Specifically, it is assumed that *K* topics are associated with a collection of topics and that each document exhibits these topics with different proportions. This is often a natural assumption to make because documents in a corpus tend to be heterogeneous, combining a subset of main ideas or themes that permeate the collection as a whole (pp. 2-3).

As explained by Blei and Lafferty (n.d.), words, documents, and corpus are the observed data, while topics are the latent topical structure. A multi-document corpus is a collection of *M* documents, denoted by $D = \{w_1, w_2, \dots, w_M\}$. A document is a sequence of *N* words, denoted by $w = (w_1, w_2, \dots, w_N)$. A word or a term is the basic unit of discrete data, defined as an item from an indexed vocabulary $\{1, \dots, V\}$ (Blei et al., 2003). As an unsupervised generative probabilistic model of a corpus (Griffiths et al., 2007), LDA seeks to reproduce the imaginary random process that is assumed to have generated the observed data (Blei & Lafferty, n.d.). Therefore, a distribution over words is drawn for each topic; a vector of topic proportions is drawn for each document; and a topic assignment is drawn for each word (Blei & Lafferty, n.d.).

The statistical formulation of the LDA topic model can be described as follows (Blei et al., 2003; Griffiths et al., 2007; "Latent Dirichlet allocation", 2017). In the vector of words $w = (w_1, \dots, w_N)$ that represents a multi-document corpus, each word w_i belongs to some document d_i. This distribution is represented in the term-document matrix with the co-occurrence of the words. The gist of each document (i.e., the core of a speech or a text), g, is a multinomial distribution over topics, with parameters $\Theta^{(d)}$. Each topic, z_i, is a multinomial distribution over the w words in the vocabulary (i.e., the set of terms in the corpus), with parameters Φ^z. This way, for a word w_i in a document d_i, $P(z|g) = \Theta_z^{(d)}$, and for a word w_i in a topic z_i, $P(w|z) = \Theta_w^{(z)}$. Then, two symmetric conjugate Dirichlet priors (Kaplan, 2014) are taken. The symmetric Dirichlet(α) prior on $\Theta^{(d)}$ for all documents, and the symmetric Dirichlet(β) prior on $\Phi^{(z)}$ for all topics. This means that $\Theta^{(d)}$ and $\Phi^{(z)}$ are obtained from posterior distributions of the words over the topics (Griffiths & Steyvers, 2004).

There are many algorithms available to identify topics, such as expectation maximization, variational expectation maximization, expectation propagation and

several forms of Markov chain Monte Carlo or MCMC (Griffiths et al., 2007). The most common inference process is Gibbs sampling (Kaplan, 2014) used in this study and available in most packages that deal with topic modeling (Chang, 2015; Grün & Hornik, 2017; Nikita, 2016b; Selivanov & Wang, 2017) in R (R Development Core Team, 2017). A detailed description of the inference process using Gibbs sampling in LDA is available at Griffiths et al. (2007) and the Wikipedia article about LDA ("Latent Dirichlet allocation", 2017). The generative process for learning topics with LDA (Figure 5) can be summarized as follows (Blei et al., 2003; Griffiths et al., 2007; "Latent Dirichlet allocation", 2017):

1. Choose $\theta^{(d)} \sim \text{Dirichlet}(\alpha)$, where α is the parameter on the Dirichlet prior on per-document topic distributions.
2. Choose $\phi^{(z)} \sim \text{Dirichlet}(\beta)$, where β is the parameter of the Dirichlet prior on the per-topic word distributions.
3. For each word w_i in the word vector $w = (w_1, \ldots, w_N)$ of a document d_i, choose:
    a. A topic $z_i | \theta^{(d_i)} \sim \text{Multinomial}(\theta^{(d_i)})$.
    b. A word from $p(w_i | z_i, \beta)$, i.e., a word from a multinomial probability conditioned on the topic z_i.

In LDA modeling, it is necessary to specify the hyperparameters α and β, which are respectively the parameters of the priors distributions on θ (i.e., per-document topic distribution) and Φ (i.e., per-topic word distribution). These hyperparameters affect the granularity (i.e., the level of detail) of the results produced by the LDA model. The number of topics, K, also needs to be specified. Usually, the strategy is to fix α and β and test different candidate number of topics, K. Some authors recommend setting $\alpha = 50/K$ and $\beta = .10$ or $\beta = \frac{200}{number\ of\ variables}$ (Griffiths, & Steyvers, 2004;

Griffiths et al., 2007; Kosinski et al., 2016). Other authors suggest setting both α and β to .10 arguing that the resulting model will generate topics that produce a few words with high probability and each document will be composed by a few topics (Priva and Austerweil, 2015).

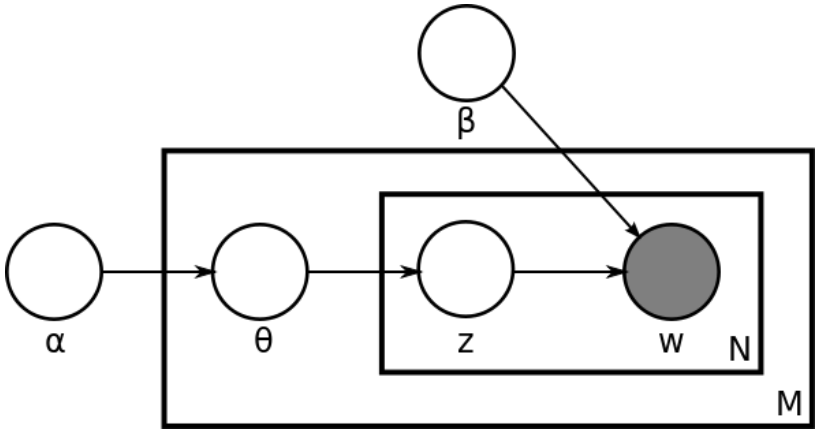

Figure 5. Graphical model representation of LDA ("Latent Dirichlet allocation", 2017). The hyperparameter α is the prior on the per-document topic distributions, and β is the prior on the per-topic word distribution. The outermost plate represents all the words belonging to document d_i, including the topic distribution for d_i, the θ_i. The M indicates that the variables are repeated M times, once per document. The inner plate represents the topics z_i associated with each w_i in d_i. The N indicates that the variables are repeated N times, once per word in d_i.

We set the LDA parameters to $\alpha = 50/K$ and $\beta = .10$, as recommended by Griffiths and Steyvers (2004) and Kosinski et al. (2016). Regarding parameter K, learning the number of topics in our corpus is the very objective of this study. To investigate this question, we employed two strategies. First, we analyzed the candidate number of topics through cross-validation techniques, in a data-driven approach. Second, in a theory-driven approach, we also examined prominent models with different numbers of candidate topics (e.g., Big Five). Subsequently, we conducted a qualitative semantic analysis. These procedures will be described next.

Selecting the number of topics: Cross-validation analyses.

Choosing the number of topics (i.e., K) is both a model selection issue and a question of interpretability. There is no single or straightforward approach to determine the "correct" number of topics in LDA and to assess the relevance and quality of the models and their underlying topics (Chuang, Manning, & Heer, 2012; Kosinski et al., 2016). A common strategy to deal with the model selection issue is to determine the optimal number of topics in a data-driven approach by comparing the fit of several models with different candidate number of topics (Grün & Hornik, 2011; Nikita, 2016a). Following this perspective, we compiled information regarding the dimensionality of the corpus through five cross-validation procedures. Four of them are implemented in the Ldatuning package (Nikita, 2016b): Griffiths and Steyvers (2004); Cao, Xia, Li, Zhang, and Tang (2009); Arun, Suresh, Madhavan, and Murthy (2010); and Deveaud, Sanjuan, and Bellot (2014).

The most known approach is the one proposed by Griffiths and Steyvers (2004). Their technique consists of computing the log-likelihood $P(w|T)$ (i.e., the posterior probability) of a set of models (i.e., with different numbers of topics) given the observed data. Then, the log $P(w|T)$ is plotted against the number of topics. In the resulting scree plot, the log-likelihood will initially increase as a function of T, flattens at optimal models, and may decrease after that, indicating a large number of topics (Griffiths & Steyvers, 2004; Kosinski et al., 2016).

Cao et al. (2009) developed a density-based method for adaptive LDA model selection. They started from the observation that when K is too small (i.e., with only a few topics), the discrimination between the topics is low, once there are words that overlap across topics. As a consequence, valuable information can be lost. On the other hand, when K is too large (i.e., with too many topics), the topics can be correlated. But,

as LDA cannot capture this correlation in its generating process, the model cannot represent the original data with accuracy. Thus, Cao et al. concluded that the "best" topic structure is correlated with distances among the topics.

The main feature of the approach of Cao et al. (2009) is that it integrates a clustering process based on density, considering a topic as equivalent to a semantic cluster. The best model will have the largest possible intra-cluster similarity, which means that the cluster (i.e., the topic) includes coherent semantic content. At the same time, the best model will also have the smallest possible between-cluster similarity (i.e., the smallest possible similarity between topics), which indicates a more stable structure.

Arun et al. (2010) proposed an approach to identify the "right" number of topics in a corpus based on symmetric Kullback-Leibler divergence measure of salient distributions. According to the authors, LDA can be interpreted as a non-negative matrix factorization method that split the term-document matrix into a topic-word matrix and a document-topic matrix. While varying the number of topics, their algorithm measures the information between two probability distributions. The first is the singular value distribution of topic-word matrix. The second is a vector of the distribution of each topic present in the corpus (i.e., the document-topic matrix).

Similarly to Arun et al. (2010), Deveaud et al. (2014) also proposed a metric for identifying the "right" number of topics based on divergence measure. Their algorithm uses the Jensen-Shannon measure, which is a symmetrized version of Kullback-Leibler. Devaud et al. named their approach as Latent Concept Modeling, using the term *latent concepts* as a synonym for LDA *topics*. It consists of computing several LDA models, seeking to maximize the information divergence (i.e., similarities or dissimilarities) between all pairs of LDA topics in each model.

The last information we compiled regarding the number of topics is the perplexity, a common strategy to evaluate an LDA fitted model. The perplexity is a metric resulting from the comparison of probability models that assess how well a probability distribution predicts a sample. In order to obtain the perplexity of a model, the dataset is split into two parts: a training part and a test part (Grün & Hornik, 2011). The perplexity evaluates the fitted model (i.e., the model fitted on training dataset) on a held-out data (i.e., the test dataset). To examine the perplexity, we conducted a 5-fold cross-validation (Flach, 2012) after randomly split the adjectives matrix in two, a training dataset with 90% of the cases and a test dataset with 10% of the cases.

In agreement with Blei et al. (2003), the perplexity is used by convention in language modeling. It is equivalent to the geometric mean per-word likelihood. The lower the perplexity, the better is the sample prediction or the generalization performance. Albeit its importance, the use of perplexity in the context of this study can be seen as a "figure of merit", as stated by Blei et al. As we are working with unigrams ("*n*-gram", 2017), we are not modeling language, which would require examining higher-order models (Blei et al., 2003).

Interpretation of the topics: The relevance of the terms.

Despite the importance of the statistical assessment of the possible number of topics, the interpretability of the uncovered dimensions is crucial in selecting a meaningful latent structure in a corpus. Poldrack et al. (2012) suggest that despite the indication of a "best" dimensionality by cross-validation techniques as the described before (Nikita, 2016a), there is significant information at several levels of topics differing in granularity. Chuang et al. (2012) alert to the presence of incoherent or insignificant topics and recommend that domain experts should verify the model outputs and eventually modify the number of topics, K, to enhance interpretability given the

domain of analysis. Experts' modifications in the statistically suggested number of topics were made in studies such as Priva and Austerweil (2015), Poldrack et al. (2012), and Hall, Jurafsky, and Manning (2008).

A common practice to interpret the content of a topic is to rank 3 to 30 of its terms by their probability to belong to the topic (Sievert & Shirley, 2014). According to Sievert & Shirley (2014), evidence suggests that this is not an optimal approach to interpret results of an LDA model. The problem with this strategy is that the terms which are most common (i.e., frequent) will often appear at the top of the ranks of different topics. Some solutions have been proposed in recent years to handle this issue (Sievert & Shirley, 2014). Bischof & Airoldi (2012) suggested the Hierarchical Poisson Convolution that examines a given topic through a measure of terms frequency and exclusivity in the topic. Chuang et al. (2012) proposed a measure of term saliency and a visualization tool, named Termite. This metric indicates how informative a given term is in determining the generation of a new topic in comparison with a randomly selected term. Sievert & Shirley (2014) combined the approaches of Bischof & Airoldi (2012) and Chuang et al. (2012) and proposed LDAvis, a metric of relevance and a visualization tool as a method for interpreting topics.

In LDAvis' relevance metric (Sievert & Shirley, 2014), the relevance of given term *w* to latent topic *k* given a weight parameter λ (where $0 \leq \lambda \leq 1$) is defined as $r(w, k \mid \lambda) = \lambda \log(\Phi_{kw}) + (1 - \lambda)\log(\frac{\Phi_{kw}}{p_w})$. In the equation, Φ_{kw} denotes the probability of term *w* under the topic *k*, p_w denotes the marginal probability of *w* in the corpus, and λ determines the weight given to Φ_{kw}. A λ=1 will decreasingly order terms according to their topic-specific probability, which tends to rank corpus' most frequent terms higher in the topic. In contrast, a λ=0 will rank terms only by their lift, which tends to rank rare words higher. Sievert and Shirley (2014) showed evidence that a λ

around .60 could be an optimal value. In this study, we adopted the relevance metric to report and interpret LDA models and set λ=.60, as suggested by Sievert and Shirley. Figures 6 and 7 illustrate the differences of ranking the terms of a given topic by ordering them according to their topic-specific probability (Figure 6; λ=1) or by their relevance (Figure 7; λ=.60).

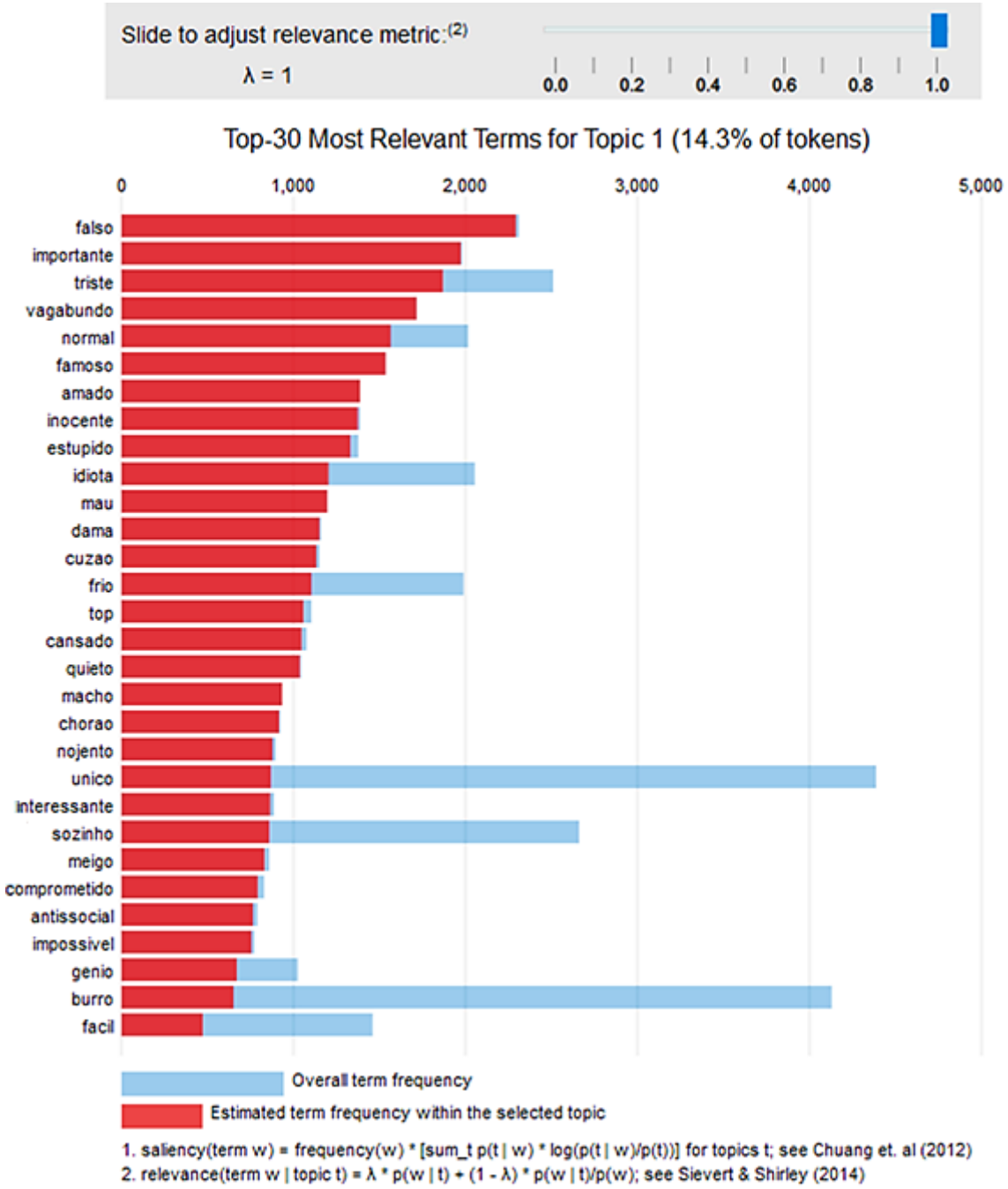

Figure 6. Order of terms within a given topic according to their topic-specific probability (relevance metric λ=1), which tends to rank corpus' most frequent terms higher in the topic. This visualization was produced by the LDAvis package (Sievert & Shirley, 2014).

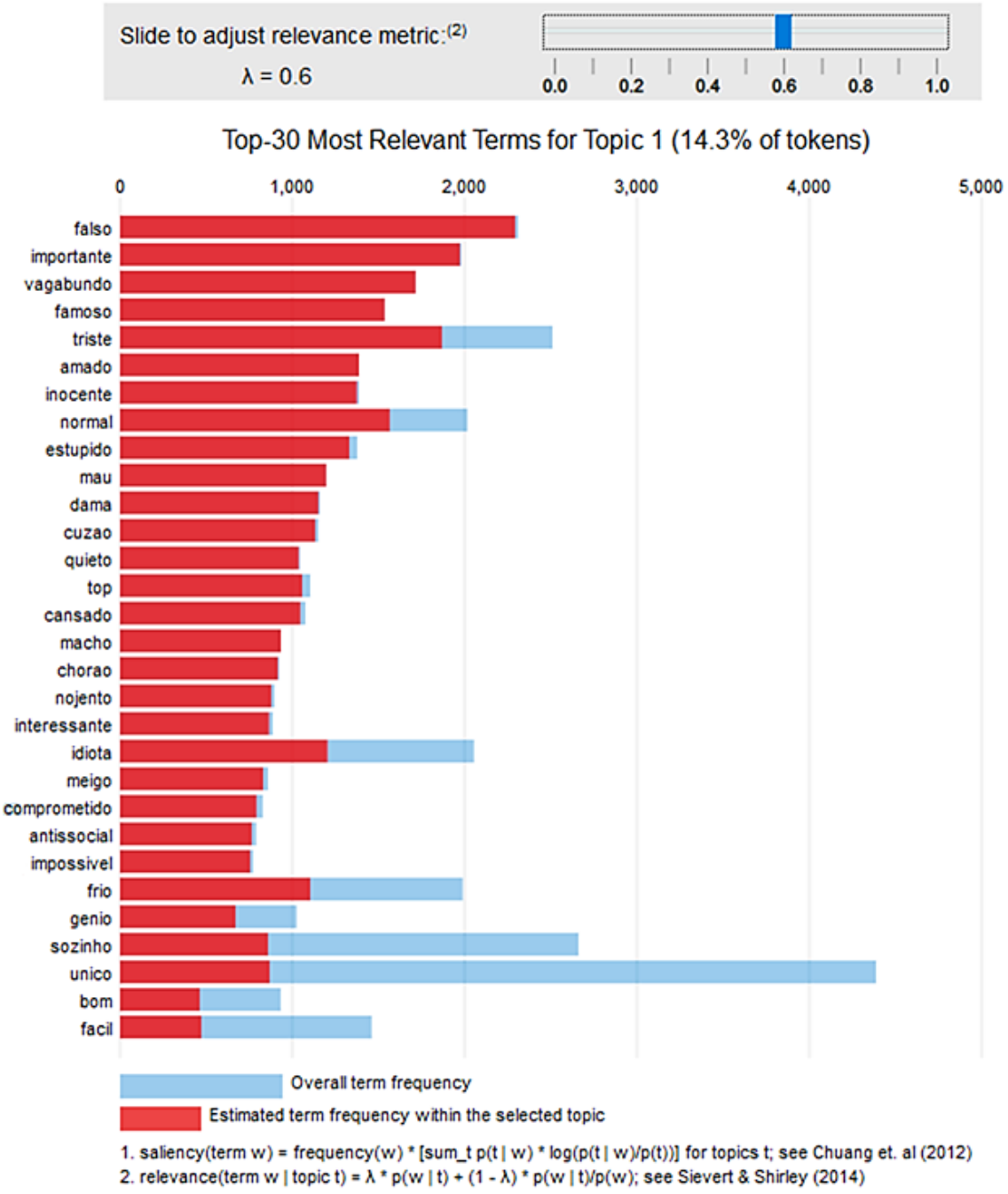

Figure 7. Order of terms within a given topic (the same of Figure 5) according to the to their relevance, with λ=.60 as suggested by Sievert and Shirley (2014). This visualization was produced by the LDAvis package (Sievert & Shirley, 2014).

Interpretation of the topics: Semantic content and coherence.

The analysis of the LDA models also requires an investigation of their interpretability and theoretical pertinence. For instance, the "best" models identified in our data by the cross-validation techniques (Nikita, 2016a) can have a latent structure different from the models found in the literature. The research with the lexical hypothesis mainly points to models with three to six or even 16 dimensions in different cultures (Eysenck, 1991; De Raad et al., 2010; Peres & Laros, 2018a; Revelle, 1995). Evidence suggests the prominence of the five-factor model (De Raad & Mlacic, 2015), the three-factors model (Revelle, 2016), the six-factor model (Ashton, Lee, & Vries, 2014), and the Cattell's sixteen-factor model or 16PF (R. B. Cattell, 1986). Once Cattell's model is composed of 15 personality factors plus a Reasoning scale, we choose the model with 15 topics as the candidate model to be compared with the 16PF. This way, in addition to the models suggested by the cross-validation analyses, we also explored models with three, five, six, and 15 topics.

Besides the latent structure, it is also necessary to examine the semantic content and coherence of each topic. We analyzed the content of each topic *vis a vis* the content of the prominent psycholexical models. We compared the correspondence of each adjective in our candidate models with its synonyms, antonyms and other related words presented in different taxonomies. Therefore, we prepared a list of the terms retrieved in our study, with their translations to English, synonyms, antonyms, and other related words organized by personality factor (see Appendix 2). It is expected that several terms may be adherent to more than one dimension (i.e., one theoretical factor or one empirical topic), due to polysemy. To analyze the semantic coherence of the content of a given topic, we considered the senses shared by most words under the topic.

For the semantic analyses, we adopted as main reference the five-factor model and Goldberg's 100 revised synonym clusters (Goldberg & Rosolack, 1994). This is justified, once theoretical and empirical comparisons between the five-factor model and the other prominent models have been investigated. Thus, we considered the correspondence between the five-factor model and Eysenck's three-factor model (Goldberg & Rosolack, 1994) and the sixteen-factor model (H. E. P. Cattell & Schuerger, 2003). Regarding the six-factor model, we additionally considered the honesty-humility factor (Ashton et al., 2014).

For elaborating the vocabulary in Appendix 2, we adopted further references, such De Raad's (2000) book, which compiled a list of 20 adjectives for each of eight five-factor taxonomies (English, Dutch, German, Polish, Czech, Hungarian, Italian Rome, and Italian Trieste). We also adopted Brazilian studies with the five-factor model as references (Andrade, 2008; Passos, 2014; Passos & Laros, 2015; Hauck Filho, Machado, Teixeira, & Bandeira, 2012; Hutz et al. 1998; Machado, Hauck Filho, Teixeira, & Bandeira, 2014; Natividade & Hutz, 2015. Additionally, we used two Portuguese dictionaries (Dicionário Houaiss da Língua Portuguesa, 2009; Dicionário Priberam da Língua Portuguesa), two English thesauri (, n.d.; Merriam-Webster.com, n.d.; Thesaurus.com, n.d.), and a translation tool (Translate.google.com, n.d.).

Reliability estimate: *Omega* total.

We calculated the reliability coefficient *Omega* total (ω_t) for each topic. According to Revelle and Zinbarg (2009), ω_t corresponds to the internal consistency or the total reliability of the test, once it refers to the "proportion of test variance due to all common factors" (p. 152). Another interpretation is that ω_t is the proportion that "indexes generalizability to the domain from which the test items are a representative sample and which may represent more than one latent variable" (p. 152). Revelle (n.d.)

argues that the Omega coefficients can be applied both to an overall test and to an individual factor.

Omega total is found through factor analysis of the correlation matrix using the Schmid-Leiman transformation (Revelle, n.d.). It is relevant to highlight that the estimation of ω_t is done in a confirmatory fashion that is different from the LDA topic model, which is an unsupervised machine learning method (Flach, 2012). Nevertheless, we believe that reporting a reliability coefficient provide complementary information to the qualitative interpretation of each topic. To estimate ω_t for each topic, we reduced the term-document matrix for the 10 most relevant terms under the topic and only the cases (i.e., users) that employed at least two of these words. This procedure was necessary to reduce the sparsity of the matrix.

Software.

We used the software R (R Development Core team, 2017) and the following packages in the analysis: *twitteR* (Gentry, 2015) to collect data from Twitter; *tm* (Feinerer & Hornik, 2017) to create corpora and to text cleaning procedures; *text2vec* (Selianov & Wang, 2017) to TF-IDF normalization and LDA analysis; *ldatuning* (Nikita, 2016b) to find the number of topics; *topicmodels* (Grün & Hornik, 2017) to estimate perplexity of candidate models; *LDAvis* (Sievert & Shirley, 2015) to create visualizations of LDA fitted models, and to calculate the relevance of terms; *psych* (Revelle, 2017) to estimate Omega reliability coefficient and to describe data.

Results

Corpus and term-document matrix

We collected tweets from 190,008 users, resulting in a total of 140,628 unique terms. After text cleaning procedures, there were 548 adjectives with an overall

frequency superior to 100. In the sequence, we subset the matrix again, maintaining only terms with overall frequency greater than 100 and users that employed at least two adjectives in their tweets. These procedures resulted in a term-document matrix with 86,899 users and 172 terms. In Appendix 1 we report the descriptive statistics of these adjectives before and after TF-IDF normalization. In Appendix 2 we present the resulting vocabulary, with a translation to English for each term, and a list of synonyms, antonyms and other related words found in studies within the lexical approach.

Number of topics: Cross-validation analyses

Figure 8 illustrates the results for the four metrics available in the *ldatuning* package (Nikita, 2010b). The Deveaud et al. (2014) metric suggested a model with 14 topics, while the Cao et al. (2009) metric indicated seven topics. The results of Griffiths and Steyvers' (2004) metric are not conclusive since there is not a flat or a decrease tendency in the metric values. Nevertheless, the rapid growth of the curve is interrupted at several points, first one occurring between 12 and 13 topics. This result is possibly an indication that an "optimal" model has a number of dimensions close to this. Arun's metric (2010) does not seem to be informative to our data since the smaller values indicates a "best" model with around 78 topics. The same happened with the 5-fold cross-validation considering the perplexity measure (Figure 9), which indicates that the perplexity is inversely related to the number of topics in data (i.e., a greater number of topics is more informative).

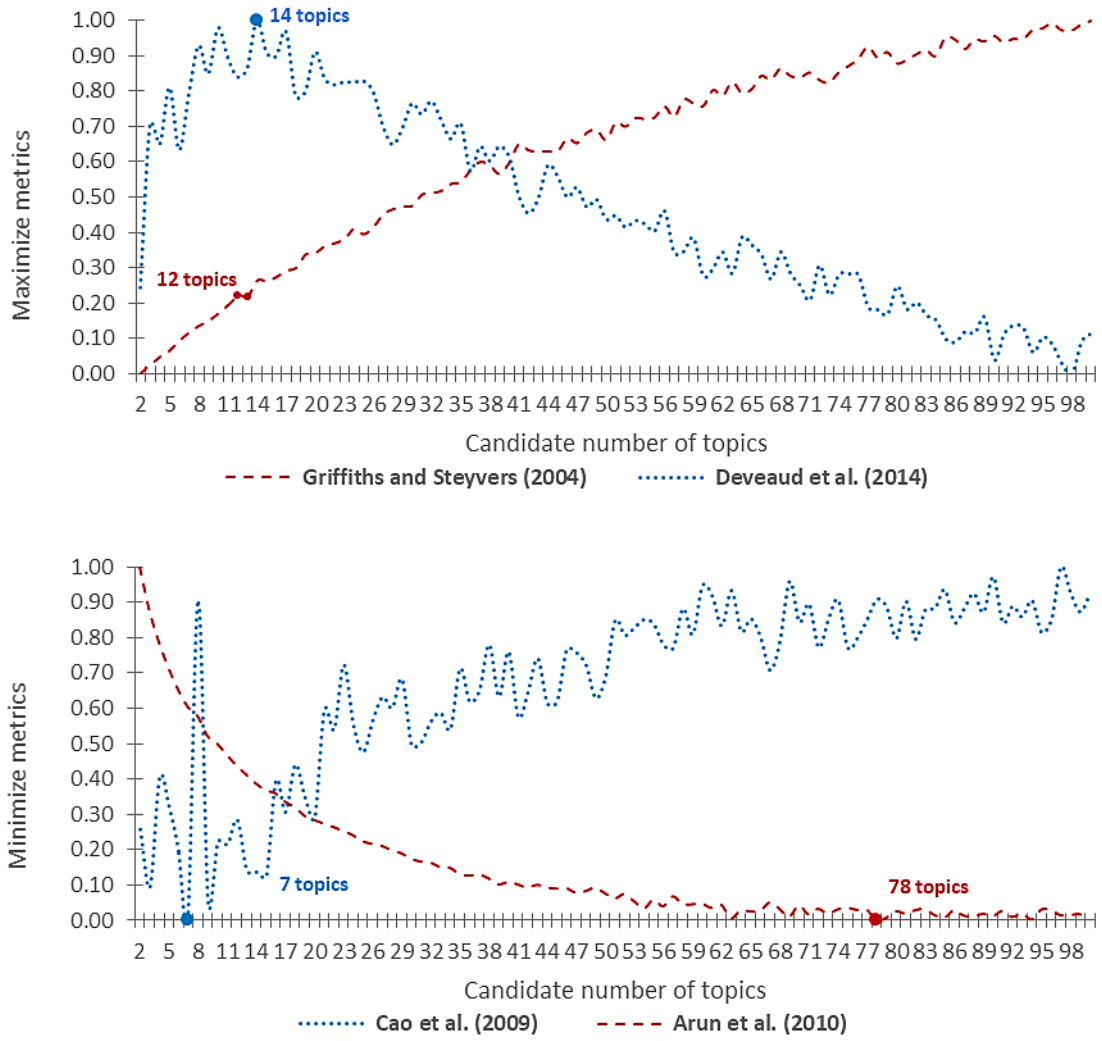

Figure 8. Number of topics indicated by four metrics. The metrics were standardized to range between zero and one.

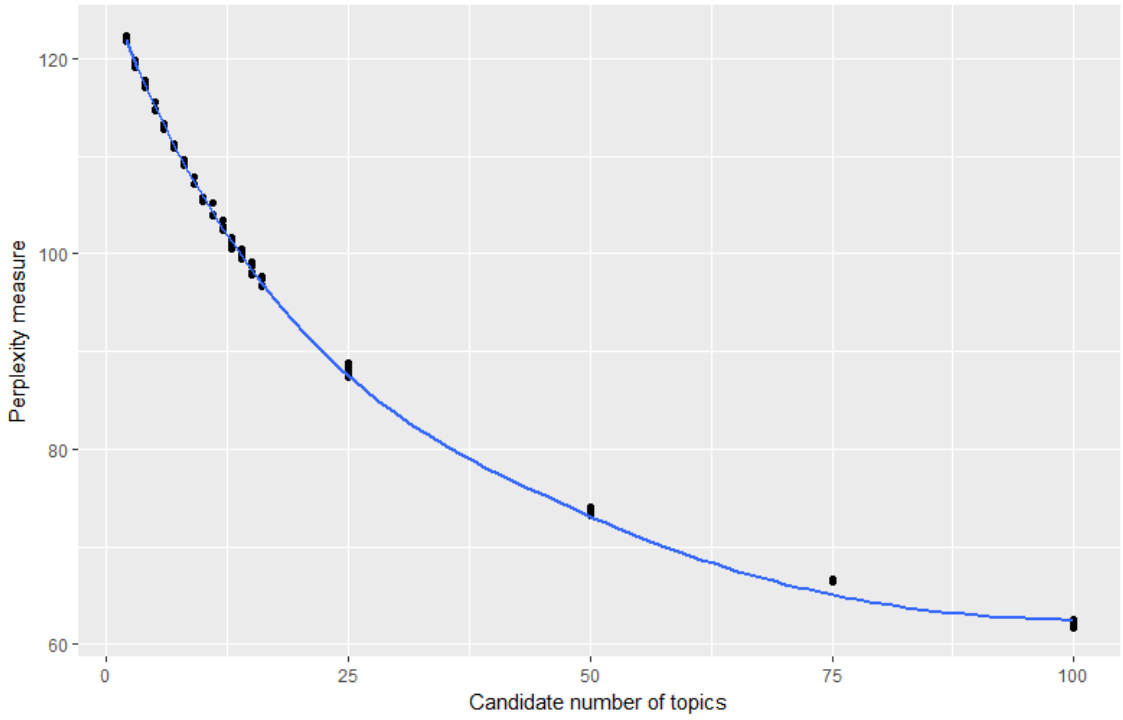

Figure 9. Five-fold cross-validation of models with different numbers of topics considering perplexity measure.

Relations between topics

We visually analyzed the relations between the topics, considering the models with a number of dimensions suggested by literature (i.e., three, five, six and 16 personality factors) and the models suggested by cross-validation analyses (i.e., seven, 12 and 14). Figure 10 presents panels of the topics models produced by LDAvis package (Sievert & Shirley, 2015). A global view of each latent topic model is displayed, illustrating both the prevalence of the topics (i.e., the circle area) and the relations between them (i.e., the intertopic distances).

The results indicate that in all models the topics assumed similar prevalence. Regarding the relations between topics, there are overlapping topics in all models, with exception of the Three-Topic Model. The overlaps are a result of the fact that various topics share some terms. Repetitions of a term in more than one topic are due to polysemy, which means that the word can assume a different sense depending on the

context of the topic (Griffiths et al., 2007). Nevertheless, if one considers only the five most relevant words in each topic, there are no duplicated words across topics. Considering the 10 most relevant words in each topic, the majority of models have no duplicated words. The Five-Topic Model has one duplicated word (i.e., *foolish*); the Fourteen-Topic Model also has one (i.e., *good*); and the Fifteen-Topic Model has four (i.e., *insane*, *normal*, *lousy*, and *foolish*). This information indicates that the overlaps are due to the less relevant words in each topic.

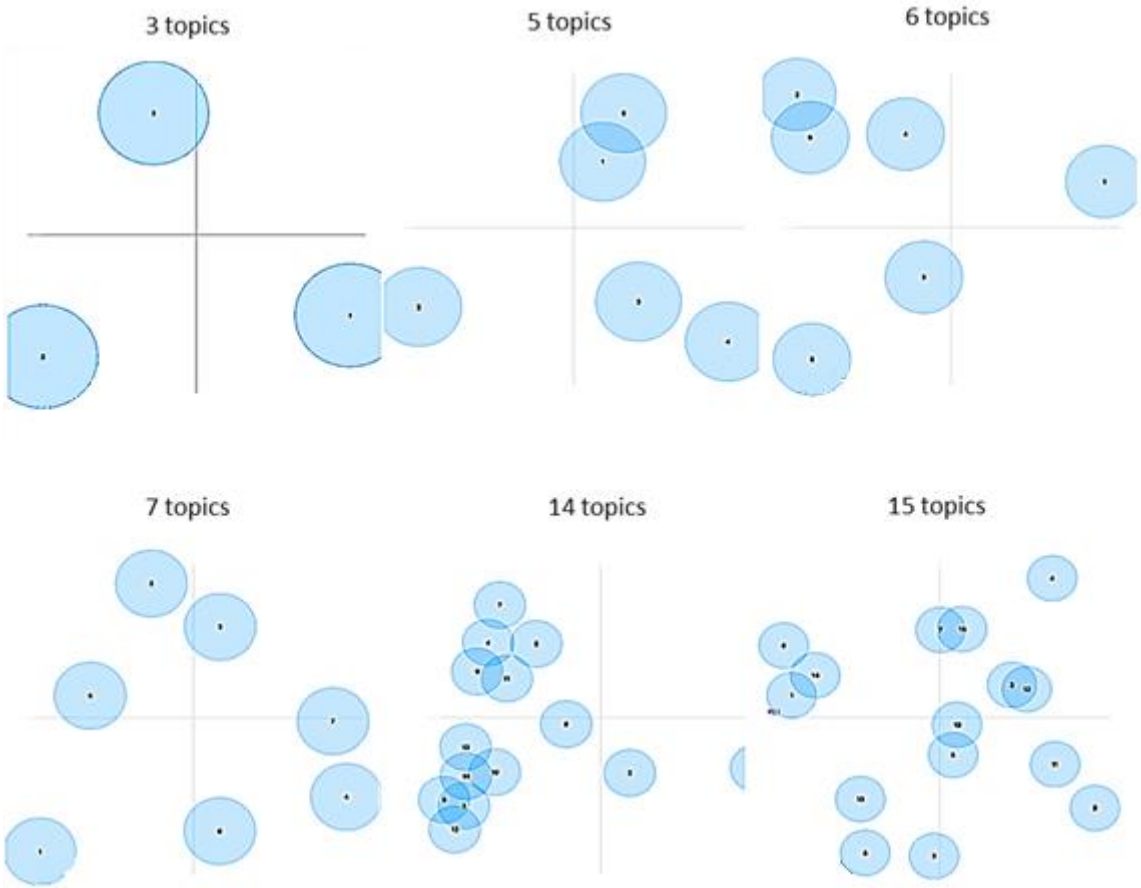

Figure 10. Intertopic distance maps for models with different numbers of topics via multidimensional scaling, considering all terms.

The content of topics: Semantic analysis

We qualitatively analyzed the semantic coherence of the content of each topic considering synonyms, antonyms, and other related words from prominent personality taxonomies (see Appendix 2), using as main reference the Goldberg's 100 revised synonym clusters (Goldberg & Rosolack, 1994). In each analysis, we considered preferably the senses shared by the terms inside each topic, seeking for an internal semantic coherence. The interpretations reported in this section are not final, but qualitative approximations. Nevertheless, from an emic-etic perspective (Cheung et al., 2011; Daouk-Öyry et al., 2016), we believe that they are able to contextualize the models in the predominant psycholexical frameworks.

Three-Topic Model.

The Three-Topic Model (Table 1) was not identified as an "optimal" model by the cross-validation analyses. Nevertheless, we investigated the presumed semantic relations between this model and the PEN model, and the results suggest that they are not similar. Topic 1 (ω_t=.83) is mostly related to Psychoticism, as a mixture of Agreeableness (four terms) and Conscientiousness (two terms). Topic 2 (ω_t=.71) is a mixture of Agreeableness and Emotional Stability and did not resemble any PEN or Big Three factors. Finally, Topic 3 (ω_t=.88) is predominantly adherent to Extraversion (six terms).

Table 1
The Three-Topic Model with the 10 most relevant terms of the topics, reliability and presumed correspondence with other psycholexical models

Topic 1 ($\boldsymbol{\omega_t}$=.83; *n*=2,022) 1.false, 7.unbearable, 9.incredible, and 6.ignorant (AGR); 4.lazy and 10.vagabond (CON); 3.silly, 5.chump, 8.foolish (EMO or INT); and 2.guilty (EMO). Presumed factor: Psychoticism.
Topic 2 ($\boldsymbol{\omega_t}$=.71; *n*=1,467) 1.rough, 9.difficult, and 2.sweet (AGR); 6.good and 8.great (AGR or INT); 3.anxious, 4.sad, and 7.happy (EMO); and 5.fool and 10.intelligent (EMO or INT). Presumed factor: mixture of Agreeableness and Emotional Stability.
Topic 3 ($\boldsymbol{\omega_t}$=.88; *n*=2,052) 1.indecisive, 10.strong, 7.weak, 5.brat, 9.damned, and 6.free (EXT); 2.annoying, 8.polite, and 3.important (AGR); and 4.cardiac (EMO). Presumed factor: Extraversion.

Legend: ω_t= McDonald's omega total reliability coefficient; *n* = subsample used to estimate the reliability; EXT (Extraversion); AGR (Agreeableness); CON (Conscientiousness); EMO (Emotional Stability); INT (Intellect). Note: the number before each term indicates its relevance within the topic.

Five-Topic Model.

The cross-validation analyses did not identify the Five-Topic Model (Table 2) as an "optimal" latent structure. Nevertheless, we compared it to the Big Five. Topic 1 (ω_t=.73) and Topic 2 (ω_t=.42) do not seem to correspondent with any Big Five factor, but they resemble the Positive/Negative Valence (NVP) dimensions of the seven-factor model (Benet-Martinez & Waller, 1997). Topic 3 (ω_t=.49) is predominantly adherent to the Psychoticism, while Topic 4 (ω_t=.90) mainly relates to Agreeableness and Topic 5 (ω_t=.90) to Emotional Stability. Finally, Topic 2 and 3 have low reliability coefficients.

Table 2
The Five-Topic Model with the 10 most relevant terms of the topics, reliability and presumed correspondence with other psycholexical models

Topic 1 (ω_t=.73; n=1,517) 1.chump, 5.useless, 7.idiot, and 9.son of a bitch (Negative Valence); 2.important, 4.perfect, 6.famous, and 8.nice (Positive Valence); 5.responsible (CON); and 3.cardiac (EMO). Presumed factor: Positive Valence.
Topic 2 (ω_t=.42; n=1,753) 1.lazy, 2.dramatic, 4.complicated, 5.crazy, 6.vile, 8.blind, 9.foolish, and 10.shameless (Negative Valence); 3.free (EXT or INT); and 7.special (Positive Valence). Presumed factor: Negative Valence.
Topic 3 (ω_t=.49; n=1,902) 4.polite, 7.ferocious, 8.sympathetic, 9.sensitive, and 10.gracious (AGR); 1.indecisive, and 5.vagabond (CON); 2.brat and 6.damned (CON or EXT); and 3.guilty (EMO). Presumed factor: Psychoticism.
Topic 4 (ω_t=.90; n=2,019) 1.false, 2.sweet 3.unbearable, 5.sentimental, 7.exaggerated, and 9.genial (AGR); 4.unique and 6.ridiculous (AGR and NPV); and 8.intelligent and 10.foolish (EMO/INT/NVP). Presumed factor: Agreeableness.
Topic 5 (ω_t=.91; n=1,731) 1.jealous, 2. anxious, 3.timid, 4.needy, 5.ignorant, 6.weak, 7.slow, and 8.silly (EMO); curious (INT); and incredible (NPV). Presumed factor: Emotional Stability.

Legend: ω_t= McDonald's omega total reliability coefficient; *n* = subsample used to estimate reliability; EXT (Extraversion); AGR (Agreeableness); CON (Conscientiousness); EMO (Emotional Stability); INT (Intellect); NVP (Negative/Positive Valence). Note: the number before each term indicates its relevance within the topic.

Six-Topic Model.

As the Three-Topic and the Five-Topic models, the cross-validation analyses did not identify the Six-Topic Model (Table 3) as an "optimal" model. Nevertheless, we compared this model with other six-factor models. However, the semantic analysis suggested that five topics (ω_t ranging from .57 to .63) in the Six-Topic Model are mostly adherent to Agreeableness, with exception of Topic 5 (ω_t=.45), which is related to Extraversion. This way, this model is not similar to other psycholexical models with six dimensions since it does not have a topic similar to the Honesty-Humility factor (Ashton et al., 2014) or Honesty/Propriety factor (Thalmayer & Saucier, 2014), for example.

Table 3
The Six-Topic Model with the 10 most relevant terms of the topics, reliability and presumed correspondence with other psycholexical models

Topic 1 ($\boldsymbol{\omega_t}$=.92; *n*=1,691) 1.nobody and 2.important (AGR or EXT); 3.brat (EXT or CON); 5.polite, 6.good-natured, 7.shameless (AGR); 4.insane, 8.stupid and 9.tranquil (AGR or EMO); and 10.strong (EXT or EMO). Presumed factor: Agreeableness.
Topic 2 ($\boldsymbol{\omega_t}$=.61; *n*=1,772) 1.weak and 3.anxious (EMO); 2.crazy (AGR/CON/EMO); 4.ferocious, 6.sensitive, 8.sympathetic, and 10.beloved (AGR); 5.exaggerated (AGR or CON); 7.curious (INT); 9.lost (CON). Presumed factor: Agreeableness.
Topic 3 ($\boldsymbol{\omega_t}$=.57; *n*=1,645) 1.sweet, 6.faithful, 7.nice, 9.easy, and 10.son of a bitch (AGR); 8.innocent (AGR or EMO); 4.chump and 5.intelligent (EMO or INT); 2.cardiac (EMO or AGR); and 3.complicated (INT). Presumed factor: Agreeableness.
Topic 4 ($\boldsymbol{\omega_t}$=.73; *n*=1,753) 1.rough, 3.false, 5.incredible, 6.famous, 7.grateful, 10.deluded (AGR); 4.guilt and 8.blind (EMO); 2.indecisive (CON); and 9.dung (EXT or NPV). Presumed factor: Agreeableness.
Topic 5 ($\boldsymbol{\omega_t}$=.45; *n*=1,846) 1.needy (EMO); 2.lazy (CON); 3.free, 5.damned, 6.dramatic, 8.useless, 9.ridiculous, and 10.marvelous (EXT); 7.special and 4.sentimental (AGR). Presumed factor: Extraversion.
Topic 6 ($\boldsymbol{\omega_t}$=.93; *n*=1,479) 1.sad (EMO or EXT); 2.ignorant, 3.unbearable, 4.vagabond, 5.vile, 6.amorous, 8.cold, 9.lady, 10.selfish (AGR); and 7.worse (EXT or NPV). Presumed factor: Agreeableness.

Legend: ω_t= McDonald's omega total reliability coefficient; *n* = subsample used to estimate reliability; EXT (Extraversion); AGR (Agreeableness); CON (Conscientiousness); EMO (Emotional Stability); INT (Intellect); NVP (Negative/Positive Valence). Note: the number before each term indicates its relevance within the topic.

Seven-Topic Model.

The Seven-Topic Model (Table 4) was identified as an "optimal" model by the Cao et al. (2009) method. To analyze the semantic content of this topic, we compared it with other seven-factor models (Almagor et al., 1995; Benet-Martinez & Waller, 1997; Church et al., 1998; Saucier, 2003) and with the Big Five (Goldberg & Rosolack, 1994).

At least three topics in this model seem to be predominantly related to Agreeableness (AGR). Topic 1 (ω_t=.94) has seven terms related to this factor, Topic 5 (ω_t=.89) has six, and Topic 6 has seven (ω_t=.66). If we consider only the terms related

to AGR in these topics, the reliability coefficient increases for Topic 5 (ω_t=.88) and for Topic 6 (ω_t=.75), and diminishes for Topic 1, but remains still high (ω_t=.89).

Topic 2 (ω_t=.30), although with a low reliability, seems also to be related to AGR, with nine terms reflecting this factor. Likewise, it is also possible to interpret the content of the Topic 2 as positive (e.g., *important, famous, beloved, innocent,* and *normal* [ω_t=.90]) vs. negative (e.g., *false, vagabond, sad, stupid* and *bad* [ω_t=.86]) valence.

Topic 3 (ω_t=.54) has seven terms compatible with the Emotional Stability (ES) factor. If we consider only these terms, the reliability of the topic increases to ω_t=.76. Topic 4 (ω_t=.39) seems to reflect the content of Eysenck's Psychoticism factor, with ω_t=.81 if we drop the term *lazy*. Finally, it was not possible to identify a clear interpretation for Topic 7 (ω_t=.59).

Table 4
The Seven-Topic Model with the 10 most relevant terms of the topics, reliability and presumed correspondence with other psycholexical models

Topic 1 ($\boldsymbol{\omega_t}$=.94; *n*=1,638) 1.happy, 2.cardiac, and 3.weak (EMO); 4.ignorant, 5.unbearable, 9.calm, (AGR); 6.special, 7.unique (AGR or NPV); 8.pure (AGR or EMO); 10.dung (AGR or EXT or NPV). *Presumed factor:* Agreeableness (seven terms, except *happy, cardiac* and *weak*; ω_t=.89; *n*=886).
Topic 2 ($\boldsymbol{\omega_t}$=.30; *n*=1,452) 1.false, 2.important, 3.vagabond, 4.famous, 6.beloved, 10.bad (AGR); 5.sad, 7.innocent, and 9.stupid (EMO or AGR); 8.normal (NPV). *Presumed factor:* Agreeableness.
Topic 3 ($\boldsymbol{\omega_t}$=.54; *n*=1,639) 1.anxious, 2.needy; 5.tranquil, 6.chump, 8.deluded (EMO); 3.ridiculous and 7.lucky (EXT); 10.simple (AGR); 4.lost, 9.responsible (CON). *Presumed factor:* Emotional Stability (seven terms, excluding *lucky, responsible* and *simples*; ω_t=.76; *n*=1,209)
Topic 4 ($\boldsymbol{\omega_t}$=.39; *n*=1,469) 1.lazy (CON); 2.crazy and 7.sick (EMO); 3.damned (EXT or AGR); and 4.ferocious, 5.dramatic, 6.vile, 8.nice, 9.trashy, and 10.marvelous (AGR). *Presumed factor:* Psychoticism (nine terms, excluding *lazy*; ω_t=.81; *n*=1,212).

Table 4 (continued)

Topic 5 ($\boldsymbol{\omega_t}$=.89; *n*=1,954) 1.indecisive (CON); 2.gracious, 3.sweet, 5.polite, 6.incredible, 8.sympathetic, and 10.son of a bitch (AGR); 4.free (EXT or INT); and 7.silly and 9.intelligent (EMO or INT). *Presumed factor:* Agreeableness (six terms; ω_t=.88; *n*=1,067).
Topic 6 ($\boldsymbol{\omega_t}$=.66; *n*=1,587) 1.rough, 3.sentimental, 4.good-natured, 5.sensitive, 7.wicked, 10.paranoid (AGR); 2.foolish (AGR or EMO or INT); 6.curious (INT); 8.confused (CON); 9.direct (EXT). *Presumed factor:* Agreeableness (seven terms; ω_t=.75; *n*=1,084).
Topic 7 ($\boldsymbol{\omega_t}$=.59; *n*=1,782) 1.guilty, 8.blockhead, 9.blind (EMO); 2.brat (EXT); 3.complicated (INT); 4.useless and 5.exaggerated (CON); 6.amorous, 7.shameless, and 10.grateful (AGR). *Presumed factor:* not identified.

Legend: ω_t= McDonald's omega total reliability coefficient; *n* = subsample used to estimate reliability; EXT (Extraversion); AGR (Agreeableness); CON (Conscientiousness); EMO (Emotional Stability); INT (Intellect); NVP (Negative/Positive Valence). Note: the number before each term indicates its relevance within the topic.

Fourteen-Topic Model.

Similarly to the Seven-Topic Model, identified as an optimal model for the data by Cao et al. (2009) method, the Fourteen-Topic Model (Table 5) was identified as an optimal model by the Deveaud et al. (2014) method. Of the 14 topics, only Topic 7 and 10 showed a reliability coefficient lower than .60. Topic 1 (ω_t=.85) and Topic 3 (ω_t=.97) are adherent to Extraversion content. The contents of Topic 2 (ω_t=.62), Topic 4 (ω_t=.70), Topic 10 (ω_t=.47), and Topic 13 (ω_t=.67) seem to be related to the dimensions of the seven-factor model of Negative Valence and Positive Valence (Almagor et al., 1995; Benet-Martinez & Waller, 1997; Saucier, 2003). Topic 5 (ω_t=.90), Topic 6 (ω_t=.88), Topic 8 (ω_t=.88), and Topic 14 (ω_t=.73 are predominantly adherent to Agreeableness. The remaining topics, Topic 7 (ω_t=.41), Topic 9 (ω_t=.93), Topic 11 (ω_t=.75), and Topic 12 (ω_t=.95), seem to be linked predominantly to Emotional Stability (ES). If one only considers the terms that are linked to ES, the reliability of Topic 7 (ω_t=.93), Topic 11 (ω_t=.80), and Topic 12 (ω_t=.96) increases. In

summary, the topics of this model reflect the content of some of the factors of the Big Five and the seven-factor models.

Table 5
The Fourteen-Topic Model with the 10 most relevant terms of the topics, reliability and presumed correspondence with other psycholexical models

Topic 1 (ω_t=.85; *n*=1,287).
1.guilty, 2.anxious, 3.great, 4.quiet, 5.warrior, 6.cocky, 7.friend, 8.maximum, 9.timid, and 10.retarded.
Presumed factor: Extraversion.

Topic 2 (ω_t=.62; *n*=1,227)
1.fool, 2.nobody, 3.innocent, 4.stupid, 5.tranquil, 6.responsible, 7.obvious, 8.silly, 9.dangerous, and 10.natural.
Presumed factor: Negative Valence.

Topic 3 (ω_t=.97; *n*=1,064)
1.vagabond, 2.horrible, 3.humane, 4.lady, 5.asshole, 6.social, 7.shit, 8.cute; 9.gothic; and 10.entangled.
Presumed factor: Extraversion (ω_t=.77; *n*=691, excluding *vagabond* and *lady*).

Topic 4 (ω_t=.70; *n*=1,440)
1.incredible, 2.damned, 3.special, 4.useless, 5.curious, 6.foolish, 7.dead, 8.genius, 9.arrogant, and 10.marvellous.
Presumed factor: Positive Valence.

Topic 5 (ω_t=.90; *n*=1,491)
1.needy, 2.lazy, 3.sentimental, 4.amorous, 5.nice, 6.perfect, 7.vacillating, 8.partner, 9.idle, and 10.joke.
Presumed factor: Agreeableness.

Topic 6 (ω_t=.88; *n*=1,348)
1.ignorant, 2.educated, 3.shameless, 4.beloved, 5.unlucky, 6.pacient, 7.soft, 8.good, 9.nervous, and 10.funny.
Presumed factor: Agreeableness.

Topic 7 (ω_t=.41; *n*=1,601)
1.indecisive, 2.unbearable, 3.sad, 4.exaggerated, 5.good-natured, 6.son of a bitch, 7.romantic, 8.paranoid, 9.fearful, and 10.annoying.
Presumed factor: Emotional Stability (seven terms, except *indecisive, unbearable* and *annoying*; ω_t=.93; *n*=1,052).

Topic 8 (ω_t=.88; *n*=1,117)
1.rough, 2.brat, 3.drunk, 4.selfish, 5.saint, 6.weary, 7.clever, 8.random, 9.tender, and 10.easy.
Presumed factor: Agreeableness.

Topic 9 (ω_t=.93; *n*=1,259)
1.trashy, 2.cardiac, 3.faithful, 4.deluded, 5.confused, 6.monster, 7.crybaby, 8.disgraced, 9.lousy, and 10.calm.
Presumed factor: Emotional Stability.

Topic 10 (ω_t=.47; *n*=1,814)
1.false, 2.important, 3.free, 4.ridiculous, 5.ferocious, 6.vile, 7.intelligent, 8.lost, 9.pure, and 10.ignored.
Presumed factor: Negative Valence.

Topic 11 (ω_t=.75; *n*=1,107)
1.happy, 2.famous, 3.bipolar, 4.douchebag, 5.normal, 6.bad, 7.stressed, 8.insane, 9.clumsy, and 10.committed.
Presumed factor: Emotional Stability (seven terms, except *famous, douchebag*, and *committed*; ω_t=.80; *n*=575).

Table 5 (continued)

Topic 12 ($\boldsymbol{\omega_t}$=.95; *n*=1,356)
1.weak, 2.sympathetic, 3.sensitive, 4.blind, 5.top, 6.direct, 7.switched on, 8.neurotic, 9.rebel, and 10.strong.
Presumed factor: Emotional Stability (seven terms, except *sympathetic*, *top*, and *direct;* ω_t=.96; *n*=907).

Topic 13 ($\boldsymbol{\omega_t}$=.67; *n*=1,156)
1.complicated, 2.blockhead, 3.buffoon, 4.proud, 5.macho, 6.crazy, 7.genial, 8.spoiled, 9.wicked, and 10.good.
Presumed factor: Negative Valence.

Topic 14 ($\boldsymbol{\omega_t}$=.73; *n*=1,379)
1.evil, 2.sweet, 3.dramatic, 4.cold, 5.lucky, 6.simple, 7.footloose, 8.difficult, 9.unique, and 10.forgotten.
Presumed factor: Agreeableness.

Legend: ω_t= McDonald's omega total reliability coefficient; *n* = subsample used to estimate reliability.
Note: the number before each term indicates its relevance within the topic.

Fifteen-Topic Model.

Similarly to the models with three, five, and six topics, the Fifteen-Topic Model (15TM) was not identified as an "optimal" model by the cross-validation analyses. We compared the semantic content of the 15TM (Table 6) with the 16PF factors (H. E. P. Cattell & Schuerger, 2003), that is composed of 15 personality factors plus a Reasoning scale. We did not identify a direct semantic correspondence between the 15TM and the 16PF since the majority of the terms of each topic contained features from more than one factor (Table 6). Regarding the reliability of the topics, Topics 9 and 11 have a ω_t bellow .60, Topics 8, 13, and 15 have a ω_t between .60 and .70, Topics 1, 2, 7, and 14 have a ω_t between .70 and .90, and the Topics 3, 4, 5, 6, 10 ad 12 have a ω_t superior to .90.

Table 6
The Fifteen-Topic Model with the 10 most relevant terms of the topics, reliability and presumed correspondence with other psycholexical models

Topic 1 (ω_t=.81; *n*=1,081)
1.cardiac (C/I/O/Q4), 2.timid (F/H/N/O), 3.faithful (G/L/Q1), 4.tranquil (C/F/L/O/Q4), 5.confused (C/O/Q3), 6.calm (C/F/L/O/Q4), 7.warrior (G/H/Q3), 8.normal, 9.friend_not (A/H/Q2), and 10.sincere (N).

Topic 2 (ω_t=.88; *n*=1,476)
1.indecisive (C/O/Q3), 2.trashy (F), 3.free (E/G/H/N/Q1), 4.amorous (A/I), 5.intelligent (M/Q1), 6.son of a bitch (E/G/H/I/Q2), 7.selfish (Q2), 8.hard (E/I/L), 9.rebel (E/G/H/N/Q1), and 10.maximum.

Topic 3 (ω_t=.95; *n*=1,496)
1.ignorant (L/Q1), 2.incredible (N/Q2), 3.vagabond (F/G/H), 4.beloved (A/Q2), 5.grateful (A/I), 6.lady, 7.romantic (A/I/M/N), 8.foolish (A/L/Q1), 9.crybaby (C/I/O/Q4), and 10.arrogant (E/Q2).

Topic 4 (ω_t=.96; *n*=1,440)
1.important (N/Q2), 2.good-natured (A/E/I/L/Q1), 3.blind (A/L/Q1), 4.douchebag (H/N), 5.buffoon (F/H/N), 6.social (H/N/Q2), 7.idle (F), 8.interesting (N/Q2), 9.committed (G/L/N/Q2), and 10.dangerous (E/G/H/L).

Topic 5 (ω_t=.95; *n*=1,156)
1.weak (C/E/H/I/L), 2.curious (M/Q1), 3.drunk (C/F), 4.top (Q2), 5.strong (C/E/H/I/L), 6.quiet (F/H/N), 7.neurotic (C/I/L/O/Q4), 8.cocky (E/H/Q2), 9.nervous (C/I/O/Q4), and 10.realistic (I/M/Q1).

Topic 6 (ω_t=.93; *n*=1,236)
1.anxious (C/O/Q4), 2.innocent (A/E/L/Q1), 3.stressed (C/L/O/Q4), 4.dead (F/L), 5.asshole (E/G/L/O), 6.macho (E/G/H/I/L/Q1), 7.disgusting, 8.gothic (F/H/I/N/Q2/Q4), 9.psychopath (E/G/H/L/Q2/Q4), and 10.sad (C/F/Q4).

Topic 7 (ω_t=.72; *n*=1,282)
1.sweet (A/I), 2.complicated (Q1), 3.polite (A), 4.blockhead (O/Q1), 5.random (Q3), 6.impossible, 7.weird (O/Q1), 8.spoiled (C/E/Q2), 9.happy (C/F/Q4), and 10.lousy (O).

Topic 8 (ω_t=.66; *n*=1,296)
1.unbearable (A), 2.sensitive (A/C/I/N), 3.lost (M/Q1/Q3), 4.deluded (L/M/Q1/Q3), 5.wicked (A), 6.horrible (A), 7.unlucky (F), 8.simple (N/Q1), 9.joke (O), and 10.idiot (L/O/Q1).

Topic 9 (ω_t=.55; *n*=1,170)
1.dramatic (I/N), 2.ridiculous (N/O), 3.responsible (F/G/Q3), 4.pacient (Q4), 5.partner (H/L/Q2), 6.ignored, 7.clumsy (H/Q3), 8.soft (E/H/I/L/N/O), 9.tender (A/E/I/L/N), and 10.forgotten (O/Q3).

Topic 10 (ω_t=.92; *n*=1,396)
1.evil (E/N), 2.sympathetic (A/I/N), 3.shameless (E/H), 4.nice (A/N), 5.bad (E/N), 6.footloose (F/Q4), 7.genius (M/Q1), 8.unique (Q2), 9.alone (Q2), and 10.insane (C/Q4).

Topic 11 (ω_t=54.; *n*=1,166)
1.chump (L/N/O/Q1), 2.needy (C/I/N/Q2), 3.exaggerated (I/L), 4.pure (I/L), 5.different, 6.switched on (Q1/Q4), 7.clever (N/Q1), 8.insecure (C/I/L/O), 9.demon, and 10.normal.

Topic 12 (ω_t=.91; *n*=1,352)
1.false (N), 2.rough (A/E/L), 3.ferocious (A/E/L), 4.vile (E), 5.lucky (F), 6.paranoid (C/L), 7.fearful (C/H/I/O/Q4), 8.disgraced, 9.proud, and 10.insane (C/Q4).

Table 6 (continued)

Topic 13 (ω_t=.63; *n*=1,353)
1.brat (F/H/Q3), 2.damned (F/H/Q4), 3.useless (Q3), 4.famous (Q2), 5.stupid (L/O/Q1), 6.vacillating (C/H/O/Q3), 7.humane (A/I), 8.monster, 9.nobody (O/Q2), and 10.great (O).

Topic 14 (ω_t=.88; *n*=1,726)
1.guilty (I/O), 2.sentimental (A/C/I/N), 3.special, 4.good (A/N), 5.silly (A/L/Q1), 6.obvious (O), 7.flacky (Q3), 8.annoying (O), 9.foolish (L/Q1), and 10.graceless (O).

Topic 15 (ω_t=.61; *n*=1,138)
1.lazy (Q3), 2.lousy (O), 3.dear (A/Q2), 4.direct (L/N), 5.weary (F), 6.jealous (L/Q2), 7.cute (A), 8.stubborn (L/Q2), 9.worse (O), and 10.gracious (A).

Legend: ω_t= McDonald's omega total reliability coefficient; *n* = subsample used to estimate reliability; A (Warmth), C (Emotional Stability), E (Dominance), F (Liveliness), G (Rule-Conscientiousness), H (Social Boldness), I (Sensitivity), L (Vigilance), M (Abstractedness), N (Privateness), O (Apprehension), Q1 (Openness to Change), Q2 (Self-Reliance), Q3 (Perfectionism), and Q4 (Tension). Note: the number before each term indicates its relevance within the topic.

Discussion

The primary objective of this study was to identify the latent structure underlying the collected data on personality descriptors obtained from Twitter. From an emic perspective, we first adopted a data-driven approach by using five cross-validation techniques to determine the models with an optimal number of latent dimensions or topics. From an emic-etic viewpoint, we also employed a theory-driven approach, exploring the most prominent factor solutions, such as models with three, five, six and seven factors and the 16PF (Almagor et al., 1995; Ashton et al., 2014; Benet-Martinez & Waller, 1997; Church et al., 1997; De Raad & Milacic, 2015; Goldberg & Rosolack, 1994; H. E. P. Cattell & Schuerger, 2003; S. B. G. Eysenck et al., 1985; Saucier, 2003; Thalmayer & Saucier, 2014). The content of the models resulting from our analyses was then semantically compared with the content of the models found in the literature.

The results from the cross-validation analyses did not converge since they provided different indications regarding the optimal model for the data. Using the technique of Deveaud et al. (2014), a model with 14 dimensions was identified.

However, a model with seven dimensions was indicated by means of the technique of Cao et al. (2009). The results from the other three techniques were not conclusive for our data. This way, in addition to the prominent models found in the literature, we also analyzed the two models indicated by the cross-validation analyses.

Three-Topic Model. The Three-Topic Model emerged from our data with one dimension resembling the Psychoticism factor, one similar to Extraversion, and one that seems to be a mixture of Agreeableness and Emotional Stability and did not reflect any PEN or Big Three factors. As this model was not identified as an optimal model by the cross-validation analyses and also lacked interpretability, we can consider that it is not a suitable latent structure for the data. This result diverges from evidence from previous studies that suggests that the three-factor model is the most cross-culturally replicable structure (De Raad et al., 2010). De Raad et al. (2010) suggested that the most typical three-factor model is composed of Extraversion, Agreeableness, and Conscientiousness.

Five-Topic Model. Numerous findings are suggestive that the five-factor model is a universal model (Allik, Realo, & McCrae, 2013; De Raad & Milacic, 2015), including evidence from Brazilian studies (Andrade, 2008; Hauck Filho et al., 2012; Hutz et al., 1998; Machado et al., 2014; Natividade & Hutz, 2015; Passos, 2014; Passos & Laros, 2015). However, the Five-Topic Model found in our study did not reflect the semantic content of the Big Five, as both the cross-validation and the semantic analyses indicated. The first two topics did not resemble any of the Big Five factors, although they show similarities with the Positive Valence and Negative Valence dimensions of the seven-factor model (Benet-Martinez & Waller, 1997). The remaining three topics were similar to Psychoticism, Agreeableness, and Emotional Stability. The content related to Conscientiousness and Intellect did not emerge, as in the Three-Topic Model.

In summary, the Five-Topic Model cannot be considered as a model that reflects the content of the Big Five.

Six-Topic Model. Like the previous models, this model did not arise from the cross-validation analyses as an optimal model. Five of the dimensions in this model seem to have semantic content predominantly related to Agreeableness, with exception of Topic 5 that shares similarities with Extraversion. This way, this model is not similar to other psycholexical models with six dimensions since it does not reflect the Big Five factors plus the sixth factor proposed in the HEXACO (Ashton et al., 2014) or in the Big Six (Thalmayer & Saucier, 2014) framework.

Seven-Topic Model. This model was indicated by cross-validation analyses as a suitable for the data. Three topics of this model seem to be predominantly congruent with Agreeableness. Of the remaining four topics, the first has similarities with the Positive Valence and Negative Valence factors from other seven-factor models (Almagor et al., 1995; Benet-Martinez & Waller, 1997; Saucier, 2003). The second and the third seem related to Emotional Stability and Psychoticism. The last topic has no clear interpretation considering the reference models (Almagor et al., 1995; Benet-Martinez & Waller, 1997; Church et al., 1998; Goldberg & Rosolack, 1994; Saucier, 2003). This way, the semantic analyses suggested that the Seven-Topic Model is not similar to other models with seven dimensions found in the literature.

In summary, at least three factors of the Seven-Topic Model seem to reflect the content of Agreeableness, while one factor does not have a straightforward interpretation. Regarding the reliability of the topics, Topic 2 (ω_t=.30), Topic 3 (ω_t=.54), Topic 4 (ω_t=.39), and Topic 7 (ω_t=.59) have all a coefficient under .60. However, maintaining only the most coherent terms within these topics, the reliability

coefficient for Topic 5 (ω_t=.88) and Topic 6 (ω_t=.75) increases, while for the remaining topics there is no increase in reliability.

Fourteen-Topic Model. This model was identified by cross-validation analyses as an optimal model for the data, together with the Seven-Topic Model. Two of its topics reflect the content of Extraversion, four of Agreeableness and four of Emotional Stability. The remaining four topics appear to be congruent with the formulations of the constructs of Negative and Positive Valence. Nevertheless, Topic 7 and 10 showed a reliability coefficient under .60. Although we did not identified in the literature a proposition of a personality model with 14 factors, the interpretation of the dimensions of the Fourteen-Topic Model suggested that this is a suitable candidate model for future research.

In comparison with the Seven-Topic Model, the topics of this model seem to be internally more coherent and, consequently, more interpretable. Although the information regarding the two models is not robust enough to indicate which one is the most suitable for the data, or if the latent dimensions of the models are valid Brazilian indigenous personality factors, it is possible to conclude that these are promising candidate models for future research.

Fifteen-Topic Model. This model did not show evidence of being a suitable model for the data, once it did not reflect the content of Cattell's 16PF factors, as both the cross-validation and the semantic analyses indicated. This result is consistent with the findings of other Brazilian research like the study of Primi et al. (2014), who proposed a factor solution with 12 dimensions for a questionnaire based on Cattell's model.

We only reported the most direct interpretation of each topic considering the Big Five framework as the primary target. This way, the presumed correspondence proposed

here between the topic models from this study and the prominent factor models (e.g., three, five, six, and seven factor models) are not final. Further research is required to empirically identify the possible similarities and discrepancies between models. Nevertheless, the results suggest that none of the most prominent psycholexical models are a suitable model for our data, considering both the topic modeling analyses and the qualitative semantic analyses. The exception is the model with seven dimensions, identified as an "optimal" model using one of the five cross-validation techniques, but with content different from theoretical models with seven factors.

Although the topic models and their interpretation are not final and need more evidence regarding its validity and reliability, they are informative clues to future research. Two possible conclusions can be hypothesized from these results. The first is that personality dimensions that can be considered autochthonous emerged from the data. The second is that not all factors identified in taxonomies of other languages are relevant to Brazilian culture. Both hypotheses are feasible, but further research will need to investigate whether the results are due to idiosyncrasies of Brazilian culture or are due to the nature of the sample and personality descriptors examined in this study.

Originated from Twitter, our sample is composed by users that freely choose which words they employ to describe their selves, if they want to evaluate publically their selves at all. As a consequence, our data collection strategy led to a sparse term-document matrix, with few terms per person. With this kind of data, it is not suitable to fit traditional psychometric models (e.g., exploratory and confirmatory factor analyses and Item Response Theory). This way, it was necessary to employ data analyses techniques developed specifically for this purposes, like the TF-IDF and the LDA topic modeling, a Bayesian machine learning approach. It is feasible to reason that the results are due to the sparsity of the term-document matrix or to the text mining techniques

employed. Other psychometric techniques applied to less sparse data could have produced more coherent semantic content within the latent dimensions.

Regarding the nature of the descriptors, our adjective list has distinct features in comparison with most studies within the psycholexical approach that relied on the examination of dictionaries to identify personality descriptors (Cheung et al., 2011; Daouk-Öyry et al., 2016; Uher, 2015). As highlighted, Twitter can be considered an open and public environment in which people choose whether and what to post. This way, one salient characteristic of the 172 adjectives investigated is that many of them can be viewed as hyperbolic, while many others are vulgar. In both cases, these terms are not easily found in other studies. For instance, many adjectives express extreme positive (e.g., *important, perfect, marvelous, special,* etc.) or extreme negative (e.g., *shameless, son of a bitch, useless, crazy,* etc.) self-evaluation.

While these terms, somewhat rare in other studies, are very relevant in our data, many common descriptors did not appear in our final list of 172 adjectives. The presence of hyperbolic and vulgar adjectives can be an indication that they are more relevant or at least more frequently used by our sample than the consecrated descriptors of factors such as Conscientiousness and Intellect. For example, common markers of Conscientiousness such as *organized, perfectionist, dedicated, efficient* and *meticulous* are not present in the present study. Nevertheless, this does not mean that the thematic underlying such factors are not relevant and these factors would not emerge as latent dimensions if the same sample answered a questionnaire of the Big Five, for example. This way, it is not possible to conclude if our data impaired the emergence of some factors or if these factors are not relevant to the Brazilian culture in general.

In summary, the issues highlighted above reinforce the pertinence of some of the criticism regarding the psycholexical approach discussed in the Introduction of this

paper. One broad critique to this approach is that using a limited set of items in test settings as the primary procedure of data collection can restrict the free expression of personality traits and the emergence of latent dimensions. A second broad critique is that research in this area frequently follows an etic imposed perspective, which can also circumscribe the personality dimensions and descriptors to those of more established models from other cultures.

From an emic-etic perspective, our study showed evidence that some factors from prominent models were not found in our data, while new latent dimensions emerged. Nevertheless, more studies are required to conclude that indigenous factors were found or that the absent factors are not relevant in Brazilian culture. Further research is required before claiming the emergence of Brazilian indigenous personality factors considering the results of this study. New studies will need to investigate the psychometric correspondence between the uncovered topics models and the factors from prominent lexical personality models.

References

Allik, J., Realo, A., & McCrae, R. (2013). Universality of the five-factor model of personality. In P. T. Costa & T. Widiger (Eds.), *Personality disorders and the Five Factor Model of Personality* (pp. 61-74). Washington: American Psychological Association.

Allik, J., Realo, A., Mottus, R., Pullmann, H., Trifonova, a., McCrae, R. R., … Korneeva, E. E. (2011). Personality profiles and the "Russian Soul": Literary and scholarly views evaluated. *Journal of Cross-Cultural Psychology*, *42*(3), 372–389. doi:10.1177/0022022110362751

Almagor, M., Tellegen, A., & Waller, N. (1995). The Big Seven model: A cross-cultural replication and further exploration of the basic dimensions of natural language of trait descriptions. *Journal of Personality and Social Psychology, 69*(2), 300-307. doi: 10.1037/0022-3514.69.2.300

Andrade, J. (2008). *Evidências de validade do Inventário dos Cinco Fatores de Personalidade para o Brasil*. Tese de Doutorado, Universidade de Brasília, Brasília.

Arun, R., Suresh, V., Madhavan, C. E. V., & Murthy, M. N. (2010). On finding the natural number of topics with Latent Dirichlet Allocation: Some observations. In Zaki, M. J., Yu, J. X., Ravindran, B., & Pudi, V. (Eds.) *Advances in knowledge discovery and data mining* (391–402). Berlin, Germany: Springer. doi: /10.1007/978-3-642-13657-3_43

Ashton, M. C., & Lee, K. (2007). Empirical, theoretical, and practical advantages of the HEXACO model of personality structure. *Personality and Social Psychology Review, 11*(2), 150–166. doi: 10.1177/1088868306294907

Ashton, M. C., Lee, K., Vries, R. E. (2014). The HEXACO honesty-humilty, agreeableness, and emotionality factors: A review of research and theory. *Personality and Social Psychology Review, 18*(2), 139-152. doi: 10.1177/1088868314523838

Benet-Martinez, V., & Waller, N. G. (1997). Further evidence for the cross-cultural generality of the Big Seven factor model: Indigenous and imported Spanish personality constructs. *Journal of Personality, 65*(3), 567–598. doi: 10.1111/j.1467-6494.1997.tb00327.x

Bischof, J. M., & Airoldi, E. M. (2012). Summarizing topical content with word frequency and exclusivity. *Proceedings of the 29th International Conference on Machine Learning, UK*. Available from https://arxiv.org/vc/arxiv/papers/1206/1206.4631v1.pdf

Blei, D. M., & Lafferty, J. D. (n.d.). *Topic Models.* Unpublished manuscript: Columbia University, New York, NY. Available from http://www.cs.columbia.edu/~blei/papers/Blei2012.pdf

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research, 3*, 993–81022. doi: 10.1162/jmlr.2003.3.4-5.993

Cao, J., Xia, T., Li, J., Zhang, Y., & Tang, S. (2009). A density-based method for adaptive LDA model selection. *Neurocomputing, 72*, 1775-1781. doi: 10.1016/j.neucom.2008.06.011.

Cattell, H. E. P. , & Schuerger, J. M. (2003). *Essentials of 16PF assessment.* Hoboken, NJ, USA: John Wiley & Sons, Inc.

Cattell, R. B., & Krug, S. E. (1986). The number of factor in the 16PF: A review of the evidence with special emphasis on methodological problems. *Educational and Psychological Measurement, 46*(3), 509-522. doi: 10.1177/0013164486463002xw

Chang, J. (2015). lda: Collapsed Gibbs Sampling Methods for Topic Models (Version 1.4.2) [Software]. Available from *The Comprehensive R Archive Network*: https://cran.r-project.org/web/packages/lda/index.html

Cheung, F. M., Van de Vijver, F., & Leong, F. (2011). Toward a new approach to the assessment of personality in culture. *American Psychologist, 66*(7), 593-603. doi: 10.1037/a0022389

Chuang, J., Manning, C. D., & Heer, J. (2012). Termite: Visualization techniques for assessing textual topic models. *Proceedings of the International Working Conference on Advanced Visual Interfaces, Italy,* 74–77. Available from http://idl.cs.washington.edu/papers/termite/

Chung, C. K., & Pennebaker, J,. W. (2008). Revealing dimensions of thinking in open-ended self-descriptions: An automated meaning extraction method for natural language. *Journal of Research in Personality, 42*(1), 96-132. doi:10.1016/j.jrp. 2007.04.006.

Church, A. T., Katigbak, M. S., & Reyes, J. A. S. (1998). Further exploration of Filipino personality structure using the lexical approach: Do the Big Five or Big-Seven dimensions emerge. *European Journal of Personality, 12*(4), 249–269. doi: 10.1002/(SICI)1099-0984(199807/08)12:4<249::AID-PER312>3.0.CO;2-T

Costa, P. T., Jr., & McCrae, R. R. (1976). Age differences in personality structure: A cluster analytic approach. *Journal of Gerontology*, *21*, 564–570.

Costa, P. T., Jr., & McCrae, R. R. (1992). Four ways five factors are basic. *Personality and Individual Differences, 13,* 653-665.

Daouk-Öyry, L., Zeinoun, P. , Choueiri, L., & Van de Vijver, F. J. R. (2016). Integrating global and local perspectives in psycholexical studies: A GloCal approach. *Journal of Research in Personality*. Advance online publication. doi: http://dx.doi.org/10.1016/ j.jrp. 2016.02.008

De Raad, B. (2000). *The big five personality factors. The psycholexical approach to personality.* Göttingen, Germany: Hogrefe & Huber Publishers.

De Raad, B., & Mlacic, B. (2015). The lexical foundation of the Big Five - Factor Model. In Widiger T. (Ed.) *The Oxford Handbook of the Five Factor Model*. Retrieved from www.oxfordhandbooks.com. doi: 10.1093/oxfordhb/9780199352487.013.12.

De Raad, B., Barelds, D., Levert, E., ... Katigbak, M. S. (2010). Only three factors of personality description are fully replicable across languages: The comparison of 14 trait taxonomies. *Journal of Personality and Social Psychology, 91*(1), 160-173. doi: 10.1037/a0017184.

Deveaud, R., Sanjuan, E., & Bellot, P. (2014). Accurate and effective latent concept modeling for ad hoc information retrieval. *Revue des Sciences et Technologies de L'Information – Série Document Numérique, 17*(1), 61-84. doi: 10.3166/dn.17.1.61-84.

Dicionário Houaiss da Língua Portuguesa. (2009). Rio de Janeiro: Objetiva.

Dicionário Priberam da Língua Portuguesa. (n.d.). Retrieved November, 2017, from https://www.priberam.pt

Digman, J. (1990). Personality structure: emergence of the five-factor model. *Annual Review of Psychology*, *41*, 417-440.

Digman, J. M. (1997). Higher-order factors of the Big Five. *Journal of Personality and Social Psychology*, *73*(6), 1246-1256. doi: 10.1037/0022-3514.73.6.1246

Eysenck, H. J. (1991). Dimensions of personality: 16, 5 or 3? – Criteria for a taxonomic paradigm. *Personality and Individual Differences*, *12*(8), 773-790. doi: 10.1016/0191-8869(91)90144-Z

Eysenck, H. J. (1992). Four ways five factors are not basic. *Personality and Individual Differences*, *13*(6), 667-673. doi: 10.1016/0191-8869(92)90237-J

Eysenck, H. J. (1997). Personality and experimental psychology: The unification of psychology and the possibility of a paradigm. *Journal of Personality and Social Psychology*, *73*(6), 1224-1237. doi: 10.1037/0022-3514.73.6.1224

Eysenck, S. B. G., Eysenck, H. J., & Barret, P. (1985). A revised version of the psychoticism scale. *Personality and individual differences, 6*(1), 21-29. doi: 10.1016/0191-8869(85)90026-1

Farahani, M. N., De Raad, B., Farzad, V., & Fotoohie, M. (2016). Taxonomy and structure of Persian personality-descriptive trait terms. *International Journal of Psychology*, *51*(2), 139-149. doi:10.1002/ijop. 12133

Feinerer, I., & Hornik, K. (2017). tm: Text mining package (Version 0.7-1) [Software]. Available from *The Comprehensive R Archive Network*: http://cran.r-project.org/ web/packages/tm/index.html

Feinerer, I., Hornik, K., & Meyer, D. (2008). Text mining infrastructure in R. *Journal of Statiscal Software*, *25*(8), 1-54.doi: 10.18637/jss.v025.i05

Fiske, D. W. (1949). Consistency of the factorial structures of personality ratings from different sources. *Journal of Abnormal and Social Psychology*, *44*, 329-344.

Flach, P. (2012). *Machine learning: The art and science of algorithms that make sense of data.* Cambridge: Cambridge University Press.

Gentry, J. (2015). twitteR: R Based Twitter Client (Version 1.1.9) [Software]. Available from *The Comprehensive R Archive Network*: http://cran.r-project.org/web/packages/twitteR/ index.html

Goldberg, L. R. (1981). Language and individual differences. The search for universals in personality lexicons. In L. Wheeler (Ed.), *Review of personality and social psychology* (Vol. 2, pp. 141–165). Beverly Hills, CA: Sage.

Goldberg, L. R. (1992). The development of markers for the Big Five factor structure. *Psychological Assessment*, *4*, 26–42.

Goldberg, L. R., & Rosolack, T. K. (1994). The Big Five Factor Structure as an integrative framework: An empirical comparison with Eysenck's P-E-N Model. In: C. F. Halverson, Jr., G. A. Kohnstamn, & R. P. Martin (Eds.), *The developing structure of temperament and personality from infancy to adulthood* (pp. 7-35). New York: Erlbaum.

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceeding of the National Academy of Sciences, USA, 101*, 5228-5235.

Griffiths, T. L., Steyvers, M., & Tenenbaum, M. (2007). Topics in semantic representation. *Psychological Review, 114*(2), 211-244. doi: 10.1037/0033-295X.114.2.211.

Grün, B., & Hornik, K. (2011). topicmodels: An R package for fitting topic models. *Journal of Statistical Software*, 40(13). doi: 10.18637/jss.v040.i13.

Grün, B., & Hornik, K. (2017). topicmodels: Topic Models (Version 0.2-6) [Software]. Retrieved from: https://cran.r-project.org/web/packages/topicmodels/index.html

Hall, D., Jurafsky, D., & Manning, C. D. (2008). Studying the history of ideas using topic models. *Proceedings of the Conference on Empirical Methods in Natural Language Processing, USA*, 363-371. Retrieved from: http://web.stanford.edu/~jurafsky/hallemnlp08.pdf

Hauck Filho, N., Machado, W., Teixeira, M., & Bandeira, D. (2012). Evidências de validade de marcadores reduzidos para a avaliação da personalidade no modelo dos cinco grandes fatores. *Psicologia: Teoria e Pesquisa*, *28*(4), 417-423.

Hutz, C. S., Nunes, C. H, Silveira, A. D., Serra, J., Anton, M., & Wieczorek, L. S. (1998). O desenvolvimento de marcadores para a avaliação da personalidade no modelo dos cinco grandes fatores. *Psicologia: Reflexão e Crítica*, *11*(2), 395-411. doi: 10.1590/S0102-79721998000200015.

Isaka, H. (1990). Factor analysis of trait terms in everyday Japanese language. *Personality and Individual Differences*, *11*(2), 115-124. doi: 10.1016/0191-8869(90)90003-A

John, O. P. , Angleitner, A., & Ostendorf, F. (1988). The lexical approach to personality: A historical review of trait taxonomic research. *European Journal of Personality*, *2*, 171-203.

Kaplan, D. (2014). *Bayesian statistics for the social sciences.* New York, USA: The Guilford Press.

Kosinski, M., Wang, Y., Lakkaraju, H., & Leskovec, J. (2016). Mining big data to extract patterns and predict real-life outcomes. *Psychological Methods*, *21*(4), 493-506. http://dx.doi.org/10.1037/met0000105

Latent Dirichlet allocation. (n.d.). In *Wikipedia.* Retrieved December 02, 2017, from https://en.wikipedia.org/wiki/Latent_Dirichlet_allocation

Lee, C., Kim, K., Seo, Y., & Chung, C. (2007). The relations between personality and language use. *The Journal of General Psychology*, *134*(4), 405-413. doi: 10.3200/GENP. 134.4.405-414.

Leskovec, J., Rajaraman, A., & Ullman, J. D. (2014). *Mining of massive datasets* (2nd ed.). Cambridge, United Kingdom: Cambridge University Press.

Liu, L., Tang, L., Dong, W., Yao, S., & Zhou, W. (2016). An overview of topic modeling and its current applications in bioinformatics. *SpringerPlus*, 5:1608, 1-22. doi: 10.1186/s40064-016-3252-8.

Machado, W. L., Hauck Filho, N., Teixeira, M. A. P. , & Bandeira, D. R. (2014). Análise de Teoria de Resposta ao Item de marcadores reduzidos da personalidade. *Psico*, *45*(4), 551-558. doi: 10.15448/1980-8623.2014.4.13138.

Merriam-Webster.com. (n.d.). Retrieved November, 2017, from https://www.merriam-webster.com/

Musek, J. (2007). A general factor of personality: Evidence for the big one in the five-factor model. *Journal of Research in Personality, 41*(6), 1213–1233. doi: doi.org/10.1016/j.jrp. 2007.02.003

Natividade, J. C., & Hutz, C. S. (2015). Escala reduzida de descritores dos cinco grandes fatores de personalidade: prós e contras. *Psico, 46*(1), 79-89. doi: 10.15448/1980-8623.2015.1.16901

Nel, J. A., Valchev, V. H., Rothmann, S., Van de Vijver, F. J. R., Meiring, D., & de Bruin, G. P. (2012). Exploring the personality structure in the 11 languages of South Africa. *Journal of Personality, 80*(4), 915-948. doi: 10.1111/j.1467-6494.2011.00751.x*n*-gram. (n.d.). In *Wikipedia.* Retrieved December 02, 2017, from https://en.wikipedia.org/wiki/N-gram

Nikita, M. (2016, October 10a). *Select number of topics for LDA model*. Retrieved from: https://cran.r-project.org/web/packages/ldatuning/vignettes/topics.html

Nikita, M. (2016b). Ldatuning: Tuning of the Latent Dirichlet Allocation Models Parameters (Version 0.2.0) [Software]. Available from *The Comprehensive R Archive Network*: https://cran.r-project.org/web/packages/ldatuning/index.html

Norman, W. T. (1963). Toward an adequate taxonomy of personality attributes. Replicated factor structure in peer nomination personality ratings. *Journal of Abnormal and Social Psychology*, *66*(6), 574–583.

Park, G., Schwartz, A., Eichstaedt, J., Kern, M., Kosinski, M., Stillwell, D., et al. (2015). Automatic personality assessment through social media language. *Journal of Personality and Social Psychology, 108*(6), 934-952. doi: 10.1037/pspp0000020.

Passos, M. F. D. (2014). *Elaboração e validação de escala de diferencial semântico para avaliação da personalidade.* Tese de Doutorado, Universidade de Brasília, Brasília.

Passos, M. F. D., & Laros, J. A. (2015). Construção de uma escala reduzida de cinco grandes fatores de personalidade. *Avaliação Psicológica*, *14*(1), 115-123.

Peres, A. J. S., & Laros, J. A. (2018a). *Lexical hypothesis, cross-cultural psychology and natural language.* Unpublished manuscript, Institute of Psychology, University of Brasilia, Brasília, Distrito Federal, Brazil.*

Peres, A. J. S., & Laros, J. A. (2018b). *The personality lexicon in the Brazilian Portuguese: Searching for descriptor terms in natural language.* Unpublished manuscript, Institute of Psychology, University of Brasilia, Brasília, Distrito Federal, Brazil.**

Poldrack, R. A., Mumford, J. A., Schonberg, T., Kalar, D., Barman, B., & Yarkoni, T. (2012). Discovering relations between mind, brain, and mental disorders using topic mapping. *PLOS Computational Biology*, *8*(10). doi: https://doi.org/10.1371/journal.pcbi.1002707

Polzehl, T. (2015). *Personality in speech. Assessment and automatic classification*. Bern: Springer.

Primi, R., Ferreira-Rodrigues, C. F., & Carvalho, L. F. (2014). Cattell's Personality Factor Questionnaire (CPFQ): Development and preliminary study. *Paidéia*, *24*(5), 29-37. doi: 10.1590/1982-43272457201405.

Priva, U. C., & Austerweil, J. L. (2015). Analyzing the history of cognition using topic models. *Cognition*, *135*, 4-9. doi: http://dx.doi.org/10.1016/j.cognition.2014.11.006

Qiu, L., Lin, H., Ramsay, J., & Yang, F. (2012). You are what you tweet: Personality expression and perception on Twitter. *Journal of Research in Personality, 46*(6), 710-718. doi: 10.1016/j.jrp. 2012.08.008.

R Development Core Team. (2017). R: A language and environment for statistical computing (Version 3.4.0) [Software]. Available from *The Comprehensive R Archive Network*: https://cran.r-project.org/

Revelle, W. (1995). Personality processes. *Annual Review of Psychology*, *46*, 295-328. doi: 10.1146/annurev.ps.46.020195.001455

Revelle, W. (2016). Hans Eysenck: Personality theorist. *Personality and Individual Differences*, *103*, 32-39. doi: 10.1016/j.paid.2016.04.007

Revelle, W. (2017). psych: Procedures for Psychological, Psychometric, and Personality Research (Version 1.7.8) [Software]. Retrieved from: https://cran.r-project.org/web/packages/psych/index.html

Revelle, W. (n.d.). Calculate McDonald's omega estimates of general and total factor saturation. Available from https://www.personality-project.org/r/psych/help/omega.html

Revelle, W., & Wilt, J. (2013). The general factor of personality: A general critique. *Journal of Research in Personality*, *47*(5), 493-504. doi:10.1016/j.jrp. 2013.04.012

Revelle, W., & Zinbarg, R. E. (2009). Coefficients Alpha, Beta, Omega, and the GLB: Comments on Sijtsma. *Psychometrika*, *74*(1), 145-154. doi: 10.1007/s11336-008-9102-z

Saucier, G. (2003). An alternative multi-language structure for personality attributes. *European Journal of Personality, 17*(3), 179–205. doi: 10.1002/per.489

Saucier, G. (2009). Recurrent personality dimensions in inclusive lexical studies: Indications for a Big Six structure. *Journal of Personality, 77*(5), 1577–1614. doi: 10.1111/j.1467-6494.2009.00593.x

Saucier, G., & Ostendorf, F. (1999). Hierarchical subcomponents of the Big Five personality factors: A cross-language replication. *Journal of Personality and Social Psychology*, 76(4), 613-627. doi: 10.1037//0022-3514.76.4.613

Saucier, G., Georgiades, S., Tsaousis, I., & Goldberg, L. R. (2005). The factor structure of Greek personality adjectives. *Journal of Personality and Social Psychology*, *88*(5), 856-875. doi: 10.1037/0022-3514.88.5.856

Selivanov, D., & Wang, Q. (2017). text2vec: Modern text mining framework for R (Version 0.5.0) [Software]. Available from *The Comprehensive R Archive Network*: https://cran.r-project.org/web/packages/text2vec/index.html

Sievert, C., & Shirley, K. (2014). LDAvis: A Method for visualizing and interpreting topics. *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces, USA*, 63–70. Available from https://nlp.stanford.edu/events/illvi2014/papers/sievert-illvi2014.pdf

Sievert, C., & Shirley, K. (2015). LDAvis: Interactive visualization of topic models (Version 0.3.2) [Software]. Retrieved from *The Comprehensive R Archive Network*: https://cran.r-project.org/web/packages/LDAvis/index.html

Silge, J., & Robinson, D. (2017). *Text mining with R. A Tidy approach.* Available from http://tidytextmining.com/index.html

Singh, J. K., Misra, G., & De Raad, B. (2013). Personality structure in the trait lexicon of Hindi, a major language spoken in India. *European Journal of Personality, 27*(6), 605-620. doi: 10.1002/per.1940

Thalmayer, A. G., & Saucier, G. (2014). The Questionnaire Big Six in 26 nations: Developing cross-culturally applicable Big Six, Big Five and Big Two inventories. *European Journal of Personality, 28*(5), 482-496. doi: 10.1002/per.1969

Thesaurus.com. (n.d.). Retrieved November, 2017, from http://www.thesaurus.com

Translate.google.com (n.d.). Retrieved November, 2017, from https://translate.google.com

Tupes, E. C., & Christal, R. E. (1961). Recurrent personality factors based on trait ratings. *USAF ASD Technical Report No. 61-97*. U.S. Air Force: Lackland Air Force Base, San Antonio, TX.

Twitter. (n.d.). In *Wikipedia.* Retrieved December 02, 2017, from https://en.wikipedia.org/wiki/Twitter

Uher, J. (2015). Developing "personality" taxonomies: Metatheoretical and methodological rationales underlying selection approaches, methods of data generation and reduction principles. *Integrative Psychological and Behavioral Science*, *49*(4), 531-589. doi: 10.1007/s12124-014-9280-4

Valchev, V., Nel, J., Van de Vijver, F., Meiring, D., Bruin, G., & Rothman, S. (2012). Similiraties and differences in implicit personality concepts across ethnocultural groups in South Africa. *Journal of Cross-Cultural Psychology, 44*(3), 365-388. doi: 10.1177/0022022112443856

Yarkoni, T. (2010). Personality in 100,000 words: A large-scale analysis of personality and word use among bloggers. *Journal of Research in Personality, 4*(3), 363-373. doi: 10.1016/j.jrp. 2010.04.001.

Zhou, X., Saucier, G., Gao, D., & Liu, J. (2009). The factor structure of Chinese personality terms. *Journal of Personality*, *77*(2), 363-400. doi: 10.1111/j.1467-6494.2008.00551.x

* The manuscript is part of this dissertation, as Manuscript 1.

** The manuscript is part of this dissertation, as Manuscript 2.

Appendix 1

Descriptive statistics of terms before and after TF-IDF normalization

Table 7 presents the frequency that each term was used by the users (*UF*) in our sample *(n= 86,899)*. The *term frequency* column shows the descriptive statistics (mean, standard deviation, minimum, maximum and range) before applying term frequency – inverse document frequency (TF-IDF) normalization. The *inverse document frequency* column shows the same statistics after normalization.

Table 7
User frequency or the number of users that used the term (UF), overall term frequency in the corpus (TF), inverse document frequency (IDF), mean (M), minimum (Min), maximum (Max), and range. Sample (n=86,899).

Term		Term frequency						Inverse document frequency					
	UF	TF	M	SD	Min	Max	Range	IDF	M	SD	Min	Max	Range
Aggressive	399	433	.00	.07	0	3	3	398.62	.00	.09	0	5.49	5.49
Alone	3758	4023	.04	.22	0	6	6	2622.12	.03	.18	0	3.23	3.23
Amorous	1459	1591	.02	.14	0	4	4	1165.38	.01	.13	0	4.17	4.17
Annoying	7806	9090	.09	.35	0	7	7	4333.78	.05	.20	0	2.53	2.53
Antisocial	491	517	.01	.08	0	3	3	551.33	.01	.11	0	5.32	5.32
Anxious	1328	1398	.01	.12	0	4	4	1292.94	.01	.15	0	4.35	4.35
Arrogant	495	574	.01	.09	0	4	4	56.42	.01	.11	0	5.26	5.26
Asshole	840	903	.01	.10	0	4	4	843.94	.01	.13	0	4.79	4.79
Bad	818	8875	.01	.11	0	9	9	2468.09	.01	.13	0	4.77	4.77
Beloved	870	968	.01	.12	0	13	13	94.74	.00	.14	0	4.75	4.75
Bipolar	1968	2123	.02	.16	0	4	4	1621.47	.02	.16	0	3.91	3.91
Blind	1030	1104	.01	.11	0	4	4	977.99	.01	.13	0	4.57	4.57
Blockhead	937	1010	.01	.11	0	4	4	93.69	.01	.13	0	4.66	4.66
Brat	1337	1471	.02	.14	0	6	6	1298.17	.01	.15	0	4.30	4.30
Buffoon	880	936	.01	.10	0	4	4	855.33	.01	.13	0	4.74	4.74
Calm	1874	2006	.02	.16	0	5	5	1512.03	.02	.15	0	3.93	3.93
Cardiac	1199	1418	.01	.14	0	6	6	125.89	.01	.16	0	4.46	4.46
Chump	4212	4808	.05	.26	0	9	9	2963.32	.03	.20	0	3.16	3.16
Clever	679	701	.01	.09	0	3	3	697.85	.01	.12	0	4.97	4.97
Clueless	245	253	.00	.05	0	3	3	266.53	.00	.07	0	4.01	4.01
Clumsy	718	740	.01	.09	0	4	4	729.55	.01	.12	0	4.97	4.97
Cocky	560	620	.01	.09	0	3	3	601.95	.01	.11	0	5.14	5.14
Cold	2709	2925	.03	.19	0	5	5	204.29	.02	.17	0	3.60	3.60
Committed	716	763	.01	.09	0	4	4	685.20	.01	.11	0	4.94	4.94
Complicated	1456	1556	.02	.14	0	5	5	1246.50	.01	.14	0	4.19	4.19
Confused	1102	1179	.01	.12	0	4	4	962.62	.01	.13	0	4.48	4.48
Crazy	2348	2587	.03	.18	0	6	6	1885.05	.02	.17	0	3.73	3.73

Table 7 (continued)

Term		Term frequency						Inverse document frequency					
	UF	TF	M	SD	Min	Max	Range	IDF	M	SD	Min	Max	Range
Crybaby	751	787	.01	.09	0	3	3	733.62	.01	.12	0	4.88	4.88
Curious	1275	1379	.01	.13	0	4	4	1092.63	.01	.13	0	4.33	4.33
Cute	927	1011	.01	.12	0	10	10	805.50	.01	.12	0	4.62	4.62
Damned	613	797	.01	.12	0	8	8	899.41	.01	.18	0	5.25	5.25
Dangerous	501	524	.01	.08	0	3	3	533.03	.01	.10	0	2.62	2.62
Dead	940	978	.01	.10	0	4	4	873.80	.01	.12	0	4.67	4.67
Dear	3816	4283	.04	.24	0	13	13	2818.93	.03	.19	0	3.27	3.27
Deluded	1015	1098	.01	.12	0	6	6	955.90	.01	.13	0	4.59	4.59
Demon	678	726	.01	.09	0	5	5	655.05	.01	.11	0	4.96	4.96
Different	2404	2529	.03	.17	0	8	8	1809.35	.02	.16	0	3.72	3.72
Difficult	3516	3780	.04	.21	0	10	10	2408.64	.03	.17	0	3.33	3.33
Direct	894	922	.01	.10	0	3	3	828.71	.01	.12	0	4.68	4.68
Disgraced	686	746	.01	.10	0	3	3	668.67	.01	.11	0	4.92	4.92
Disgusting	781	828	.01	.10	0	5	5	734.08	.01	.11	0	4.80	4.80
Douchebag	917	1000	.01	.11	0	5	5	88.88	.01	.13	0	4.67	4.67
Dramatic	1474	1592	.02	.14	0	5	5	126.73	.01	.14	0	4.19	4.19
Drunk	960	1040	.01	.11	0	4	4	905.10	.01	.12	0	4.60	4.60
Dung	2050	2335	.02	.18	0	6	6	1683.64	.02	.16	0	3.86	3.86
Easy	2212	2344	.02	.17	0	4	4	1724.04	.02	.16	0	3.77	3.77
Entangled	441	460	.00	.07	0	4	4	508.06	.01	.10	0	2.73	2.73
Exaggerated	984	1123	.01	.12	0	5	5	1051.59	.01	.15	0	4.62	4.62
Faithful	1053	1149	.01	.12	0	4	4	1035.23	.01	.14	0	4.54	4.54
False	1704	1989	.02	.18	0	14	14	1492.89	.02	.16	0	4.05	4.05
Famous	1296	1394	.01	.13	0	10	10	1094.95	.01	.13	0	4.32	4.32
Fearful	667	703	.01	.09	0	4	4	676.85	.01	.11	0	4.99	4.99
Ferocious	1142	1240	.01	.12	0	3	3	1099.83	.01	.14	0	4.44	4.44
Fool	2701	2956	.03	.19	0	6	6	2145.55	.02	.18	0	3.59	3.59
Foolish	9433	11514	.12	.42	0	17	17	5375.45	.06	.24	0	2.39	2.39
Footloose	742	801	.01	.10	0	6	6	805.33	.01	.12	0	4.84	4.84
Forgotten	597	615	.01	.08	0	3	3	567.90	.01	.10	0	3.41	3.41
Free	1361	1502	.02	.14	0	6	6	1244.02	.01	.15	0	4.28	4.28
Friend	16303	19983	.21	.54	0	15	15	7077.98	.08	.22	0	1.79	1.79
Friend_not	195	196	.00	.05	0	2	2	223.07	.00	.07	0	3.10	3.10
Funny	2286	2442	.03	.17	0	4	4	173.91	.02	.15	0	3.73	3.73
Genial	8012	9396	.10	.37	0	16	16	4463.10	.05	.21	0	2.52	2.52
Genius	604	627	.01	.08	0	3	3	676.43	.01	.12	0	5.10	5.10
Good	33020	12060	.13	.41	0	27	27	6755.34	.06	.21	0	2.24	2.24
Good-natured	1202	1263	.01	.12	0	4	4	1088.88	.01	.14	0	4.41	4.41
Gothic	608	660	.01	.09	0	3	3	613.92	.01	.11	0	5.10	5.10
Graceless	337	344	.00	.06	0	2	2	36.79	.00	.09	0	3.76	3.76
Gracious	4110	4813	.05	.28	0	30	30	2742.00	.03	.19	0	3.17	3.17
Grateful	2179	2476	.02	.20	0	20	20	187.25	.02	.17	0	3.84	3.84
Great	4434	4801	.05	.24	0	7	7	2949.42	.03	.18	0	3.09	3.09
Guilty	1366	1433	.01	.13	0	5	5	1291.24	.01	.15	0	4.32	4.32
Happy	7857	9033	.09	.35	0	9	9	4478.93	.05	.22	0	2.55	2.55
Hard	590	611	.01	.08	0	3	3	596.60	.01	.11	0	5.12	5.12
Horrible	2348	2628	.03	.19	0	7	7	177.93	.02	.16	0	3.70	3.70
Humane	2482	2680	.03	.18	0	11	11	1966.24	.02	.17	0	3.65	3.65

Table 7 (continued)

Term		Term frequency						Inverse document frequency					
	UF	TF	M	SD	Min	Max	Range	IDF	M	SD	Min	Max	Range
Idiot	5014	5620	.06	.27	0	7	7	3341.13	.04	.20	0	2.99	2.99
Idle	821	926	.01	.29	0	81	81	766.45	.01	.11	0	3.18	3.18
Ignorant	1406	1532	.02	.14	0	4	4	1276.63	.01	.14	0	4.23	4.23
Ignored	587	627	.01	.09	0	5	5	635.75	.01	.12	0	5.10	5.10
Impolite	206	211	.00	.05	0	3	3	283.64	.00	.08	0	4.09	4.09
Important	1527	1618	.02	.14	0	10	10	1309.14	.02	.14	0	2.77	2.77
Impossible	522	532	.01	.08	0	3	3	552.82	.01	.10	0	3.48	3.48
In love	16178	20704	.21	.58	0	16	16	7545.94	.09	.25	0	1.86	1.86
Incredible	1312	1404	.01	.14	0	9	9	117.88	.01	.14	0	4.29	4.29
Indecisive	1896	2029	.02	.16	0	10	10	158.41	.02	.16	0	4.01	4.01
Innocent	985	1088	.01	.12	0	6	6	974.97	.01	.14	0	4.61	4.61
Insane	3604	4188	.04	.26	0	26	26	2611.92	.03	.19	0	3.31	3.31
Insecure	771	832	.01	.10	0	6	6	703.68	.01	.11	0	4.82	4.82
Intelligent	1296	1370	.01	.13	0	4	4	1103.99	.01	.14	0	4.33	4.33
Interesting	547	580	.01	.08	0	4	4	612.54	.01	.11	0	5.14	5.14
Jealous	3681	4064	.04	.23	0	5	5	2472.84	.03	.17	0	3.29	3.29
Joke	806	859	.01	.10	0	6	6	721.40	.01	.11	0	4.78	4.78
Lady	789	859	.01	.11	0	4	4	869.40	.01	.12	0	4.73	4.73
Lazy	1607	1692	.02	.14	0	6	6	1328.22	.02	.14	0	4.13	4.13
Loco	14177	17636	.18	.52	0	17	17	6822.11	.08	.24	0	1.95	1.95
Lost	1088	1147	.01	.11	0	2	2	997.04	.01	.13	0	4.54	4.54
Lousy	4337	4752	.05	.24	0	8	8	2894.76	.03	.19	0	3.12	3.12
Lucky	933	973	.01	.10	0	5	5	909.03	.01	.13	0	4.72	4.72
Macho	796	873	.01	.10	0	5	5	742.86	.01	.12	0	4.81	4.81
Marvelous	3752	4169	.04	.23	0	10	10	2568.47	.03	.18	0	3.26	3.26
Maximum	558	572	.01	.08	0	5	5	541.01	.01	.10	0	3.44	3.44
Monster	761	800	.01	.09	0	3	3	739.67	.01	.12	0	4.89	4.89
Natural	480	491	.00	.07	0	2	2	501.60	.01	.10	0	2.67	2.67
Needy	1674	1825	.02	.15	0	5	5	1377.21	.02	.14	0	4.04	4.04
Nervous	766	796	.01	.10	0	5	5	712.01	.01	.11	0	4.81	4.81
Neurotic	518	556	.01	.08	0	3	3	586.29	.01	.11	0	5.22	5.22
Nice	837	1086	.01	.14	0	8	8	94.60	.01	.14	0	4.72	4.72
Nobody	12145	14222	.15	.45	0	30	30	5653.08	.07	.21	0	2.08	2.08
Normal	2973	3212	.03	.20	0	5	5	2119.19	.02	.17	0	3.48	3.48
Obvious	741	763	.01	.09	0	2	2	715.05	.01	.12	0	4.88	4.88
Paranoid	904	953	.01	.10	0	3	3	821.40	.01	.12	0	4.69	4.69
Partner	630	663	.01	.09	0	7	7	683.02	.01	.12	0	5.07	5.07
Patient	654	676	.01	.08	0	2	2	675.37	.01	.12	0	5.04	5.04
Perfect	2822	3044	.03	.19	0	4	4	2144.78	.02	.18	0	3.58	3.58
Polite	1157	1300	.01	.13	0	9	9	1181.46	.01	.16	0	4.44	4.44
Proud	2333	2509	.03	.17	0	9	9	1806.30	.02	.16	0	3.76	3.76
Psychopath	643	687	.01	.09	0	4	4	641.46	.01	.11	0	5.00	5.00
Pure	966	1039	.01	.11	0	4	4	901.84	.01	.12	0	4.60	4.60
Quiet	909	946	.01	.10	0	3	3	833.15	.01	.12	0	4.65	4.65
Random	573	602	.01	.08	0	3	3	607.19	.01	.11	0	5.15	5.15
Realistic	545	563	.01	.08	0	2	2	571.77	.01	.11	0	5.19	5.19
Rebel	508	548	.01	.09	0	10	10	564.01	.01	.11	0	5.29	5.29
Responsible	717	778	.01	.10	0	13	13	747.80	.01	.13	0	5.02	5.02

Table 7 (continued)

Term		Term frequency						Inverse document frequency					
	UF	TF	M	SD	Min	Max	Range	IDF	M	SD	Min	Max	Range
Retarded	2316	2548	.03	.18	0	5	5	1819.60	.02	.16	0	3.74	3.74
Ridiculous	1487	1612	.02	.14	0	5	5	1205.65	.01	.14	0	4.17	4.17
Romantic	929	992	.01	.11	0	4	4	859.45	.01	.12	0	4.64	4.64
Rough	3799	4420	.05	.25	0	9	9	2793.55	.03	.19	0	3.24	3.24
Sad	3815	4579	.05	.28	0	17	17	2532.13	.03	.17	0	3.21	3.21
Saint	3415	3771	.04	.22	0	14	14	2495.45	.03	.18	0	3.37	3.37
Selfish	841	896	.01	.10	0	3	3	825.91	.01	.12	0	4.75	4.75
Sensitive	1370	1459	.02	.13	0	7	7	1128.60	.01	.13	0	4.25	4.25
Sentimental	1391	1464	.02	.13	0	5	5	1227.29	.01	.14	0	4.22	4.22
Serious	5726	6442	.07	.29	0	7	7	3316.81	.04	.18	0	2.78	2.78
Shameless	988	1074	.01	.12	0	5	5	998.71	.01	.13	0	4.56	4.56
Shit	5781	6951	.07	.37	0	50	50	3564.51	.04	.20	0	2.82	2.82
Sick	1961	2150	.02	.16	0	5	5	1641.13	.02	.16	0	3.89	3.89
Silly	9251	11004	.11	.40	0	14	14	5005.42	.06	.22	0	2.38	2.38
Simple	861	893	.01	.10	0	3	3	847.08	.01	.12	0	4.72	4.72
Sincere	2244	2491	.03	.27	0	60	60	1691.08	.02	.15	0	3.76	3.76
Slow	5244	6175	.06	.30	0	8	8	341.55	.04	.20	0	2.93	2.93
Social	649	669	.01	.09	0	3	3	686.87	.01	.12	0	5.01	5.01
Soft	607	638	.01	.09	0	2	2	631.08	.01	.11	0	3.37	3.37
Son of a bitch	954	1002	.01	.11	0	3	3	928.25	.01	.13	0	4.64	4.64
Special	1087	1148	.01	.12	0	4	4	1115.98	.01	.14	0	4.48	4.48
Spoiled	530	575	.01	.08	0	4	4	541.33	.01	.10	0	5.20	5.20
Stressed	954	1010	.01	.11	0	4	4	898.60	.01	.13	0	4.62	4.62
Strong	2061	2205	.02	.16	0	5	5	1694.22	.02	.16	0	3.87	3.87
Stubborn	707	754	.01	.09	0	4	4	646.14	.01	.11	0	4.94	4.94
Stupid	1007	1159	.01	.13	0	4	4	977.16	.01	.13	0	4.57	4.57
Sweet	1465	1561	.02	.13	0	3	3	1304.69	.02	.15	0	4.21	4.21
Switched on	750	772	.01	.09	0	3	3	75.82	.01	.12	0	4.88	4.88
Sympathetic	1260	1376	.01	.13	0	4	4	1164.30	.01	.14	0	4.32	4.32
Tender	527	572	.01	.08	0	4	4	592.51	.01	.11	0	5.22	5.22
Timid	2835	3315	.04	.22	0	7	7	2112.89	.02	.17	0	3.51	3.51
Top	923	1056	.01	.13	0	10	10	839.72	.01	.12	0	4.68	4.68
Tranquil	998	1047	.01	.11	0	5	5	95.03	.01	.13	0	4.58	4.58
Trashy	3070	4080	.04	.29	0	15	15	2361.90	.03	.18	0	3.45	3.45
Unbearable	1575	1684	.02	.14	0	5	5	1283.09	.01	.14	0	4.11	4.11
Unique	10276	11364	.11	.36	0	6	6	5199.61	.06	.22	0	2.29	2.29
Unlucky	834	882	.01	.10	0	4	4	845.42	.01	.13	0	4.86	4.86
Useless	1247	1371	.01	.13	0	5	5	1124.17	.01	.14	0	4.36	4.36
Vacillating	620	688	.01	.10	0	6	6	674.97	.01	.12	0	5.06	5.06
Vagabond	2114	2317	.02	.17	0	7	7	1822.49	.02	.17	0	3.82	3.82
Vile	1189	1320	.01	.13	0	6	6	1099.12	.01	.14	0	4.39	4.39
Warrior	543	592	.01	.09	0	9	9	618.53	.01	.12	0	5.26	5.26
Weak	1355	1463	.01	.13	0	4	4	1231.04	.01	.15	0	4.30	4.30
Weary	803	830	.01	.09	0	3	3	801.40	.01	.12	0	4.84	4.84
Weird	2827	3103	.03	.20	0	10	10	2009.75	.02	.16	0	3.54	3.54
Wicked	5357	5892	.06	.27	0	5	5	3393.22	.04	.20	0	2.93	2.93
Worse	4851	5154	.05	.24	0	5	5	2964.04	.03	.18	0	3.01	3.01

Appendix 2

List of adjectives in English with their original form in Portuguese (in parenthesis), synonyms, antonyms, and/or other related words organized by personality factor

Legend:

- **Personality factors names:** Extra. (Extroversion or Surgency), Agree. (Agreeableness), Consc. (Conscientiousness), Emo. Stab. (Emotional Stability or Neuroticism), Intel. (Intellect or Openess), Psyc. (Psychoticism), Hum. (Humility-Modesty), Neg. Val. (Negative Valence), Pos. Val. (Positive Valence).
- **References of Brazilian Portuguese taxonomies:** And (Andrade, 2008); Hau (Hauck Filho et al., 2012); Hut (Hutz et al., 1998); Mac [Machado et al., 2014]; Nat [Natividade & Hutz, 2015]; Pas [Passos, 2014].
- **References of other taxonomies:** Ben (Benet-Martinez & Waller, 1997); Cze (Czech taxonomy [De Raad, 2000]); Eng (American English taxonomy [De Raad, 2000]); Eys (Eysenck's P-E-N model [Eysenck, Eysenck & Barrett, 1985]); Dut (Dutch taxonomy [De Raad, 2000]); Ger (German taxonomy [De Raad, 2000]); Gol (Goldberg & Rosolack, 1994); Hun (Hungarian taxonomy [De Raad, 2000]); Pol (Polish taxonomy [De Raad, 2000]); Rom (Italian-Rome taxonomy [De Raad, 2000]); Sau (Saucier, 2003); Tri (Italian-Trieste taxonomy [De Raad, 2000]).
- **Other observations:** * adjectives adapted from original items or facets names; ** *palavra-ônibus* (i.e., word with multiple meanings in Portuguese); *** pronoum or noum.

Aggressive (Agressivo): Extra. (aggressive [Eng], hot-blooded [Ger], fiery [Ger], unaggressive [Eng]); **Agree.** (affectionate [Eng], aggressive [Hun/Rom/Tri], argumentative [Cze], belligerent [Cze], choleric [Rom/Tri], conciliating [Rom], cordial [Dut/Tri], delicado [Hut], discutidor [And*], domineering [Cze/Dut/Ger/Rom/Tri], explosive [Hun], frio [And/Hut], genial [Dut], grosseiro [And], hotheaded [Dut], impetuous [Hun], indulgent [Dut], intolerant [Dut/Rom], intolerante [Pas], irascilbe [Tri], irritable [Rom/Tri], leninent [Pol], mild [Dut], peaceful [Dut/Hun/Rom/Tri], quarrelsome [Cze/Rom/Tri], rough [Cze], rude [And/Pas/Eng], ruthless [Dut/Pol], tempestuous [Hun], tolerant [Cze/Dut/Rom/Tri], tolerante [Pas], touchy [Rom], unaggressive [Cze]); **Emo. Stab.** (calmo [And/Pas], estressado [And*], explosive [Pol], fretful [Eng], gruff [Pol], impaciente [Pas], impetuous [Pol], irritable [Cze/Eng], nervoso [And/Pas], paciente [Pas], relaxado [And], ruthless [Rom], short-tempered [Ger], tenso [And], touchy [Cze/Eng/Ger/Gol], tranquilo [Nat/Pas]); **Intel.** (rough [Tri], rude [Tri]).

Alone (Sozinho): Extra. (detached [Gol], reservado [And], reserved [Eng/Dut/Ger/Gol/Tri], seclusive [Gol], social [Cze/Eng], sociable [Cze/Eng/Ger/Hun/Tri], sociável (And), solitary [Tri], unsociable [Gol], withdrawn [Gol]); **Agree** (sociável [Hut]); **Emo. Stab.** (reserverd [Cze], solitário [Hut]).

Amorous (Carinhoso): Agree. (affectionate [Eng], agradável [Hut], agreeable [Cze/Pol], amável [And/Hau/Hut/Mac/Nat], amigável [Nat/Pas], amoroso [Pas], antipático [Nat/Pas], charitable [Ger/Gol], cordial [Dut/Tri], genial [Dut], gentil [Hau/Hut/Mac/Pas], gentle [Hun/Tri], hearty [Pol], inconsiderate [Eng], kind [Eng/Gol], meek [Rom], mild [Dut], simpático [Hau/Hut/Mac/NatPas], sympathetic [Eng], warm [Eng], warm-hearted [Hun]); **Emo. Stab.** (antipático [Hut]); **Intel.** (sweet [Tri]).

Annoying (Chato): Extra. (dull [Hun], witty [Tri]); **Agree.** (agreeable [Cze/Pol], amável [And/Hut], amigável [Pas/Hau/Hut/Nat], antipático [Nat/ Pas], conciliating [Rom], cordial [Tri], friendly [Hun], genial [Dut], simpático [Hau/Hut/Mac/Nat/Pas], sympathetic [Eng], unsympathetic [Eng]); **Consc.** (playful [Ger], wishy-washy [Ger]); **Intel.** (dull [Cze/Pol], prosaico [Pas]).

Antisocial (Antissocial): Extra. (calado [And/Hut/Nat/Pas], detached [Gol], introverted [Eng/Dut/Rom/Tri], introvertido [Hut], quiet [Cze/Eng/Gol/Pol], quieto [And/Hau/Hut/Mac], reservado [And], reserved [Eng/Dut/Ger/Gol/Tri], seclusive [Gol], secretive [Cze/Gol], silent [Cze/Dut/Eng/Ger/Gol/Hun/Tri], social [Cze/Eng], sociable [Cze/Eng/Ger/Hun/Tri], sociável (And), somber [Dut/Gol], solitary [Tri], taciturn [Hun/Rom], timid [Dut/Eng/Ger/Gol/Pol/Tri], tímido [And/Hau/Hut/Mac/Nat/Pas], untalkative [Cze/Eng/Ger/Gol/Hun], unsociable [Gol], withdrawn

[Gol]); **Agree** (agreeable [Cze/Pol], amável [And/Hau/Hut/Mac/Nat], antipático [Nat/ Pas], considerate [Eng/Ger/Gol], compassionate [Pol], inconsiderate [Eng], simpático [Hau/Hut/Mac/Nat/Pas], sympathetic [Eng], uncharitable [Eng], unsympathetic [Eng]); **Emo. Stab.** (antipático [Hut], reserverd [Cze]). **Intel.** (secretive [Hun]).

Anxious (Ansioso): Extra. (enthusiastic [Dut], fearful [Pol], quiet [Cze/Eng/Gol/Pol]); **Agree.** (cold [Eng], frio [And/Hut], calm [Rom/Tri], patient [Rom/Tri]); **Emo. Stab.** (ansioso [Hau/Hut/Mac/Nat/Pas], anxious [Cze/Dut/Hun/Tri], assured [Cze/Dut/Tri], calm [Dut/Hun], calmo [And/Pas], cold [Rom], confident [Cze], excitable [Cze/Pol], fearful [Hun/Tri], impaciente [Pas], imperturbable [Dut/Eng/Ger/Rom], insecure [Tri], inseguro [Hau/Hut/Mac], nervoso [And/Pas], nervous [Cze/Dut/Hun/Pol], paciente [Pas], panicky [Dut], patient [Eng/Pol], peaceful [Dut/Hun/Rom/Tri], preocupado [And], relaxado [And], relaxed [Eng], restless [Cze], self-assured [Hun], tenso [And], tranquil [Cze], tranquilo [Nat/Pas], tenso [And], uncertain [Dut], unexcitable [Eng/Pol], worrying [Hun]); **Intel.** (philosophical [Eng/Dut]).

Arrogant (Arrogante): Agree. (arrogant [Dut], conceited [Pol], egoistic [Ger], egoistical [Ger/Pol], humble [Gol], modest [Gol], self-opinionated [Ger], smug [Gol], unassuming [Gol]); **Consc.** (modesto [Pas]); **Emo. Stab.** (self-doubting [Ger]); **Intel.** (conceited [Hun], overbearing [Hun], pretending [Hun], swollen-headed [Hun]).

Asshole (Cuzão): Extra. (cowardly [Pol], fearful [Pol]); **Agree.** (inconsiderate [Eng], mercenary [Pol], moral [Cze/Gol]); **Consc.** (conscientious [Cze/Dut/Ger/Hun/Pol], dependable [Eng], honesto [Hut], honrado [Hut], immoral [Dut/Hun], inconsiderate [Hun], lax [Dut], scrupulous [Pol], trustful [Eng], unconscientious [Cze]); **Emo. Stab.** (assured [Cze/Dut/Tri], confident [Cze], dishonest [Rom], fearful [Hun/Tri], insecure [Tri], inseguro [Hau/Hut/Mac], self-assured [Hun]); **Intel.** (disloyal [Tri], perfidious [Tri], reliable [Hun/Tri], truthful [Hun]); **Neg. Val.** (filthy [Ben], idiotic [Ben]).

Bad (Mau): see good (bom).

Beloved (Amado): Agree. (agreeable [Cze/Pol], amável [And/Hau/Hut/Mac/Nat], antipático [Nat/ Pas], genial [Dut], simpático [Hau/Hut/Mac/Nat/Pas], sympathetic [Eng], unsympathetic [Eng]); **Emo. Stab.** (antipático [Hut]).

Bipolar: see anxious (ansioso).

Blind (Cego): Consc. (irrational [Rom], rational [Rom]); **Emo. Stab.** (astute [Tri], bright [Cze], brilliant [Cze], crafty [Hun], cunning [Tri], gullible [Gol], impressionable [Tri], naïve [Gol], rational [Pol], suggestible [Gol/Tri], uncritical [Eng], wily [Hun]); **Intel.** (acute [Dut], clever [Cze/Ger], critical [Dut], dull [Pol], highly intelligent [Ger], imperceptive [Eng], intelligent [rel -Cze/ Eng/Ger], retarded [Ger], slow-witted [Pol], stupid [Ger], uncritical [Dut], unintelligent [Cze/Eng/Ger], unperspicacious [Pol], unreflective [Eng]).

Blockhead (Besta): see fool (bobo).

Brat (Moleque): Extra. (bold [Cze/Eng/Pol], brisk [Pol], cheerful [Dut/Rom/Tri], dinâmico [Pas], dull [Hun], dynamic [Ger/Rom], dynamical [Pol], enterprising [Cze/Pol], extroverted [Eng/Rom/Rom/Tri], extrovertido [And], exuberant [Dut], full of life [Hun], hyperactive [Hun], impulsive [Ger], introverted [Dut/Eng/Hun/Ro/Trim], jovial [Dut], laughing [Hun], lively [Ger/Rom], merry [Dut], vivacious [Dut/Hun/Ger/Pol], witty [Tri]); **Agree** (moral [Cze/Gol]); **Consc.** (confiável [And], conscientious [Cze/Dut/Pol], disciplined [Hun/Tri], frivolous [Dut], hard-working [Ger], immature [Hun], inconstante [Pas], indisciplinado [Nat], indolent [Dut], industrious [Cze/Dut/Ger/Rom], irresponsável [Mac/Nat/Hau/Hut/Pas], irresponsible [Dut], lax [sHun], rash [Pol], responsável [Pas], responsible [Eng], scatterbrained [Dut], sloppy [Eng], unconscientious [Cze], unruly [Rom/Tri], workshy [Ger]); **Intel.** (audacioso [Hau/Mac], aventureiro [Hau/Hut/Mac/Pas], curioso [And/Nat/Pas], engraçado [Hau], impulsive [Cze]).

Buffoon (Palhaço):** see funny (engraçado) and idiot (idiota).

Calm (Calmo): see anxious (ansioso).

Cardiac (Cardíaco): Agree. (frio [And/Hut], calm [Rom/Tri], insensitive [Eng], patient [Rom/Tri]); **Emo. Stab.** (ansioso [Nat/Mac/Hau/Hut/Pas], anxious [Cze/Dut/Hun/Tri], calm [Dut/Hun], calmo [And/Pas], excitable [Pol], impaciente [Pas], impressionable [Tri], insensitive [Ger/Rom], nervoso [And/Pas], oversensitive [Hun], paciente [Pas], patient [Eng/Pol], preocupado [And], relaxado [And], sensitive [Dut/Eng/Ger/Rom], tenso [And], tranquilo [Nat/Pas], unexcitable [Pol]); **Intel.** (philosophical [Dut/Eng]).

Chump (Otário): see fool (bobo).

Clever (Esperto): see intelligent (inteligente).

Clueless (Sem-noção): Extra. (helpless [Pol]); **Agree.** (helpful [Eng/Ger/Pol]); **Consc.** (careful [Dut/Hun], careless [Eng/Tri], chaotic [Pol], consistent [Cze/Rom/Pol], cuidadoso [Hut/Mac], descuidado [And], neglectful [Hun], negligent [Eng/Pol], sloppy [Eng]); see confused (confuso) and lost (perdido); **Neg. Val.** (idiotic [Ben]).

Clumsy (Desastrado): Consc. (careful [Dut/Hun], careless [Eng/Tri], cuidadoso [Hut/Mac], descuidado [And], neglectful [Hun], negligent [Eng/Pol], sloppy [Eng]).

Cocky (Metido): see arrogant (arrogante).

Cold (Frio): Extra. (reservado [Pas], reserved [Eng/Dut/Ger/Tri]); **Agree.** (agreeable [Cze/Pol], amável [And/Hau/Hut/Mac/Nat], antipático [Nat/ Pas], callous [Dut], calm [Rom/Tri], cold [Eng/Gol], distante [And], frio [And/Hut], genial [Dut], hard [Eng], hearty [Pol], impersonal [Gol], insensitive [Cze/Gol], patient [Rom/Tri], unsympathetic [Eng], warm [Eng], warm-hearted [Hun]); **Consc.** (nonchalant [Dut]); **Emo. Stab.** (antipático [Hut], calm [Dut/Hun], calmo [And/Pas], cold [Rom], emotional [Dut/Eng/Ger/Gol/Rom], emotivo [And], excitable [Cze/Pol], hard-boiled [Ger], imperturbable [Dut/Eng/Ger/Rom], indiferente [Pas], indifferent [Rom], insensitive [Ger], nervoso [And/Pas], nervous [Cze/Dut/Hun/Pol], oversensitive [Hun], paciente [Pas], panicky [Dut], passionate [Cze], patient [Eng/Pol], peaceful [Dut/Hun/Rom/Tri], passionate [Cze], reserved [Pol], romantic [Rom], sentimental [And/Ger/Rom], sensitive [Dut/Ger/Rom], thick-skinned [Ger], unemotional [Pol], unexcitable [Eng/Pol]); **Intel.** (insensitive [Tri], romantic [Tri], sensitive [Tri], sentimental [Tri]).

Committed (Comprometido): see faithful (fiel) and responsible (responsável).

Complicated (Complicado): Extra. (enigmatic [Tri], inscrutable [Dut]); **Agree** (compreensível [Hut], hard [Eng]); **Consc.** (chaotic [Pol], consistent, [Cze/Rom/Pol], extravagant [Ger], inconsistent [Cze/Eng/Rom/Tri]); **Emo. Stab.** (erratic [Pol], unbalanced [Dut], uncertain [Dut], unstable [Cze/Dut]); **Intel.** (complex [Eng], deep [Dut/Eng], extravagant [Rom], inacessível [Pas], shalow [Dut/Eng], simple [Cze/Eng], sofisticado [And], unsophisticated [Cze/Eng]).

Confused (Confuso): Extra. (enigmatic [Tri], inscrutable [Dut]); **Agree** (compreensível [Hut]); **Consc.** (chaotic [Pol], consistent, [Cze/Rom/Pol], discontinuous [Tri], desorganizado [And/Hut/Nat/Pas], disorganized [Cze/Eng/Tri], disorderly [Eng/Tri], extravagant [Ger], fickle [Ger], haphazard [Eng], inconsistent [Cze/Eng/Rom/Tri], imprecise [Rom], strong-minded [Ger], unstable [Cze/Ger], unsystematic [Eng]); **Emo. Stab.** (erratic [Pol], unbalanced [Dut], uncertain [Dut], unstable [Cze/Dut]); **Intel.** (complex [Eng], deep [Dut/Eng], extravagant [Rom], inacessível [Pas], shalow [Dut/Eng], simple [Cze/Eng], sofisticado [And], unsophisticated [Cze/Eng]).

Crazy (Maluco): see insane (doido).

Crybaby (Chorão): see sensitive (sensível) and spoiled (mimado).

Curious (Curioso): Intel. (curioso [Ant/Hau/Hut/Nat/Pas], desinteressado [Pas], inquieto [Pas], uninquisitive [Eng]).

Cute (Fofinho): see sweet (doce).

Damned (Danado): Extra. (active [Gol/Pol/Rom], ativo [Pas], bold [Cze/Eng/Pol], brisk [Pol], cheerful [Dut/Rom/Tri], dinâmico [Pas], dull [Hun], dynamic [Ger/Rom], dynamical [Pol], energetic [Cze/Gol/Pol/Rom], enterprising [Cze/Pol], extroverted [Eng/Rom/Rom/Tri], extrovertido [And], hyperactive [Hun], introverted [Dut/Eng/Hun/Ro/Trim], quiet [Cze/Eng/Gol/Pol], quieto [And/Hau/Hut/Mac], resourceful [Pol], witty [Tri]); **Agree.** (impetuous [Hun], patient [Rom/Tri]); **Consc.** (prompt [Dut]); **Emo. Stab.** (alert [Cze], astute [Tri], bold [Hun], bright [Cze], brilliant [Cze], calmo calmo [And/Pas], courageous [Cze/Tri], crafty [Hun], cunning [Tri], impaciente [Pas], patient [Eng/Pol], paciente [Pas], tranquilo [Nat/Pas], wily [Hun]); **Intel.** (acute [Dut], audacioso [Hau/Mac], aventureiro [Hau/Hut/Mac/Pas], bright [Pol], clever [Cze/Ger], curioso [And/Nat/Pas], dull [Cze/Pol], engenhoso And], highly intelligent [Ger], imperceptive [Eng], ingenious [Ger], inquieto [And], intelligent [Cze/Eng/Ger], inventivo [And], retarded [Ger], silly [Pol], simple [Eng], slow-witted [Pol], stupid [Ger], unintelligent [Cze/Eng/Pol], unperspicacious [Pol]).

Dangerous (Perigoso): see aggressive (agressivo) and good (bom).

Dead (Morto): Extra. (brisk [Pol], cheerful [Dut/Gol/ Rom/Tri], energetic [Cze/Pol], full of life [Hun], hyperactive [Hun], grey [Hun], jovial [Dut], lively [Ger/Rom], melancholic [Gol/Rom/Tri], merry [Dut/Gol], sparkling [Rom], somber [Dut/Gol], vivacious [Dut/Ger/Hun/Pol/Tri]); **Emo. Stab.** (alegre [Pas]).

Dear (Querido): see beloved (amado).

Deluded (Iludido): Consc. (impractical [Eng]); **Emo. Stab.** (down-to-earth [Dut], gullible [Gol], naïve [Gol], impressionable [Tri], realistic [Dut], suggestible [Gol/Tri]); **Intel.** (imperceptive [Eng], unperspicacious [Pol], unreflective [Eng]).

Demon (Demônio*):** see good (bom) and brat (moleque).

Different (Diferente): Extra. (enigmatic [Tri]); **Consc.** (extravagant [Ger]); **Intel.** (complex [Eng], deep [Dut/Eng], extravagant [Rom], shalow [Dut/Eng], simple [Cze/Eng], sofisticado [And], unsophisticated [Cze/Eng]); **Neg. Val.** (weird [Sau]).

Difficult (Difícil): Extra. (free and easy [Rom]); **Agree.** (adaptable [Hun], agreeable [Cze/Pol], amável [And/Hut], amigável [Pas/Hau/Hut/Nat], antipático [Nat/ Pas], conciliating [Rom], cordial [Tri], dócil [Hut], easygoing [Gol], friendly [Hun], genial [Dut], good-natured [Ger/Tri], hard [Eng], polemical [Tri], simpático [Hau/Hut/Mac/Nat/Pas], sympathetic [Eng], unsympathetic [Eng]); **Emo. Stab.** (antipático [Hut]); **Intel.** (complex[Eng], docile [Dut], servile [Dut], simple [Cze/Eng]).

Direct (Direto): Extra. (assertivo [And], candid [Dut], frank [Ger], direct [Hun]); **Agree.** (fair [Cze], frio [And/Hut], insincere [Eng]); **Consc.** (honesto [Hut], honrado [Hut]); **Emo. Stab.** (afirmativo [Hut], dishonest [Rom], insincere [Rom]); **Intel.** (autêntico [Pas], loyal [Tri], fingido [Pas], hypocritical [Hun], insincere [Tri], truthful [Hun]).

Disgraced (Desgraçado): see happy (feliz), asshole (cuzão), and son of a bitch (filho da puta).

Disgusting (Nojento): see vile (escroto).

Douchebag (Babaca): see fool (bobo) and ignorant (ignorante).

Dramatic (Dramático): Extra. (contido [Pas], dull [Hun], retraído [Pas], spontaneous [Dut], unrestrained [Eng], unspontaneous [Ger]); **Agree** (apaixonado [Hut], frio [And/Hut], romântico [Hut], sentimental [Gol/Hut]); **Consc.** (extravagant [Ger], wishy-wahsy [Ger]); **Emo. Stab.** (emotional [Dut/Eng/Ger/Gol/Rom], emotivo [And], excitable [Cze/Pol], impressionable [Tri], passionate [Cze], romantic [Rom], sentimental [And/Ger/Rom], unemotional [Pol], unexcitable [Eng/Pol], unselfconscious [Eng]); **Intel.** (autêntico [Pas], conservative [Dut/Rom], extravagant [Hun], fingido [Pas], natural [Hun], pretending [Hun], romantic [Tri], sentimental [Tri], simple [Eng], theatrical [Hun]).

Drunk (Bêbado): see brat (moleque) and confused (confuso).

Dung (Bosta): see nobody (ninguém) and useless (inútil).

Easy (Fácil): see difficult (difícil).

Entangled (Enrolado): see confused (confuso) and responsible (responsável).

Evil (Mal*):** see good (bom).

Exaggerated (Exagerado): Agree. (reasonable [Rom]); **Consc.** (extravagant [Ger]); **Emo. Stab.** (down-to-earth [Dut], realistic [Dut]); **Intel.** (autêntico [Pas], convencional [Nat], prosaico [Pas], conventional [Eng/Rom], extravagant [Rom], natural [Hun], simple [Eng], theatrical [Hun], unconventional [Dut/Gol]).

Faithful (Fiel): Agree. (honest (Gol), insincere (Eng), moral [Cze/Gol], sincere [Gol]); **Consc.** (confiável [And], conscientious [Cze/Dut/Ger/Pol/Rom], dependable [Eng], honesto [Hut], honrado [Hut], immoral [Dut], lax [Hun], scrupulous [Pol], unconscientious [Cze]); **Emo. Stab.** (dishonest [Rom], insincere [Rom]); **Intel.** (disloyal [Tri], fingido [Pas], insincere [Tri], loyal [Tri], perfidious [Tri], reliable [Hun/Tri], truthful [Hun]).

False (Falso): Extra. (candid [Dut], direct [Hun], frank [Ger], secretive [Cze/Gol]); **Agree.** (insincere [Eng], moral [Cze/Gol]); **Consc.** (confiável [And], conscientious [Cze/Dut/Ger/Pol/Rom], dependable [Eng], immoral [Dut], lax [Hun], scrupulous [Pol], unconscientious [Cze]); **Emo. Stab.** (insincere [Rom], slippery [Ger]); **Intel.** (autêntico [Pas], disloyal [Tri], fingido [Pas], hypocritical

[Hun], insincere [Tri], just [Hun], loyal [Tri], perfidious [Tri], pretending [Hun], reliable [Hun/Tri], secretive [Hun], truthful [Hun]).

Famous (Famoso): see special (especial) and marvelous (maravilhoso).

Fearful (Medroso): Extra. (cowardly [Pol], fearful [Pol], passive [Cze/Pol/Rom]); **Agree.** (); **Consc.** (indecisive [Cze], indeciso [Pas], foolhardy [Rom], reckless [Dut/Ger/Pol]); **Emo. Stab.** (assured [Cze/Dut/Tri], confident [Cze], courageous [Tri], fearful [Hun/Tri], imperturbable [Eng/Dut/Rom], insecure [Tri], inseguro [Hau/Hut/Mac], nerves of steel [Hun], panicky [Dut], self-assured [Hun], weak [Tri], worrying [Hun]); **Intel.** (audacioso [Hau/Mac], aventureiro [Hau/Hut/Mac], corajoso [Hau]); see anxious (ansioso) and timid (tímido).

Ferocious (Bravo): see aggressive (agressivo).

Fool (Bobo): Extra. (dull [Hun]); **Consc.** (frivolous [Dut/Ger/Hun], irrational [Rom], rational [Rom], scatterbrained [Dut/Ger], thoughtless [Cze/Dut/Rom], wishy-washy [Ger]); **Emo. Stab.** (astute [Tri], bright [Cze], brilliant (Cze), crafty [Hun], cunning [Tri], gullible [Gol], naïve [Gol], impressionable [Tri], rational [Pol], suggestible [Gol/Tri], wily [Hun]); **Intel.** (acute [Dut], bright [Pol], clever [Cze/Ger], dull [Cze/Pol], highly intelligent [Ger], ingenious [Ger], intelligent [Cze/Eng/Ger], retarded [Ger], slow-witted [Pol], silly [Pol], simple [Eng], sofisticado [And], stupid [Ger], thoughtful [Tri], unperspicacious [Pol], unintelligent [Cze/Eng/Pol], unsophisticated [Cze/Eng]); **Neg. Val.** (idiotic [Ben], stupid [Sau]).

Foolish (Trouxa): see fool (bobo).

Footloose (Leve): see anxious (ansioso).

Forgotten (Esquecido): see confused (confuso) and alone (sozinho).

Free (Livre): Extra. (contido [Pas], free and easy [Rom], inibido [And/Hut/Mac], retraído [Pas], spontaneous [Dut], unrestrained [Eng], unspontaneous [Ger]); **Consc.** (disobedient [Tri], dutiful [Pol], unruly [Rom/Tri]); **Emo. Stab.** (dependent [Dut], independent [Dut], obedient [Dut/Rom]); **Intel.** (autonomous [Gol], docile [Dut], independent [Gol], individualistic [Gol], natural [Hun], narrow-minded [Dut], nonconformistic [Rom], original [Dut/Rom], rebellious [Rom], revolutionary [Rom], servile [Dut/Rom]).

Friend (Amigo): Extra. (agreeable [Cze/Pol], amável [And/Hau/Hut/Mac/Nat], genial [Dut], social [Cze/Eng], sociable [Cze/Eng/Ger/Hun/Tri], sociável (And); **Agree** (agreeable [Cze/Pol], amável [And/Hau/Hut/Mac/Nat], considerate [Eng/Ger/Gol], compassionate [Pol], simpático [Hau/Hut/Mac/Nat/Pas], sympathetic [Eng]); see good (bom) and nice (giro).

Funny (Engraçado): Extra. (brisk [Pol], cheerful [Dut/Rom/Tri], comunicativo [And/Pas/Hau/Hut/Mac/Nat], dinâmico [Pas], dull [Hun], dynamic [Ger/Rom], dynamical [Pol], enterprising [Cze/Pol], extroverted [Eng/Rom/Rom/Tri], extrovertido [And], exuberant [Dut], full of life [Hun], introverted [Dut/Eng/Hun/Ro/Trim], jovial [Dut], laughing [Hun], lively [Ger/Rom], merry [Dut], vivacious [Dut/Hun/Ger/Pol], witty [Tri]); **Agree** (agreeable [Cze/Pol], amável [And/Hau/Hut/Mac/Nat]); **Consc.** (extravagant [Hun], playful [Hun]); **Intel.** (engraçado [Hau], extravagant [Rom], ironical [Rom]).

Genial (Legal):** see nice (giro).

Genius (Gênio): see intelligent (inteligente).

Good (Bom): Agree. (agreeable [Cze/Pol], amável [And/Hau/Hut/Mac/Nat], benevolent [Gol], bondoso [Hau/Hut/Mac], charitable [Ger/Gol], fair [Cze], considerate [Eng/Ger/Gol], compassionate [Pol], dishonet [Gol], ethical [Gol], genial [Dut], gentil [Hau/Hut/Mac/Pas], gentle [Hun/Tri], good-hearted [Dut/Ger], good-natured [Ger/Tri], humane [Ger], humanitarian [Hun], inconsiderate [Eng], kind [Eng/Gol], kind-hearted [Dut/Hun], moral [Cze/Gol], soft-hearted [Cze/Eng], uncharitable [Eng], unkind [Eng], unscrupulous [Gol], warm-hearted [Ger]); **Consc.** (bright [Cze], brillian [Cze], conscientious [Cze/Dut/Ger/Pol/Rom], immoral [Dut], unconscientious [Cze], scrupulous [Cze]); **Emo. Stab.** (humane [Rom]); **Intel.** (gifted [Ger/Pol], humane [Tri], humanitariam [Tri], talented [Ger/Pol], untalented [Cze/Ger]); **Neg. Val.** (cruel [Ben], evil [Sau], filthy [Ben], horrible [Ben], vandalic [Ben]).

Good-natured (Bonzinho, diminutive of bom [good]): see bom (good), doce (sweet), and inocente (innocent).

Gothic (Gótico): Extra. (bashful [Cze/Dut/Eng/Ger/Gol/Rom/Pol], calado [And/Hut/Nat/Pas], depressive [Rom], detached [Gol], grey [Hun], introverted [Eng/Dut/Rom/Tri], introvertido [Hut], joyless [Gol], lethargic [Gol], melancholic [Gol/Rom/Tri], negativistic [Gol], pessimistic [Gol], quiet [Cze/Eng/Gol/Pol], quieto [And/Hau/Hut/Mac], reservado [And], reserved [Eng/Dut/Ger/Gol/Tri], seclusive [Gol], secretive [Cze/Gol], shy [Cze/Eng/Dut/Ger/Gol/Rom/Pol], silent [Cze/Dut/Eng/Ger/Gol/Hun/Tri], somber [Dut/Gol], taciturn [Hun/Rom], timid [Dut/Eng/Ger/Gol/Pol/Tri], tímido [And/Hau/Hut/Mac/Nat/Pas], untalkative [Cze/Eng/Ger/Gol/Hun], unsociable [Gol], vivacious [Dut/Ger/Hun/Pol/Tri], withdrawn [Gol]); **Agree** (romântico [Hut], sentimental [Gol/Hut]); **Emo. Stab.** (depressivo [And/Pas], deprimido [Hut], emotional [Dut/Eng/Ger/Gol/Rom], emotivo [And], infeliz [Hut], reserverd [Cze], romantic [Rom], sentimental [And/Ger/Rom], triste [And/Hut/Mac/Pas]). **Intel.** (romantic [Tri], secretive [Hun], sentimental [Tri], unconventional [Dut/Gol]).

Graceless (Sem-graça): see ridiculous (ridículo) and timid (tímido).

Gracious (Fofo): see sweet (doce).

Grateful (Grato): Agree. (rude [And/Eng/Pas]); **Consc.** (thoughtless [Cze/Dut/Rom]); **Intel.** (rude [Tri], ungrateful [Tri]).

Great (Ótimo): see bom (good), especial (special), and marvelous (maravilhoso).

Guilty (Culpado): Emo. Stab. (guilt feelings [Eys]).

Happy (Feliz): Extra. (cheerful [Dut/Gol/ Rom/Tri], depressive [Rom], full of life [Hun], grey [Hun], laughing [Hun], lively [Ger/Rom], melancholic [Gol/Rom/Tri], merry [Dut/Gol], somber [Dut/Gol], vivacious [Dut/Ger/Hun/Pol/Tri]); **Emo. Stab.** (alegre [Pas], depressivo [And/Pas], deprimido [Hut], feliz [Hut], infeliz [Hut], triste [And/Hut/Mac/Pas]).

Hard (Duro): Extra. (intrasigente [Pas], stiff [Dut]); **Agree.** (accommodating [Dut], callous [Dut], compreensível [Hut], compreensivo [Mac], flexible [Dut], frio [And], hard [Eng], harsh [Eng], indulgent [Dut], intolerant [Dut/Rom/Tri], intolerante [Pas], lenient [Ger], ruthless [Ger], soft-hearted [Cze/Eng], tolerant [Cze/Dut/Rom/Tri], tolerante [Pas]); **Consc.** (firm [Ger], steady [Cze/Rom/Tri]); **Emo. Stab.** (afirmativo [Hut], hard-boiled [Ger], ruthless [Rom], solid [Ger], steady [Dut/Ger], thick-skinned [Ger]); **Intel.** (flexível [Pas], rígido [Pas]).

Horrible (Horrível): see vile (escroto) and good (bom).

Humane (Humano): see good (bom).

Idiot (Idiota): see fool (bobo) and ignorante (ignorant).

Idle (À toa): see lazy (preguiçoso).

Ignorant (Ignorante): Extra. (aggressive [Eng], fiery [Ger], unaggressive [Eng]); **Agree.** (affectionate [Eng], aggressive [Hun/Rom/Tri], argumentative [Cze], belligerent [Cze], choleric [Rom/Tri], conciliating [Rom], cordial [Dut/Tri], delicado [Hut], discutidor [And*], domineering [Cze/Dut/Ger/Rom/Tri], explosive [Hun], frio [And/Hut], genial [Dut], grosseiro [And], hostil [Pas], hotheaded [Dut], impetuous [Hun], indulgent [Dut], intolerant [Dut/Rom], intolerante [And], irascilbe [Tri], irritable [Rom/Tri], leninent [Pol], mild [Dut], peaceful [Dut/Hun/Rom/Tri], quarrelsome [Cze/Rom/Tri], rough [Cze], rude [And/Pas/Eng], ruthless [Dut/Pol], tempestuous [Hun], tolerant [Cze/Dut/Rom/Tri], tolerante [And], touchy [Rom], unaggressive [Cze]); **Consc.** (frivolous [Dut/Ger/Hun], inconsiderate [Hun], irrational [Rom], rational [Rom], scatterbrained [Dut/Ger], thoughtless [Cze/Dut/Rom]); **Emo. Stab.** (astute [Tri], bright [Cze], brilliant [Cze], calm [Dut/Hun], calmo [And/Pas], crafty [Hun], cunning [Tri], fretful [Eng], gruff [Pol], impaciente [Pas], impetuous [Pol], irritable [Cze/Eng], nervoso [And/Pas], nervous [Cze/Dut/Hun/Pol], paciente [Pas], patient [Eng/Pol], rational [Pol], ruthless [Rom], short-tempered [Ger], touchy [Cze/Eng/Ger/Gol], wily [Hun]); **Intel.** (cultured [Cze], deep [Dut/Eng], dense [Pol], ignorant [Ger], intellectual [Eng], knowledgeable [Cze/Ger/Pol], reflexivo [And*], rough [Tri], rude [Tri], shallow [Dut/Eng], simple [Eng], sofisticado [And], uninformed [Ger], unintellectual [Eng], unperspicacious [Pol], unreflective [Eng], unsophisticated [Cze]).

Ignored (Ignorado): see alone (sozinho).

Impolite (Mal-educado): see educado (polite)

Important (Importante): see especial (special) and marvelous (maravilhoso).

Impossible (Impossível): see complicado (complicated).

In love (Apaixonado): Extra. (); **Agree** (apaixonado [Hut], romântico [Hut], sentimental [Gol/Hut]); **Consc.** (); **Emo. Stab.** (emotional [Dut/Eng/Ger/Gol/Rom], emotivo [And], excitable [Cze/Pol], passionate [Cze], romantic [Rom], sensitive [Dut/Ger/Rom], sentimental [And/Ger/Rom]); **Intel.** (romantic [Tri], sentimental [Tri], sensitive [Tri]).

Incredible (Incrível): see especial (special) and marvelous (maravilhoso).

Indecisive (Indeciso): Extra. (candid [Dut], enigmatic [Tri], inscrutable [Dut]); **Agree** (compreensível [Hut]); **Consc.** (accurate [Dut], balanced [Rom], chaotic [Pol], compenetrado [Hut], consistent [Cze/Pol/Rom], decidido [Pas], deliberate [Gol/Hun], discontinuous [Tri], fickel [Ger], haphazard [Eng], imprecise [Rom], inconsistent [Eng/Rom/Tri], indecisive [Cze], indeciso [Pas], innacurate [Pol/Tri], precise [Dut/Eng/Hun/Pol/Rom], steady [Pol/Rom/Tri], strong-minded [Ger], unstable [Cze/Ger/Rom], wishy-wahsy [Ger]) ; **Emo. Stab.** (ansioso [Hau/Hut/Mac/Nat/Pas], anxious [Cze/Dut/Hun/Tri], assured [Cze/Dut/Tri], confident [Cze], decisive [Dut/Tri], indecisive [Tri], inseguro [Hau/Hut/Mac], preocupado [And], resolute [Dut/Tri], self-assured [Hun], solid [Ger], stable [Dut], steady [Dut/Ger], unbalanced [Dut], uncertain [Dut], unstable [Cze/Dut], weak [Tri], well-balanced [Hun]); **Intel.** (hesitante [Pas]).

Innocent (Inocente): Extra. (candid [Dut]); **Agree.** (meek [Rom], mild [Dut]) **Emo. Stab.** (gullible [Gol], naïve [Gol], impressionable [Tri], suggestible [Gol/Tri]); **Intel.** (meek [Hun], puritan [Rom], silly [Pol], simple [Eng], unperspicacious [Pol]); see also culpado (guilty).

Insane (Doido): Extra. (fiery [Ger], hot-blooded [Ger]); **Agree** (impulsivo [Pas], prudente [Pas], reasonable [Rom]); **Consc.** (balanced [Rom], chaotic [Pol], consequent [Ger], disciplined [Hun/Tri], extravagant [Ger], foolhardy [Rom], indisciplinado [Nat], irrational [Rom], judicious [Tri], rational [Rom], steady [Pol/Rom/Tri], unstable [Cze/Ger/Rom]); **Emo. Stab.** (down-to-earth [Dut], equilibrado [Pas], erratic [Pol], estável [And/Nat/Pas], instável [Pas], oscilante [Pas], poised [Cze/Ger], rational [Pol], realistic [Dut], solid [Ger], stable [Dut], steady [Dut/Ger], unbalanced [Dut], unstable [Cze/Dut], well-balanced [Hun]); **Intel.** (aventureiro [Hau/Hut/Mac/Pas], concencional [Nat], extravagant [Hun], prosaico [Pas]); **Neg. Val.** (crazy [Sau], insane [Sau]).

Insecure (Inseguro): see fearful (medroso) and anxious (ansioso).

Intelligent (Inteligente): Extra. (dinâmico [Pas], dynamic [Ger/Rom], dynamical [Pol], dull [Hun], resourceful [Pol]); **Consc.** (estudioso [Hut], irrational [Rom], rational [Rom], scatterbrained [Dut/Ger]); **Emo. Stab.** (astute [Tri], bright [Cze], brilliant (Cze), crafty [Hun], cunning [Tri], rational [Pol], wily [Hun]); **Intel.** (acute [Dut], bright [Pol], clever [Cze/Ger], dull [Cze/Pol], engenhoso [And], highly intelligent [Ger], ingenious [Ger], intelligent [Cze/Eng/Ger], unintelligent [Cze/Eng/Pol], retarded [Ger], slow-witted [Pol], silly [Pol], simple [Eng], stupid [Ger], thoughtful [Tri], unperspicacious [Pol]).

Interesting (Interessante): see ridiculous (ridículo).

Jealous (Ciumento): Extra. (fearful [Pol]); **Agree.** (demanding [Dut], domineering [Cze/Dut/Ger/Rom/Tri], intolerant [Dut/Rom], tolerant [Cze/Dut/Rom/Tri], trustful [Eng], understanding [Eng/Rom/Tri]); **Emo. Stab.** (ansioso [Hau/Hut/Mac/Nat/Pas], anxious [Cze/Dut/ Hun/Tri], assured [Cze/Dut/Tri], confident [Cze], envious [Eng], fearful [Hun/Tri], insecure [Tri], inseguro [Hau/Hut/Mac], jealous [Eng], preocupado [And], self-assured [Hun], undemanding [Eng], unenvious [Eng], worrying [Hun]); **Intel.** (understanding [Cze]).

Joke (Piada): see ridiculous (ridículo) and funny (engraçado).

Lady (Dama*):** see polite (educado) and nice (giro).

Lazy (Preguiçoso): Extra. (active [Rom/Pol], ativo [Pas], brisk [Pol], cheerful [Dut/Rom/Tri], dinâmico [Pas], dynamic [Ger/Rom], dynamical [Pol], energetic [Cze/Gol/Pol/Rom], enérgico [And*], enterprising [Cze/Pol], exuberant [Dut/Tri], full of life [Hun], hyperactive [Hun], lively [Ger/ Rom], vivacious [Ger]); **Agree.** (willing [Ger]); **Consc.** (careful [Dut/Hun], careless [Eng/Tri], cuidadoso [Hut/Mac], descuidado [And] dedicado [Hau/Hut/Mac], diligent [Dut/Ger/Hun/Tri], esforçado [Hau/Hut/Mac/Nat], hard-working [Ger], indolent [Dut], industrious [Cze/Dut/Ger/Rom], lax [Hun], lazy [Cze/Dut/Hun], neglectful [Hun], negligent [Eng/Pol], painstaking [Dut], preguiçoso [And], prompt [Dut], sloppy [Eng], workshy [Ger]); **Intel.** (aventureiro [Hau/Hut/Mac/Pas], enérgico [Hau], apático [Pas], dull [Cze/Pol]).

Loco (Louco): see insane (doido).

Lost (Perdido): Extra. (helpless [Pol]); **Agree.** (helpful [Eng/Ger/Pol]); **Consc.** (chaotic [Pol], consistent, [Cze/Rom/Pol], desmotivado [Pas], discontinuous [Tri], fickle [Ger], frivolous [Dut/Ger/Hun], haphazard [Eng], incoherent [Tri], inconsistent [Cze/Eng/Rom/Tri], indecisive [Cze], indeciso [Pas], motivado [Pas], obstinado [Pas], persistente [And*/Pas], purposeful [Cze/Ger], steady [Pol/Rom/Tri], strong-minded [Ger], unstable [Cze/Ger/Rom], wishy-wasy [Ger]); **Emo. Stab.** (decisive [Dut/Tri], equilibrado [Pas], estável [And/Nat/Pas], indecisive [Tri], instável [Pas], oscilante [Pas], resolute [Dut/Tri], solid [Ger], stable [Dut], steady [Dut/Ger], unbalanced [Dut], uncertain [Dut], unstable [Cze/Dut], well-balanced [Hun]).

Lousy (Péssimo): see good (bom).

Lucky (Sortudo): Extra. (azarado [Pas], otimista [Pas], pessimista [Pas]); **Emo. Stab.** (optimistic [Eng], pessimista [Hau/Hut/Mac], unselfconscious [Eng]).

Macho: Agree. (autocratic [Dut/Ger], bossy [Dut/Ger], domineering [Cze/Dut/Ger/Tri], imperious [Dut]); **Emo. Stab.** (masculine [Eng]); **Intel.** (conservative [Dut/Rom], traditional [Rom]); see aggressive (agressivo).

Marvelous (Maravilhoso): Extra. (dull [Hun], exuberant [Dut/Tri], flamboyant [Gol]); **Agree.** (arrogant [Dut], conceited [Pol], egoistic [Ger], egoistical [Ger/Pol], humble [Gol], modest [Gol], pompous [Gol], self-opinionated [Ger], smug [Gol], unassuming [Gol]); **Consc.** (extravagant [Ger], frivolous [Dut/Ger/Hun], modesto [Pas], wishy-wahsy [Ger]); **Emo. Stab.** (self-doubting [Ger]); **Intel.** (conceited [Hun], dull [Cze/Pol], original [And/Dut/Rom], overbearing [Hun], pretending [Hun], prosaico [Pas], simple [Eng], swollen-headed [Hun]); **Pos. Val.** (amazing [Ben], favorite [Ben], formidable [Ben], marvelous [Ben]).

Maximum (Máximo): see marvelous (maravilhoso).

Monster (Monstro): see vile (escroto) and good (bom).

Natural: Agree. (casual [Gol], easygoing [Gol], informal [Gol], natural [Gol]); **Intel.** (natural [Hun]).

Needy (Carente): Extra. (helpless [Pol], solitary [Tri]); **Agree.** (demanding [Dut]); **Emo. Stab.** (dependent [Dut], fragile [Tri], independent [Dut], solitário [Hut], undemanding [Eng], vulnerable [Dut/Hun/Ger/Rom], weak [Tri], whining [Hun]); **Intel.** (autonomous [Gol], independent [Gol], individualistic [Gol]).

Nervous (Nervoso): see anxious (ansioso) and aggressive (agressivo).

Neurotic (Neurótico): see insane (doido) and sensitive (sensível).

Nice (Giro): Extra.** (free and easy [Rom]); **Agree** (accommodating [Dut], agradável [Hut], agreeable [Cze/Pol], amável [Mac/Hau/Hut/And], amigável [Nat/Hut/Pas], antipático [Nat/Pas], cordial [Dut/Tri], friendly [Hun], genial [Dut], good-natured [Ger/Tri], simpático [Nat/Mac/Hau/Hut/Pas], sympathetic [Eng]); **Emo. Stan.** (antipático [Hut]).

Nobody (Ninguém*): Extra.** (self-critical [Gol], self-pitying [Gol]), **Consc.** (frivolous [Dut/Ger/Hun], wishy-wahsy [Ger]); **Emo. Stab.** (self-doubting [Ger], self-pitying [Eng]); see especial (special) and marvelous (maravilhoso).

Normal: Neg. Val. (normal [Sau]); see natural.

Obvious (Óbvio): see ridiculous (ridículo).

Paranoid (Paranóico): see insane (doido) and sensitive (sensível).

Partner (Parceiro*):** see friend (amigo).

Patient (Paciente): see anxious (ansioso).

Perfect (Perfeito): see special (especial) and marvelous (maravilhoso).

Polite (Educado): Agree. (agreeable [Cze/Pol], gentil [Hau/Hut/Mac/Pas], gentle [Hun/Tri], grosseiro [And], kind [Eng/Gol], polite [Cze], prestativo [And], rough [Cze], rude [And/Eng/Pas], understanding [Eng/Rom/Tri], unkind [Eng]); **Consc.** (estudioso [Hut]); **Intel.** (cultured [Cze], educated [Ger], ignorant [Ger], knowledgeable [syn -Cze/ Ger/Pol], rough [Tri], rude [Tri], sofisticado [And], undereducated [Pol], understanding [Cze], uneducated [Ger/Pol/Tri], uninformed [Ger], unsophisticated [Cze]).

Proud (Orgulhoso): see arrogant (arrogante) and marvellous (maravilhoso).

Psychopath (Psicopata): see antisocial (antissocial), insane (doido), and sensitive (sensível).

Pure (Puro): see innocent (inocente).

Quiet (Quieto): Extra. (energetic [Cze/Gol/Pol/Rom], hyperactive [Hun], quiet [Cze/Eng/Gol/Pol], quieto [And/Hau/Hut/Mac]); **Agree.** (calm [Rom/Tri]); **Emo. Stab.** (calm [Dut/Hun], calmo [And/Pas], tranquil [Cze], tranquilo [Nat/Pas]); **Intel.** (inquieto [And]).

Random (Aleatório): see confused (confuso) and lost (perdido).

Realistic (Realista): see deluded (iludido).

Rebel (Rebelde): see free (livre).

Responsible (Responsável): Consc. (dependable [Eng], diligent [Dut/Ger/Hex/Hun], inconstante [Pas], indisciplinado [Nat], irresponsável [Mac/Nat/Hau/Hut/Pas], irresponsible [Dut/Eng/Hun], responsável [Pas], responsible [Eng], scatterbrained [Dut/Ger], steady [Pol]); **Emo. Stab.** (estável [And/Pas], dishonest [Rom], uncertain [Dut], solid [Ger], steady (Dut/Ger], unstable [Cze/Dut]); **Intel.** (devoted [Rom], disloyal [Tri], loyal [Tri], perfidious [Tri], reliable [Hun/Tri], truthful [Hun]).

Retarded (Retardado): see fool (bobo).

Ridiculous (Ridículo): Extra. (dull [Hun]); **Agree.** (arrogant [Dut], conceited [Pol], egoistic [Ger], egoistical [Ger/Pol], self-opinionated [Ger]); **Consc.** (extravagant [Ger], frivolous [Dut/Ger/Hun], modesto [Pas], wishy-wahsy [Ger]); **Emo. Stab.** (self-doubting [Ger], self-pitying [Eng]); **Intel.** (conceited [Hun], dull [Pol], extravagant [Rom], original [And/Dut/Rom], overbearing [Hun], pretending [Hun], prosaico [Pas], simple [Eng], swollen-headed [Hun]); **Neg. Val.** (horrible [Ben], idiotic [Ben], unimportant [Ben], weird [Sau]).

Romantic (Romântico): see in love (apaixonado) and sensitive (sensível).

Rough (Grosso): Extra. (aggressive [Eng], fiery [Ger], unaggressive [Eng]); **Agree.** (affectionate [Eng], aggressive [Hun/Rom/Tri], argumentative [Cze], belligerent [Cze], choleric [Rom/Tri], conciliating [Rom], cordial [Dut/Tri], delicado [Hut], discutidor [And*], domineering [Cze/Dut/Ger/Rom/Tri], explosive [Hun], frio [And/Hut], genial [Dut], grosseiro [And], hostil [Pas], hotheaded [Dut], impetuous [Hun], indulgent [Dut], intolerant [Dut/Rom], intolerante [And], irascilbe [Tri], irritable [Rom/Tri], leninent [Pol], mild [Dut], peaceful [Dut/Hun/Rom/Tri], quarrelsome [Cze/Rom/Tri], rough [Cze], rude [And/Pas/Eng], ruthless [Dut/Pol], tempestuous [Hun], tolerant [Cze/Dut/Rom/Tri], tolerante [And], touchy [Rom], unaggressive [Cze]); **Consc.** (inconsiderate [Hun], thoughtless [Cze/Dut/Rom]); **Emo. Stab.** (calm [Dut/Hun], calmo [And/Pas], fretful [Eng], gruff [Pol], impaciente [Pas], impetuous [Pol], irritable [Cze/Eng], nervoso [And/Pas], nervous [Cze/Dut/Hun/Pol], paciente [Pas], patient [Eng/Pol], ruthless [Rom], short-tempered [Ger], touchy [Cze/Eng/Ger/Gol]); **Intel.** (cultured [Cze], deep [Dut/Eng], dense [Pol], ignorant [Ger], rough [Tri], rude [Tri], shallow [Dut/Eng], simple [Eng], sofisticado [And], uninformed [Ger], unsophisticated [Cze]).

Sad (Triste): see happy (feliz).

Saint (Santo): see good (bom).

Selfish (Egoísta): Agree. (altruísta [Pas], benevolent [Cze], cooperador [And*], cooperative [Eng], egocentric [Dut], egoistic [Ger], egoistical [Ger/Pol], generous [Cze], greedy [Gol], individualista [Pas], magnanimous [Dut/Pol], philantropic [Hun], selfish [Ger/Gol/Pol], self-indulgent [Gol], self-seeking [Ger]); **Emo. Stab.** (altruistic [Rom], egoísta [Hut], generous [Rom], individualistic [Rom]); **Intel.** (altruistic [Hun], self-seeking [Hun], unselfish [Hun]).

Sensitive (Sensível): Agree. (callous [Dut], cold [Eng], considerate [Eng/Ger/Gol], frio [And/Hut], hard [Eng], inconsiderate [Eng], insensitive [Cze/Eng], romântico [Hut], sentimental [Gol/Hut], touchy [Rom]); **Consc.** (nonchalant [Dut]); **Emo. Stab.** (cold [Rom], emotional [Dut/Eng/Ger/Rom], emotivo [And], excitable [Cze/Pol], hard-boiled [Ger], impressionable [Tri], indifferent [Rom], insensitive [Ger], moody [Eng/Ger], oversensitive [Hun], romantic [Rom], sensitive [Dut/Ger/Rom], sentimental [Ger/Rom], thick-skinned [Ger], touchy [Cze/Eng/Ger/Gol], unemotional [Pol], unexcitable [Eng/Pol]); **Intel.** (insensitive [Tri], romantic [Tri], sentimental [Tri], sensitive [Tri]).

Sentimental: see sensitive (sensível).

Serious (Sério): see responsible (responsável) and free (livre).

Shameless (Safado): Extra. (acanhado [Hut], bashful [Cze/Eng/Pol/Rom], candid [Dut], contido [Pas], desembaraçado [Hau/Hut/Mac], desinibido [Pas], envergonhado [Hut], fiery [Dut], hot-blooded [Dut], inhibited [Cze], inibido [And/Hut/Mac], retraído [Pas], shy [Cze/Dut/Eng/Ger/Pol/Rom], timid [Dut/Ger/Pol/Tri], uninhibited [Dut]); **Agree** (mercenary [Pol], moral [Cze/Gol]); **Consc.** (confiável [And], conscientious [Cze/Dut/Ger/Pol/Rom], dependable [Eng], dishonest [Rom], honesto [Hut], honrado [Hut], immoral [Dut], inconsiderate [Hun], lax [Dut/Hun], neat [Eng], scrupulous [Pol], unconscientious [Cze]); **Emo. Stab.** (dishonest [Rom]); **Intel.** (devoted [Rom], disloyal [Tri], insincere [Tri], loyal [Tri], perfidious [Tri], puritan [Rom], reliable [Hun/Tri], truthful [Hun]).

Shit (Merda): see dung (bosta).

Sick (Doente): see insane (doido) and dead (morto).

Silly (Burro): see fool (bobo).

Simple (Simples): see complicado (complicated).

Sincere (Sincero): see direct (direto) and faithful (fiel).

Slow (Lerdo): Extra. (acomodado [Pas], active [Rom/Pol], ativo [Pas], brisk [Pol], cheerful [Dut/Rom/Tri], dinâmico [Pas], dull [Hun], dynamic [Ger/Rom], dynamical [Pol], energetic [Cze/Hex/Rom/Pol], enérgico [And*], enterprising [Cze/Pol], entusiasmado [And*], exuberant [Dut/Tri], full of life [Hun], hyperactive [Hun], jovial [Dut], lively [Ger/Rom], merry [Dut], passivo [Pas], vivacious [Ger]); **Agree.** (willing [Ger]); **Consc.** (diligent [Dut/Ger/Hun/Tri], esforçado [Hau/Hut/Mac/Nat], frivolous [Dut/Ger/Hun], hard-working [Ger], indolent [Dut], industrious [Cze/Dut/Ger/Rom], lax [Hun], lazy [Cze/Dut/Hun], motivado [Pas], obstinado [Pas], painstaking [Dut], persistente [And*/Pas], preguiçoso [And], prompt [Dut], punctual [Dut], scatterbrained [Dut/Ger], workshy [Ger]); **Emo. Stab.** (alert [Cze], astute [Tri], bright [Cze], crafty [Hun], cunning [Tri], impressionable [Tri], rational [Pol], suggestible [Gol/Tri], wily [Hun]); **Intel.** (acute [Dut], bright [Pol], clever [Cze/Ger], curioso [Ant/Hau/Hut/NatPas], dull [Cze/Pol], engenhoso [And], highly intelligent [Ger], imperceptive [Eng], ingenious [Ger], intelligent [Cze/Eng/Ger], retarded [Ger], rough [Tri], rude [Tri], silly [Pol], simple [Eng], slow-witted [Pol], sofisticado [And], stupid [Ger], thoughtful [Tri], unintelligent [Cze/Eng/Pol], unperspicacious [Pol]).

Social: see antisocial (antissocial) and friend (amigo).

Soft (Mole): see hard (duro).

Son of a bitch (Filho da puta): Agree. (accommodating [Dut], agradável [Hut], agreeable [Cze/Pol], amável [Mac/Hau/Hut/And], amigável [Nat/Hut/Pas], antipático [Nat/Pas], considerate [Eng/Ger], good-hearted [Dut/Ger], good-natured [Ger/Tri], inconsiderate [Eng], mercenary [Pol], moral [Cze/Gol]); **Consc.** (conscientious [Cze/Dut/Ger/Hun/Pol], dependable [Eng], honesto [Hut], honrado [Hut], immoral [Dut/Hun], inconsiderate [Hun], lax [Dut], scrupulous [Pol], unconscientious [Cze]); **Emo. Stab.** (dishonest [Rom]); **Intel.** (devoted [Rom], disloyal [Tri], loyal [Tri], perfidious [Tri], reliable [Hun/Tri], truthful [Hun]); **Neg. Val.** (corrupt [Sau], cruel [Ben], filthy [Ben], horrible [Ben], vandalic [Ben]).

Special (Especial): Extra. (dull [Hun]); **Agree.** (arrogant [Dut], conceited [Pol], egoistic [Ger], egoistical [Ger/Pol], self-opinionated [Ger]); **Consc.** (extravagant [Ger], frivolous [Dut/Ger/Hun], modesto [Pas], wishy-wahsy [Ger]); **Emo. Stab.** (self-doubting [Ger]); **Intel.** (conceited [Hun], dull [Cze/Pol], original [And/Dut/Rom], overbearing [Hun], pretending [Hun], prosaico [Pas], simple [Eng], swollen-headed [Hun]); **Pos. Val.** (not special [Ben], mediocre [Ben], favorite [Ben], formidable [Ben], super [Ben]).

Spoiled (Mimado): Agree. (demanding [Dut]); **Emo. Stab.** (dependent [Dut], fragile [Tri], independent [Dut], undemanding [Eng], vulnerable [Dut/Hun/Ger/Rom], weak [Tri], whining [Hun]); **Intel.** (autonomous [Gol], independent [Gol]).

Stressed (Estressado): see anxious (ansioso) and aggressive (agressivo).

Strong (Forte): Extra. (active [Gol/Pol/Rom], assured [Gol], ativo [Pas], bold [Cze/Eng/Pol], brisk [Pol], cowardly [Pol], courageous [Gol], energetic [Cze/Gol/Pol/Rom], ernérgico [And*], helpless [Pol], passive [Cze/Gol/Rom/Tri], passivo [Pas], shy [Cze/Eng/Dut/Ger/Gol/Rom/Pol]); **Agree.** (bold [Hun], courageous [Cze/Tri], dócil [Hut], hard [Eng], meek [Rom], mild [Dut]); **Consc.** (disobedient [Tri], firm [Ger], lax [Dut/Hun], obedient [Tri], purposeful [Ger], sloppy [Eng], unruly [Rom/Tri], wishy-washy [Ger]); **Emo. Stab.** (courageous [Tri], fearful [Hun/Tri], fragile [Tri],

masculine [Eng], obedient [Dut/Rom], resistente [Pas], solid [Ger], strong [Tri], vulnerable [Dut/Hun/Ger/Rom], vulnerável [Pas], weak [Tri], whining [Hun]); **Intel.** (docile [Dut], meek [Hun], servile [Dut/Rom]).

Stubborn (Teimoso): Extra. (intrasigente [Pas]); **Agree.** (adaptable [Hun], accommodating [Dut], argumentative [Cze], compreensível [Hut], compreensivo [Mac], flexible [Dut], headstrong [Hun], indulgent [Dut], intolerant [Dut/Rom/Tri], intolerante [Pas], lenient [Ger], obstinate [Hun], polemical [Tri], tolerant [Cze/Dut/Rom/Tri], suggestible [Tri], tolerante [Pas]); **Consc.** (disobedient [Tri], unruly [Rom/Tri]); **Emo. Stab.** (afirmativo [Hut], obedient [Dut/Rom], obstinate [Ger]); **Intel.** (bull-headed [Cze], docile [Dut], flexível [Pas], rígido [Pas]).

Stupid (Estúpido): see fool (bobo) and ignorant (ignorante).

Sweet (Doce): Agree. (affectionate [Eng], agreeable [Cze/Pol], amável [And/Hau/Hut/Mac/Nat], amoroso [Pas], antipático [Nat/Pas], considerate [Eng/Ger/Gol], genial [Dut], gentil [Hau/Hut/Mac/Pas], gentle [Hun/Tri], hard [Eng], harsh [Eng], inconsiderate [Eng], kind [Eng/Gol], meek [Rom], mild [Dut], simpático [Hau/Hut/Mac/NatPas], sympathetic [Eng], unkind [Eng], unsympathetic [Eng]); **Emo. Stab.** (antipático [Hut]); **Intel.** (sweet [Tri]).

Switched on (Ligado): see intelligent (inteligente) and damned (danado).

Sympathetic (Simpático): Extra. (dull [Hun]); **Agree.** (agreeable [Cze/Pol], amável [And/Hau/Hut/Mac/Nat], antipático [Nat/ Pas], considerate [Eng/Ger/Gol], friendly [Hun], genial [Dut], inconsiderate [Eng], kind [Eng/Gol], simpático [Hau/Hut/Mac/Nat/Pas], sympathetic [Eng], unkind [Eng], unsympathetic [Eng]); **Emo. Stab.** (antipático [Hut]).

Tender (Meigo): see sweet (doce).

Timid (Tímido): Extra. (acanhado [Hut], bashful [Cze/Dut/Eng/Ger/Gol/Rom/Pol], bold [Cze/Eng/Pol], calado [And/Hut/Nat/Pas], comunicativo [Hau/Hut/Mac/NatPas], contido [Pas], cowardly [Pol], desembaraçado [Hau/Hut/Mac], envergonhado [Hut], expansivo [Pas], extroverted [Eng/Dut/Rom/Tri], extrovertido [Pas/Hut], fearful [Pol], inhibited [Cze], inibido [And/Mac], introverted [Eng/Dut/Rom/Tri], introvertido [Hut], quiet [Cze/Eng/Gol/Pol], quieto [And/Hau/Hut/Mac], reservado [And], reserved [Eng/Dut/Ger/Gol/Tri], shy [Cze/Eng/Dut/Ger/Gol/Rom/Pol], silent [Cze/Dut/Eng/Ger/Gol/Hun/Tri], taciturn [Hun/Rom], talkative [Eng], timid [Dut/Eng/Ger/Gol/Pol/Tri], tímido [And/Hau/Hut/Mac/Nat/Pas], uninhibited [Dut], untalkative [Cze/Eng/Ger/Gol/Hun], verbal [Eng], verbose [Cze]); **Emo. Stab.** (assured [Cze/Dut/Tri], audacioso [Hau/Mac], bold [Hun], confident [Cze], corajoso [Hau], courageous [Tri], decisive [Dut/Tri], fearful [Hun/Tri], indecisive [Tri], insecure [Tri], inseguro [Hau/Hut/Mac], ousado [Pas], reserverd [Cze], resolute [Dut], self-assured [Hun], temeroso [Pas]); **Intel.** (hesitante [Pas]).

Top: see special (especial) and marvelous (maravilhoso).

Tranquil (Tranquilo): Extra. (aggressive [Eng], energetic [Cze/Gol/Pol/Rom], fearful [Pol], hyperactive [Hun], quiet [Cze/Eng/Gol/Pol], quieto [And/Hau/Hut/Mac], unaggressive [Eng]); **Agree.** (aggressive [Hun/Rom/Tri], belligerent [Cze], calm [Rom/Tri], choleric [Rom/Tri], explosive [Hun], frio [And/Hut], hotheaded [Dut], impetuous [Hun], irascilbe [Tri], irritable [Rom/Tri], mild [Dut], patient [Rom/Tri], peaceful [Dut/Hun/Rom/Tri], quarrelsome [Cze/Rom/Tri], tempestuous [Hun], unaggressive [Cze]); **Emo. Stab.** (ansioso [Hau/Hut/Mac/Nat/Pas], anxious [Cze/Dut/Hun/Tri], assured [Cze/Dut/Tri], calm [Dut/Hun], calmo [And/Pas], cold [Rom], confident [Cze], estressado [And*], excitable [Cze/Pol], explosive [Pol], fearful [Hun/Tri], fretful [Eng], impaciente [Pas], impetuous [Pol], imperturbable [Dut/Eng/Ger/Rom], insecure [Tri], inseguro [Hau/Hut/Mac], irritable [Cze/Eng], nervoso [And/Pas], patient [Eng/Pol], nervous [Cze/Dut/Hun/Pol], panicky [Dut], patient [Eng/Pol], paciente [Pas], peaceful [Dut/Hun/Rom/Tri], preocupado [And], relaxado [And], relaxed [Eng], restless [Cze], self-assured [Hun], short-tempered [Ger], tenso [And], touchy [Cze/Eng/Ger/Gol], tranquil [Cze], tranquilo [Nat/Pas], unexcitable [Eng/Pol], worrying [Hun]); **Intel.** (philosophical [Eng/Dut], inquieto [And]).

Trashy (Lixo): see ridiculous (ridículo) and vile (escroto).

Unbearable (Insuportável): Extra. (dull [Hun], free and easy [Hun]); **Agree.** (accommodating [Dut], agradável [Hut], agreeable [Cze/Pol], amável [Mac/Hau/Hut/And], amigável [Nat/Hut/Pas], antipático [Nat/Pas], cordial [Dut/Tri], friendly [Hun], genial [Dut], simpático [Nat/Mac/Hau/Hut/Pas], sympathetic [Eng]); **Intel.** (dull [Cze/Pol]).

Unique (Único): see special (especial).

Unlucky (Azarado): see lucky (sortudo).

Useless (Inútil): Extra. (resourceful [Pol]); **Agree.** (helpful [Eng/Ger/Pol]); **Consc.** (desmotivado [Pas], diligent [Dut/Ger/Tri], efficient [Eng], eficaz [Pas], eficiente [And/Hut/Pas], esforçado [Hau/Hut/Mac/Nat], frivolous [Dut/Ger/Hun], hard-working [Ger], indolent [Dut], industrious [Cze/Dut/Ger/Rom], inefficient [Eng], ineficaz [Pas], ineficiente [Pas], lazy [Cze/Dut/Hun], motivado [Pas], obstinado [Pas], persistente [And*/Pas], preguiçoso [And], purposeful [Cze/Ger], wishy-wasy [Ger], workshy [Ger]); **Emo. Stab.** (self-doubting [Ger], self-pitying [Eng]); **Intel.** (efficient [Cze], gifted [Ger/Pol], highly gifted [Ger], inefficient [Cze], overbearing [Hun], prosaico [Pas], swollen-headed [Hun], talented [Ger/Pol], untalented [Cze/Ger]); **Neg. Val.** (good-for-nothing [Sau]).

Vacillating (Vacilão): see asshole (cuzão) and son of a bitch (filho da puta).

Vagabond (Vagabundo): Extra. (enterprising [Cze/Pol]); **Agree.** (mercenary [Pol], moral [Cze/Gol]); **Consc.** (confiável [And], conscientious [Cze/Dut/Ger/Pol/Rom], dependable [Eng], diligent [Dut/Ger/Hun/Tri], disciplined [Hun/Tri], frivolous [Dut], hard-working [Ger], immoral [Dut], inconsiderate [Hun], inconstante [Pas], indisciplinado [Nat], indolent [Dut], industrious [Cze/Dut/Ger/Rom], irresponsável [Mac/Nat/Hau/Hut/Pas], irresponsible [Dut], lax [Hun], lazy [Cze/Dut/Hun], painstaking [Dut], preguiçoso [And], responsável [Pas], responsible [Eng], scrupulous [Pol], unconscientious [Cze], workshy [Ger]); **Intel.** (aventureiro [Hau/Hut/Mac/Pas], devoted [Rom], disloyal [Tri], loyal [Tri], perfidious [Tri], reliable [Hun/Tri], rough [Tri], rude [Tri], truthful [Hun]).

Vile (Escroto): Agree** (agradável [Hut], agreeable [Cze/Pol], amável [Hau/Hut/Mac], amigável [Hut/Nat/Pas], confiável [And], considerate [Eng], good-natured [Ger/Tri], grosseiro [And], inconsiderate [Eng], indiferente [Pas], kind [Eng], mercenary [Pol], moral [Cze/Gol], rough [Cze], rude [And/ Eng/Pas], unkind [Eng]); **Consc.** (conscientious [Cze/Dut/Ger/Pol/Rom], dependable [Eng], honesto [Hut], honrado [Hut], immoral [Dut], inconsiderate [Hun], lax [Dut/Hun], neat [Eng], negligent [Eng/Pol], neglectful [Hun], nonchalant [Dut], scrupulous [Pol], unconscientious [Cze]); **Emo. Stab.** (dishonest [Rom], impetuous [Pol], gruff [Pol], ruthless [Rom]); **Intel.** (disloyal [Tri], loyal [Tri], perfidious [Tri], puritan [Rom], reliable [Hun/Tri], rough [Tri], rude [Tri]); **Neg. Val.** (corrupt [Sau], cruel [Ben], filthy [Ben], horrible [Ben], vandalic [Ben]).

Warrior (Guerreiro): Extra. (active [Gol/Pol/Rom], assured [Gol], ativo [Pas], brave [Gol], bold [Cze/Eng/Pol], confident [Gol], cowardly [Pol], courageous [Gol], docile [Gol], energetic [Cze/Gol/Pol/Rom], passive [Cze/Gol/Rom/Tri], passivo [Pas], submissive [Gol], vigorous [Gol]); **Agree.** (bold [Hun], courageous [Cze/Tri], dócil [Hut]); **Consc.** (aimless [Gol], decidido [Pas], decisive [Gol], dedicado [Hut/Mac/Nat], deliberate [Gol/Hun], desistente [Pas], esforçado [Hau/Hut/Mac/Nat], firm [Ger/Gol], hard-working [Ger], indecisive [Cze], indeciso [Pas], lazy [Cze/Dut/Hun], obstinado [Pas], persistent [Gol], persistente [And*/Pas], preguiçoso [And], purposeful [Gol], strong-minded [Ger], tenacious [Gol], unruly [Rom/Tri], wishy-washy [Ger], workshy [Ger]); **Emo. Stab.** (assured [Cze/Dut/Tri], confident [Cze], courageous [Cze/Tri], fearful [Hun/Tri], fragile [Tri], masculine [Eng], resistente [Pas], strong [Tri], vulnerable [Dut/Hun/Ger/Rom], vulnerável [Pas], weak [Tri], whining [Hun]); **Intel.** (docile [Dut], servile [Dut]).

Weak (Fraco): see strong (forte).

Weary (Cansado): see dead (morto).

Weird (Estranho): see different (diferente) and ridiculous (ridículo).

Wicked (Ruim): see good (bom).

Worse (Pior): see useless (inútil) and ridiculous (ridículo).

Final Considerations

The lexical approach can be summarized by the axiom sustaining that a taxonomy of personality can be obtained from natural language since the most significant individual differences for quotidian social interactions become eventually encoded in the language people use. According to this approach, the more relevant a difference is in the relations between persons, the more likely a culture will conceive one or more specific words to represent such difference. The idea that personality traits can be found in languages lexicons led to the development of some of the most renowned theoretical models in the field of personality psychology, such as the Big Five and Cattell's 16 primary personality factors.

In the first manuscript of this dissertation, we demonstrate how research under the lexical hypothesis perspective have been developed since the initial ideas of Francis Galton and Ludwig Klages, and the pioneering work of Gordon Allport, Raymond Cattell, and many other scientists. Throughout the three manuscripts, but especially in the first, we also seek to synthesize some of the criticism to the lexical approach to the study of the personality. We also discussed two broad limitations regarding methodological issues reviewed in the literature of cross-cultural psychology and psychological study of natural language.

The criticisms originating from cross-cultural psychology concern the predominance of *etic* imposed research in personality. As we reviewed, the taxonomic models of personality were substantially developed in western, educated, industrialized, rich, and developed countries, such as the United States, The Netherlands, and Germany. In a universalistic perspective, most research after the initial development of such models was concerned with the replicability of the supposedly universal

personality models in different cultural contexts. Albeit the universality of trait descriptive terms and personality models is a corollary of the lexical hypothesis, a purely *etic* approach has the potential to compromise the emergence of autochthonous traits and dimensions of personality.

The criticisms from the perspective of the psychological study of natural language are related to the most frequently explored sources of personality-descriptive terms and data. Traditionally, researchers make use of dictionaries as the primary source to select personality trait-descriptive terms. The restrict use of dictionaries is one of the major criticisms to the lexical approach since they may not be synchronized with the current social use of the words. Regarding the data sources, the common practice of developing instruments with a restrict set of words retrieved from dictionaries and of collect data in test settings are also a major criticism concerning the psycholexical research. For the critics, these strategies are strongly dependent on the researchers' decisions regarding which are the most relevant traits to be investigated and how to interpret them and the research results. Therefore, these strategies can restrict the free expression of personality traits and circumscribe the findings to the limited set of investigated items.

In the fields of cross-cultural and natural language psychology different methodological approaches were proposed to address the highlighted issues. An integrative emic-etic approach was recommended to combine universal aspects of personality with unique or culturally specific aspects. Regarding the question of data sources, we reviewed studies that made use of alternative sources to obtain personality descriptors and of distinct data collection strategies. Examples of alternative traits sources to the dictionaries are several text types (e.g., literary, scholarly, journalistic, and biographical), recordings of written and oral personality descriptions made by

laypersons (e.g., interviews, conversations, and essays, etc.), and registers of behavior in social media (e.g., blogs, social networks, apps, etc.).

With a focus on the Brazilian culture and with the general objective of contributing to overcome the issues mentioned above from a methodological perspective, we conducted the studies reported in the second and third manuscript of this dissertation. In both studies, our main strategy was to use a natural language source to identify personality descriptors and to collect data for dimensionality analyses: the public and spontaneous messages posted (i.e., tweets) in the online social network Twitter. In the second study, we examined 6,303 posts from 5,493 unique Twitter users. As the main result, we obtained a list of 1,118 adjectives and 332 nouns, many of them absent in other Brazilian compilations of personality traits. We believe that this list can be useful for the selection of personality descriptors in subsequent research in Brazil.

With the feasibility regarding the use of a social network as data and trait source assessed in the second study, we designed a third study with the main objective of investigating the dimensionality of the data obtained from Twitter. We examined the data concerning 86,899 users and 172 adjectives. To assess the dimensionality of the data, we employed a topic modeling technique designed for text mining, called Latent Dirichlet Allocation. The semantic content of the latent models examined was interpreted using as reference the most prominent theoretical psycholexical models, such as the models with three, five, six, and seven factors, and Cattell's 16PF. Cross-validation analyses indicated models with seven and 14 dimensions as the most appropriate for the data. Besides these two models, we also examined another four models with a latent structure similar to the theoretical models (e.g., Big Five). The results regarding the models with seven and 14 dimensions are promising, but the second model has shown more evidence of interpretability when considering the

semantic internal coherence of the content of each of its topics. Concerning the remaining four models examined, the semantic content of their latent structure was not congruent with the formulations of the correspondent theoretical models.

The results from the second and third study suggested that some traits and latent dimensions from prominent theoretical models found in the international literature were not recovered in our data, while new latent dimensions and traits emerged. These results, however, should be complemented with new evidence from further studies before sustaining conclusions regarding the presence of Brazilian autochthonous personality factors or the lack of relevance in Brazilian culture of some factors of potentially universal models of personality (e.g., Big Five). These questions require further investigation of the psychometric correspondence between the topics models uncovered and the factors from theoretical personality models.

With these three studies, we advocate that an approach combining the study of natural language in an integrative *emic-etic* cross-cultural perspective is a promising and already feasible strategy to be explored to advance both theoretical and methodological aspects of personality research under the postulates of the lexical hypothesis. The big data with available records of a diversity of human behaviors and the rapid development of data science and computational technology represent an open path for personality and psychometric research, in particular, and for the psychological science in general. A path in which developing countries are welcome.