Frequency of use metrics for American English person descriptors:

Extensions of Roivainen's internet search methodology

David M. Condon[1] and Sarah McDougald[1]

[1] Department of Psychology, University of Oregon

Draft dated May 2, 2022.
Note: this version has not been peer reviewed, and
may not be the authoritative document of record.

Author Note

David M. Condon https://orcid.org/0000-0002-8406-783X
Sarah McDougald https://orcid.org/0000-0002-1794-3705

Author contributions: David M. Condon – Conceptualization, Methodology, Software, Validation, Formal Analysis, Investigation, Resources, Data Curation, Writing - Original Draft, Writing - Review & Editing. Sarah McDougald – Validation, Formal Analysis, Writing - Review & Editing.

We have no known conflicts of interest to disclose. Data, analysis code, and other supplemental materials are available at https://pie-lab.github.io/tdafrequency/.

Correspondence concerning this article should be addressed to David M. Condon, Department of Psychology, University of Oregon, 1451 Onyx Street, Eugene, Oregon, USA, 97403. Email: dcondon@uoregon.edu

Abstract

Personality traits are often measured using person-descriptive terms, but data are limited regarding the frequency of usage for these terms in everyday language. This project reports on the relative frequency of usage for a large pool of American English terms (N = 18,240) using count estimates from search engine results and in books cataloged by Google. These estimates are based on the ngrams formed when each descriptor is combined with a common person-related noun (person, woman, man, girl, boy). Results are reported for each noun form and a frequency index in an online database that can be sorted, searched, and downloaded. We report on associations among the different noun forms and data types and propose recommendations for the use of these data in conjunction with other resources. In particular, we encourage collaborative approaches among research teams using large language models in psycholexical research related to personality structure.

Keywords: personality descriptors, trait descriptive adjectives, ngrams, psycholexical, personality structure, language modeling

Frequency of use metrics for American English person descriptors:

Extensions of Roivainen's internet search methodology

The primary aim of this project is to identify the relative frequency of usage of person-descriptive terms in American English. Though multiple methods have been proposed (Leising et al., 2014; Motschenbacher & Roivainen, 2020; Wood, 2015), the current approach focuses on the procedure of using the counts estimated in the Google Books™ Ngram Viewer (Michel et al., 2011) and Google Web Search™ results for terms containing two or more words (ngrams) composed of a person-descriptive term and a common noun indicating an individual person. For example, "happy person" or "funny woman". The person-descriptive terms used in this work include a diverse range of qualifiers – many describe features of personality, while many more relate to other psychological and non-psychological differences.

This project expands on the search results approach introduced by Roivainen (2013) by evaluating a larger pool of terms than previously reported (from 432 to 18,240) and using several additional ngrams. While thousands of research projects have made use of Google Books Ngram data (Michel et al., 2011), counts from search engine results have been less widely reported in scientific research. The motivation for including both sources in this work was based on the intention of tapping into knowledge graphs (Hogan et al., 2021) with broader domain coverage than published books (Paulheim, 2017; Pechenick, 2015).

It has previously been noted that many of the descriptors used to identify personality structure in the context of the "lexical hypothesis" (Goldberg, 1992, 1993) are uncommon words (Roivainen, 2013, 2015). This problem is particularly evident in longer lists of personality descriptors such as the list developed by Allport and Odbert (1936) with nearly 18,000 terms, the

2,797-term list used by Norman (1967), and the 2,818-term list evaluated by Condon, Coughlin, and Weston (2022). For example, the latter list includes – for the sake of continuity – many terms from Goldberg and Norman (Goldberg, 1982; Ashton et al., 2004) that were originally chosen for their expected utility as bipolar markers (Goldberg, 1992) rather than their frequency of use in everyday descriptions of personality (e.g., "uncautious", "unfaltering", "unmaidenly"). In fact, some of these are unfamiliar to most native English speakers ("unmercenary", "unimpressible"; Condon et al., 2022). At the same time, many terms (e.g., "joking", "discouraging", "nerdy") have been excluded from large-scale personality studies of adjective ratings despite frequent usage in everyday language.

Roivainen (2013) proposed that internet search results and frequencies in books are useful indices of the relative frequency of usage across descriptors, and reported these values for the 435-term list previously used by Goldberg (1992) and Saucier and Goldberg (1996). Though search engine results and books frequencies were highly correlated (r = .58 for search results in 2012 and counts in the books cataloged in 2000), these approaches reflect different contexts of language use. Frequencies in books should be expected to reflect usage among more literate, published authors, and – for Google Books – this approach has the benefit of being permanently archived. Search engine results, by comparison, should be expected to capture frequencies of usage in less formal contexts, on average. That said, these results are more difficult to interpret and track, as search engines use unique, opaque, evolving, and proprietary algorithms. Additional resources are available for evaluating the frequency of use of specific words (Davies et al., 2010; Brysbaert et al., 2019), but these are less useful for evaluating the frequency of usage in specific contexts, such as personality description.

The need to characterize longer lists of terms stems from the potential to use language models (aka natural language processing techniques) in personality structure research (Cutler & Condon, 2022; Jackson et al., 2021). These models include, for example, transformer-encoder models such as BERT (Devlin et al., 2019), DeBERTa (He et al., 2021), and GPT-3 (Brown et al., 2020). Unlike traditional survey-based approaches that require ratings of each descriptor from individual raters, these models are not constrained by the attention and fatigue limits of human raters. This introduces the possibility of including many thousands of trait descriptive adjectives in analyses of personality structure (Cutler & Condon, 2022). In turn, this possibility increases the benefit of characterizing the full universe of person descriptors on various features, including the extent to which they are commonly used. While it is also important to characterize the descriptors according to other criteria (i.e., the extent to which each is relevant to psychology), frequency of usage estimates are particularly useful for identifying a subset of terms that can be considered reasonably comprehensive.

Thus, the current work seeks to estimate the frequencies of use for an overly inclusive set of person descriptors. This is done by generating two indices of frequency estimates that are considerably more extensive than prior reports; we use five "descriptor + noun" forms instead of one and 42 times as many descriptors.

Hypotheses

Though the primary aim of this work was descriptive, several hypotheses were posited and pre-registered (https://osf.io/9br67). First, we expected all forms of the descriptor + noun ngrams to be highly correlated in the Google Books results and, separately, in the Google Web Search™ results. Specifically, we expected all correlations to be above .7 for "[descriptor] + person", "+ woman", "+ man", "+ girl", and "+ boy". Second, we expected the search results to be moderately to highly correlated with the frequencies in Google Books™ (*r*s > .5; Michel et al., 2011), and for both types to be highly correlated with the frequencies reported by Roivainen (2013) for the overlapping terms (*r*s > .7). Third, we expected to find differences among the correlations between ngrams such that (1) “man” and “woman” would be more closely associated to “person” than “boy” and “girl” are associated with “person”; (2) “man” and “woman” would be more closely associated to one another than they are to either “boy” or “girl” and vice versa; (3) the terms would be more highly correlated within gender (i.e., “man” to “boy”) than across (“woman” to “boy”); and (4) each of the five forms would be most highly correlated with the overall index derived using a “leave one out" average. For the last of these, the index was calculated by dropping the ngram with the most counts (leave one out), and this was done to reduce the influence of outlier ngrams (e.g., those that formed proper nouns or works of art).

We also pre-registered less well-specified, directional expectations that ordering of the ngram variables by z-score difference relative to the index for each descriptor would reflect stereotypical age and gender social roles (for extended review of this topic, see Motschenbacher and Roivainen, 2020). For example, “man” and “woman” were expected to have higher z-scores for “dangerous” and “experienced”; “boy” and “girl” were expected to have higher relative z-scores for “youthful” and “innocent.” However, we generally expected similar averages of search results (across all terms) for the “woman”, “man”, “girl”, and “boy” forms.

Finally, and perhaps most importantly, we expected that many of the descriptors would have consistently low counts for 4 or 5 of the ngram forms. Of the 2,818 terms in Condon et al. (2022), we expected 10% to 30% would produce relatively few search results and/or no occurrences in recent books. This would suggest that these terms are rarely, if ever, used to describe personality in everyday language. To clarify the rationale for specifying 4 *or* 5 of the 5 possible ngram forms, we expected that some ngrams would have an unpredictably large number of search results for only one form due to the unexpected formation of culturally meaningful ngrams, such as proper nouns or works of arts (e.g., song titles, fictional characters).

Note that data collection deviated from the pre-registration in one important respect. Specifically, the initial scope of the project included only the 2,818 descriptors reported in Condon, Coughlin, and Weston (2022). However, the results of a related project (Cutler & Condon, 2022) demonstrated that future research on personality structure need not be limited to relatively small sets of person-descriptive terms, and this prompted the decision to proceed with data collection for a much more comprehensive list, as described below.

Method

Materials

A total of 18,241 person-descriptors were used to form ngrams with each of 5 nouns indicating individual persons. The nouns were “person”, “woman”, “man”, “girl”, and “boy”. The person descriptors were aggregated from several resources, though the majority of content overlapped with the large pool of descriptors published by Allport and Odbert (1936), who provided “a tabulation of all the trait-names in the English language, — all at least that are included in Webster's [1925] unabridged New International Dictionary” (p. vi, Allport & Odbert, 1936). Note that the number of unique terms in this list – 17,913 – is slightly less than the count claimed in the original publication (17,953). To account for the possibility that the Allport and Odbert list was incomplete, many additional lists were considered, including the overlapping lists of Norman (2,797 terms; 1967), Anderson (555 terms; 1968); Goldberg (1,710 terms; 1982), Chandler (1,042 terms; 2018), and Condon et al. (2,818 terms; 2022). Collectively, these lists contained approximately 600 terms that were not included in the Allport and Odbert list, but only 359 additions remained after removing alternate spellings, type-nouns, and slang or vulgar terms. Similarly, 31 terms from the Allport and Odbert list were deprecated because they were alternative spellings of a single descriptor (i.e., only one form of a hyphenated and non-hyphenated version of the descriptor was kept), or because they were no longer widely accepted for use as descriptors (i.e., derogatory or excessively inappropriate descriptors).

It is also important to emphasize the over-inclusive nature of this list. Despite stating that “each single term specifies in some way a form of human behavior” (p. vi, Allport & Odbert, 1936), the authors later clarify (and close inspection confirms) that the main criteria for inclusion were based on "the capacity of any term to distinguish the behavior of one human being from that of another" (p. 24). The difference is slight but meaningful, as few of the terms *specify* behavior. Most could be classified as qualifiers or descriptors of behavior, though a substantial minority of the terms are non-psychological (e.g., demographic or occupational classifiers, physical attributes). Similarly, most of the terms are adjectives (specifically, descriptive adjectives, including many past and present participles), though there are also many “type” nouns (e.g., martyr, slob, clown). Several of the terms are not typically considered part of American English (e.g., acharné, auld-farrant, concitato, dégagé).

In addition, a large proportion of the terms are uncommon and/or unfamiliar. In an attempt to address the cumbersome length of this list, the original authors separated the terms into four groupings based on familiarity and expected utility, though they acknowledge that their procedure relied on several arbitrary decisions. Rather than subset from this list based on these arbitrary criteria or some other method, we used the full list to collect frequency estimates in the current work with the expectation that these estimates will facilitate less arbitrary procedures for subsetting in the future.

Procedure

Data collection procedures generally followed those described by Roivainen (2013). For search engine results, frequency of usage was operationalized as the number of results shown for internet searches for each ngram. Correspondence with Roivainen suggested that commercial/proprietary features of the search engine algorithm may alter the number of search results returned based on attributes of the client. As Google Web Search™ is a proprietary tool, the method by which it indexes web content is opaque, though the search results are known to be dependent on more factors than just semantic frequency (Pechenick, 2015; Paulheim, 2017). This was confirmed with pilot data collection, as inconsistent results were produced when using different combinations of browsers, operating systems, networking equipment, locations, and ngram forms. Pilot data collection (involving approximately 25 descriptors) also highlighted a tendency of the search engine to redirect searches for uncommon or potentially misspelled ngrams.

To address these concerns, we introduced three deviations from the procedures described by Roivainen (2013). The first involved extensive use of quotation marks. Specifically, all hyphens were replaced with spaces (causing some bigrams to become 3- or 4- grams), quotes were added around all individual words to ensure that no alternate spellings would be introduced, and additional quotes were included around all phrases to reduce the incidence of results being returned for reordered forms of the phrase. For example, the exact search entry for the bigram “self-reliant person” was:

""self" "reliant" "person""

Second, all searches were made from a novel browsing profile (without a search history) that was set to limit search results to the United States. Finally, we extended Roivainen's approach beyond using only the noun "person". This was primarily done to improve the signal/noise ratio produced when using only one noun. However, we also incorporated this change to evaluate the effect of using other common nouns referring to people, in a manner similar to Motschenbacher & Roivainen (2020).

For the Google Books™ Ngram Viewer, frequency of usage was operationalized as the proportion of occurrences of the ngram in the total corpus of words cataloged for each year (Michel et al., 2011). Using the same example given above, the ngram search entry was:

"self - reliant person"

Note that ngram searches require the use of spaces around the hyphen in hyphenated terms. We used the average of the most recently available 10 year period (2010 to 2019) in the American English corpus. Though data are available for prior years as far back as 1500, changes in frequency over time were not a focus of the current work.

Analyses

The data were collected in January 2022. The analyses included reporting of descriptive statistics based on the raw data for both the search engine and books results. The hypotheses related to mean differences by noun form (e.g., "person", "woman") were evaluated with pairwise t-tests within the two types (books or search engine). These results are presented with and without adjustments for multiple comparisons. All remaining analyses (and data made publicly available) were based on z-score transformations of the raw data within type and noun form. Indices of frequency were created for both types (books and search engine results) by averaging z-scores across the noun forms after removing the maximum z-score value across all 5 forms. This method for creating an index was used instead of the simple arithmetic mean to reduce the influence of arbitrarily inflated results that might occur if a specific ngram has meaning beyond the context of personality (i.e., in popular culture or media).

Correlational analyses were used to evaluate many of the hypotheses, including the associations between all of the descriptor + noun forms within the books results, within the search engine results, across the books and search engine results, and with the results reported in Roivainen (2013). Pearson correlations were reported along with 95% confidence intervals; statistically significant differences in correlations were identified based on the absence of overlapping confidence intervals. Hypotheses about the organization of descriptors by noun form relative to the index were evaluated by sorting the z-score differences. For example, differences between the index of frequencies and the z-scores for “[descriptor] + man” were sorted, and the largest differences were evaluated qualitatively for evidence of consistency with stereotypes or other biases.

Results

The supplemental materials for this project include databases containing all frequency estimates for both types in a format that is searchable and sortable. These materials also include documentation of the analytic code and provide access to the data. The supplemental materials are posted at https://pie-lab.github.io/tdafrequency.

Following collection of the data, one of the descriptors – “self-harming” – was dropped. This decision was prompted by its placement as the most frequently used descriptor in the search engine results. Further inspection indicated that the number of results was high for all noun forms, but especially for “girl” and “boy”. Given that prior research related to online searches about self-harming behavior (Lewis et al., 2014; Stanicke, 2021) has pointed to the importance of providing access to unbiased information, and the existence of policies about self-harm-related content at Google, it seemed likely that the results for this descriptor were affected by factors unrelated to frequency estimates. All of the other 18,240 descriptors were retained.

Tests for significant differences in means by noun form indicated only one significant pairwise difference in means after correcting for multiple comparisons: the “boy” and “girl” forms in the search engine results (p =.002). None of the means were significantly different in the books results. For the 432 descriptors overlapping with Roivainen (2013), both data types were highly correlated with the data collected here: the search results were correlated .86 (95% CI [.83-.88]) and the books results were correlated .89 (95% CI [.84, .92]). The Roivainen search results were collected in 2012; the books results were based on books published in 2000. These results are based only on the noun form common across studies (“+ person”).

Evaluations of the associations among the various noun forms and data types are shown in Table 1. The table shows correlations and 95% confidence intervals after Holm-adjustment for multiple comparisons. Correlations among the search engine results were all statistically significant, though the magnitudes of the correlations varied widely (rs ranged from .03 to .92). Within the search engine results, ngrams using the noun “person” were weakly associated with all other forms (rs from .03 to .07). The remaining correlations among all other noun forms in the search results were moderately to highly correlated, but only the woman-man correlation was above the hypothesized threshold (r > .7). For the remaining pairwise hypotheses among forms

Table 1: Correlations and 95% confidence intervals among noun forms in books and search results

Variable	1	2	3	4	5	6	7	8	9	10	11
1. “person” search											
2. “woman” search	.06 [.05, .08]										
3. “man” search	.04 [.02, .05]	.71 [.70, .72]									
4. “girl” search	.04 [.02, .05]	.65 [.64, .66]	.53 [.52, .54]								
5. “boy” search	.03 [.02, .05]	.51 [.50, .52]	.38 [.37, .39]	.58 [.57, .59]							
6. Frequency Index search	.07 [.05, .08]	.92 [.92, .93]	.76 [.76, .77]	.75 [.74, .75]	.70 [.69, .70]						
7. “person” books	.16 [.14, .19]	.38 [.36, .41]	.29 [.27, .32]	.23 [.21, .26]	.27 [.24, .29]	.43 [.41, .45]					
8. “woman” books	.05 [.02, .08]	.91 [.90, .91]	.84 [.83, .85]	.56 [.54, .58]	.47 [.45, .49]	.89 [.88, .89]	.37 [.34, .40]				
9. “man” books	.04 [.02, .07]	.81 [.81, .82]	.90 [.90, .91]	.53 [.51, .54]	.47 [.45, .49]	.84 [.84, .85]	.34 [.31, .36]	.96 [.96, .96]			
10. “girl” books	.02 [-.01, .06]	.41 [.38, .44]	.39 [.37, .42]	.49 [.46, .52]	.83 [.82, .84]	.57 [.54, .59]	.20 [.16, .23]	.45 [.43, .48]	.47 [.44, .50]		
11. “boy” books	.02 [-.01, .06]	.45 [.43, .48]	.56 [.54, .58]	.52 [.49, .54]	.81 [.80, .82]	.62 [.60, .64]	.21 [.18, .25]	.56 [.53, .58]	.61 [.59, .63]	.96 [.95, .96]	
12. Frequency Index books	.05 [.03, .08]	.82 [.81, .83]	.81 [.80, .82]	.61 [.59, .62]	.70 [.69, .72]	.90 [.90, .91]	.41 [.38, .43]	.92 [.92, .92]	.93 [.93, .93]	.73 [.71, .74]	.80 [.79, .82]

Note. Values in square brackets indicate the 95% confidence interval for each correlation.

of the search engine results, the results supported only two: (1) “woman” and “man” (r = .71) were more closely associated than either was with “girl” ($r_{\text{girl-woman}}$ = 65; $r_{\text{girl-man}}$ = .53) or “boy” ($r_{\text{boy-woman}}$ = .51; $r_{\text{boy-man}}$ = .38); and (2) all of the search engine noun forms were most closely associated with the search engine index.

Correlations among the books results were also all statistically significant, and the magnitude of the correlations were higher than those for the search results, on average ($r_{\text{books mean}}$ = .59 vs $r_{\text{search mean}}$ = .45). The “person” ngram was again less closely associated with the other terms but the difference was less pronounced than with the search results (rs from .20 to .41). Again, “woman” and “man” were more closely associated (r = .96) than either was with “girl” ($r_{\text{girl-woman}}$ = .45; $r_{\text{girl-man}}$ = .47) or “boy” ($r_{\text{boy-woman}}$ = .56; $r_{\text{boy-man}}$ = .61), and “girl” and “boy” were more highly associated (r = .96) than either was with “woman” or “man”. Only “person”, “man”, and “boy” were most highly associated with the books index relative to other ngram forms. Correlations of the same noun forms across the books and search results were generally strong (rs > .80), with the exception of the “person” (r = .16) and “girl” ngram forms (r = .49).

These correlational results are consistent with evidence from analyses of the most exclusive descriptors for each noun form. Table 2 shows the top 10 most exclusive terms for each form and data type. The lists are generally consistent among the same noun forms across data types, especially for the most highly correlated noun forms. For the “woman” ngrams, 6 of the 10 words are the same in the books and search results; 4 are the same for “man”, “boy”, and “person”. However, only 2 of the terms were shared for the “girl” ngram. Across all forms, demographic attributes are the most common type of descriptors — young, old, black, white, big, little, small. For the female noun forms (woman, girl), there are also many physical attributes; the descriptor “beautiful” was among the top 4 for the woman and girl ngrams of both types.

Table 2: The 10 most exclusive descriptors for each ngram form and data type

	ngram form ([descriptor] + …)					Frequency
Data type	person	woman	man	girl	boy	Index
Books results	"first"	"young"	"old"	"little"	"little"	"young"
	"single"	"old"	"young"	"pretty"	"small"	"black"
	"human"	"beautiful"	"white"	"beautiful"	"old"	"old"
	"average"	"pregnant"	"dead"	"sweet"	"big"	"little"
	"last"	"first"	"black"	"dear"	"bad"	"white"
	"reasonable"	"black"	"big"	"nice"	"dear"	"beautiful"
	"right"	"elderly"	"rich"	"poor"	"good"	"first"
	"second"	"middle-aged"	"great"	"lovely"	"golden"	"sexy"
	"whole"	"white"	"tall"	"silly"	"pretty"	"good"
	"particular"	"attractive"	"wise"	"smart"	"stable"	"hot"
Search results	"athletic"	"young"	"old"	"hot"	"little"	"young"
	"first"	"mature"	"iron"	"cute"	"bad"	"old"
	"single"	"old"	"black"	"beautiful"	"young"	"little"
	"right"	"beautiful"	"young"	"little"	"big"	"good"
	"overdrinking"	"pregnant"	"dog"	"sexy"	"game"	"black"
	"opiate"	"pretty"	"punch"	"black"	"gay"	"white"
	"human"	"first"	"good"	"old"	"out"	"poor"
	"normal"	"black"	"gay"	"horny"	"good"	"first"
	"better"	"sexy"	"dead"	"new"	"fat"	"dead"
	"specific"	"addicted"	"last"	"skinny"	"small"	"new"

Discussion

The primary contribution of this project is to make the books and search results available for researchers, especially personality psychologists, who seek information about the frequency of usage for person-descriptive terms. Most prominently, these data can contribute to the long arc of psycholexical research that began in earnest in 1936 (Allport & Odbert) and remains ongoing (Cutler & Condon, 2022). Prior work in this area has led to the identification of several multi-dimensional structural models of personality, including the Big Five (Goldberg, 1992), the HEXACO (Ashton et al., 2004) and the High Dimensional 20 (Saucier & Iurino, 2020). The motivation for more detailed characterization of person-descriptive terms at this stage stems from (1) the availability of novel methods for identifying structure (i.e., language models such as

BERT [Devlin et al., 2019] and DeBERTa [He et al., 2021]), and (2) renewed interest in the development of a bottom-up taxonomy of personality traits (Condon et al., 2021).

Our hypothesis-driven results also suggest lines of further inquiry, especially with respect to the differential use of person-descriptors. For example, the results indicated that person-descriptors are notably less correlated across noun forms by age. Among books results, the correlations were very high (.96) for both the woman and man forms and the girl and boy forms, though the correlations across age forms were much lower (.45-.61). The differences in correlations were larger and more consistent across ages than gender. Of note, the moderate-to-high associations among all of the age and gender specific noun forms did not match their associations to the more generic “person” noun form. These findings generally suggest that structural analyses of subsets of these terms may be affected by the inclusion of terms that insinuate affiliation with a demographic or other group type, but more work is needed to consider the effects of other types of group affiliation and the extent to which these differences are present among subsets of the terms (especially the most personality-relevant terms). In the interim, a reasonable recommendation is to use the frequency indices.

The results provided further evidence that the books and search engine results are generally highly similar, with one exception. All the noun forms were highly correlated (.81-.91) across the books and search results, but the girl noun form was much less (.49). Further work is needed to identify factors contributing to the difference, and to determine whether this difference remains evident in the personality-relevant subset of descriptors.

These suggestions for future research highlight a limitation of these data for personality applications: the list of terms is highly over-inclusive. A non-trivial proportion of the terms seem irrelevant as person-descriptors; for example, “car” or “elk”. Further, a large proportion of these terms are unrelated to psychological attributes. Even among the descriptors that *may* be related to psychological attributes, there is considerable variability with respect to (1) the extent of psychological relevance (consider: “injured”, “overdressed”, and “unclean”); (2) the extent to which the term describes a stable or passing attribute (“flustered”, “giddy”); and (3) the extent to which the term is unambiguously defined or operationalized (“owlish”, “compelling”, “hurting”). Thus, for research on personality structure specifically, it is expected that only a fraction of the terms in this list would have utility – the subset of psychologically relevant terms that are unambiguously used to describe stable attributes.

Identifying this subset of terms is a priority. As both Google Web Search™ and Google Books™ Ngram Viewer are imperfect means of indexing frequency, we do not recommend that either be used independently when considering personality descriptors for subsequent structural analyses. Just as the frequency indices presented herein are intended to provide more stable and reliable estimates of frequency than any single descriptor + noun form, we advise using these values together and in conjunction with alternative frequency estimates. These include tools like the Corpus of Contemporary American English (Davies, 2010) and word prevalence norms (Brysbaert et al., 2019). Frequency estimates should also be considered in relation to the familiarity, knowledge, and ambiguity of the meaning of each descriptor. To evaluate the frequency of usage in selected populations, survey-based sampling methods would be more appropriate (Leising et al., 2014; Wood, 2015).

We encourage readers to use and improve upon these tools collaboratively, helping the field move closer towards the development of a comprehensive personality taxonomy.

References

Allport, G. W. and Odbert, H. S. (1936). Trait-names: A psycho-lexical study. *Psychological Monographs: General and Applied, 47*(1):1-170. https://doi.org/10.1037/h0093360

Anderson, N. H. (1968). Likableness ratings of 555 personality-trait words. *Journal of personality and social psychology, 9*(3), 272. https://doi.org/10.1037/h0025907

Ashton, M. C., Lee, K., & Goldberg, L. R. (2004). A hierarchical analysis of 1,710 English personality- descriptive adjectives. *Journal of Personality and Social Psychology , 87*(5), 707. https://doi.org/10.1037/0022-3514.87.5.707

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. *arXiv*. https://arxiv.org/pdf/2005.14165v4.pdf

Brysbaert, M., Mandera, P., McCormick, S. F., & Keuleers, E. (2019). Word prevalence norms for 62,000 English lemmas. *Behavior research methods, 51*(2), 467-479. https://doi.org/10.3758/s13428-018-1077-9

Chandler, J. (2018). Likeableness and meaningfulness ratings of 555 (+ 487) person-descriptive words. *Journal of Research in Personality, 72*, 50-57. https://doi.org/10.1016/j.jrp.2016.07.005

Condon, D. M., Coughlin, J., & Weston, S. J. (2022). Personality Trait Descriptors: 2,818 Trait Descriptive Adjectives characterized by familiarity, frequency of use, and prior use in psycholexical research. *Journal of Open Psychology Data, 10*(1), 1. https://doi.org/10.5334/jopd.57

Cutler, A., & Condon, D. M. (2022). Deep Lexical Hypothesis: Identifying personality structure in natural language. *PsyArXiv*. https://psyarxiv.com/gdm5v/

Davies, M. (2010). The Corpus of Contemporary American English as the first reliable monitor corpus of English. *Literary and Linguistic Computing 25*(4): 447–64. https://doi.org/10.1093/llc/fqq018

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv*. https://arxiv.org/abs/1810.04805

Goldberg, L. R. (1982). From Ace to Zombie: Some explorations in the language of personality. In Spielberger, C. D. and J. N. Butcher, J. N., editors, *Advances in personality assessment: Vol. 1*, pages 203-234. Erlbaum, Hillsdale, NJ.

Goldberg, L. R. (1992). The development of markers for the Big-Five factor structure. *Psychological Assessment*, *4*(1):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

He, P., Liu, X., Gao, J., & Chen, W. (2021). DeBERTa: Decoding-enhanced BERT with disentangled attention. *International Conference on Learning Representations.* https://openreview.net/forum?id=XPZIaotutsD

Hogan, A., Blomqvist, E., Cochez, M., d'Amato, C., Melo, G., Gutierrez, C., Kirrane, S., Gayo, J., Navigli, R., Neumaier, S., Ngomo, A., Polleres, A., Rashid, S., Rula, A., Schmelzeisen, L., Sequeda, J., Staab, S., Zimmermann, A. (2021). Knowledge Graphs. *Synthesis Lectures on Data, Semantics, and Knowledge, Vol.12* (2), p.1-257. https://doi.org/10.2200/S01125ED1V01Y202109DSK022

Jackson, J., Watts, J., List, J.-M., Drabble, R., & Lindquist, K. (2021). From text to thought: How analyzing language can advance psychological science. *Perspectives on Psychological Science.* https://doi.org/10.1177/17456916211004899

Leising, D., Scharloth, J., Lohse, O. and Wood, D. (2014). What types of terms do people use when describing an individual's personality? *Psychological Science 25*(9): 1787–94. https://doi.org/10.1177/0956797614541285

Lewis, S. P., Mahdy, J. C., Michal, N. J., & Arbuthnott, A. E. (2014). Googling Self-injury: the state of health information obtained through online searches for self-injury. *JAMA pediatrics*, *168*(5), 443-449. https://doi.org/10.1001/jamapediatrics.2014.187

Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Team, G. B., Pickett, J. P., Hoiberg, D., Clancy, D., Norvig, P., Orwant, J., Pinker, S., Nowak, M. A., & Aiden, E. L. (2011). Quantitative analysis of culture using millions of digitized books. *Science, 331*(6014):176-182. https://doi.org/10.1126/science.1199644

Motschenbacher, H. and Roivainen, E. (2020). Personality traits, adjectives and gender: integrating corpus linguistic and psychological approaches. *Journal of Language and Discrimination*, *4*(1):16-50. https://doi.org/10.1558/jld.40370

Norman, W. T. (1967). *2800 Personality Trait Descriptors: Normative Operating Characteristics for a University Population.* Unpublished manuscript, University of Michigan, Department of Psychology, Ann Arbor.

Pechenick, E. A., Danforth, C. M., & Dodds, P. S. (2015). Characterizing the Google Books corpus: Strong limits to inferences of socio-cultural and linguistic evolution. *PloS one, 10*(10), e0137041. https://doi.org/10.1371/journal.pone.0137041

Paulheim, H. (2017). Knowledge graph refinement: A survey of approaches and evaluation methods. *Semantic web, 8*(3):489-508. https://doi.org/10.3233/SW-160218

Roivainen, E. (2013). Frequency of the use of English personality adjectives: implications for personality theory. *Journal of Research in Personality 47*(4): 417–20. https://doi.org/10.1016/j.jrp.2013.04.004

Roivainen, E. (2015). The Big Five factor marker adjectives are not especially popular words: are they superior descriptors? *Integrative Psychological and Behavioral Science 49*(4): 590–9. https://doi.org/10.1007/s12124-015-9311-9

Saucier, G. and Goldberg, L. R. (1996). Evidence for the Big Five in analyses of familiar English personality adjectives. *European Journal of Personality 10*(1): 61–77. https://doi.org/10.1002/(SICI)1099-0984(199603)10:1<61::AID-PER246>3.0.CO;2-D

Saucier, G., & Iurino, K. (2020). High-dimensionality personality structure in the natural language: Further analyses of classic sets of English-language trait-adjectives. *Journal of Personality and Social Psychology, 119*(5), 1188. https://doi.org/10.1037/pspp0000273

Stänicke, L. I. (2021). The punished self, the unknown self, and the harmed self–toward a more nuanced understanding of self-harm among adolescent girls. *Frontiers in psychology, 12*, 967. https://doi.org/10.3389/fpsyg.2021.543303

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*(2): 317–35. https://doi.org/10.1037/a0038343