European Journal of Personality, *Eur. J. Pers.* **34**: 917–943 (2020)
Published online 5 June 2020 in Wiley Online Library (wileyonlinelibrary.com) **DOI**: 10.1002/per.2270

Tracing Personality Structure in Narratives: A Computational Bottom-Up Approach to Unpack Writers, Characters, and Personality in Historical Context

RONALD FISCHER[1,2]*, JOHANNES ALFONS KARL[1], MARKUS LUCZAK-ROESCH[3,4], VELICHKO H. FETVADJIEV[5,6] and ADAM GRENER[7]

[1] *School of Psychology, Victoria University of Wellington, Wellington, New Zealand*
[2] *Instituto D'Or de Pesquisa e Ensino, Rio de Janeiro, Brazil*
[3] *School of Information Management, Victoria University of Wellington, Wellington, New Zealand*
[4] *Te Pūnaha Matatini-Centre of Research Excellence (CoRE) for Complex Systems, New Zealand*
[5] *Department of Social Psychology, University of Amsterdam, The Netherlands*
[6] *Department of Psychology, University of Limerick, Ireland*
[7] *Faculty of Humanities and Social Sciences, Victoria University of Wellington, New Zealand*

Abstract: We present a new method for personality assessment at a distance to uncover personality structure in historical texts. We focus on how two 19th century authors understood and described human personality; we apply a new bottom-up computational approach to extract personality dimensions used by Jane Austen and Charles Dickens to describe fictional characters in 21 novels. We matched personality descriptions using three person-description dictionaries marker scales as reference points for interpretation. Factor structures did not show strong convergence with the contemporary Big Five model. Jane Austen described characters in terms of social and emotional richness with greater nuances but using a less extensive vocabulary. Charles Dickens, in contrast, used a rich and diverse personality vocabulary, but those descriptions centred around more restricted dimensions of power and dominance. Although we could identify conceptually similar factors across the two authors, analyses of the overlapping vocabulary between the two authors suggested only moderate convergence. We discuss the utility and potential of automated text analysis and the lexical hypothesis to (i) provide insights into implicit personality models in historical texts and (ii) bridge the divide between idiographic and nomothetic perspectives. © 2020 European Association of Personality Psychology

Key words: personality; five-factor model; idiographic analysis; automated text analysis; transcendental information cascades

INTRODUCTION

The lexical hypothesis has been central for the description of human personality. Humans have judged others in terms of their personality for at least the last 3000 years (Mayer, Lin, & Korogodsky, 2011). Based on extensive survey studies in psychology, a consensus has emerged that personality traits in Western literate populations are best described by five major factors (Goldberg, 1981, 1993, 1990; Costa & McCrae, 1992). The assumption is that traits organized along these five dimensions provide sufficient information for individuals to describe themselves and others at a relatively granular level within a social context. The prevalence and the success of the lexical approach to capture five distinct factors open the intriguing possibility to capture and describe personality information in other textual sources. One of the intriguing questions is what implicit personality models may have been used in historical times by storytellers when describing others.

It is not possible to directly question deceased persons about their personality traits or to conduct a psychological analysis of personality traits across historical time periods using modern diagnostic tools (e.g. a person responding to a Big Five questionnaire). One option to overcome this temporal distance problem is to analyse texts produced by individuals to discover or reconstruct information about their implicit personality characteristics (Rosenberg & Jones, 1972). However, such analyses can be very time consuming if they require hand coding large amounts of texts produced by individuals (e.g. letters, novels, manuscripts, and articles). Using human raters might also introduce various biases, including biases driven by the personalities of the individual coding the information (Srivastava, Guglielmo, & Beer, 2010; Wood, Harms, & Vazire, 2010), individual differences in the coded accuracy of personality judgements (Hall, Goh, Mast, & Hagedorn, 2016), stereotypes about social categories (Uher & Visalberghi, 2016), and contemporary definitions biasing the interpretation of meanings in historical texts (Pagel, Beaumont, Meade, Verkerk, & Calude, 2019). It is

*Correspondence to: Ronald Fischer, School of Psychology, Victoria University of Wellington, Kelburn, Wellington 6012, New Zealand.
E-mail: ronald.fischer@vuw.ac.nz

This article earned Open Data and Open materials badges through Open Practices Disclosure from the Center for Open Science: https://osf.io/tvyxz/wiki. The data are permanently and openly accessible at https://osf.io/8bh3e/ and http://www.gutenberg.org/. Author's disclosure form may also be found at the Supporting Information in the online version.

Handling editor: John Rauthmann
Received 7 August 2019
Revised 29 April 2020, Accepted 29 April 2020

therefore reasonable to suggest that (i) there is currently no canonical method for historical analysis of human personality from text and (ii) there is a gap in our understanding how to develop such an approach so that it is not superimposing contemporary assumptions and biases about the linkage between language and human personality.

It is often argued that big data approaches show much promise for personality research (Bleidorn, Hopwood, & Wright, 2017). However, computational analysis methods are not free from potential human biases, and they may even amplify those. One of the core limitations of current big data approaches, for example, is that they have been used in a deductive, theory-driven way focused on prediction (e.g. predicting survey responses from textual analyses or digital traces). This embeds a contemporary understanding of human language and personality within the algorithmic approach and therefore falls short of reflecting historical context and language use. We thus suggest that the potential of big data for an inductive understanding of personality theory, such as testing whether lexical structures are time and source invariant, has been underutilized.

We take up this challenge and present a text-mining approach that is aimed at extracting and dimensionalizing person-relevant information in large corpora of text. Our focus is on extracting possible implicit personality models used by authors, that is, perceived characteristics or interrelations of characteristics underlying people's behaviours (similar to the way implicit personality models have been studied in psycholexical and indigenous studies (e.g. Nel et al., 2012; Rosenberg & Jones, 1972). These models are 'implicit' because they are inferred from the author's descriptions of characters and groups, without the author explicitly stating or organizing them into a formal, coherent, and parsimonious theory of personality (Rosenberg & Jones, 1972). Our text-mining approach may provide insights into the implicit personality models that authors use when constructing and describing their fictional characters, as expressed through the choice of trait terms and phrases. We describe the applicability and promise of this approach by analysing novels by two well-known English authors: Jane Austen and Charles Dickens. By comparing the work of two authors in this case study, we also show the utility for a broader reconnection between idiographic and nomothetic personality studies. While idiographic approaches aim to identify patterns within single individuals across various processes and situations, nomothetic approaches aim to extract regularities of behaviour across a population of individuals (Barenbaum & Winter, 2008; Conner, Tennen, Fleeson, & Barrett, 2009). The personality structure evident in the work of individual authors facilitates insights into the implicit personality models used by those authors (an idiographic perspective), which can then be compared across authors for the emergence of possible common structures representative of a larger social context (a nomothetic perspective).

Our work is a transdisciplinary endeavour into the feasibility of devising computational tools to capture psychological meaning from text in an inductive fashion. It triangulates the capabilities of contemporary big data analytics, in particular natural language processing, sequential data mining, and dynamical systems theory, with insights from both personality psychology and literary studies. In particular, we shift the attention from the view prevalent in most psycholexical work that there is a single statistical regime underlying the use of person-descriptive terms to a view that is empirically informed by dynamical systems research (Altmann, Pierrehumbert, & Motter, 2009; Gerlach & Altmann, 2013) indicating that careful attention must be paid to nonlinear properties, recurrences, and long-range dependencies. This article reports one first step of this ambitious journey and contributes important insights that have the potential to lead to completely new ways to construct psycholexical studies as a dialectic of big data analytics and human interpretative practice.

STORYTELLING AND PERSON INFORMATION

Fictional narratives provide the opportunity to analyse implicit personality models within distinct historical periods. Works of art are in one sense the creation of individual minds, and research on language and personality has demonstrated that what we tell others reveals a lot about the person communicating (e.g. Hirsch & Peterson, 2009; Pennebaker & King, 1999; Watzlawick, Beavin, & Jackson, 1967). To this extent, then, recurring features within an artistic work might be considered to be an expression of the creator's personality (Robinson, 1985), as the creator's preoccupations, fears, and inclinations are transmitted to audiences through the words he or she chooses to describe people and events. The analysis of historical texts can thus provide information about individual authors, resulting in a rich idiographic analysis from a distance.

More importantly, historicist approaches predominant within literary studies have emphasized for many decades that texts must be situated within their historical and cultural contexts (Hamilton, 1996). Literary works are not mere reflections of their historical moment, but the ideas, meanings, and values they convey cannot be divorced from the cultural currents that led to their production. We would assume that a writer's descriptions of persons and personality might be richer and more nuanced than most others within the writer's milieu. To convey meaning, however, those descriptions must be comprehensible to the audience (see Vermeule, 2010). If writers offer new or unique depictions of personhood, those representations circulate within and shape the culture in which they are received. Dickens, for example, was one of the first novelists whose characters were widely commercialized (e.g. in figurines, clothing items, etc.) and thus attained a life of their own within his contemporary popular culture (John, 2010), speaking to the relevance and appeal of these fictional characters. An analysis of the texts produced by specific authors within a historical period can provide us with insights into what aspects were central for describing persons in that period. Although the current study focuses on just two authors, applied more widely (e.g. different books, authors, or genres), the method can provide rich insight into the linguistically mediated implicit models of the person within a historical period.

DOI: 10.1002/per

Descriptions of characters in fictional worlds differ in important ways from individuals' descriptions of themselves or other people. Characters in fictional works are not real people, and they function beyond being mere representations of people to embody and further the plot and themes of a narrative (Phelan, 1989). Literary theorists have emphasized alternatively how characters are mere linguistic constructs and how they elicit affective investments that prompt us to relate to them in the same way that we relate to real people. For example, recent cognitivist approaches to fiction have argued (Vermeule, 2010; Zunshine, 2006) that fictional texts activate and exercise our cognitive capacity to interpret the intentions and beliefs of others. Thus, if fictional characters are merely words on a page, those words draw on conventional constructions of types and roles within a given culture to elicit the cognitive and affective engagements of readers: 'Both fictional characters and kinds of persons are models of an aspect of the world, schemata that generalize and simplify human being in conventional ways and make it available to understanding and action' (Frow, 2018, p. 111). Analyses of fiction should thus allow one of the clearest and cleanest tests of the lexical hypothesis because the analyses of person descriptions tells us something about what behavioural information is deemed relevant and important to be passed on through words to audiences in order to stimulate their interest and attention to particular stories about fictional characters.

USING A TEXT-MINING APPROACH TO HISTORICAL NOVELS

There are two major approaches that are possible for extracting personality-relevant information from written (or transcribed) text. The first is content coding along pre-established psychometric categories, for example, using marker terms that indicate specific personality traits and to then interpret their frequency for specific entities in terms of existing theory (e.g. Chung & Pennebaker, 2008; Passakos & De Raad, 2009). We label this method top-down because the classification is done along pre-existing, theory-driven categories. The alternative is to start with a bottom-up analysis of the themes and topics that emerge around specific characters. At the extreme end, no theoretical grounding is presupposed and any combination of words can be analysed in terms of their coherence and usefulness for describing specific characters.

These two different approaches should in theory be achievable both via human coders and using algorithmic processes. With respect to the top-down approach, there is little debate that it is practically possible to develop and apply automatic text coding methods using pre-established categories and dictionaries, the most widely used system being the Linguistic Inquiry and Word Count (LIWC; Pennebaker et al., 2015). This dictionary-based approach allows the categorization of texts in terms of valence, emotions, attentional focus, thinking styles, and other psychologically relevant characteristics. It uses both pre-assigned semantic meaning of words in the form of dictionaries and standard linguistic features of the analysed text (grammatical tense, use of specific word categories such as stop words, pronouns, etc.) to classify text in psychological terms (Tausczik & Pennebaker, 2010). Classification of text in terms of the Big Five is possible and has been widely used for online content (e.g. Schwartz et al., 2013; Yarkoni, 2010). Similarly, Passakos and De Raad (2009) manually content coded the characters in Homer's Iliad in terms of the Big Five factors, using the characters' epithets. They were able to classify these epithets in terms of the five factors, with a predominance of terms capturing Agreeableness and Conscientiousness. The fewest references were made to Openness/Intellect. Passakos and De Raad observed that the majority of epithets in the Iliad did not capture traits but rather other person-relevant descriptors such as physical features, reputation, skills, or social evaluations.

However, the situation looks different with respect to bottom-up approaches. While one may intuitively assume that developing a computational approach can be easily achieved by algorithmically mimicking human assessment of psychological constructs, in fact only little progress has been made in this regard over the last 40 years, because human assessment or classification of free-response person descriptions in terms of psychological constructs has not been fully understood (Goldberg, 1982; Kim & Klinger, 2018; Rosenberg & Sedlak, 1972).

The most commonly used computational bottom-up approaches available at the moment are structural topic models that can be used to identify salient terms or topics with high frequencies in open-ended responses in a bottom-up fashion. Topic modelling approaches are relatively widespread in social and political science to identify common themes in speeches, open responses, or social media texts (Farrell, 2016; Roberts et al., 2014; Tvinnereim & Fløttum, 2015). These approaches provide summaries of themes across a whole text (e.g. with assigned probabilities that a specific text belongs to a thematic category or not) but do not allow fine-grained differentiation of change or variation within texts that would allow for an analysis of the dimensional properties of implicit personality models. Topic models cannot be used to identify personality dimensions because they classify text into distinct classes rather than continua. These approaches are also often difficult to interpret because the associations might be highly source and content specific, limiting replicability across text sources.

A second recently developed technique is diachronic (cross time) word embeddings, which is promising because the approach incorporates semantic change (Garg, Schiebinger, Jurafsky, & Zou, 2018; Hamilton, Leskovec, & Jurafsky, 2016). Word embeddings in general make use of a diverse range of linguistic features that can be assessed for words (e.g. their co-occurrence with other words) to mathematically map each word onto a coordinate system (in technical terms called a vector space) (Devlin, Chang, Lee, & Toutanova, 2018; Vaswani et al., 2017; Zhang et al., 2019). This vector space can have a very large number of dimensions; the number of coordinates can be much larger than the two or three dimensions that humans are comfortable dealing with visually. This word embedding representation then allows the usage of established mathematical methods to assess the semantic distance between words, which has

DOI: 10.1002/per

been proven useful in a variety of natural language understanding (e.g. sentiment analysis) and natural language generation tasks. The diachronic word embedding approach allows tracing the transition of words (and the variation of their distances to other words) through such a vector space via temporal snapshots. However, the word embedding approach works at the aggregate level: it relies on large corpora of texts to construct the mapping of words to an overall coordinate system and does not allow the matching of specific fictional characters within novels to specific trait terms that enables a dimensional analysis of personality descriptions.

Our computational bottom-up approach aims to move away from an analysis of fixed aggregate semantics and instead focus on the temporal patterns and dynamics of human expression that are reflective of personality and can be recovered in text. In our view, this requires a three-step approach. First, we need to develop an unsupervised approach to capture person-descriptive language artefacts from text and compare the emerging structure from associations between words (e.g. trait terms) with contemporary knowledge about prevalent models of personality. Do factor structures derived from trait co-occurrences in text replicate factors derived from survey responses? If not, are those factor structures interpretable, given what we know about psychological concepts of personality within socioecological contexts? The co-occurrence of words and named characters in their linguistic context is the starting point and is therefore the focus of the work presented here (for an earlier human coding approach, see Rosenberg & Jones, 1972). Second, once we have a better understanding of the feasibility and dynamics of this co-occurrence structure, it will become possible to study the recurrences of linguistic markers extracted from text in order to understand the significance of variance at various levels, from the depiction of characters and story lines within a novel to the level of the author producing different works across a career to the macroscopic scale of human collectives producing (and consuming) works of literature. Finally, we aim to study both co-occurrences and recurrences in conjunction with other data (e.g. biographical data and socio-economic data) to try to derive an integrated theory of human personality from invariant properties and patterns within such an expanded model.

The method that we describe below has the computational capabilities for the three-step integration. We present a first application to personality data by focusing on the co-occurrences of trait terms and how bottom-up constructed factor structures from literary texts match contemporary survey-based responses as well as converging between two different authors within the same historical epoch. By turning to trait terms and co-occurrence patterns first, we are building on and extending the foundational psycholexical approach.

THE PSYCHOLEXICAL APPROACH

One of the foundations of the five-factor model has been the lexical hypothesis, which states that the most important personality characteristics tend to be encoded in language as single terms (Goldberg, 1981). The hypothesis has been central for the description of personality traits via self-ratings and other-ratings using curated sets of adjectives. Allport & Odbert (1936) created the first comprehensive list of terms that was thought to capture important psychological attributes. The original list contained nearly 18 000 terms and was not further analysed in terms of their underlying structure. Subsequent studies applying refined and redacted sets in student and adult population samples, where participants rated themselves or others on these lists, ultimately led to a first consensus that there are probably five major factors that are sufficient to describe personality traits in broad stroke dimensions (Goldberg, 1993). The most commonly used terms to describe these five dimensions are Conscientiousness (C), Agreeableness (A), Neuroticism (N), Openness/Intellect (O), and Extraversion (E). Following the lexical hypothesis, the Big Five model captures the basic dimensions that people use to communicate important information about themselves or others.

Despite the widespread use in mainstream psychology, research in various languages has suggested that five dimensions may or may not be sufficient. Alternative models with larger number of factors in addition to the Big Five (Ashton, Lee, & de Vries, 2014; Cheung, Fan, & To, 2008; Cheung, Van de Vijver, & Leong, 2011; Nel et al., 2012; Valchev, Van de Vijver, Nel, Rothmann, & Meiring, 2013), smaller number of conceptually similar factors to the Big Five (De Raad et al., 2010, 2014; Saucier, Thalmayer, & Bel-Bahar, 2014; Saucier et al., 2014), as well as different factor structures that do not resemble the Big Five (Saucier, Thalmayer, & Bel-Bahar, 2014; Saucier, Thalmayer, Payne, et al., 2014) have been identified across a number of languages and cultural samples.

The psycholexical approach has been employed much less in text analysis. One line of research is exemplified by Chung and Pennebaker's (2008) study of self-descriptive narratives. These authors identified several dimensions such as Sociability, Evaluation, and Negativity, which only partly resembled the Big Five. A different approach can be found in some indigenous personality research, where fiction literature has been used as one of the sources for the identification of implicit personality concepts in a given culture. Such studies have found additional interpersonal factors in Chinese (e.g. Cheung et al., 2011) or only partially replicable Big Five factors with additional social factors in Hindi (Singh, Misra, & De Raad, 2013).

In summary, a long line of research has suggested that there is an underlying dimensionality when individuals use trait terms to describe themselves or others. Factors related to social relationships appear to be more malleable and unstable across languages and cultural contexts. Although mainstream personality research has adopted the five basic factors of personality as the basis for describing personality, both the exact number of dimensions and how to best rotate the content to factors remain under debate. Furthermore, the psycholexical approach has been underused in the study of large, naturally occurring texts that do not have an explicit focus on self-ratings or other-ratings. Expanding the psycholexical approach in the direction of text analysis of fiction literature, our study aims to reconstruct personality

DOI: 10.1002/per

dimensions in the works of two major English authors of the 19th century.

THE SOCIOCULTURAL CONTEXT AND POSSIBLE IMPLICATIONS FOR PERSONALITY TRAITS

The social context of England during the 19th century differed markedly from modern social conditions, with greater social hierarchies and lower individualism compared with modern times. The lexical hypothesis specifies that important traits are encoded in single terms. However, the combination of terms may still vary depending on the information that needs to be transmitted during social interactions (including storytelling). It is this co-association of behaviorally important single terms that has been analysed through factor analysis, and we may or may not find strong resemblance of those factors when analysing associations of described characters. The Big Five model is supposed to map onto some basic biological architecture (e.g. DeYoung, 2015; McCrae & Costa, 1999; McAdams & Pals, 2006). At the same time, the current literature on personality traits in non-Western contexts suggests that some factors may not emerge consistently and other factors, especially social or moral dimensions might be more differentiated, probably due to the greater importance of making more nuanced differentiations of such traits in more highly interconnected social settings (Fischer, 2017).

The writing by Austen was set during the Industrial Revolution in England, which marked great social transformations. Social hierarchies within traditional moral orders were slowly but steadily changing. Dickens already inhabited a world transformed by those changes, with many unskilled workers working long hours in factories or suffering in poverty in densely populated slums. As noted by literary scholars such as Vermeule (2010), literature provides readers with an opportunity to sample socially relevant information at a low cost because it allows us to reason about social contracts without having perfect access to relevant information yet not suffering the consequences of that imperfect knowledge. Hence, the social order and status hierarchies about to be transformed would form an important backdrop to character descriptions. A number of contemporary studies have indicated that social dimensions are more differentiated in more tightly organized interdependent cultural contexts (e.g. Cheung et al., 2011; Nel et al., 2012; Valchev et al., 2013). Since the 19th century was still more communitarian and hierarchical compared with contemporary English society, we could speculate that traits describing Agreeableness and Extraversion might have been highly relevant and differentiated. At the same time, it is not clear whether traits related to Openness or Conscientiousness would be as clearly visible or structured. Openness/Intellect is highly relevant in a society with a universal education system open to all members. Traits related to curiosity and exploration are particularly relevant to inform others about if you and your interlocutor have choices. In Georgian and Victorian times, those choices were restricted for most members of society. Even such obvious choices today such as choosing one's profession were dictated by family traditions, gender, and inheritance order rather than by ability or interest. Similarly, Conscientiousness captures traits related to personal efficiency and drive as well as being reliable and responsible. Modern trait dictionaries often apply these terms to work and education settings, which are most likely not applicable to conditions in historical England. Considering the greater importance of social etiquette and order, as well as the absence of a universal education system, it would be plausible that Conscientiousness-relevant terms emerge more broadly in relation to moral and social virtues (e.g. being interpersonally responsible or reliable and following social norms and traditions). Hence, one important question that has not been addressed previously is whether person descriptions in historical texts follow modern descriptive personality models.

JANE AUSTEN AND CHARLES DICKENS

Jane Austen (1775–1817) and Charles Dickens (1812–1870) are writers central to the development of the realist mode that came to define the 19th century novel in its rich depictions of the everyday, contemporary society, and individual psychology (Eagleton, 2005; Williams, 1973; Woloch, 2003). Although the two writers are only a generation apart, their novels are distinct in mode, style, and representational techniques. These differences can be attributed in part to the rapid transformation of British society in the two decades that separated the publication of Austen's final works following her death in 1818 and that of Dickens's first novel in 1836. Austen's novels are set in and around the country estates that provided the political and economic scaffolding of British society at the beginning of the 19th century. Her cast of characters is populated by the landed gentry and respected professionals such as lawyers and clergy, with individuals from higher (aristocratic) and lower (labourer and lesser professionals) ranks in the social hierarchy occupying more marginal positions within her character systems. Austen's novels register the shifting foundations of British society, with her final novel *Persuasion* elevating the virtues of the navy in favour of a declining aristocracy. However, the books are largely conservative in their orientation insofar as the marriage plot of each novel consolidates and strengthens the existing social hierarchy through the wedding of social authority to moral virtue (Butler, 1975; Duckworth, 1971).

Dickens, in contrast, is the great novelist of London, and his novels document a rapidly evolving urban landscape with its attendant social problems such as crime, sanitation, and poverty, as well as technologies such as the railway. As such, his novels feature a much more diverse range of characters, from rogues and orphans to a broadening array of middle-class professionals to aristocrats. Noted (and occasionally derided) for their strong sentimentality, Dickens's novels often make an appeal to social and individual moral reform by emphasizing interconnection and selflessness in opposition to the alienating effects of a world shaped increasingly by industrial capitalism. His work proved highly influential at a popular level, inspiring commercialized merchandise such as figurines, dolls, etc.

DOI: 10.1002/per

Clear differences are also seen in the two writers' styles and methods of characterization. Austen helped advance the representation of human interiority and psychology by transforming the epistolary novel—the novel of letters—through narrative techniques that depicted individual thought within third-person narration. Her novels offer rich depictions of the mind within its social setting and pay scrupulous attention to the dynamics of social interaction within a milieu where those interactions are shaped by protocols of decorum and propriety (Ferguson, 2000). Because these protocols generally promote forms of civility necessary to maintain social cohesion within a hierarchical order, individuals' true feelings or desire are often masked or obscured. As the original title to *Pride and Prejudice*—'First Impressions'—famously indicates, her heroines' narrative trajectories are structured around discovery or recognition, both of others and of themselves. Dickens' modes of characterization offer a stark contrast to this rich internal world, as his novels are notable for their general lack of attention to or depiction of the rich interiority of characters. Instead, his novels are populated by a larger cast of characters that capture society at a more expansive scale through multiple plots and numerous minor characters (Miller, 1965). While some have seen individuals in his novels as closer to caricatures than characters, the grotesque or deformed nature of many of the characters—exemplified by, for example, repetitive tics or idiosyncratic behaviours—can be interpreted as a manifestation and consequence of the economic order his novels dissect (Woloch, 2003). Hence, his person descriptions are more focused on externally visible behavioural characteristics.

THE CURRENT STUDY

We explore an automated text-mining approach to examine the personality dimensions that two 19th century authors, Jane Austen and Charles Dickens, used to describe their fictional characters. We describe the application of a recently developed method for unsupervised extraction of information from sequential data called Transcendental Information Cascades (TICs, Luczak-Roesch, Tinati, Van Kleek, & Shadbolt, 2015; Luczak-Roesch, Tinati, & Shadbolt, 2015; Luczak-Roesch, O'Hara, Dinneen, & Tinati, 2018), which seeks to overcome problems encountered by heavily contextualized automated text analysis approaches. For example, many machine learning methods use broad linguistic feature sets that work well when trained on and applied to contemporary data sources but may add contemporary bias when applying the methods to historic or non-English textual data (Da, 2019) or not perform as well as on modern texts (Daelemans & Hoste, 2002; Tahmasebi, Niklas, Theuerkauf, & Risse, 2010; Yang & Eisenstein, 2016). In the first instance, TICs treat all parts-of-speech (POS) features (e.g. all words or n-grams and POS token types such as nouns, verbs, adjectives, or pronouns) as low-level information tokens (i.e. they treat these simply as symbols that occur in sequential order over the course of a text) and then generate a temporally ordered network of the recurrence of tokens. For example, in this sentence from Pride and Prejudice: 'Mr. Bennet was so odd a mixture of quick parts, sarcastic humour, reserve, and caprice, that the experience of three-and-twenty years had been insufficient to make his wife understand his character', each word and grammatical feature would be given information value: 'Mr. Bennet' is the subject (coded in 1st and 2nd position), 'was' is the verb and encoded as 'to be' coded in the third position, 'quick' would be identified as an adjective and containing information in relation to the semantic meaning of 'quick' in 9th position within this sentence, etc. This creates a unique link between analyses that rely on higher order semantics derived from structural features of language (such as term co-occurrence; 'quick' and 'sarcastic' appear together in this sentence) and the low-level analysis of individual tokens' occurrences (e.g. frequency and periodicity; how many times is 'quick' used and how frequently within a specific segment) when any linguistic context is removed from the analytical view.

The work we present provides initial insights into the application of TICs to the problem of extracting psychological meaning from text in a bottom-up fashion. We focus our analysis and elaboration on dictionary-based psycholexical approaches to the problem of studying human personality from historical texts. We use a text-mining approach with the aim to identify core dimensions of personality that Austen and Dickens used when describing their characters. Instead of using human coders and content analysis based on psychological theory that may introduce contemporary bias, we use a bottom-up approach that identifies all trait terms from three different dictionaries as applied to fictional characters. We analyse the frequencies and co-occurrences of terms associated with each of the fictional characters as if these characters were human subjects in a factor analytic study, treating the co-occurrence of adjectives across characters as an input to factor analysis similar to survey studies that use the rating responses to survey items as an input. By running the analyses separately for the two authors, we can identify the implicit personality models that the authors used when describing their fictional characters.

Our research questions are descriptive in that we want to uncover the personality structure evident in novels. Specifically, our first research question is:

> *Research question 1*: What terms are used by these two authors?

What do the most frequent terms used by each author tell us about the personalities of the main characters (and possibly the author)? Because we are analysing a larger number of works, we also examine the relative stability of those frequencies of trait terms used across novels. Our central question is analysed next, namely:

> *Research question 2*: What is the emerging personality structure of fictional characters in the novels of Austen and Dickens?

There are two main steps that we need to address: (i) we need to decide how many factors should be extracted, and (ii) these structures need to be interpreted. We use a number of different statistical approaches to examine possible factor numbers. To interpret the structure, we compare the extracted

DOI: 10.1002/per

literature factors with a reference structure derived from student ratings (Ashton, Lee, & Goldberg, 2004), to marker scales (Goldberg, 1992), and via an interpretation of the unfolding factor structure. Our third question is:

> *Research question 3*: How similar or different are the factor structures for the two authors?

Do authors writing at different times within the 19th century and for different audiences use similar implicit personality dimensions? This question provides some first approximation of the idiographic vs. nomothetic question that becomes tractable with larger scale textual analyses.

PRE-REGISTRATION STATEMENT

Our project describes an exploratory application of a new data analysis approach to personality data in novels. Our study relies on textual analysis, therefore, we do not use sampling in a traditional sense. The authors and books were selected as they represent two important exemplars of 19th century literature. Austen has been credited with a rich internal description of characters, whereas Dickens has been argued to provide rich contextual descriptions of fictional characters. Our sample included all recognizable characters appearing in 21 novels by the two authors. Therefore, our sample encompasses the whole population of characters available to study within those sources.

Open material statement

We provide all the relevant information on procedures and measures in the Method section. The code, sources and dictionaries are available via https://osf.io/8bh3e/

Open data statement

The source data are publicly available via Gutenberg.org. The dictionaries and extracted raw data frames are available at https://osf.io/8bh3e/.

Reproducible script statement

Both the data processing scripts for the analyses reported here as well as more extended experimental applications to the data reported here are available at https://osf.io/8bh3e/.

Effects statement

We report descriptive statistics on word frequencies below. Our analysis is exploratory and given the statistical approach used, we do not report statistical significance levels. Tables 8 and 10 with 95% confidence intervals are included in the supporting information.

METHOD

Data sources

We used 6 novels by Jane Austen (*Sense and Sensibility*, *Pride and Prejudice*, *Emma*, *Mansfield Park*, *Northanger Abbey*, and *Persuasion*) and 15 novels by Charles Dickens (*The Pickwick Papers*, *Oliver Twist*, *Nicholas Nickleby*, *The Old Curiosity Shop*, *Barnaby Rudge*, *Great Expectations*, *A Tale of Two Cities*, *Dombey and Son*, *David Copperfield*, *Bleak House*, *Hard Times*, *Little Dorrit*, *Our Mutual Friend*, *Martin Chuzzlewit*, and *The Mystery of Edwin Drood*). The books were downloaded from Gutenberg.com in text format.

Dictionaries

We used three dictionaries (see https://osf.io/8bh3e/). The Allport and Odbert (1936) trait list is arguably the oldest and most comprehensive list of person-relevant terms in the English language (17 953 trait terms, full transcription of the list provided by Parker, Karl, Fischer, Luczak-Roesch, & Grener, 2019). This list has not been consistently evaluated using either self-reports or automated text analysis. We call this list ‘Allport’ from now on.

Saucier (1997) developed a set of 500 terms to capture broad and widely used person descriptors. These included dispositions, temporary conditions, social and reputational terms, and terms describing the appearance and physical characteristics of individuals. A three-factor structure provided the most robust solution across a community and smaller observer student sample. We call this the ‘500 list’ from now on.

Ashton et al. (2004) relied on an earlier dictionary of 1710 trait terms and data collected by Goldberg (1982), which presents the most comprehensive list of trait and disposition terms in English. In contrast to the list by Saucier (1997), the Goldberg list captures only trait-focused terms. Ashton et al. (2004) reanalysed a combined data set of Australian and U.S. self-reports and found seven factors, six of which resembled the HEXACO—the Big Five plus Honesty-Humility; and the seventh factor captured religiosity. Because this dictionary is the most comprehensive in-depth dictionary of trait terms and because we had access to the full factor loading results for all the solutions (Gerard Saucier, personal communication, Jan 10, 2019), we will use this dictionary as the reference solution (called ‘1,710 list’ from now on). We used the spelling as provided by the dictionary compilers.

Data processing

Fundamental to our analysis is the construction of TICs (Luczak-Roesch, 2016; Luczak-Roesch, Grener, Fenton, & Goldfinch, 2016; Luczak-Roesch, Tinati, Van Kleek, & Shadbolt, 2015; Luczak-Roesch, Tinati, & Shadbolt, 2015). TICs are a sequential data mining approach that transforms any kind of sequential data (e.g. the ordered sentences in a text) into a directed network of information token recurrence. An information token is any kind of distinct information that

DOI: 10.1002/per

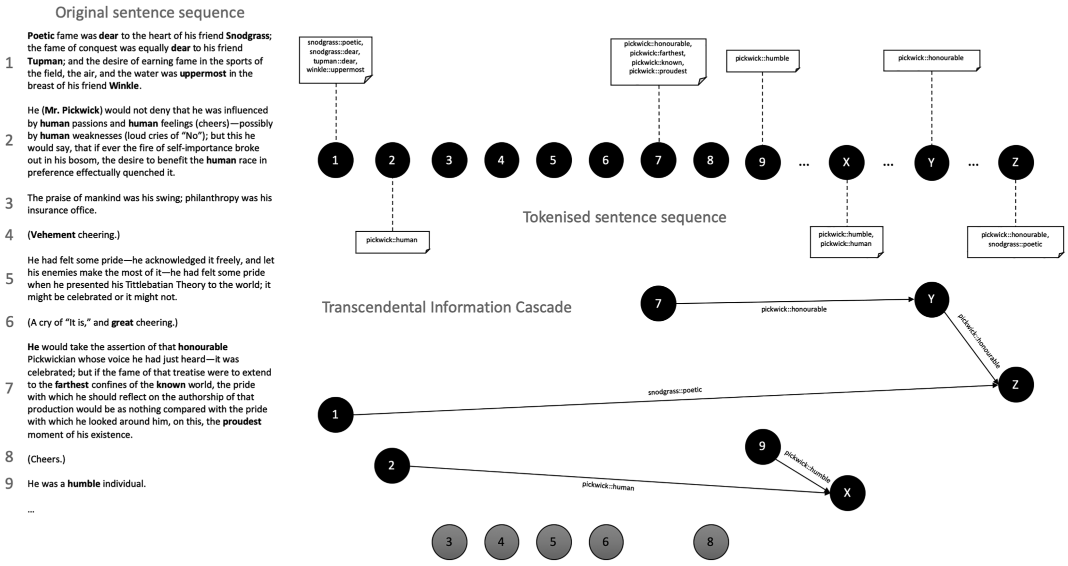

Figure 1. A schematic explanation of the Transcendental Information Cascade process. The example features nine consecutive sentences that occur in *The Pickwick Papers* by Charles Dickens. The Transcendental Information Cascade approach will represent each sentence as a network node (black and grey circles). In this example, *X*, *Y*, and *Z* are hypothetical sentences that may occur later in the book and which feature person-adjective tokens that were also found in any of the sentences 1–9. For each recurring token, an edge from the earlier to the later network node is inserted. Network nodes 3,4,5,6, and 8 do not feature any edges forward because they either did not contain any token match at all or a token match that does not recur at any time later in the text.

can be identified in source data using an information extraction method (e.g. matching adjectives using POS tagging).

To construct TICs, any element of a source sequence is analysed using a predefined information extraction method in chronological order from oldest to newest. In our case, we use natural language processing algorithms to extract trait adjectives matched to fictional characters. Any extracted tokens are kept in a distinct token set per sequence element. Then a network is created where each sequence element is represented by a unique vertex (vertices in networks are sometimes also referred to as nodes) that holds the sequential index as well as the token set as metadata attributes. A directed edge (link) is created between any two vertices that share a particular token as part of their token set but only if there is no vertex that occurs in between these two vertices in the overall sequence (i.e. has a sequential index between the two vertices to be linked). In other words, any edge allows to trace the path from the occurrence of an information token to the next occurrence of that exact token (see Figure 1 for a representation—the matched trait terms to fictional characters are used to create edges across the occurrences of references to fictional characters).

The motivation and strength behind this approach that combines token co-occurrence (via the token sets) and token recurrence (via the edges of the network) is that it preserves a relationship between low-level variances of information token co-occurrences and recurrences at the resolution of sentences, for example, and the macroscopic view to the aggregate state of an entire text or even corpus. For our work, this is promising because we suggest that significant fine-grained nuances in personality may be overridden when using highly aggregate approaches such as word embeddings or topic models because these approaches may penalize rare variance due to their statistical insignificance or because they do not allow to trace the path to the local context of the origin of a particular information token in the source data (aggregated patterns are not traceable to the location of their origin within a text; a common problem with text analysis algorithms).

We constructed different TICs for the same source data stemming from the aforementioned 21 English novels published by Charles Dickens and Jane Austen. Those novels were first transformed into sequential data by slicing the text into distinct sentences. For each sentence, we then examined three different tokenizations that can be realized using the state-of-the-art POS tagger available as part of the spaCy natural language processing library. spaCy is a python library for common natural language processing tasks that performs well in comparison with other state-of-the-art solutions (Al Omran & Treude, 2017; Spacy, 2019). Because our TIC toolchain is implemented in the R programming language, we instantiated spaCy through the spacyr interface (Benoit & Matsuo, 2018; Honnibal & Montani, 2017; see Figure 1 for an example).

First, we simply tokenized all adjectives matched in the respective text slices, identifying all adjectives in each sentence in each novel. Negations are detected using POS typing and the syntactic dependency tree so that we are able to tokenize negated adjectives with a prepended '#-'.

Second, we paired all adjectives with all the fictional characters that occur in the same slice as the adjectives. To detect fictional characters, we relied on the entity extractor of the spaCy library, which returns specific labels for named person entities. According to benchmark results, the validity of entity extraction of spaCy in reference texts is above 92% (Spacy, 2019). Whenever we detected explicitly named person entities in a slice, we stored these in a state object and resolved any personal pronouns in succeeding slices where no characters are mentioned by their names to this set of fictional characters in the state object. This processing allows to link adjectives to fictive characters across the boundary of a particular slice (sentence in our case) in which they were explicitly named. Despite the high level of validity, the entity extraction is not perfect, and we identified a few inconsistencies. For example, a number of entities were extracted that did not map onto name characters from novels (e.g. 'the darkness') or different entities were identified for identical characters (e.g., 'pickwick' and 'mr.pickwick'). To resolve these inconsistencies, the data were cleaned by a team of English literature scholars under supervision by the last author to identify overlapping entities and exclude entities that did not map onto named characters. We removed the entities that were not named characters and merged overlapping entities. This yielded cleaned data sets for each author in which only entities were included that were clearly identifiable as characters of the books. In the end, we had a total of 250 characters from Austen's novels and 1021 characters from the novels by Dickens. Therefore, in this step, we identified all fictional characters in each sentence as well as cross references to characters across sentences.

Third, we used the syntactic dependency tree for resolving which entities are linked to which adjectives. In case that a sentence did not match a named entity but contained personal pronouns, the algorithm linked the adjective to the last character in a preceding sentence that has been stored in the state object. In the case that no character-adjective matching could be achieved, the adjective was dropped from further analysis in this slice (sentence). Figure 1 depicts an abstract example of this approach.

We want to note at this point that dependency parsing and co-reference resolution are known to be technically very challenging, in particular for written language that does not follow a consistent and contemporary style (Peng, Khashabi, & Roth, 2015). To develop a dependency parsing and coreference resolution algorithm that is specifically focused on literary texts from the 19th century is beyond the scope of our line of inquiry here. Our preprocessing, therefore, uses intuitive heuristics for dependency parsing and coreference resolution, which are simpler but transparent (i.e. no complex modelling and additional inferences influence the output of this computational step) and scalable (i.e. it does not require the training of a model for any new corpus it is applied to).

At the end of the slicing and tokenization, our preprocessing resulted in a total of 63 distinct TICs represented by the respective nodes and links (edges) stored as CSV files (see Tables 1, and 2 for an example of this data representation in the first 19 sentences of *Great Expectations*). That is, we obtained one TIC per book (21 books) in a particular slicing (one in our case because we used sentence level slicing) and tokenization (three tokenizations). We only report results for adjectives matched to fictional characters because this is the conceptual equivalent to standard self-rated or observer-rated personality scales: the association of trait terms with fictional characters within each novel. Each TIC is materialized as (i) a file containing the sequence of vertices (one vertex/node per slice) with the matched tokens as a context attribute; (ii) a file containing the edges between vertices (i.e. an adjacency list) with the token that caused the edge to be created (i.e. the recurrence of an adjective); and (iii) an undirected network of co-occurring adjectives (i.e. adjectives that occur together within slices) as a single adjacency matrix with the number of co-occurrences as the values within that matrix. For our analysis, we are using the undirected network of co-occurring adjectives matched to characters as input to be processed for the factor analyses (the files with nodes and links for the 21 novels can be found at https://osf.io/8bh3e/).

Table 1. Example of the nodes output created by the Transcendental Information Cascade constructions algorithm from the first 19 sentences of the novel Great Expectations by Charles Dickens

Node ID	Extracted tokens
1	pip::christian, pip::explicit
2	
3	
4	
5	
6	
7	joe gargery::childish
8	joe gargery::little, joe gargery::neat, joe gargery::sacred, joe gargery::universal
9	
10	joe gargery::vivid
11	pip::afraid, pip::certain, pip::distant
12	abel magwitch::terrible
13	abel magwitch::little
14	abel magwitch::coarse, abel magwitch::fearful, abel magwitch::great
15	abel magwitch::old
16	
17	
18	
19	abel magwitch::quick
…	

Note: The table shows the tokens extracted when using the person-centric tokenizer.

Filtering for shared author dictionaries

We filtered the resultant matrices to only retain words shared with the three dictionaries. As discussed previously, we extracted all adjectives from the data set and matched them to fictional characters at the sentence level. This resulted in 1174 terms for Austen and 3387 terms for Dickens (including negations).[1] At this step, we now dropped all adjective-character matches that were not included in one of the three psychologically relevant dictionaries. For terms shared across authors, we found the greatest overlap between authors when using the 500 list: 1,710 (k = 198, 11.58%), 500 (k = 216, 43.20%), and Allport (k = 651, 3.65%). For the author-unique lists, we found similar results: 1,710 (Dickens: k = 516, 30.18%; Austen: k = 227, 13.27%); 500 (Dickens: k = 331, 66.20%; Austen: k = 221, 44.20%); and Allport (Dickens: k = 1913, 10.73%; Austen: k = 763, 4.28%). This low frequency rate of matching used adjectives to dictionaries indicates that contemporary dictionaries are not well represented in 19th century novels. It also shows

[1]The complete lists of adjective-character matches in both authors can be found at https://osf.io/8bh3e/.

DOI: 10.1002/per

Table 2. Example of the links output created by the Transcendental Information Cascade constructions algorithm from the first 19 sentences of the novel Great Expectations by Charles Dickens

Source node ID	Target node ID	Recurring token
8	47	joe gargery::little
63	64	joe gargery::young
64	65	joe gargery::young
65	66	joe gargery::young
66	67	joe gargery::young
67	68	joe gargery::young
68	69	joe gargery::young
27	74	abel magwitch::young
68	105	joe gargery::great
92	106	joe gargery::smooth
59	111	joe gargery::old
105	114	joe gargery::great
69	142	joe gargery::young
89	146	mrs joe gargery::great
93	148	joe gargery::good
148	153	joe gargery::good
153	154	joe gargery::good
47	164	joe gargery::little
146	165	mrs joe gargery::great
99	167	joe gargery::alone
…	…	…

Note: The table shows example 20 links that associate nodes for the *Great Expectations* example when using the person-centric tokenizer.

that Austen used a less diverse vocabulary to describe her characters.

Data analysis

Construction of correlation matrices

Based on the extracted data matrices containing term occurrences per character, we constructed Pearson correlation matrices for all terms in the selected dictionary. This is equivalent to constructing correlation matrices based on self-reported or other-reported ratings of persons in psychology. Instead of using ratings, we used the frequency of adjectives to create the correlation matrix.

Determining factor numbers

To identify the statistically optimal number of factors that should be extracted from our data, we (i) inspected the scree plot, (ii) ran parallel analysis, and (iii) ran Velicer's MAP (Velicer, 1976) on the correlation matrix of each author/dictionary combination. We used the *fa.parallel* and *vss* functions from the *psych* package (Revelle, 2018). Due to computational constraints given the data complexity, we performed 20 iterations for each parallel analysis.

Factor analysis

We used the *fa* function with a varimax rotation and a minimum residuals factoring from the *psych* package (Revelle, 2018).[2] We extracted one to seven factor solutions for each author and dictionary, as described in the Factor Structures section of the Results section.

[2]When using oblique rotation, the extracted factor structures were identical to the varimax solutions (congruence coefficients above.98 in all cases).

Factor cascades

To compute the factor cascades within each author/dictionary combination, we correlated the factor loading matrices of each n dimensional factor solution with the $n - 1$ dimensional solution.

Validation steps

Validation of bottom-up structures is challenging. Reviewing the literature, Boyd and Pennebaker (2017) noted that the verification of language-based personality models is typically done against self-ratings (or other-ratings) of personality traits in terms of the Big Five or some other measure (e.g. Schwartz et al., 2013; Yarkoni, 2010). For example, the classification of an individual based on textual analysis is compared with self-ratings on standard personality scales. These studies have shown some moderate convergence of text-based classifications using pre-existing dictionaries and standard psychometric tests. With unavailability of self-ratings for deceased individuals or fictional characters, observer (reader) ratings might be usable (Johnson, Carroll, Gottschall, & Kruger, 2011). The problem with this approach is that the structure and relevance of questionnaire-based trait dimensions cannot be taken for granted due to known problems with lack of self-insights, response biases, imprecise measurement procedures, and problems with applicability of those trait dimensions to specific individuals and cultural contexts (e.g. Boyd & Pennebaker, 2017; Cheung et al., 2011; Fischer, 2017; Uher & Visalberghi, 2016). Using pre-existing meaning categories also increases the risk of applying pre-existing biases in the interpretation of text. Boyd and Pennebaker (2017) noted as an alternative for verifying language-based analyses to use textual insights for prediction of personality-relevant real-world behaviours. This second approach has been shown to provide novel and nuanced insights for both describing and predicting specific behaviour (Boyd & Pennebaker, 2017) but is certainly limited when applied to fictional characters in novels or when examining textual descriptions of individuals who have long died. We therefore decided to use three steps to examine the validity of our structures.

Procrustes analysis

First, we examined the similarity with previously reported factor structures (Ashton et al., 2004). We used Procrustes factor rotation to rotate the factor structure extracted from the text to the factor structures identified in rating studies of real individuals and computed Tucker's phi coefficients (Fischer & Fontaine, 2010; Fischer & Karl, 2019; Van de Vijver & Leung, 1997) to examine factorial similarity. In psycholexical research, a threshold of .80 has been suggested as indicating sufficient similarity when dealing with structures that come from different sets of terms (De Raad et al., 2010).

Random data structures

Second, to evaluate the robustness of our solutions, we used a random data approach. Based on our previously generated data set, we generated two additional data sets containing

Table 3. Overlap between the Big Five marker lists and dictionaries by author

Austen					Dickens				
A	N	C	E	O	A	N	C	E	O
1,710									
Agreeable	Anxious	Careful	Active	Bright	Agreeable	Anxious	Careful	Active	Bright
Generous	Fearful	Steady	Bold	Deep	Generous	Fearful	Steady	Bold	Deep
Kind	Jealous	Practical	Talkative	Intellectual	Helpful	Jealous	Practical	Energetic	Intellectual
Pleasant	Nervous	Prompt			Kind	Moody	Prompt	Talkative	Imaginative
Warm	Envious	Thorough			Pleasant	Nervous	Systematic	Unrestrained	Philosophical
Considerate	Irritable				Warm	Envious	Conscientious	Vigorous	
					Considerate	Irritable	Thorough	Daring	
					Trustful	Fretful			
					Sympathetic	Emotional			
500									
Agreeable	Anxious	Careful	Active	Bright	Agreeable	Anxious	Careful	Active	Bright
Generous	Jealous	Neat	Bold	Intellectual	Generous	Jealous	Neat	Bold	Intellectual
Kind	Nervous	Practical	Talkative		Helpful	Moody	Practical	Energetic	Imaginative
Pleasant	Irritable	Prompt			Kind	Nervous	Prompt	Talkative	
Warm		Thorough			Pleasant	Irritable	Conscientious	Daring	
Considerate					Warm	Emotional	Thorough		
					Considerate				
					Sympathetic				
Allport									
Agreeable	Anxious	Careful	Active	Bright	Agreeable	Anxious	Careful	Active	Bright
Generous	Fearful	Neat	Bold	Deep	Generous	Fearful	Neat	Bold	Deep
Kind	Jealous	Steady	Talkative	Intellectual	Helpful	Jealous	Steady	Energetic	Intellectual
Pleasant	Nervous	Practical			Kind	Moody	Practical	Talkative	Imaginative
Warm	Envious	Prompt			Pleasant	Nervous	Prompt	Unrestrained	Philosophical
Considerate	Irritable	Thorough			Warm	Envious	Systematic	Daring	
					Considerate	Irritable	Conscientious		
					Trustful	Fretful	Thorough		
					Sympathetic	Emotional			

randomized data for each of the entities. First, we re-ran the extraction procedure of the initial data set but replaced the extracted entities across books randomly with an entity of the final data set without weighting for occurrence probability of the entity, resulting in a non-probabilistic randomized data set. Therefore, any fictional character in any book had an equal probability to be replaced. Second, we ran the same randomization weighted by the probability that an entity occurs matched with a word. In other words, less frequently mentioned characters were less likely to be chosen. This resulted in a probabilistic random character data set. This second data set preserves author specific choices about centrality of characters, which may carry over specific biases instead of a completely random approach as used for the first data set. We use these random factor structures for comparison purposes, both for determining optimal factor numbers and to compare the Procrustes analysis results.

Marker scales

Third, to aid in the interpretation of our factors, we created marker scores using the positive terms for each of the Big Five domains from Goldberg's (1992) list. It should be acknowledged that this is conceptually problematic as the marker scales only capture a narrow subset of the meaning of a factor (De Raad & Peabody, 2005); still, marker scales are informative for contextualizing these factors from a modern factor analytic perspective. Table 3 shows the overlap between each of the marker terms and the dictionaries by author. We created marker scores for each character by summing up the number of occurrences of all terms in a marker scale (e.g. active, bold, and talkative for Extraversion in the 500 dictionary in Austen's novels). We used only the positive terms to reduce the potential ambiguity of combining the frequency of positive and negative terms.

RESULTS

Research question 1: Descriptive analysis of the trait terms used

To examine the distribution of trait terms across data sets and dictionary, we extracted the 100 most common terms for each author. We show the results of both authors for comparison in Table 4 (if ties for the 100th spot occurred between term frequencies, all tied terms are shown). As can be seen there, the dictionaries show quite different terms; the choice of dictionary matters if the aim is to characterize individuals via salient terms.

To examine the overall psychological characteristics of the two authors' novels, we analysed the frequency of terms from the 1,710 dictionary. We adjusted the frequency of each word by the overall frequency of all terms from this dictionary. A first observation of these standardized frequencies

DOI: 10.1002/per

Table 4. List of the 100 most frequent trait terms per author and dictionary

Austen			Dickens		
1,710	500	Allport	1,710	500	Allport
Anxious	Little	Sure	Quiet	Little	Little
Kind	Good	Dear	Proud	Old	Dear
Agreeable	Happy	Little	Certain	Good	Old
Determined	Glad	Good	Natural	Young	Good
Natural	Poor	Happy	Cold	Glad	Sure
Silent	Afraid	Glad	Silent	Poor	Better
Certain	Pleased	Better	Kind	Afraid	Right
Capable	Satisfied	Poor	Anxious	Happy	Young
Cold	Anxious	Able	Bright	Alone	Glad
Serious	Young	Afraid	Pleasant	Great	Poor
Amiable	Ashamed	Sorry	Agreeable	Quiet	Afraid
Eager	Great	First	Curious	Proud	Sorry
Indifferent	Old	Impossible	Proper	Certain	Happy
Proud	Kind	Ready	Particular	Surprised	Last
Pleasant	Agreeable	Present	Honest	Natural	First
Proper	Determined	Possible	Gentle	Hard	Alone
Quick	Surprised	Last	Quick	Open	Late
Impatient	Natural	Pleased	Gracious	Short	Dead
Ignorant	Tired	Satisfied	Faithful	Strong	Great
Particular	Delighted	Anxious	Humble	Cold	Best
Foolish	Angry	Best	Cheerful	Difficult	Ready
Lively	Fair	Young	Dull	Tired	True
Cheerful	Certain	Least	Bold	Pretty	Quiet
Quiet	Handsome	Right	Wild	Well	New
Civil	Sensible	Ashamed	Warm	Bad	Necessary
Earnest	Alone	Fond	Cruel	Kind	Full
Reasonable	Bad	Aware	Childish	Beautiful	Whole
Warm	Comfortable	Likely	Jealous	Strange	Dark
Wild	Capable	Great	Deep	Anxious	Proud
Humble	Cold	Old	Friendly	Bright	Able
Confident	Serious	Necessary	Calm	Pleasant	Impossible
Intimate	Well	Ill	Careful	Fair	Easy
Rational	Eager	Kind	Generous	Rich	Present
Fearful	Open	Agreeable	Slow	Weak	Possible
Jealous	Proud	Determined	Affectionate	Innocent	Certain
Unreasonable	Unhappy	Surprised	Earnest	Ashamed	Surprised
Affectionate	Disappointed	Wrong	Serious	Handsome	Natural
Clever	Fortunate	Natural	Foolish	Plain	Hard
Vain	Pretty	Tired	Determined	Angry	Open
Constant	Strange	Delighted	Constant	Sensible	Short
Unkind	Difficult	Mistaken	Amiable	Satisfied	Strong
Generous	Pleasant	Angry	Patient	Agreeable	Cold
Gentle	Useful	Silent	Thoughtful	Curious	Silent
Simple	Plain	Fair	Tight	Busy	Free
Independent	Strong	Late	Indifferent	Comfortable	Least
Scrupulous	Charming	Certain	Steady	Sweet	Black
Wise	Proper	True	Timid	Worthy	Difficult
Obliging	Short	Easy	Confused	Proper	Tired
Calm	Delightful	Handsome	Gloomy	Delighted	Likely
Careless	Impatient	Sensible	Ignorant	Pleased	Aware
Cruel	Hard	Equal	Rough	Private	Ill
Faithful	Ignorant	Alone	Simple	Honest	Pretty
Selfish	Foolish	Bad	Nervous	Fine	Safe
Prudent	Secure	Comfortable	Original	Gentle	Well
Sincere	Uncomfortable	Full	Confident	Grateful	Bad
Steady	Rich	Real	Haughty	Unhappy	Kind
Modest	Concerned	Capable	Moral	Gracious	Clear
Punctual	Lively	Cold	Restless	Interested	Fond
Loud	Cheerful	General	Considerate	Faithful	Beautiful
Ungenerous	Interested	New	Capable	Thankful	Close
Elegant	Quiet	Serious	Clever	Cheerful	Small

(Continues)

Table 4. (Continued)

Austen			Dickens		
1,710	500	Allport	1,710	500	Allport
Severe	Thankful	Willing	Cool	Awake	Strange
Timid	Stupid	Safe	Merry	Remarkable	Anxious
Awkward	Fine	Amiable	Sensitive	Bold	Bright
Eloquent	Reasonable	Miserable	Fearful	Blind	White
Nervous	Warm	Well	Intimate	Familiar	Wrong
Shy	Excellent	Eager	Firm	Warm	Alive
Impertinent	Pleasing	Indifferent	Stern	Professional	High
Confused	Confident	Open	Modest	Cruel	Pleasant
Honest	Desirable	Proud	Brave	Childish	Conscious
Just	Extraordinary	Unhappy	Reasonable	Devoted	Pale
Original	Important	Disappointed	Expressive	Jealous	Fair
Solemn	Lucky	Absent	Solemn	Friendly	Rich
Unsuspicious	Nice	Fortunate	Unreasonable	Calm	Weak
Active	Jealous	Pretty	Mild	Careful	Innocent
Cautious	Unreasonable	Whole	Responsible	Generous	Ashamed
Deep	Affectionate	Strange	Wise	Lovely	Round
Dependent	Clever	Odd	Prudent	Affectionate	Handsome
Hearty	Generous	Superior	Tender	Important	Heavy
Imprudent	Gentle	Difficult	Selfish	Serious	Plain
Noisy	Private	Pleasant	Worldly	Foolish	Mad
Rude	Sweet	Unable	Dependent	Determined	Angry
Chatty	Tall	Useful	Graceful	Useful	Sensible
Excessive	Independent	Astonished	Loving	Lonely	Satisfied
Careful	Lovely	Fit	Respectful	Thin	Agreeable
Compassionate	Wise	Plain	Awkward	Youthful	Curious
Curious	Attentive	Strong	Practical	Ridiculous	Large
Expressive	Wonderful	Charming	Immovable	Charming	Red
Lazy	Beautiful	Proper	Eager	Patient	Unable
Reserved	Busy	Quick	Hearty	Tall	Busy
Thoughtful	Calm	Short	Impatient	Thoughtful	Comfortable
Dull	Careless	Sick	Polite	Uncomfortable	Sweet
Firm	Cruel	Delightful	Cordial	Excited	Worthy
Formal	Faithful	Impatient	Boyish	Confused	Proper
Merry	Grateful	Free	Genteel	Fortunate	Low
Mild	Selfish	Hard	Narrow	Hopeful	Particular
Tender	Interesting	Ignorant	Orderly	Ignorant	Delighted
Contrary	Sincere	Particular	Hospitable	Rough	Frightened
Indulgent	Unlucky	Personal	Independent	Delightful	Pleased
Polite	Modest	Small	Severe	Nervous	Private
Smart	Punctual	Unwilling	Benevolent		
Spirited	Ridiculous	Chief			
Violent		Foolish			
Affected		Secure			
Hasty		Uncomfortable			
Suspicious					
Thoughtless					
Extravagant					
Insolent					

Note: If ties for the 100th spot occurred between term frequencies, all tied terms are shown, leading to uneven numbers across dictionaries and authors.

is that *anxious* was the most frequently used word in Austen's novels when matched to characters (4.8%), followed by *kind* (3.5%), *agreeable* (3.2%), *determined* (3.2%), and *natural* (2.9%). Hence, there was a noticeable gap between the most frequent term *anxious* and subsequent terms. In contrast, the distribution of word frequencies was somewhat more even for Dickens, with the most frequently used word being *quiet* (3.6%), followed by *proud* (3.3%), *certain* (2.8%), and *natural* (2.8%). To provide more context, Figure 2 shows the standardized index for the 20 most frequently used terms per author (28 terms in total). A number of terms occur with similar frequencies in novels by both authors (e.g. *natural*, *silent*, *certain*, *cold*, *pleasant*, and *proper, quick*). However, there are also very clear distinctions. Austen makes considerably more use of words like *anxious*, *kind*, *agreeable*, *determined*, *capable*, *serious*, *amiable*, and *eager*. In contrast, Dickens uses *proud*, *quiet*, *bright*, *honest*, *curious*, and *gracious* considerably more often than Austen. The terms used by Austen have a more emotional and interpersonal orientation but also references to

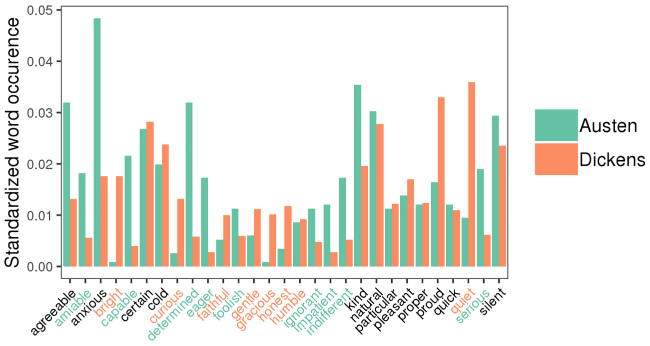

Figure 2. The 20 most frequently used terms from the 1,710 dictionary by author. Because the ranking of words was different, the total number of terms is 29. Words are coloured depending whether they appeared in the Top 20 for Austen (green), Dickens (orange) or were shared by both (black).

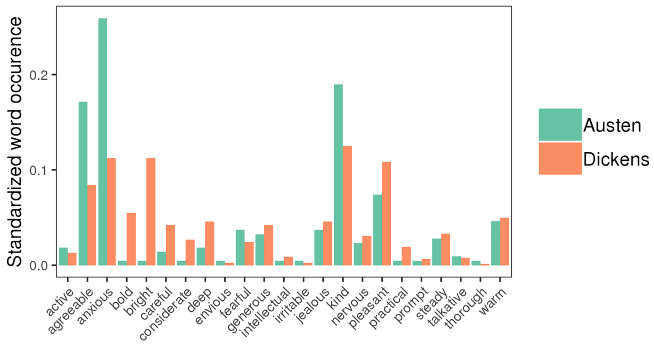

Figure 3. Relative frequency of the positive Big Five marker terms (Goldberg, 1992) matched across the two authors.

competence. In contrast, the most frequently used descriptors used by Dickens imply characters differing in dominance, trust, and mental capability.

Descriptive analysis of marker scales per author

To examine the psychological characteristics of the two authors' novels in the Big Five framework, we analysed the positive marker terms (Goldberg, 1992) used by the two authors across the 21 novels (the character parsed lists, independent of any dictionaries). We created a list that was unique to each author (e.g. capturing all marker terms) and a list that was based on terms shared between the two authors (e.g. only marker terms that were used by both authors were included). These lists were then standardized within each author, creating two new standardized lists: one list standardized for the relative frequency of these marker terms across all dictionaries and one list standardized for the relative frequency of all used marker terms by author.

Overall, Dickens used more terms from the marker list (38) compared with Austen (24). The most frequent terms for both authors were for Agreeableness and Neuroticism, whereas relatively few terms were used from the Openness and Extraversion lists.

When examining the relative term frequencies overall, the psychological make-up of characters in Austen novels was geared towards descriptions in terms of Agreeableness (51.4%) and Neuroticism (36.2%), with Conscientiousness (6.4%), Extraversion (3.2%), and Openness (2.8%) being of minor importance. Dickens' descriptions were also strongly biased towards Agreeableness (41.9%) but with less emphasis on Neuroticism (21.5%) and relatively greater importance for Openness (16.3%), Conscientiousness (12.2%), and Extraversion (8.1%) compared with Austen. These overall classifications mark interesting differences in specific word usage and may be related to the smaller number of characters in Austen compared with Dickens. Figure 3 shows the relative frequency of the shared terms. First, similar to the results noted above, it is noteworthy that fewer terms are used with greater frequency overall by Austen (*anxious*, *kind*, and *agreeable*), whereas there are less obvious distinctions between term usage in Dickens' novels. The most frequently used words by Dickens from the jointly matched list are *kind*, *pleasant*, *bright*, and *anxious*. However, as outlined above, a few words that are hardly used by Austen show some greater prominence in Dickens' vocabulary when describing characters (e.g. *bright*, *bold*, *deep*, and *careful*). In summary, Austen's descriptions matched to the positive Big Five marker scales show a preponderance of Agreeableness and Neuroticism; on the other hand, Agreeableness is the primary concern for Dickens, but there are also relatively frequent descriptions of characters in terms of all the other of the Big Five.

Stability of results per author

We next tested how stable the term frequencies were per author across books. We extracted the terms that are common to all books and then compared the rank order of each book with the overall rank order. Overall, the 500 list showed the highest stability (Austen r = .73, Dickens r = .74), followed by the Allport list (Austen r = .65; Dickens r = .58) and the 1,710 list (Austen r = .57; Dickens r = .54). Among Austen's books, Emma (r range from .63 to .78) and Pride and Prejudice (r range from .68 to .76) typically showed the highest similarity to the overall frequency distribution; whereas Persuasion showed the lowest similarity to the other works combined (range from .46 to .63). Among Dickens' works, Bleak House showed the highest similarity in term frequencies to the overall work (range from .64 to .86; but David Copperfield also showed high similarity for the 1,710 term frequencies with the overall corpus: r = .67); whereas Mystery was the least similar to other books (range from .37 to .63).

Research question 2: Factor structures

Parallel analysis, Velicer's MAP, and scree test

We ran factor analyses separately across all books for each author. Figure 4 shows the eigenvalues up to 20 factors for the three dictionaries for the two authors separately. For the 500 and 1,710 dictionaries, no clear bends were visible beyond a very strong first factor for both authors, with EVs levelling off after seven factors. For the Allport dictionary, a four-factor solution might be indicated for Austen. For both authors, the EVs level off after eight factors.

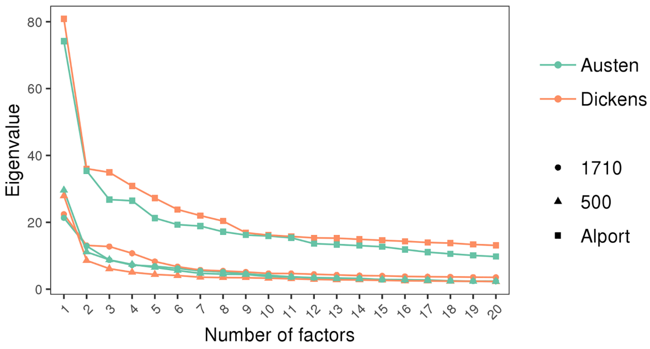

Figure 4. Scree test for the three dictionaries by author.

Parallel analysis suggested extracting 11 factors for the Austen novels when using both the 500 and the 1,710 dictionaries; for the Allport dictionary, a total of 22 factors were above a random threshold. Using Velicher's MAP, the analysis of the 500 and 1,710 dictionaries for Austen suggested six factors, whereas for the Allport dictionary, it suggested two factors. In contrast, for the novels by Dickens, parallel analysis suggested two factors for the 500 dictionary, four factors for the 1,710 dictionary, and five factors for the Allport dictionary. With Velicer's MAP, the results suggested extracting eight factors for all the dictionaries for Dickens.

For the random structures, parallel analyses typically suggested no factors or a single factor. Hence, our character-based factor structures do not appear to be driven by random word associations.

Procrustes analysis with the 1,710 dictionary self-rating data

We used the factor structure reported by Ashton et al. (2004) as target for Procrustes rotation. We compared our three, five, six, and seven factor loadings with the respective factors from Ashton et al., excluding trait terms not observed in our data. Given the inability to adequately capture negations, we used the absolute values of the loadings in the target matrices for examining factor similarity. Table 5 shows the overall results. The highest congruence coefficients did not exceed .54 for Austen and .41 for Dickens. These indices of factor similarity are well below the suggested lower threshold of .80 (De Raad et al., 2010). Hence, our textual factor analyses did not recover a structure directly corresponding to the common lexical solutions of three to seven factors from self-rating data.

At the same time, it is important to examine how similar a random structure would be to the self-rating-based factor structures. When comparing our observed congruence

Table 5. Congruence coefficients between the self-rating solution (Ashton et al., 2004) used as a target and the Procrustes-rotated factors based on the character analysis, random samples, and random probability samples using the 1,710 trait dictionary

	Austen				Dickens			
	Three factors	Five factors	Six factors	Seven factors	Three factors	Five factors	Six factors	Seven factors
Factors Character analysis								
1	.46	.47	.48	.45	.42	.40	.42	.41
2	.39	.54	.48	.40	.40	.41	.40	.40
3	.49	.38	.40	.49	.43	.41	.40	.41
4		.41	.44	.45		.33	.34	.33
5		.30	.31	.33		.24	.22	.22
6			.33	.30			.28	.24
7				.35				.28
Mean	.45	.42	.41	.40	.42	.36	.34	.33
Random sample								
1	.20	.19	.19	.19	.10	.08	.07	.09
2	.13	.18	.20	.18	.05	.06	.07	.08
3	.13	.18	.18	.20	.04	.07	.08	.06
4		.05	.08	.08		.03	.08	.08
5		.08	.15	.16		.07	.04	.05
6			.09	.14			.09	.18
7				.15				.11
Mean	.15	.14	.15	.16	.06	.06	.07	.09
Random probability sample								
1	.37	.35	.39	.41	.21	.27	.26	.29
2	.39	.41	.37	.41	.23	.29	.29	.24
3	.38	.40	.41	.38	.22	.23	.24	.31
4		.32	.33	.36		.20	.20	.22
5		.23	.21	.20		.24	.17	.16
6			.27	.20			.18	.12
7				.28				.19
Mean	.38	.34	.33	.32	.22	.25	.22	.22

DOI: 10.1002/per

Table 6. Highest loading terms in the seven-factor solution in Austen’s novels based on the set of 1,710 terms (227 terms in analysis)

[Factor 1] Civility
ingenious (.99), brutal, clownish, envious, fanciful, graceful, irritable, womanly, unfeeling, prejudiced, particular, obliging, kind, quick, unreasonable, hasty, insolent, eager, rational, compassionate, contrary, fearful, merry, amiable, impatient, determined, ignorant, natural, unsuspicious, original, confused, just, cautious, nervous, cruel, cold, gentle, jealous, vain, confident, humble (.31)

[Factor 2] Intelligence
biased (.98), practical, philanthropic, unsophisticated, timid, deep, charitable, uninformed, curious, silent, serious, dull, tender, mild, hasty, unreasonable, solemn, dependent, noisy, steady, foolish, eloquent, nervous, ungenerous, lively, sincere, negligent, modest, cruel, capable, affectionate, simple, earnest, independent, rational (.30)

[Factor 3] Approachability
haughty (.97), indirect, social, consistent, dignified, intelligent, proud, affected, civil, cautious, honest, original, ungenerous, selfish, punctual, faithful, generous, impertinent, wise, amiable, certain, jealous, silent, clever, ignorant (.31)

[Factor 4] Vigour
bright (.98), cool, explicit, feminine, invariable, slow, independent, simple, lazy, suspicious, active, pleasant, serious, quiet (.34)

[Factor 5] Egocentrism
greedy (.95), resentful, unguarded, impertinent, stubborn, rude, extravagant, honest, smart, noisy, dull, solemn, constant, unsuspicious, anxious, ungenerous, faithful, careless, quick, simple, humble, wild (.30)

[Factor 6] Sternness
reasonable (.65), severe, confused, reserved, formal, mercenary, lively, intimate, expressive, polite, certain, proper, indifferent, dependent, elegant, natural, determined, capable, agreeable, serious, amiable, foolish, generous, cheerful, quiet, anxious, earnest, constant, silent, faithful, eloquent, wild, humble, eager, nervous, genteel, straightforward, illiterate, affectionate (.30)

[Factor 7] Prudence
moderate (.92), sensitive, speedy, manly, prudent, unkind, uninformed, contrary, wise, generous, selfish, calm, affectionate, certain, anxious (.33)

Note: N = 250 character entities. Terms with loadings at or above .30 are presented. The terms are listed in decreasing order of their loadings; the highest and lowest loading in each factor are indicated in parentheses.

coefficients with those obtained after rotating the random character data and random probability character data to the rating-based structure, the coefficients from the two random data sets were considerably lower than the observed ones (see Table 6). For example, for the seven-factor solutions, we found average congruence coefficients for Austen and Dickens, respectively, of .40 and .33 in observed data, .16 and .09 in random data, and .32 and .22 in random probability data. These indices suggest that although the character-based structure did not meet structural equivalence with rating data, randomly created character sets showed considerably lower congruence coefficients with rating data, suggesting that the character-based textual analysis is different from randomly constructed trait matrices and shows higher but not sufficient similarity with rating-based factor structures.

Factor structure description

Because both the 1,710 and 500 dictionaries had shown interpretable (although different) seven-factor structures with rating data in previous studies, we extracted one to seven factors for each of the dictionaries separately by author. In order to compare the emerging structures, we examined the similarity of factor loadings across dictionaries for matching terms. We used a criterion for correlations of .80 to identify factors that showed high similarity with each other across dictionaries and correlations of .50 to identify factors that showed fair similarity across dictionaries. For purposes of better presentation, we use the 1,710 dictionary again as the reference solution and then compare the 1,710 dictionary separately with both the 500 and Allport dictionaries.

We base our presentation on the seven-factor solution for the sake of comprehensiveness. Judging from the scree test, seven factors may be too many for the 1,710 and 500 dictionaries; still, it is informative to see what factors emerge if we allow over-extraction. In general, the meaning of the factors changed only slightly with the extraction of additional factors. In the following section, we trace the sequence of emergence of factors from solutions with lower dimensionality.

Table 7. Correlations of the factors of the seven-factor solution in the 1,710 dictionary with the factors in the 500 and Allport dictionaries in Austen’s novels

	5.2	A.2	5.5	A.1	5.4	A.5	5.3	A.7	5.6	A.4	5.1	A.3	5.7	A.6
Civility	**1**	**1**	.07	.10	.01	.01	−.12	−.13	−.11	−.04	.13	.09	−.15	.01
Intelligence	.06	.07	**.93**	**.99**	−.18	−.20	.00	−.06	−.02	.06	.48	.33	−.13	.02
Approachability	.02	.01	−.27	−.20	**.97**	**.97**	−.12	−.09	−.04	.02	.08	.10	.00	.12
Vigour	−.16	−.19	.02	.02	−.14	−.20	**1**	**.96**	−.16	−.10	−.11	−.14	.01	−.23
Egocentrism	−.03	−.04	−.04	−.03	−.02	.01	−.07	−.12	**.92**	**.99**	.02	−.02	−.21	.12
Sternness	.10	.13	.11	.23	−.05	−.04	−.12	−.05	.05	.02	**.86**	**.93**	.23	.10
Prudence	−.05	−.04	−.18	−.05	.19	.17	−.14	−.19	.07	.03	.42	.42	−.21	**.67**

Note. Factor labels starting with 5 indicate the factors in the 500 dictionary and those with A indicate factors in the Allport dictionary (see text for dictionary description). The factors of the 1,710 dictionary are presented in order of extraction. Correlations over .50 are in bold.

DOI: 10.1002/per

Table 8. Highest loading terms in the seven-factor solution in Dickens' novels based on the set of 1,710 terms (516 terms in analysis)

[Factor 1] Approachability
arrogant (.98), inconstant, reproachful, icy, derogatory, haughty, disdainful, overbearing, audacious, sullen, extravagant, unapproachable, uninformed, complimentary, judicious, elegant, dignified, discreet, taciturn, natural, sensitive, moody, rugged, timid, proud, blunt, dull, social, intellectual, cold, submissive, distant, severe, observant, jealous (.31)

[Factor 2] Dominance
argumentative (.99), negative, humorous, humane, meditative, heartless, treacherous, relaxed, philosophical, fraudulent, philanthropic, jovial, faithful, productive, hospitable, fickle, awkward, cunning, invariable, prompt, inconsistent, intelligent, speedy, talkative, considerate, eloquent, exact, certain, aloof, restless, inflexible, amiable, smart, spirited, merciful, careful (.31)

[Factor 3] Sociability
forgiving (.99), frolicsome, tolerant, venturesome, unobtrusive, unpretending, pious, condescending, sociable, pensive, jovial, humble, lavish, noisy, frivolous, judicious, angelic, industrious, constant, moral, lenient, gentle, ambitious, moody, zealous, dutiful, agreeable, tight, blunt, pleasant, proud, harsh, thoughtful, vindictive, loud, affected, unmindful, severe, bold, careful (.30)

[Factor 4] Civility
unsocial (.98), guileless, urbane, wilful, chatty, relentless, whimsical, extravagant, inconsiderate, simple, shrewd, lenient, taciturn, brave, intelligent, serious, careful, conversational, worldly, mild, dissatisfied, warm, bashful, nervous, honest, ignorant, stern, pleasant, genial, contented, gallant, lively, harsh, amiable, certain, gracious, deep (.31)

[Factor 5] Integrity
changeable (.94), deceitful, demure, vivacious, moderate, courageous, unvarying, mischievous, industrious, complacent, wise, manly, inward, serene, merry, spirited, wild, unreasonable, punctual (.30)

[Factor 6] Dynamism
impressible (.56), quiet, anxious, prejudiced, generous, sly, rational, vain, patient, earnest, silent, wild, bright, kind, rigid, inexhaustible, proud, certain, determined, natural, independent, childish, steady, compassionate, sincere, clumsy, affectionate, unapproachable, grumpy, lawless, proper, pleasant, inflexible, mild, spirited, speedy, warm, insolent, quick (.30)

[Factor 7] Activity
adventurous (.82), defiant, forbearing, girlish, chatty, kind, angelic, restless, ingenuous, serene, resentful, gentle, earnest, cruel, smart, loving, fierce, sullen, solemn (.30)

Note. N = 1021 character entities. Terms with loadings at or above .30 are presented. The terms are listed in decreasing order of their loadings; the highest and lowest loading in each factor are indicated in parentheses.

Table 9. Correlations of the factors of the seven-factor solution in the 1,710 dictionary with the factors in the 500 and Allport dictionaries in Dickens' novels

	5.4	A.2	5.2	A.4	5.3	A.3	5.5	A.6	5.6	A.1	5.1	A.7	5.7	A.5
Approachability	**.96**	**.99**	−.06	−.07	.02	.06	−.07	.03	.14	.18	.10	.14	.11	.11
Dominance	−.09	−.09	**1**	**1**	.00	−.01	.08	.10	−.02	.03	−.01	−.01	.11	.18
Sociability	.04	.06	.00	.01	**.98**	**1**	−.14	.06	.01	.06	−.05	−.05	.11	.00
Civility	.00	.03	.17	.16	.07	.03	**.79**	**.96**	.17	.36	.31	.17	.33	.22
Integrity	.16	.01	.05	.04	.01	−.03	.33	.14	.00	.37	.26	.19	.18	.22
Dynamism	.27	.14	−.01	−.01	.02	−.03	.35	.30	**.55**	**.64**	**.80**	**.78**	.28	.29
Activity	.16	.08	.00	−.02	.03	−.04	.01	−.19	.28	**.74**	.03	−.20	**.55**	.19

Note. Factor labels starting with 5 indicate the factors in the 500 dictionary and those with A indicate factors in the Allport dictionary (see text for dictionary description). The factors of the 1,710 dictionary are presented in order of extraction. Correlations over .50 are in bold.

The highest loading terms per factor in the seven-factor solution in the 1,710 dictionary in both authors are presented in Tables 7 and 9, respectively; the terms from the solutions in the other two dictionaries are presented in the supporting information. The correlations of matching terms in the 1,710 dictionary with those in the other two dictionaries are presented in Tables 8 and 10. To interpret the structure, we use the associations with the positive terms from Goldberg's (1992) marker scales to provide an interpretation in light with contemporary associations. We use as a criterion marker-scale correlations of .20 and above for both authors.

For the novels by Austen, the first factor captured Civility-related terms. High loading terms on this factor included *graceful*, *ingenious*, *clownish*, and *irritable* (Table 6). The highest correlations of the factor were with the marker scales for Neuroticism (r = .52) and Agreeableness (r = .49). This factor replicated as Factor 2 in the 500 dictionary (r = 1) and Factor 2 (r = 1) in the Allport dictionary (Table 7).

The second factor suggested Practical Intelligence and Social Maturity. High loading terms included *practical*, *philanthropic*, *unsophisticated*, and *uninformed*. The highest correspondence with the Big Five marker scales were for Openness (r = .68) and Conscientiousness (r = .52), and there were smaller correlations with Neuroticism (r = .25) and Agreeableness (r = .24). This factor replicated as Factor 5 in the 500 dictionary (r = .93) and Factor 1 in the Allport dictionary (r = .99).

DOI: 10.1002/per

The third factor captured terms related to Social Humility/Approachability. High loading terms included *haughty*, *dignified*, *proud*, and *social*. This factor's highest marker-scale correlation was r = .25 with Agreeableness. This factor closely approximated Factor 4 in the 500 dictionary (r = .97) and Factor 5 in the Allport list (r = .97).

The fourth factor was defined by terms that indicated Vigour. High loading terms included *bright*, *explicit*, *cool*, and *slow*. The highest correlation was with Extraversion marker scales (r = .41), followed by correlations with Openness (r = .35) and Agreeableness (r = .29). This factor corresponded to Factor 3 (r = 1) and Factor 7 (r = .96) in the 500 and Allport dictionaries, respectively.

The fifth factor was characterized by terms indicating Egocentrism and Selfishness, possibly with a Negative Valence dimension. High loading terms included *resentful*, *greedy*, *impertinent*, and *rude*. The highest correlations with marker scales were with Neuroticism (r = .36) and Agreeableness (r = .21). This factor corresponded to Factor 6 in the 500 list (r = .92) and Factor 4 in the Allport list (r = .99).

The sixth factor was best described as Sternness in the 1,710 dictionary. High loading terms included *severe*, *reasonable*, *reserved*, and *formal*. The highest correlations were with the Agreeableness (r = .36) and Neuroticism marker scales (r = .33). This factor corresponded to Factor 3 in the Allport list (r = .93) and Factor 1 in the 500 list (r = .86).

The seventh factor was best described as Prudence. High loadings included *moderate*, *speedy*, *sensitive*, and *prudent*. It correlated with the Neuroticism marker scale (r = .28). This factor had weak similarity with Factor 1 of the 500 list (r = .42) and both Factors 6 (r = .67) and 3 (r = .42) from the Allport dictionary.

Examining unique factors in the other two dictionaries that were not apparent in the 1,710 dictionary, one factor in the 500 dictionary was found (Factor 7) that captured primarily terms related to Social Energy/Extraversion (e.g., *prominent*, *joyful*, *friendly*, *devoted*, and *shy*). However, this factor did not correlate with any of the Big Five marker scales.

To summarize the factor structure of the Austen novels, the factors showed some conceptual overlap with five factor marker scales but suggested a primary focus on social-normative aspects of personality. Even the Practical Intelligence factor, which could be considered as related to intellect, had content dealing with social orientation, such as *philanthropic* and *charitable*. Six of the seven factors showed quite strong convergence across the dictionaries. One factor of the 1,710 dictionary, Prudence, was unrelated to any of the marker scales but showed some commonality with factors in the other two dictionaries.

For Dickens, the first factor in the 1,710 list captured trait terms describing arrogance with high Social Power vs. Social Humility/Approachability. High loading terms included *arrogant*, *derogatory*, *icy*, and *reproachful* (Table 8). Correlations with the marker scales suggested a weak correlation with Openness (r = .23). This factor corresponded to Factor 4 of the 500 dictionary (r = .96) and Factor 2 of the Allport dictionary (r = .99; Table 9).

The second factor suggested aspects of Power and Dominance, captured in terms expressing both negative and positive interpersonal effects. High loading terms included *argumentative*, *negative*, *humorous*, and *humane*. Correlations with marker scales suggested a weak association with Conscientiousness (r = .23) and Openness (r = .21). This factor corresponded to Factor 2 in the 500 dictionary (r = 1) and Factor 4 in the Allport dictionary (r = 1).

The third factor showed high loadings from terms indicating Sociability both in the sense of gregariousness and benevolence. High loading terms included *forgiving*, *tolerant*, *frolicsome*, and *venturesome*. This factor correlated with markers for Agreeableness (r = .23) and Extraversion (r = .20). The factor corresponded to Factor 3 in the 500 dictionary (r = .98) and Factor 3 in the Allport dictionary (r = 1).

The fourth factor captured trait terms related to Civility. High loading terms included *urbane*, *guileless*, *wilful*, and *inconsiderate*. This factor correlated with Agreeableness (r = .32), Conscientiousness (r = .27), and Openness (r = .20) markers. The factor shared some similarity with Factor 5 in the 500 dictionary (r = .79) and strongly resembled Factor 6 in the Allport dictionary (r = .96).

The fifth factor consisted of terms indicating a sense of Integrity and personal stability. Some of the high loading terms were *deceitful*, *changeable*, *demure*, and *moderate*. This factor showed no correlation with any of the marker scales. The factor was also unique to the 1,710 dictionary and showed no similarity with factors extracted in the 500 and Allport dictionaries.

The sixth factor resembled a (lack of) Dynamism factor. It included high loadings of terms such as *impressible*, *quiet*, *anxious*, and *sly*. This factor showed correlations with markers from all of the Big Five (Neuroticism = .37; Agreeableness = .39; Conscientiousness = .29; Openness = .26; Extraversion = .21). This factor showed similarity with Factors 1 (r = .80) and 6 (r = .55) in the 500 dictionary as well as Factors 7 (r = .78) and 1 (r = .64) in the Allport dictionary. Hence, the terms of this factor spread across two separate factors in the other two dictionaries. We labelled this factor '(lack of) Dynamism' because the factors in the other two dictionaries had a stronger flavour of success and dynamic externality, whereas in the 1,710 factor, there was an element of reservedness or restraint, with both an emotional and potentially manipulative connotation.

Finally, the seventh factor captured terms that resembled Activity. High loading terms included *adventurous*, *chatty*, *forbearing*, and *defiant*. This factor showed no correlation with the Big Five marker scales and had some weak similarity with Factor 7 (r = .55) in the 500 list and stronger similarity with Factor 1 (r = .74) in the Allport list.

Examining the data from the view of the other two dictionaries, Factor 5 in the Allport dictionary and Factor 7 in the 500 dictionary shared some similarity (r = .81) that was not shared with any factor of the 1,710 structure. Shared terms included *confusing*, *skilled*, *cheap*, *selfish*, *disappointed*, and *bitter*. Some of the other highest loading items of the Allport list included *infected*, *minor*, *weakened*, *undoubted*, and *theatrical*; this set of terms suggested a Social Status or Social Success dimension. This factor correlated mostly with the marker scales for Neuroticism (r = .37 and .46) and

DOI: 10.1002/per

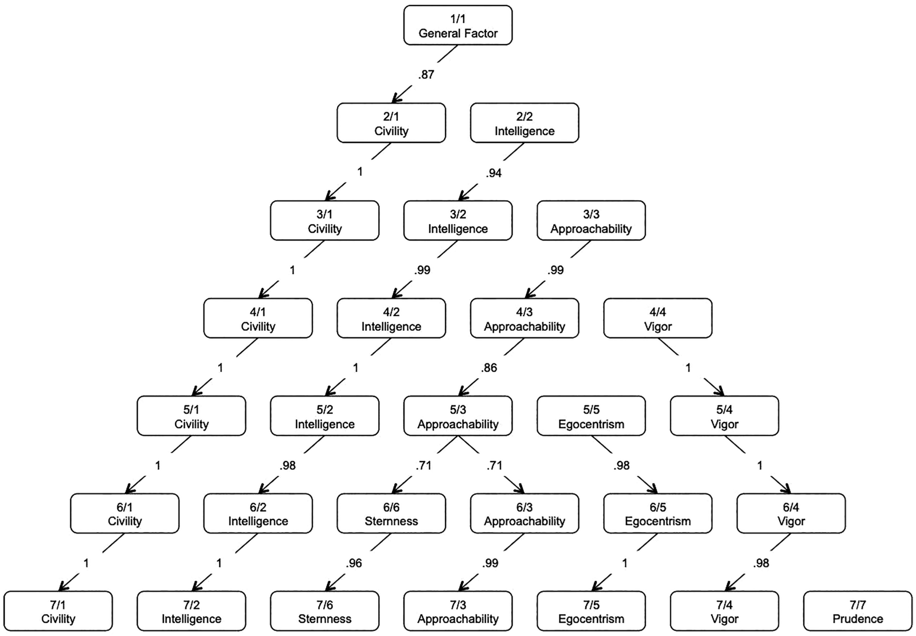

Figure 5. Factor cascades for the Austen novels, using the 1,710 dictionary. Correlations above .50 are shown.

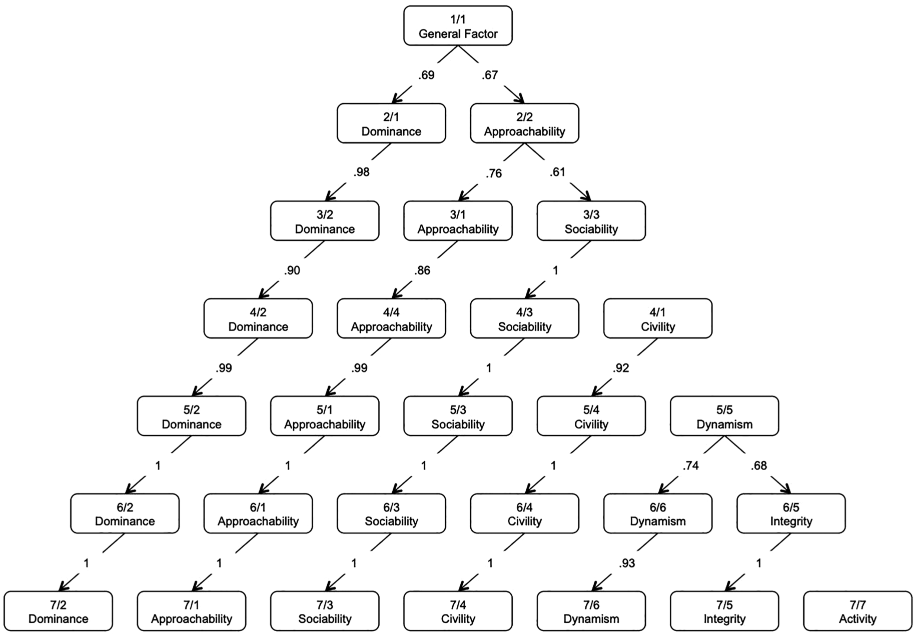

Figure 6. Factor cascades for the Dickens novels, using the 1,710 dictionary. Correlations above .50 are shown.

Agreeableness (r = .21 and .34 in the Allport and 500 dictionaries, respectively).

In summary, similarly to the findings for Austen, the factors identified in Dickens' novels had some relations to combinations of markers of the Big Five but had their own idiosyncratic content revolving around themes of social relations such as arrogance, dominance, sociability, and civility. Five of the seven factors within the 1,710 dictionary for Dickens showed high correspondence with factors in the other two dictionaries. One factor, Integrity, was unique to the 1,710 factor structure.

Factor cascades

A complementary perspective to understand the overall factor structure is to examine the unfolding of factors. This differentiation for the 1,710 dictionary is visually shown as factor cascades in Figures 5 and 6. Starting with Austen, the general factor turned into Civility, with the Practical Intelligence/Maturity factor emerging as a second factor in the two-factor solution. These two factors remained stable for the rest of the other solutions. With three factors, the Social Humility/Approachability factor emerged. At the level of four factors, Vigour emerged and remained a stable factor for

DOI: 10.1002/per

Table 10. Congruence coefficients (Tucker's Φ) between the Austen and Dickens solutions for the one-factor to seven-factor structures after Procrustes rotation

Dictionary	1	2	3	4	5	6	7
1,710	.70	.57	.56	.57	.24	.25	.24
		.43	.30	.26	.37	.33	.35
			.48	.39	.54	.42	.41
				.32	.36	.52	.44
					.32	.35	.27
						.18	.50
							.20
500	**.83**	.73	.68	.59	.55	.53	.51
		.55	.58	.50	.64	.42	.62
			.43	.63	.58	.60	.34
				.40	.42	.34	.38
					.34	.56	.52
						.44	.48
							.38
Allport	**.81**	.69	.67	.53	.47	.41	.34
		.50	.48	.44	.56	.32	.33
			.36	.53	.33	.33	.36
				.35	.37	.38	.37
					.40	.55	.59
						.44	.36
							.34
Mean							
1,710	.70	.50	.45	.39	.37	.34	.34
500	**.83**	.64	.56	.53	.51	.48	.46
Allport	**.81**	.60	.50	.46	.43	.41	.38

Note. Congruence coefficients above .80 are bolded.

the remaining solutions. When extracting five factors, the Egocentrism factor was added and remained stable thereafter. With six factors, Sternness split out from Social Humility/Approachability. Finally, the last factor to emerge in the seven-factor solution was Prudence. Similarly, the factor that was relatively unique to the 500 dictionary (Social Energy/Extraversion) emerged last when extracting seven factors.

For Dickens, the emergence of factors followed a different trajectory. The general factor first split into Dominance and Social Humility/Approachability. This social factor further split into Social Humility/Approachability and Sociability when three factors are extracted. These three factors then stayed relatively unaltered when more factors are extracted. With four factors, the Civility factor emerged. When extracting five factors, a new factor emerged, broadly interpreted as (lack of) Dynamism, which subsequently split into (lack of) Dynamism and Integrity in the six factor structure. The final factor to emerge was Activity.

Research question 3: Similarity of structures across authors

As suggested by the parallel descriptions in the previous sections, the factor structures based on the novels by the two authors appeared to differ substantively, yet some common themes seem to have emerged. To examine the similarity more quantitatively, we extracted common terms from each dictionary that were used by both authors, ran a new factor analysis on the shared terms, and then assessed the similarity of terms included in both analyses using Procrustes rotation. The solutions within each author (e.g. the complete list vs. the list that was shared across the two authors) remained highly similar, mean congruence Austen: 1,710 (M = .96, SD = .08), 500 (M = 1, SD = 0), Allport (M = 1, SD = .01); Dickens: 1,710 (M = .91, SD = .09), 500 (M = .97, SD = .04), and Allport (M = .98, SD = .02). Across the two authors, the factor similarity was statistically low, especially when extracting more than three factors (Table 10). The single-factor solutions showed relatively high similarity overall, except for the 1,710 dictionary. One likely explanation for this strong convergence might be that the similarity is driven by overall word frequency effects. The first factor for both authors strongly correlated with word frequency, ranging from .50 (Austen: 1,710 dictionary) to .62 (Dickens: 500 dictionary). For the final seven-factor solution, the average congruence coefficients were .34, .46, and .38 for the 1,710, 500, and Allport dictionary, respectively. The highest congruence between factors did not exceed .62 (which was found for the 500 dictionary).

It is interesting to compare descriptively two factors that had some similarity in broad meaning between the two authors and yet were defined using different terms: Civility (Factor 1 in Austen and 4 in Dickens) and Approachability (Factor 1 in Dickens and 3 in Austen, both in the seven-factor solution in the 1,710 dictionary). Civility in both authors described the general quality of interaction of the individual in his/her social environment and included, apart from terms that would fall under agreeableness in a contemporary framework, terms related to emotions and to

DOI: 10.1002/per

intellect. Focusing on the 10 highest loading terms, Civility in Austen included terms dealing with elegance and disruptiveness (e.g. *graceful*, *ingenious* vs. *clownish*, and *irritable*); Civility in Dickens, on the other hand, included terms focusing on the intentionality of the individual (e.g. *wilful*, *relentless*, and *guileless*). In turn, Approachability was defined in both authors to some extent by terms indicating lack of approachability. However, in Austen, the factor included a high-loading positive term (*social*, .97) as well as several terms with nuanced or ambivalent implications for approachability (*dignified*, *proud*, *cautious*, and *indirect*). In Dickens, in contrast, there was a clear prevalence of terms signifying unapproachability (e.g. *arrogant*, *icy*, and *disdainful*) as well as elements of dominance (*derogatory* and *overbearing*). In summary, although these two factors had common respective meanings between the two authors in the abstract, the composition of the factors suggested clear author differences in focus and nuance.

DISCUSSION

We presented a computerized bottom-up approach that allows capturing of trait terms included in established trait dictionaries and applied this algorithm to novels by Jane Austen and Charles Dickens to examine the implicit personality models that these authors used when describing their characters. This approach is a first transdisciplinary step towards theory development with big data approaches as called for by Bleidorn et al. (2017). One of the main options for conducting psychological analyses of personality at a distance across temporal epochs is to examine written records. The key insight from our analysis is that we did find meaningful factors across dictionaries for each author, but these did not follow the Five Factor Model as currently popular. At the same time, the factor structures did not converge across the two authors, although there seemed to be some qualitative similarity.

Personality dimensions across cultures and time

We present a first systematic historical analysis of personality structure based on an empirical analysis of textual data. Both with respect to the most frequently used terms (Research Question 1) and the underlying factors of their interrelations (Research Question 2), our findings suggest personality concepts that differ from currently prevalent models. A critical reader of the current literature describing lexical and psychometric studies across contemporary cultures may not be surprised by these divergent findings considering the social context of 19th century England. Austen placed her novels mainly in the countryside, describing a period in which social hierarchies were still intact. Dickens describes an urban world in transition, commenting on the social and economic upheaval in the wake of the industrial revolution. There is mounting evidence that the Five Factor Model works well in highly educated and affluent populations, but the model does not describe personality structure well when studying less affluent and more culturally diverse populations (Gurven, 2018; Lajaaj et al., 2019; Lukaszewski, Gurven, von Rueden, & Schmitt, 2017; for a general review: Fischer, 2017). This divergence might actually be compatible with a revised model of the lexical hypothesis. Important traits are likely to be encoded in single terms. However, how individuals combine and communicate relevant information to others are shaped by the contingencies within the social, economic, and ecological environment in which a community is living. In other words, traits are important across cultures and time (Mayer et al., 2011), but the relevant information that needs to be communicated is adapted to the local context. Therefore, in some contexts information on social skills, abilities and virtues might be more relevant, especially if social hierarchies are weakening and the negotiation of social position becomes more paramount; hence, finer distinctions in traits capturing social domains are being made when individuals talk about each other (Nel et al., 2012). In other contexts, the ability to provide for others and be trustworthy might be most important to communicate to others (Gurven, von Rueden, Massenkoff, Kaplan, & Lero Vie, 2013). Hence, specific cultural and historical periods may require different information packages to be communicated to others. The Georgian and Victorian periods were considerably more hierarchical than contemporary Western society, with fewer personal choices available to individuals in those earlier times. Individuals were still tightly integrated into social cleavages of family and class but with increasing uncertainty about the stability and legitimacy of those social hierarchies. In turn, this appears to have resulted in different associations between trait terms. In line with contemporary psycholexical literature, the agreeableness-related terms were most prevalent in both authors. Furthermore, our factors overall had a more social connotation and resembled previous descriptions of virtues (De Raad & van Oudenhouven, 2011), which suggests that personality descriptions were more socially focused than even today (see Wood, 2015). Of course, as the correlations with the marker terms show, there is some semblance with modern understandings of the five factors. However, the specific connotation of how these terms are arranged in the factor structures was more complex and diverged from contemporary factor structures. The packaging of information is context dependent, even if the individual bits of information are universal. In other words, the social context determines what implicit personality models might be most relevant.

A second point that we would like to highlight, which will be important as we start discussing the findings in relation to the two authors, is the fact that all personality trait research currently is cognitively mediated. Trait descriptions as currently studied in psychology rely on a complex set of cognitive processes (Fischer, 2017). Individuals need to perceive a relevant behaviour (in themselves or others), integrate that observation into short-term memory, and integrate the representation into abstract categories that can then be stored in long-term memory. When probed by a researcher or a communication partner to provide information about another person (or themselves) in relation to a specific stimulus (e.g. survey item or question about the target person in a specific context), the respondent then needs to retrieve relevant

DOI: 10.1002/per

information from long-term memory, encode it in language that is appropriate for the relevant context, and evaluate whether the information was received and interpreted appropriately by the person asking for that information. Traits capture a field of meaning instead of precise categories (Uher & Visalberghi, 2016). This fuzzy set structure within a cognitively demanding social interaction process requires more theoretical attention. In particular, cognitive limitations such as short-term memory constraints as well as top-down preprocessing of information (in line with predictive coding models of consciousness (Clark, 2013) suggest that there is no unlimited set of reliable factors that could be recovered through linguistic means. People might have very complex internal models of each other, but when needing to respond to interaction stimuli and communicate relevant information about each other, our human cognitive limitations to only be able to process and store 5 ± 2 pieces of information might be a constraining factor for how much information can be simultaneously communicated and processed (Cowan, 2001; Miller, 1956). In other words, our cognitive limitations during interactions might constrain the number of factors that can be lexically communicated. An indirect indication of this relevance of cognitive capacity overall for the structuring of personality is that the number of factors that can be reliably distinguished is linearly dependent on the cognitive ability of the population studied (Bowler, Bowler, & Phillips, 2009; Bowler, Bowler, & Cope, 2012). When people have time to formulate their perceptions of others (e.g. when trying to write a novel), they might be able to make more fine-grained distinctions and engage in more complex simulation exercises (Vermeule, 2010). In specific interpersonal interactions, the person may not have the ability and motivation to make those fine-grained distinctions. This may also apply to survey studies of personality traits where researchers instruct participants to not process information too much and respond quickly.

As a corollary of this, communicators need to consider the expectations and mindsets of their interaction partners. This is a central area of research under the term of Theory of Mind—humans are able to attribute mental states (including personality relevant information such as motivation, desires, and emotions) to oneself and others, and our hypersocial environment requires that we consistently attempt to understand the cognitive states of others (e.g. what are our interlocutors' current beliefs, intentions, motivations, and emotions) (Premack & Woodruff, 1978). Hence, humans may tailor their communication about others (the implicit personality models adopted) to the specific audience and situations that they are dealing with. Our interpretations of novels as a form of sharing valuable social information (including personality information) are compatible with contemporary cognitive approaches to literature (Vermeule, 2010; Zunshine, 2006).

Factor structure divergence by author

We found that the factor structures diverged between the two authors (Research question 3). There are a number of plausible reasons for this. First, the factor structures might have been influenced by the personality of the authors themselves. As indicated by the types of terms used, Austen and Dickens wrote in very different ways and used differently charged trait terms to describe their characters. Austen described characters predominantly in terms of Agreeableness and Neuroticism. Her character world is socially and emotionally nuanced. Dickens in contrast used more diverse sets of trait terms overall, and the character descriptions covered the Big Five dictionaries more broadly. The full factors extracted for Austen suggest finer distinctions along social and reputational lines, whereas the three to five factors suggested by parallel analysis for Dickens captured the factors described above (differentiating power and social dimensions). The word choices and character descriptions may indicate that Austen was more socially and emotionally centred, whereas Dickens as a person and writer could have been characterized by less emotional concern and more emphasis on Conscientiousness and issues related to Power and exploration of opportunities. Beyond the idiosyncratic profiles of Austen and Dickens, these differences might also reflect more general gender differences (Schmitt et al., 2017).

As a second possible reason, the different structures may reflect the different social context of the two authors. The two authors published in different formats and for different audiences. From a Theory of Mind perspective, an author has to anticipate how a reader will interact with the writing. Dickens' novels were first published in serial instalments, often within journals or magazines. This will have required setting up characters that might be more stereotypical in their externality, easy to relate to, or associate with specific social and personal categories. Dickens had a larger cast of characters and used a broader vocabulary, which may make the stories more lively and entertaining, character descriptions less redundant but also requires simplification and diversification of the depths of characters, and less complexity in the inner workings. In other words, Dickens may have written towards a broad audience with a fuzzier field of personality trait systems, which required simpler structures with more diverse vocabulary (see the lower number of factors suggested by parallel analysis). In contrast, Austen wrote and edited her novels over a long period of time and published them anonymously due to social constraints. Her audience was considerably more limited and selected (a relatively small circle of upperclass readers). Hence, the description and psychological details of a smaller set of characters took priority, as may be indicated by the richer differentiation of factors suggested by parallel analyses, even with the more restricted set of vocabulary used in her novels. Examining the commonalities in overall factor similarities as attempted in our qualitative comparison of the Civility and Approachability factors suggests that there were shared social dimensions of personality prevalent during this period. However, the two authors tapped into those common dimensions of personality using different terminology. The difference in production and audience as well as the personal personality predispositions of the two authors are likely to have impacted how the two authors communicated person-relevant information to their readers.

DOI: 10.1002/per

The potential for idiographic–nomothetic integration

The tension between nomothetic and idiographic approaches is as old as personality psychology (Barenbaum & Winter, 2008; Conner et al., 2009), and there have been various attempts to integrate the two perspectives (for a recent example, see Beck & Jackson, 2017). Our approach offers a promising avenue for integration. The current study suggests implicit models of personality that are meaningfully related both to the characteristics of the two authors we studied and to their social context. Beck and Jackson (2017) reiterated the importance of idiographic analyses, especially recommending the use of network models. TIC is ideally suited for those analyses because its underlying architecture is a temporally structured network of information tokens. Our goal here was to examine the dimensional structure of personality trait words; hence, we only used the first result of the TIC process and ignored the richer directed (recurrence) and undirected (co-occurrence) network of the underlying model. However, even the simple structural analysis of the body of work by two authors reinforces the potential for idiographic–nomothetic integration in working with big data. Using factor analysis, we were able to extract personality dimensions that appear relevant when understanding the two individual authors and their work. Current personality factors are based on inter-individual, population-level associations. We show that by using textual analyses of work by individual authors, it becomes possible (i) to analyse the personality trait terms used and to compare them with pre-established reference dictionaries (e.g. marker scales and nomothetic analysis) as well as (ii) to examine the dimensional structure based on the co-occurrence of trait terms used by the authors (idiographic analysis). Once a number of such models have been developed for specific individuals, it will be possible to examine their uniqueness (idiographic component) and convergence (nomothetic component) and to integrate these two perspectives within a multilevel framework. Similarly, to research comparing personality structure across cultures (Cheung et al., 2011; De Raad et al., 2014), such a study would allow the comparison of shared and unique aspects of personality models across individual authors and across time periods. This would be a true manifestation of an integrated idiographic–nomothetic representation of personality.

Limitations

One significant challenge is that current character parsing is based on modern English and language common in newspapers and social media blogs. The application and validity to character parsing in novels and historic texts need innovative computational approaches and methods as well as experimental validation. We used state-of-the-art methods that nonetheless remain relatively imprecise; refining those parsing algorithms is beyond the scope of the current study. However, we stayed at the sentence level for extracting trait relevant information. This minimizes misattributions to some extent, but we may have missed out important trait information that is elaborated on in subsequent sentences after a character was first mentioned.

There are also significant challenges associated with using pre-selected contemporary dictionaries, either in the form of psychologically derived adjective dictionaries that were developed in the 20th century as in our case or using contemporary language usage dictionaries such as LIWC (Pennebaker et al., 2015). The use of such contemporary adjective lists presents a number of challenges. First, as demonstrated by the usage statistics, historically different terms may have been used for describing individuals. Even the 500 list, which was selected to represent frequently used broad person descriptors (Saucier, 1997) showed only limited relevance in the context of Dickens (66% usage) and Austen (44% usage) novels. A cursory look at high loading terms (e.g. unfeeling, impertinent, and prudent) against word usage statistics from Google n-gram suggests that those highest loading terms were frequently in use at the time of Austen and Dickens' writing but are less frequently used today. Hence, pre-selected dictionaries from a different time or social period may not capture the most relevant personality descriptors.

A second and related drawback of using pre-selected dictionaries is that the biases and assumptions in constructing those dictionary sets appear to influence the factor structures to some extent. In our analysis, there were a few dictionary specific factors that emerged once a core number of factors were extracted. Furthermore, the relative interpretation of factors, even though there was some convergence across dictionaries, may differ depending on the combination of unique high loading terms. In other words, interpretational nuances due to added or missed marker terms within specific dictionaries could be substantive, although there is high convergence across the common terms. Our automated approach of extracting and co-relating information was not biased by contemporary usage and meanings, but the interpretation of the factors certainly was influenced by modern English. One option forward is to construct semantic networks of high-loading terms to contextualize the meaning of key personality terms in historical texts.

A third drawback of using adjectives is that much personality-relevant information might be captured in actions in context. This is a crucial point of social-cognitive theories (Mischel & Shoda, 1995): personality is most informative when examining the actions (verbs) in context. The act of walking 3 miles across a wet field to visit a sick sister is a significant marker of personality characteristics of Austen's Elizabeth Bennet in *Pride and Prejudice*. However, the problem with these highly contextualized behavioural accounts is that human coders often draw very different inferences from behaviour (Uher & Visalberghi, 2016). From an automated text-mining perspective, the extraction of verbs in context creates significant additional inference problems. However, with a focus only on the characters, an adjective-based trait perspective might be insufficient to capture core personality traits.

Methodological advances and possibilities for further development

We presented a new approach and the first systematic attempt on this scale to extract a personality model (i) from literature

DOI: 10.1002/per

and (ii) from a past period. Our work goes beyond previous analyses where contemporary models were retrospectively applied to historical texts (e.g. Johnson et al., 2011; Passakos & De Raad, 2009) or current top-down text-mining approaches (e.g. LIWC) by simultaneously parsing character descriptions of targets (fictional characters) and examining the factor structure, while also allowing some information recovery about the authors. Our bottom-up unsupervised approach therefore facilitates an examination of personality structure that is less biased by contemporary personality models. We believe that this is a promising technique for larger scale analyses that may increase the scope for describing whole time periods.

We used co-occurrence of terms associated with fictional characters. We have not yet exploited the full potential of the TIC approach. There are unique features available through TICs that go beyond the personality description that we are focusing on in this study (for other applications to historical narratives, see Luczak-Roesch, Grener, & Fenton, 2018). The combination of the co-occurrence and recurrence patterns together with external data (e.g. biographical, historical, and physiological) opens up possibilities for much broader applications and insights. The bottom-up algorithm and approach we presented here can be used in flexible ways to analyse time series data such as textual productions to identify implicit personality structures across individuals and historical time periods, complementing and extending other computational methods that aim at quantifying linguistic change at a macroscopic scale (Hamilton et al., 2016).

Our analyses suggested that there are possible psychological differences between the two authors. With an analysis of two authors, we are certainly limited in drawing generalizations about the model of personality in 19th century England, while also not allowing us to draw any definite conclusions about the impact of psychological differences between the two authors given the different audiences, writing formats and social and economic conditions that they experienced. In order to develop a more holistic idiographic analysis, it might be possible to use automated text analysis of all outputs (including letters and personal correspondence) and to then extract key terms used by the two authors to define their personality. A full analysis of the nodes by book or period available via TIC may also allow a closer analysis of the authors in terms of their psychological states (e.g. depression) and intellectual maturity over successive works separated in time. If the technique is used in the context of biographies written about the same target person by multiple authors, it might be possible to disentangle both author and target characteristics, leading to more nuanced insights about historical individuals and their biographers. The further development of this approach would allow the idiographic analysis of individuals from the past to move beyond the current reliance on labour-intensive psychobiographical analysis (e.g. Giammarco, 2013) and assessment methods that may impose a contemporary interpretation model (e.g. Ritzler & Singer, 1998). Applying temporally sensitive methods taking into account both co-occurrence and recurrence can help to derive an integrated theory of human personality from invariant properties and patterns of artistic expression.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Table S1 *Seven Factors in Austen's Novels Based on the Set of 500 Terms (221 Terms in Analysis)*
Table S2 *Seven Factors in Austen's Novels Based on the Set of Allport & Odbert's Terms (763 Terms in Analysis)*
Table S3 *Seven Factors in Dickens' Novels Based on the Set of 500 Terms (331 Terms in Analysis)*
Table S4 *Seven Factors in Dickens' Novels Based on the Set of Allport & Odbert's Terms (1913 Terms in Analysis)*

REFERENCES

Al Omran, F. N. A., & Treude, C. (2017, May). *Choosing an NLP library for analyzing software documentation: A systematic literature review and a series of experiments*. In *2017 IEEE/ACM 14th International Conference on Mining Software Repositories (MSR)* (pp. 187-197). IEEE.

Allport, G., & Odbert, H. (1936). Trait-names: A psycho-lexical study. *Psychological Monographs*, *47*, 1–171. https://doi.org/10.1037/h0093360.

Altmann, E. G., Pierrehumbert, J. B., & Motter, A. E. (2009). Beyond word frequency: Bursts, lulls, and scaling in the temporal distributions of words. *PLoS ONE*, *4*. https://doi.org/10.1371/journal.pone.0007678.

Ashton, M., Lee, K., & Goldberg, L. (2004). A hierarchical analysis of 1,710 English personality-descriptive adjectives. *Journal of Personality and Social Psychology*, *87*, 707–721. https://doi.org/10.1037/0022-3514.87.5.707.

Ashton, M. C., Lee, K., & de Vries, R. E. (2014). The HEXACO Honesty-Humility, Agreeableness, and Emotionality Factors: A review of research and theory. *Personality and Social Psychology Review*, *18*, 139–152. https://doi.org/10.1177/1088868314523838.

Barenbaum, N. B., & Winter, D. G. (2008). History of modern personality theory and research. In O. P. John, R. W. Robins, & L. A. Pervin (Eds.), *Handbook of personality: Theory and research* (3rd ed., pp. 3–23). New York, NY: Guilford Press.

Beck, E. D., & Jackson, J. J. (2017). The search for a bridge: Idiographic personality networks. *European Journal of Personality*, *31*, 530–532. https://doi.org/10.1002/per.

Benoit, K., & Matsuo, A. (2018). *Spacyr: Wrapper to the spaCy NLP library*. Retrieved from https://CRAN.R-project.org/package=spacyr

Bleidorn, W., Hopwood, C. J., & Wright, A. G. C. (2017). Using big data to advance personality theory. *Current Opinion in Behavioral Sciences*, *18*, 79–82. https://doi.org/10.1016/j.cobeha.2017.08.004.

Bowler, M. C., Bowler, J. L., & Cope, J. G. (2012). Further evidence of the impact of cognitive complexity on the five-factor model. *Social Behavior and Personality*, *40*, 1083–1098. https://doi.org/10.2224/sbp.2012.40.7.1083.

Bowler, M. C., Bowler, J. L., & Phillips, B. C. (2009). The Big-5 ± 2? The impact of cognitive complexity on the factor structure of the five-factor model. *Personality and Individual Differences*, *47*, 979–984. https://doi.org/10.1016/j.paid.2009.08.002.

Boyd, R. L., & Pennebaker, J. W. (2017). Language-based personality: A new approach to personality in a digital world. *Current Opinion in Behavioral Sciences*, *18*, 63–68. https://doi.org/10.1016/J.COBEHA.2017.07.017.

Butler, M. (1975). *Jane Austen and the war of ideas*. Oxford, UK: Oxford University Press.

Cheung, F., Fan, W., & To, C. (2008). The Chinese Personality Assessment Inventory as a culturally relevant personality measure in applied settings. *Social and Personality Psychology Compass*, *2*, 74–89. https://doi.org/10.1111/j.1751-9004.2007.00045.x.

Cheung, F. M., Van de Vijver, F. J. R., & Leong, F. T. L. (2011). Toward a new approach to the assessment of personality in culture. *American Psychologist*, *66*, 593–603. https://doi.org/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*, 96–132. https://doi.org/10.1016/j.jrp.2007.04.006.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, *36*, 181–204. https://doi.org/10.1017/S0140525X12000477.

Conner, T. S., Tennen, H., Fleeson, W., & Barrett, L. F. (2009). Experience sampling methods: A modern idiographic approach to personality research. *Social and Personality Psychology Compass*, *3*, 292–313. https://doi.org/10.1111/j.1751-9004.2009.00170.x.

Costa, P. T., & McCrae, R. R. (1992). *Revised NEO Personality Inventory (NEO-PI-R) and NEO Five-Factor Inventory (NEO-FFI) professional manual*. Odessa, FL: Psychological Assessment Resources.

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, *24*, 87–114. https://doi.org/10.1017/S0140525X01003922.

Da, N. Z. (2019). The computational case against computational literary studies. *Critical Inquiry*, *45*, 601–639. https://doi.org/10.1086/702594.

Daelemans, W., & Hoste, V. (2002). *Evaluation of machine learning methods for natural language processing tasks*. In 3rd International conference on Language Resources and Evaluation (LREC 2002). European Language Resources Association (ELRA).

De Raad, B., Barelds, D. P., Levert, E., Ostendorf, F., Mlačić, B., Blas, L., … 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*, *98*, 160–173. https://doi.org/10.1037/a0017184.

De Raad, B., Barelds, D. P., Timmerman, M. E., Roover, K., Mlačić, B., & Church, T. A. (2014). Towards a pan-cultural personality structure: Input from 11 psycholexical studies. *European Journal of Personality*, *28*, 497–510. https://doi.org/10.1002/per.1953.

De Raad, B., & Peabody, D. (2005). Cross-culturally recurrent personality factors: Analyses of three factors. *European Journal of Personality*, *19*, 451–474. https://doi.org/10.1002/per.550.

De Raad, B., & Van Oudenhouven, J. P. (2011). A psycholexical study of virtues in the Dutch language, and relations between virtues and personality. *European Journal of Personality*, *25*, 43–52. https://doi.org/10.1002/per.777.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). *Bert: Pre-training of deep bidirectional transformers for language understanding*. *arXiv preprint arXiv:1810.04805*.

DeYoung, C. G. (2015). Cybernetic Big Five Theory. *Journal of Research in Personality*, *56*, 33–58. https://doi.org/10.1016/J.JRP.2014.07.004.

Duckworth, A. (1971). *The improvement of the estate: A study of Jane Austen's novels*. Baltimore, MD: Johns Hopkins University Press.

Eagleton, T. (2005). *The English novel: An introduction*. Oxford, UK: Wiley-Blackwell.

Farrell, J. (2016). Corporate funding and ideological polarization. *Proceedings of the National Academy of Sciences*, *113*, 92–97. https://doi.org/10.1073/pnas.1509433112.

Ferguson, F. (2000). Jane Austen, Emma and the impact of form. *Modern Language Quarterly*, *61*, 157-180.

Fischer, R. (2017). *Personality, values, culture: An evolutionary perspective*. Cambridge, UK: Cambridge University Press.

Fischer, R., & Fontaine, J. (2010). Methods for investigating structural equivalence. In D. Matsumoto, & F. van de Vijver (Eds.), *Cross-cultural research methods in psychology* (pp. 179–215). New York, NY: Oxford University Press.

Fischer, R., & Karl, J. A. (2019). A primer to (cross-cultural) multi-groupinvariance testing possibilities in R. *Frontiers in Psychology*, *10*, 1507. https://doi.org/10.3389/fpsyg.2019.01507.

Frow, J. (2018). *Character. The Cambridge Companion to Narrative Theory.*

Garg, N., Schiebinger, L., Jurafsky, D., & Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, *115*, E3635–E3644.

Gerlach, M., & Altmann, E. G. (2013). Stochastic model for the vocabulary growth in natural languages. *Physical Review X*, *3*, 021006. https://doi.org/10.1103/PhysRevX.3.021006.

Giammarco, E. (2013). Edgar Allan Poe: A psychological profile. *Personality and Individual Differences*, *54*, 3–6. https://doi.org/10.1016/j.paid.2012.07.027.

Goldberg, L.R. (1981). Language and individual differences: The search for universals in personality lexicons. *Review of Personality and Social Psychology*, *2*, 141–165.

Goldberg, L. R. (1982). From ace to zombie: Some explorations in the language of personality. *Advances in personality assessment*, *1*, 203–234.

Goldberg, L. R. (1990). An alternative "description of personality": The Big-Five factor structure. *Journal of Personality and Social Psychology*, *59*, 1216-1229. https://doi.org/10.1037/0022-3514.59.6.1216

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

Goldberg, L.R. (1993). The structure of phenotypic personality traits. *American Psychologist*, *48*, 26–34. https://doi.org/10.1037/0003-066x.48.1.26.

Gurven, M. D. (2018). Broadening horizons: Sample diversity and socioecological theory are essential to the future of psychological science. *Proceedings of National Academy of Sciences*, *115*(45), 11420-11427. https://doi.org/10.1073/pnas.1720433115

Gurven, M., von Rueden, C., Massenkoff, M., Kaplan, H., & Lero 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*, 354–370. https://doi.org/10.1037/a0030841.

Hall, J. A., Goh, J. X., Mast, M. S., & Hagedorn, C. (2016). Individual differences in accurately judging personality from text. *Journal of Personality*, *84*, 433–445. https://doi.org/10.1111/jopy.12170.

Hamilton, P. (1996). *Historicism*. London: Routledge.

Hamilton, W. L., Leskovec, J., & Jurafsky, D. (2016). *Diachronic word embeddings reveal statistical laws of semantic change*. *arXiv preprint arXiv:1605.09096*.

Hirsch, J. B., & Peterson, J. B. (2009). Personality and language use in self-narratives. *Journal of Research in Personality*, *43*, 524–527. https://doi.org/10.1016/j.jrp.2009.01.006.

Honnibal, M., & Montani, I. (2017). *SpaCy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing*. https://doi.org/10.5281/zenodo.1212304

John, J. (2010). *Dickens and mass culture*. Oxford: Oxford University Press.

Johnson, J. A., Carroll, J., Gottschall, J., & Kruger, D. (2011). Portrayal of personality in Victorian novels reflects modern research findings but amplifies the significance of agreeableness. *Journal of Research in Personality*, *45*, 50–58. https://doi.org/10.1016/j.jrp.2010.11.011.

DOI: 10.1002/per

Kim, E., & Klinger, R. (2018). *Who feels what and why? Annotation of a literature corpus with semantic roles of emotions*. In Proceedings of the 27th International Conference on Computational Linguistics (pp. 1345-1359).
Lajaaj, R., Macours, K., Hernandez, D. A. P., Arias, O., Gosling, S., Potter, J., Rubio-Codina, M., et al. (2019). Challenges to capture the Big Five personality traits in non-WEIRD populations. *Science Advances*, *5*, eeaw5226. https://doi.org/10.1126/sciadv.aaw5226.
Luczak-Roesch, M. (2016). *TICJ: An application to generate transcendental information cascades (Version v1.1)*. *Zenodo*. https://doi.org/10.5281/zenodo.165116
Luczak-Roesch, M., Grener, A., & Fenton, E. (2018). Not-so-distant reading: A dynamic network approach to literature. *it-Information Technology*, *60*, 29–40. https://doi.org/10.1515/itit-2017-0023.
Luczak-Roesch, M., Grener, A., Fenton, E., & Goldfinch, T. (2016, November 6). *Creating transcendental information cascades from English literature: An intuitive approach in R (version 0.1)*. Zenodo. https://doi.org/10.5281/zenodo.164949
Luczak-Roesch, M., O'Hara, K., Dinneen, J. D., & Tinati, R. (2018). What an entangled web we weave: An information-centric approach to time-evolving socio-technical systems. *Minds and Machines*, *28*, 709–733. https://doi.org/10.1007/s11023-018-9478-1.
Luczak-Roesch, M., Tinati, R., & Shadbolt, N. (2015). *When resources collide: Towards a theory of coincidence in information spaces*. In *Proceedings of the 24th international conference on world wide web* (pp. 1137-1142). ACM.
Luczak-Roesch, M., Tinati, R., Van Kleek, M., & Shadbolt, N. (2015). *From coincidence to purposeful flow? Properties of transcendental information cascades*. In *2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining* (ASONAM) (pp. 633-638). IEEE.
Lukaszewski, A. W., Gurven, M., von Rueden, C. R., & Schmitt, D. P. (2017). What explains personality covariation? A test of the socioecological complexity hypothesis. *Social Psychological and Personality Science*, *8*, 943–952. https://doi.org/10.1177/1948550617697175.
Mayer, J. D., Lin, S. C., & Korogodsky, M. (2011). Exploring the universality of personality judgments: Evidence from the Great Transformation (1000 BCE–200 BCE). *Review of General Psychology*, *15*, 65–76. https://doi.org/10.1037/a0020711.
McAdams, D., & Pals, J. (2006). A new Big Five: Fundamental principles for an integrative science of personality. *American Psychologist*, *61*, 204–217. https://doi.org/10.1037/0003-066X.61.3.204.
McCrae, R. R., & Costa, P. T. (1999). A five-factor theory of personality. In L. A. Pervin, & O. P. John (Eds.), *Handbook of personality: Theory and research* (2nd ed., pp. 139–153). New York, NY: Guilford Press.
Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, *63*, 81–97. https://doi.org/10.1037/h0043158.
Miller, J. H. (1965). *Charles Dickens: The world of his novels*. Cambridge, MA: Harvard University Press.
Mischel, W., & Shoda, Y. (1995). A cognitive-affective system theory of personality: Reconceptualizing situations, dispositions, dynamics, and invariance in personality structure. *Psychological Review*, *102*, 246–268. https://doi.org/10.1037/0033-295x.102.2.246.
Nel, J., Valchev, V. H., Rothmann, S., Van de Vijver, F., Meiring, D., & De Bruin, G. (2012). Exploring the personality structure in the 11 languages of South Africa. *Journal of Personality*, *80*, 915–948. https://doi.org/10.1111/j.1467-6494.2011.00751.x.
Pagel, M., Beaumont, M., Meade, A., Verkerk, A., & Calude, A. (2019). Dominant words rise to the top by positive frequency-dependent selection. *Proceedings of the National Academy of Sciences of the United States of America*, *116*, 7397–7402. https://doi.org/10.1073/pnas.1816994116.
Parker, I., Karl, J. A., Fischer, R., Luczak-Roesch, M., & Grener, A. (2019, July 24). *Allport's trait-word list digitized*. Doi:https://doi.org/10.17605/OSF.IO/K6RWJ
Passakos, C. G., & De Raad, B. (2009). Ancient personality: Trait attributions to characters in Homer's Iliad. *Ancient Narrative*, *7*, 75–95.
Peng H, Khashabi D, Roth D (2015) *Solving hard coreference problems*. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 809–819
Pennebaker, J. W., Booth, R. J., Boyd, R. L., & Francis, M. E. (2015). *Linguistic inquiry and word count: LIWC2015*. Austin, TX: Pennebaker Conglomerates (www.LIWC.net).
Pennebaker, J. W., & King, L. A. (1999). Linguistic styles: Language use as an individual difference. *Journal of Personality and Social Psychology*, *77*, 1296–1312. https://doi.org/10.1037/0022-3514.77.6.1296.
Phelan, J. (1989). *Reading people, reading plots: character, progression, and the interpretation of narrative*. Chicago: University of Chicago Press.
Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, *1*, 515–526. https://doi.org/10.1017/S0140525X00076512.
Revelle, W. (2018). *Psych: Procedures for personality and psychological research*. Evanston, IL: Northwestern University.
Ritzler, B., & Singer, M. (1998). MMPI-2 by proxy and the Rorschach: A demonstration assessment of the commandant of Auschwitz. *Journal of Personality Assessment*, *71*, 212–227. https://doi.org/10.1207/s15327752jpa7102_9.
Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S. K., Albertson, B., et al. (2014). Structural topic models for open-ended survey responses. *American Journal of Political Science*, *58*, 1064–1082. https://doi.org/10.1111/ajps.12103.
Robinson, J. M. (1985). Style and personality in the literary work. *The Philosophical Review*, *94*, 227–247.
Rosenberg, S., & Jones, R. (1972). A method for investigating and representing a person's implicit theory of personality: Theodore Dreiser's view of people. *Journal of Personality and Social Psychology*, *22*, 372–386.
Rosenberg, S., & Sedlak, A. (1972). Structural representations of implicit personality theory. In L. Berkowitz (Ed.), *Advances in experimental social psychology 6*. New York: Academic Press.
Saucier, G. (1997). Effects of variable selection on the factor structure of person descriptors. *Journal of Personality and Social Psychology*, *73*, 1296–1312. https://doi.org/10.1037/0022-3514.73.6.1296.
Saucier, G., Thalmayer, A., Payne, D., Carlson, R., Sanogo, L., Ole-Kotikash, L., … Zhou, X. (2014). A basic bivariate structure of personality attributes evident across nine languages. *Journal of Personality*, *82*, 1–14. https://doi.org/10.1111/jopy.12028.
Saucier, G., Thalmayer, A. G., & Bel-Bahar, T. S. (2014). Human attribute concepts: Relative ubiquity across twelve mutually isolated languages. *Journal of Personality and Social Psychology*, *107*, 199–216. https://doi.org/10.1037/a0036492.
Schmitt, D. P., Long, A. E., McPhearson, A., O'Brien, K., Remmert, B., & Shah, S. H. (2017). Personality and gender differences in global perspective. *International Journal of Psychology*, *52*, 45–56. https://doi.org/10.1002/ijop.12265.
Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., … Ungar, L. H. (2013). Personality, gender, and age in the language of social media: The open-vocabulary approach. *PLoS ONE*, *8*, e73791. https://doi.org/10.1371/journal.pone.0073791.
Singh, J. K., Misra, G., & De Raad, D. B. (2013). Personality structure in the trait lexicon of Hindi, a major language spoken in India. *European Journal of Personality*, *27*, 605–620. https://doi.org/10.1002/per.1940.
Spacy (2019). *Facts and figures*. Retrieved July 30, 2019 from https://spacy.io/usage/facts-figures.

DOI: 10.1002/per

Srivastava, S., Guglielmo, S., & Beer, J. S. (2010). Perceiving others' personalities: Examining the dimensionality, assumed similarity to the self, and stability of perceiver effects. *Journal of Personality and Social Psychology*, *98*, 520–534. https://doi.org/10.1037/a0017057.

Tahmasebi, N., Niklas, K., Theuerkauf, T., & Risse, T. (2010, June). *Using word sense discrimination on historic document collections*. In *Proceedings of the 10th annual joint conference on digital libraries* (pp. 89-98). ACM.

Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, *29*, 24–54. https://doi.org/10.1177/0261927X09351676.

Tvinnereim, E., & Fløttum, K. F. (2015). Explaining topic prevalence in answers to open-ended survey questions about climate change. *Nature Climate Change*, *5*, 744–747. https://doi.org/10.1038/nclimate2663.

Uher, J., & Visalberghi, E. (2016). Observations versus assessments of personality: A five-method multi-species study reveals numerous biases in ratings and methodological limitations of standardised assessments. *Journal of Research in Personality*, *61*, 61–79. https://doi.org/10.1016/j.jrp.2016.02.003.

Valchev, V. H., Van de Vijver, F. J. R., 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*, 1077–1091. https://doi.org/10.1037/a0032276.

Van De Vijver, F. J. R., & Leung, K. (1997). *Methods and data analysis for cross-cultural research*. Thousand Oaks, CA: SAGE.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., … & Polosukhin, I. (2017). *Attention is all you need*. In *Advances in Neural Information Processing Systems* (pp. 5998–6008).

Velicer, W. F. (1976). Determining the number of components from the matrix of partial correlations. *Psychometrika*, *41*, 321–327.

Vermeule, B. (2010). *Why do we care about literary characters?* Baltimore, MD: Johns Hopkins University Press.

Watzlawick, P., Beavin, J. H., & Jackson, D. D. (1967). *Pragmatics of human communication: A study of interactional patterns, pathologies, and paradoxes*. New York, NY: W. W. Norton & Company.

Williams, R. (1973). *The country and the city*. London, UK: Chatto & Windus.

Woloch, A. (2003). *The one vs. the many: Minor characters and the space of the protagonist in the novel*. NJ: Princeton UP.

Wood, D. (2015). Testing the lexical hypothesis: Are socially important traits more densely reflected in the English lexicon? *Journal of Personality and Social Psychology*, *108*, 317–335. https://doi.org/10.1037/a0038343.

Wood, D., Harms, P., & Vazire, S. (2010). Perceiver effects as projective tests: What your perceptions of others say about you. *Journal of Personality and Social Psychology*, *99*, 174–190. https://doi.org/10.1037/a0019390.

Yang, Y., & Eisenstein, J. (2016). *Part-of-speech tagging for historical English*. arXiv preprint arXiv:1603.03144.

Yarkoni, T. (2010). Personality in 100,000 words: A large-scale analysis of personality and word use among bloggers. *Journal of Research in Personality*, *44*, 363–373. https://doi.org/10.1016/j.jrp.2010.04.001.

Zhang, Z., Han, X., Liu, Z., Jiang, X., Sun, M., & Liu, Q. (2019). *ERNIE: Enhanced language representation with informative entities*. *arXiv preprint arXiv:1905.07129*.

Zunshine, L. (2006). *Why we read fiction: Theory of mind and the novel*. Columbus, OH: Ohio State University Press.

DOI: 10.1002/per