Automated Dictionary Creation for Analyzing Text:

An Illustration from Stereotype Content

Gandalf Nicolas[a]

Xuechunzi Bai

Susan T. Fiske

Department of Psychology, Princeton University, Princeton NJ 08544

Author Note

[a]Corresponding author: Department of Psychology, Princeton University

330 Peretsman-Scully Hall, Princeton, NJ 08540

Email: gnf@princeton.edu

Data and code available at: https://osf.io/yx45f/?view_only=570a9017944d4ecfa35a88e690f081d2

The authors declare no conflicts of interest with respect to the authorship or the publication of this article.

Abstract

Recent advances in natural language processing provide new approaches to analyze psychological open-ended data. However, many of these methods require translating to the needs of psychologists working with text. Here, we introduce automated methods to create and validate extensive dictionaries of psychological constructs using Wordnet and word embeddings. Specifically, we first expand an initial list of seed words by using Wordnet to obtain synonyms, antonyms, and other semantically related terms. Next, we evaluate dictionary reliability by using word embeddings trained on independent sources. Finally, we evaluate the dictionaries' convergent validity against traditional scale ratings and human judgments. We illustrate these innovations by creating stereotype content dictionaries, a construct in social psychology that lacks specialized and validated dictionaries for coding open-ended data. These dictionaries achieved over 80% coverage of new responses, compared to 20% coverage by the seed-word-only approach. Cosine similarity with word embeddings confirmed that the dictionaries are more similar within the same concept than across concepts. Moreover, open-ended responses predicted both traditional scale ratings and human judgments about the dictionary topic. The R package *Automated Dictionary Creation for Analyzing Text* (ADCAT; ***https://github.com/gandalfnicolas/ADCAT***) allows anyone to create novel dictionaries for constructs of interest and to access the stereotype content dictionaries.

Keywords: Dictionaries, Text Analysis, WordNet, Word Embeddings, Stereotype Content Model

Automated Dictionary Creation for Analyzing Text:

An Illustration from Stereotype Content

Text data are everywhere. Researchers may obtain text data from sources such as the internet, literary collections, archival entries, and experimental psychology's open-ended responses. Recently, advances in natural language processing in machine learning allow easier extraction of information about psychological processes and content. Unlike traditional response scales in psychological research, embracing text data allows unobtrusive and unconstrained approaches to measurement. For example, online data provide information about participants' responses, free from demand characteristics associated with some laboratory studies.

Using open-ended (v. forced-choice) responses in controlled settings also enables more ecologically-valid and data-driven study of psychological processes and content. These benefits appear in studying emotion (Gendron et al., 2015) and racial categorization (Nicolas, Skinner, & Dickter, 2018), challenging previously held findings by employing free-response measures that circumvent researcher constraints on participants' responses. Despite the advantages, creating and validating text analysis instruments such as dictionaries differs considerably from developing traditional scales, and currently not many appropriately reviewed guidelines exist. As a result, many areas have yet to fully incorporate text analysis methods into their repertoire. An example is stereotyping, which despite being one of the largest research areas within social psychology, suffers from a dearth of specialized text analysis methods and literature that may support new avenues of research.

In this paper, we aim to provide a guide on how to develop dictionaries using a novel approach that largely automates the development process and provides high levels of coverage, internal reliability, and validity. At the same time, we illustrate this process by developing

widely-applicable novel dictionaries of stereotype content dimensions. We provide accompanying R code for all procedures.

Current Approaches to Text Analysis

The most common method to analyze text data in psychology has traditionally been human coding. In this approach, each text is evaluated by a group of human judges in terms of how much it reflects a construct of interest. Measures of agreement between human judges often document reliability. Evidently, however, this approach is time-consuming and resource-demanding, and these limitations rapidly worsen, the more data that need to be coded (Iliev, Dehghani, & Sagi, 2014). Furthermore, this approach for text analysis lacks standardization—that is, judges coding may vary across studies or laboratories.

An increasingly popular alternative to per-study human coding of text are dictionaries[1] (see Iliev, Dehghani, & Sagi, 2014). Dictionaries list words that are indicators of the construct of interest. Once created, dictionaries are a standardized approach for coding text data, across studies, without additional human judge intervention. For this reason, they are also less resource-intensive and time-consuming for users. Dictionaries are also easy to use in analysis (vs. some more advanced natural language processing methods). The analysis process most often consists of counting the number of words in a text that are included in the dictionary. The larger the number of words from the dictionary that are present in the text, the higher the score for the construct of interest measured by the instrument. To illustrate, if evaluating the positivity of a

[1] Other approaches to text analysis include topic modeling (Blei, 2012) and word embeddings (e.g., Mikolov, 2013), both techniques developed in the natural language processing and machine learning fields. Here, we do not deal with topic modeling, and we use word embeddings only as a novel method to study the internal consistency of dictionaries. However, dictionaries can also be used in conjunction with word embeddings to obtain measures of the construct in a continuous scale (rather than the all-or-none nature of dictionaries). This application is illustrated in Garten et al (2018) and Nicolas, Bai, and Fiske (in press.).

particular text (e.g., a self-description, or a diary entry) is, a researcher would count the number of words that fall into a positive valence dictionary (e.g., “good,” “nice,” “amazing”) as a measure of the constructs.

The most widely used set of dictionaries in psychology and akin areas is the Linguistic Inquiry and Word Count (LIWC; Pennebaker, Booth, Boyd, & Francis, 2015). LIWC has been the benchmark for studying text related to content as varied as emotion, social relationships, social hierarchies, thinking styles, among others (see Tausczik & Pennebaker, 2010). The original creation and wide usage of the LIWC dictionaries highlights both that available dictionaries are cheaper and easy to implement, and that the cost of creating new dictionaries in the first place may be prohibitively expensive and time-consuming for many researchers. The latest LIWC dictionaries (Pennebaker, Booth, Boyd, & Francis, 2015) contain almost 6,400 words covering a variety of topics. The dictionaries typically develop through an iterative process heavily reliant on human coders. For many dictionaries, the creation involved 2-6 judges generating word lists, followed by brain-storming sessions of 4-8 judges attempting to expand the dictionaries, and subsequent rating phases by 4-8 judges to evaluate goodness-of-fit. In addition, more advanced steps for dictionary expansion involved additional human judges (4-8) to determine whether high-frequency words from natural language text could fit existing dictionaries, and final revision stages also involving human judges (Pennebaker, Booth, Boyd, & Francis, 2015). This illustrates the high burden on human judges needed to create dictionaries with current techniques, resulting in increased costs and high reliance on coder decisions (i.e., creation and evaluation is based on judgments from different groups of judges). Given that the constructs measured by existing dictionaries is inevitably limited compared to the diversity of constructs studied by psychologists, new methods to facilitate dictionary creation can be useful.

This paper provides a complementary way to automatize many processes, in order to facilitate new dictionaries that are also less coder-reliant, may handle more words, and address distinctive topics, among other benefits. Additionally, we provide novel approaches to evaluate the reliability of dictionaries, as well as suggestions to evaluate their coverage and validity.

Illustration: Stereotype Content

To illustrate the development of dictionaries, we focus on the field of social psychology and the topic of stereotype content. The study of stereotypes has one of the longest traditions within social psychology (Bergsieker et al., 2012, Study 4; Katz & Braly, 1933). However, to date, no comprehensive instruments for the analysis of text data have been developed in the area.[2] This paper provides such an instrument for measuring several relevant dimensions of content.

The stereotype content model (SCM; Fiske, Cuddy, Glick, & Xu, 2002), a well-known current framework, proposes that the two main dimensions of content are warmth (i.e., whether the social group is a friend or a foe) and competence (i.e., whether the social group can act on their intentions). A large body of research has corroborated that evaluations along these dimensions occur cross-culturally and predict multiple other perceptions of target groups (behavior, emotions, status, interdependence; Fiske, 2018).

More recent models of stereotype content have either defined different facets of warmth and competence, or proposed novel, distinct dimensions of stereotype content. For example, working from the Dual Perspective Model (DPM; Abele & Wojciszke, 2014), Abele and colleagues (2016) found evidence that warmth (or communion) could subdivide into friendliness

[2] To be sure, the Katz-Braly method allowed participants to choose from 84 adjectives, allowing some spontaneity. In addition, other instruments have been developed to measure related content in person perception (e.g., see http://markallenthornton.com/blog/3daffect/).

and morality facets and that competence (or agency) could subdivide into ability and assertiveness. Indeed, the centrality of the morality facet as a dimension of person perception appears in numerous studies (e.g., see Ellemers, 2017; Goodwin, 2015). The most recent stereotype content model, the Agency-Beliefs-Communion (ABC) model introduces beliefs (i.e., religious-secular beliefs and political orientation) as a dimension and combines competence with status. While other social psychological theories (including the SCM) had discussed status and beliefs as relevant social information, the ABC model suggests that these two dimensions are the most relevant in describing social groups, followed by warmth (communion) as a more idiosyncratic dimension (Koch et al., under review). Thus, stereotype content dimensions are still contested, and open-ended data could shed some light on this issue (see also Abele, Ellemers, Fiske, Koch, & Yzerbyt, under review).

The stereotype content literature has so far largely relied on traditional metrics of measurement, in particular Likert-type scales measuring how much a social group allegedly possesses a particular dimension of content. In the broader field of person perception, these dimensions, under the names of communion (warmth) and agency (competence), have been studied occasionally using more open-ended approaches. A variety of methods appear in the literature, asking participants to write: recalled positive or negative episodes (Wojciszke, 1994; see also 1998), self-descriptions (Diehl et al., 2004; Uchronski, 2008), or descriptions of acquaintances (Abele & Bruckmüller, 2011). Human judges' coding of the open-ended responses in terms of communion and agency accounted for 75-99% of responses. Granted, all these studies may have overestimated the prevalence of warmth and competence as they used human judges trained specifically to identify these topics. But the data do suggest these dimensions are central to social cognitive content.

A couple of studies (Decter-Frain & Frimer, 2016; Dupree & Fiske, in press) have used some LIWC dictionaries to measure warmth (e.g., the family and friend dictionaries) and competence (e.g., the work and achievement dictionaries), but because these dictionaries were not designed to measure those constructs, they may cover both a small subset of appropriate words and correlated constructs rather than the target concepts. In addition, most LIWC dictionaries, including the ones used in these studies, cover mostly high directional words for the construct. Thus, for example, the LIWC affiliation dictionary includes words such as *friend* or *friendly*, but not antonyms (e.g., *enemy* or *unfriendly*). On the other hand, the achievement dictionary includes both words about high (e.g., *success*) and low (*fail*) achievement, but no indicators separate them. In absence of a specialized indicator or separate dictionary for the antonyms of these dimensional constructs, text data that include responses along the whole dimension (such as stereotypes) will suffer from lack of coverage, depending on the application. Finally, a recent study (Pietraszkiewicz, Formanowicz, Gustafsson Sendén, Boyd, Sikstrom, & Sczesny, 2018) developed dictionaries of communion (similar to warmth) and agency (similar to competence) using the LIWC development approach. In the supplement, we further discuss these dictionaries and how they compare to the subset of dictionaries developed here for these constructs.

Thus, despite the promising trajectory in stereotype content research, a more data- and discovery-driven examination seems desirable, and text analyses through an instrument such as the one developed here could address this issue. In the field of emotion, for instance, a movement toward more expansive inclusion of open-ended data has allowed researchers to revisit well-established theories of emotion universality. For example, free-labeling found differences in the spontaneous use of emotion words across cultures, and even Westerners often

responded differently to patterns believed to be universal on the basis of more constrained tasks (see Gendron et al., 2015). In the area of racial categorization, a recent study (Nicolas, Skinner, & Dicker, 2018) compared open-ended categorization tasks to more commonly used constrained-choice tasks that force participants to categorize Black-White mixed-race targets as either Black, White, or Multiracial (i.e., both Black and White). Their results indicated that the extant literature had greatly overestimated the probability of Multiracial categorizations by using forced-choice tasks. Given that the typical response options are researcher-selected, they prevented participants from categorizing mixed-race targets as Hispanic and Middle Eastern, which were at least as or more common than any other single categorization.

For an area such as stereotype content, where responses go beyond limited to a set of categories, to a large number of possible nouns and adjectives, developing a more comprehensive instrument becomes even more vital to examining spontaneous cognition. Potentially, this instrument could expand current theoretically-derived models of social cognition, in addition to examining stereotype content in multiple untapped sources of text data on and offline.

Dictionary creation aimed to achieve three indicators of quality: coverage, internal reliability, and convergent validity. We make available helper functions used to create dictionaries using this approach in the R package *Automated Dictionary Creation for Analyzing Text*, available at ***https://github.com/gandalfnicolas/ADCAT***. The package also contains functions to code text into the stereotype content dictionaries developed here. All data and code for the analyses presented here are also available at https://osf.io/yx45f/?view_only=570a9017944d4ecfa35a88e690f081d2.

Coverage

Coverage refers to the number of relevant words included in the dictionaries. A traditional psychological scale can measure a construct with a few items sampled from a larger pool of intercorrelated items without wasting any data. However, with open-ended measures, where participants choose the items (i.e., words) they wish to convey about the construct, a larger pool of items is needed to code the participants' responses. Coverage refers to the proportion of possible participant responses that is covered by the dictionary (i.e., the pool of items). Coverage will be domain-dependent, so our aim was to obtain good coverage for stereotype content as a broad, relevant, and commonly studied area in social psychology. Thus, we aimed to explain a majority of participants' responses when prompted to provide characteristics they associated with social groups (i.e., stereotypes).

Although one could manually gather many words using suggestions from field experts, that labor- and time-consuming method would be limiting. Wordnet offers one automated way to obtain a large pool of items by adding words that are semantically associated to a smaller pool of words (i.e., seed dictionaries obtained from the literature). Wordnet (Miller, 1995) is a large lexical database for the English language. The database contains metadata about English words, including part-of-speech (i.e., noun, adjective, verb, and adjective), glosses (i.e., short definitions), and usage examples in sentences. Most importantly, Wordnet distinguishes words' different senses (e.g., warmth may refer to both psychological warmth and temperature), and these senses then associate with other words/senses through several relations.

The most relevant relations for our purposes are hypo/hypernymy, syno/antonymy, derivational forms, as well as generally related terms. Hyponymy refers to the property of being a more specific instance of another term (e.g., dog is a hyponym of the concept canine, but also

of domestic animal). Hypernymy is the property of being a more general instance of another term (e.g., canine is a hypernym of dog). Both hyponymy and hypernymy are available only to nouns, which are organized hierarchically. Synonyms denote the same concept and are largely interchangeable (e.g., canine and canid). Antonyms are semantically opposite terms (e.g., domesticated and wild). Derivational forms connect different parts-of-speech (e.g., the verb domesticate to the adjective domesticated). Other relational forms include "similar" and "see also," which denote semantically related words connected indirectly in the Wordnet network (i.e., words that are used similarly in context). In addition, we used relations such as "attribute" (i.e., adjectives that define a noun), "part meronym" (denoting something that is a part of the target, e.g., "wing" is a meronym of "bird"), and "part holonym" (opposite of meronym, e.g., "bird" is a holonym of "wing"), where applicable. For a more in-depth explanation of Wordnet, see Fellbaum (1999).

Traditionally, dictionary development has been constrained to a small pool of items derived from the literature or expanded manually through consensus between the researchers or other human coders. Wordnet provides an automated, standardized route to expand dictionaries, using a thoroughly validated resource. Furthermore, Wordnet allows researchers to obtain not only words, but specific senses of words, allowing more accurate information from additional tools developed for Wordnet. For example, SentiWordnet (Baccianella, Esuli, & Sebastiani, 2010) allows researchers to obtain the valence or sentiment of senses. This has the advantage over other sentiment coders that it is based on the specific sense of the word (e.g., sentiment scores for the word "warm" can be based on the specific sense referring to psychological warmth, rather than the word itself which contains multiple meanings, including the physical sense of warmth). Furthermore, obtaining senses along with the words in a dictionary allows the

use of translation tools (e.g., Babelnet; Navigli & Ponzetto, 2012) that can automate obtaining the corresponding appropriate *sense* in other languages. This is vital to expand a body of research cross-culturally, helping ensure generalizability of findings. All of these are advantages that the method provided here offers over traditional, manual dictionary creation approaches.

Internal Reliability

Internal reliability refers to the consistency and intercorrelations of the pool of items that make up the dictionaries. In other words, reliability measures whether words within a dictionary bear higher semantic similarity than words in different dictionaries. To measure text data reliability, numeric values need to represent non-numeric characters (text data), which enables the necessary calculations. Here, recent methods in natural language processing generate vector representations of text data and semantic similarities between words. In particular, we use word embeddings, which are based on large corpora of natural language text, as the vector representation and cosine similarity as the similarity metrics (see Figure 1). The specific word embeddings used here are Word2Vec (Mikolov, Chen, Corrado, & Dean, 2013) and Glove (Pennington, Socher, & Manning, 2014), trained on two independent corpora.

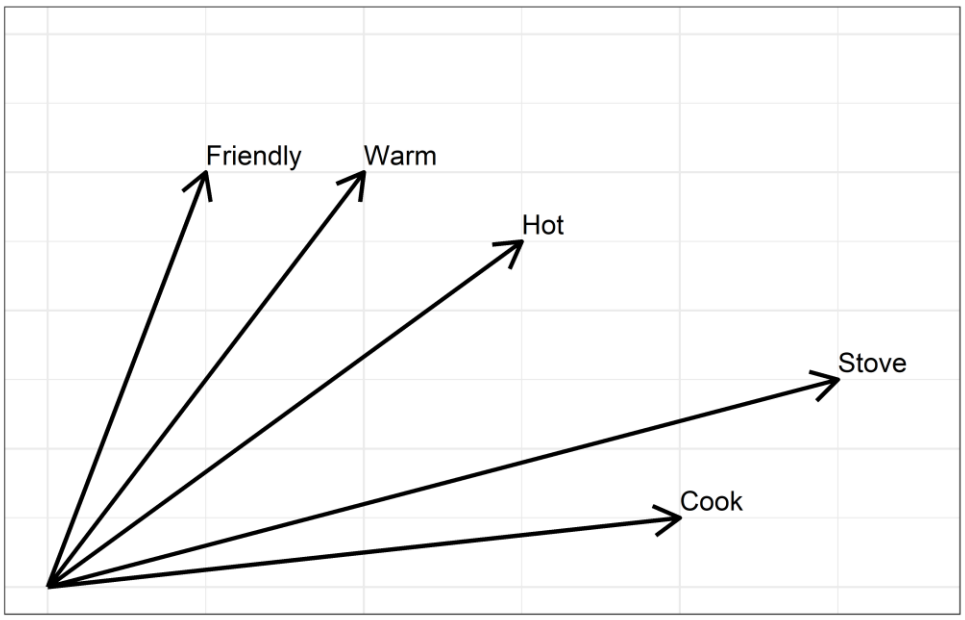

Figure 1. Hypothetical two-dimensional word space, with vectors representing different words. Cosine similarity is measured as the angle between vectors, and as shown, words used in similar contexts such as friendly and warm are more similar to each other than to words used in other contexts (e.g., stove and cook).

The background computations for these models are complex. As a simplified explanation of the process underlying the calculation of similarities between words, the intuition is that similar words occur in similar contexts (Miller & Charles, 1991). In order to determine if two contexts are similar, we could look at the co-occurrences of words in large corpora of text. Co-occurrence measures how often two words appear close to each other (say, at most three words apart) in the text corpora. For example, we would obtain high co-occurrence between *swim* and *fish*, as these words tend to be used often together, but low co-occurrence for words such as *stove* and *swim*. For each word in our dictionaries we can obtain its measure of co-occurrence with every other word in large corpora of text. For example, if we have 50 words in our dictionaries and a corpus of 1,000 words, we could obtain a 50 x 1,000 matrix where each row is a dictionary word, each column is a word in the corpus, and each cell is their pairwise co-occurrence.

Thus, we have a vector for each word in our dictionary that indicates its context, that is with which other words in the corpus our dictionary word tends to co-occur. If we then take two of these vectors representing two words in our dictionaries, we can measure their similarity by calculating the cosine of their angle. Cosine similarity is a normalized dot product of the two vectors, indicating the extent to which the two word-vectors point in the same direction in the multidimensional space. In theory, cosine similarity can range from -1 to 1, but the actual range depends on the specifics of the vectors generated by the algorithm. Cosine similarity differs from Pearson correlation in that the vectors are not mean-centered before norming, but it can be interpreted similarly, with larger numbers indicating larger similarity. Using this metric, we may obtain high similarity for words such as *river* and *ocean*, as these words often co-occur with the same words, such as *swim* and *fish*, and both rarely co-occur with the same words, such as *stove*.[3]

Obtaining pairwise similarities enables calculating traditional metrics of internal reliability, such as the average "correlation" (in this case average cosine similarity) or Cronbach's alpha. However, dictionaries tend to have very large numbers of items, an issue that demonstrably affects Cronbach's alpha. Furthermore, cosine similarities are not equivalent to Pearson correlations (although they are related), so interpretations of average similarity and alpha differ from traditional measures. Nonetheless, here we provide a variety of metrics, offering convergent evidence of the internal reliability of our dictionaries. Others have recently

[3] As mentioned, the actual algorithms employ more complex variations of this theme. Word2vec and Glove use natural language texts to train an algorithm that results in word-vectors. Word2vec uses neural networks to obtain these word-vectors, while Glove, most true to the details above, uses co-occurrence metrics, but includes additional details and applies dimensionality-reduction techniques. The Word2vec word embeddings used here was trained using part of the Google News dataset (~ 100 billion words), and it has word-vectors with 300 dimensions for 3 million words and phrases (available at https://drive.google.com/file/d/0B7XkCwpI5KDYNlNUTTlSS21pQmM/edit). The Glove word embeddings used here were trained using around 840 billion words from the common crawl (a very large database of web text), and it has word-vectors with 300 dimensions for 2.2 million words (available at https://nlp.stanford.edu/projects/glove/).

noted the novel technique described here for internal reliability as a measure of semantic coherence (Garten, Hoover, Johnson, Bogharti, Iskiwitch, & Dehghani, 2018), but not directly for the purpose of evaluating dictionary quality.

Validity

Validity is relatively straightforward, in being most similar to scales. Creating the dictionaries allows some measure of the construct from participants' responses, which may then correlate with other constructs expected to be theoretically related (i.e., convergent validity) or unrelated (i.e., divergent validity). Here we test validity in two ways. First, we test validity in relation to scale ratings, that is, whether stereotypes measured with our dictionaries correlate with stereotypes measured by scales. Then, we test validity in relation to human ratings, that is, whether human coders identify the semantic meaning of each dictionary from a small subset of its items.

Dictionary Creation Overview

Dictionaries evolved through an iterative process that subsequent sections will explain. To anticipate: First, we collected development data (i.e., "test data," used exploratorily to improve the dictionaries) to test our dictionaries for coverage and validity. Then, in a theory-driven step, we collected seed words covering relevant stereotype and person perception dimensions. We tested how much these initial seed words accounted for our development data (coverage). Subsequently, we used Wordnet to expand the seed words to a larger dictionary. We iterated the process of testing coverage and adding words until we reached a good proportion of dictionary coverage. After completing the dictionary, we tested dictionary reliability using various similarity metrics discussed above. Finally, we tested the validity with scale ratings from

development data and human ratings from a new data pool. Other papers validate the dictionaries by using them to test competitive theories (Nicolas et al., under review).

Development Data

Development data allowed for initial tests of coverage and validity of the dictionaries. The development data consisted of a survey (N = 201)[4] asking for participants' spontaneous thoughts about characteristics that different social groups would have. We used a total of 20 social groups (e.g., "Asian," "Elderly," "Wealthy"), sampled from the literature, and showed five to each participant, in random order. Participants provided 10 open-ended single-word responses for each target. Next, participants saw the same social groups again and rated them on warmth (items: friendly, sincere) and competence (items: efficient, competent) using a scale ranging from 1 (*not at all*) to 5 (*extremely*), as well as a measure of familiarity with the social group. Finally, participants completed some demographic questions.

The open-ended responses were preprocessed in the following manner. First, we deleted extraneous spaces in the responses, deleted grammatical signs that were not of interest (e.g., dashes, dots), and transformed all letters to lower case. Then, we lemmatized the responses, which standardizes words related to the same lexical unit by removing, for example, inflectional endings (using R's korpus treetager; Michalke, 2017). Thus, words such as *running*, *ran*, and *run* all get transformed to *run*.

Subsequently, we used a custom spellchecker to correct for misspelled words. For responses not in the dictionary, our spellchecker first used the "wordnet" (Feinerer & Hornik, 2017) and "hunspell" (Németh L. Hunspell. <http://hunspell.sourceforge.net/> [accessed 25.02.13]) R packages to check whether a response was correctly spelled and, if not, to provide

[4] Under simplified assumptions and depending on the within-subject variance, power analyses for all studies reveal over 80% power to detect small effects of *r* or *f* between .1 and .2 in our main tests.

suggestions for correct spellings. Our spellchecker went through the top five suggestions, in order, and it had several arguments to choose a suggested correction. First, if the current suggestion was the participants' response before preprocessing, the preprocessed response was used. Second, if the current suggestion was mentioned by any participants for any of their responses, that suggestion (preprocessed) was used. The logic behind this step is that a suggestion mentioned by others ought to be more relevant to the topic being evaluated, and thus should be prioritized over other suggestions. If no suggestion conformed to the previous priorities, the first suggestion was preprocessed and used. Responses that had spaces (or hyphens) in them before preprocessing were not spellchecked (multiple words are unlikely to be properly corrected). In addition, the spellchecker marked words that indicated lack of knowledge (e.g., "idk", "na", "don't know") to code these into a separate dictionary. The following step deleted the "s" in words with that ending, to reduce duplicate words by removing plural forms.

Theory-driven Seed Dictionaries

For the collection of theory-driven seed words, we reviewed the literature (Abele, Uchronski, Suitner, & Wojciszke, 2008; Abele et al., 2016; Fiske et al., 2002; Koch et al., 2016; Todorov & Oosterhof, 2008; Wojciszke, Baryla, Parzuchowski, Szymkow, & Abele, 2011) for lists of words used to measure sociability, morality/trustworthiness, ability, assertiveness/dominance, status, political beliefs, and religious beliefs in relation to social groups. For every word, if not already included, we also obtained its antonym (using Wordnet). The final seed list consisted of 341 distinct words, with their corresponding theoretical direction (i.e., high or low on their corresponding dimension, based on how they were labeled in the reviewed literature (see Table 1 for example words; for full list see online repository). Because words can have multiple senses, the researchers independently went through the list of seed

words and decided on the most appropriate sense(s), based on their part of speech, definition, and example sentences, which resulted in a list of 455 senses[5].

[5] Final senses were those that the two researchers agreed on, 90% of the total senses selected by either researcher.

Table 1. Example words for each seed dictionary

Dimension	Term	Direction	Dimension	Term	Direction
Sociability	sociable	high	Status	wealthy	High
Sociability	unsociable	low	Status	poor	Low
Sociability	friendly	high	Status	powerful	High
Sociability	unfriendly	low	Status	powerless	Low
Sociability	warm	high	Status	superior	High
Sociability	cold	low	Status	inferior	Low
Sociability	liked	high	Status	influential	High
Sociability	disliked	low	Status	uninfluential	low
Sociability	outgoing	high	Status	successful	high
Sociability	shy	low	Status	unsuccessful	low
Morality	moral	high	Politics	traditional	high
Morality	immoral	low	Politics	modern	low
Morality	trustworthy	high	Politics	conventional	high
Morality	untrustworthy	low	Politics	unconventional	low
Morality	sincere	high	Politics	conservative	high
Morality	insincere	low	Politics	liberal	low
Morality	fair	high	Politics	republican	high
Morality	unfair	low	Politics	democrat	low
Morality	tolerant	high	Politics	narrow-minded	high
Morality	intolerant	low	Politics	open-minded	low
Ability	competent	high	Religion	religious	high
Ability	incompetent	low	Religion	irreligious	low
Ability	competitive	high	Religion	christian	high
Ability	uncompetitive	low	Religion	muslim	high
Ability	intelligent	high	Religion	jewish	high
Ability	unintelligent	low	Religion	atheist	low
Ability	able	high	Religion	secular	low
Ability	unable	low	Religion	believer	high
Ability	educated	high	Religion	nonbeliever	low
Ability	uneducated	low	Religion	skeptic	low
Assertiveness	confident	high			
Assertiveness	diffident	low			
Assertiveness	assertive	high			
Assertiveness	unassertive	low			
Assertiveness	independent	high			
Assertiveness	dependent	low			
Assertiveness	active	high			
Assertiveness	inactive	low			
Assertiveness	determined	high			
Assertiveness	doubtful	low			

Dictionaries were mostly balanced in terms of high and low senses, but bad-good morality (57.5%) and conservative-liberal politics (57.1%) were biased toward senses coded as low on their spectrum ("bad morality" and "conservative politics" respectively). On the other hand, low-high status (61.7%) and progressive-traditional religious beliefs (64.2%) were biased towards senses coded as high on their spectrum ("high status" and "traditional beliefs"). Note that by high and low we do not mean valence: it is simply an indicator of which end of the antonymy dimension the word refers to; whether one or the other antonym is coded as high vs. low is arbitrary. For example, we coded beliefs as ranging from progressive to traditional, and thus high direction in this dictionary means that the word is more about traditional beliefs than progressive beliefs. Specific dictionaries ranged from 28 senses (Religion and Politics) to 120 (Morality; Ability: 79; Assertiveness: 81; Sociability: 85; Status: 34).

Seed Dictionary Expansion

As expected, the words used in the existing literature to describe content dimensions were not a good measure of the diversity of open-ended responses, accounting for only 20.2% of our development data (6.2% of distinct responses). Mapping the content of spontaneous stereotypes requires accounting for most of the responses. However, open-ended responses allow for any number of synonymous terms that have not been exhaustively listed in previous studies. For instance, even though we were able to find words such as *thief* in the literature referring to morality, other synonyms such as *robber* were absent. Thus, in order to cover synonymy, antonymy, and other forms of semantic associations, we used Wordnet (Miller, 1995) as an objective source for dictionary expansion.

After the first round of dictionary words expansion, cluster analysis of the unaccounted-for words identified some potential additional topics. Specifically, given clustering results, we

identified topics and chose the higher-order noun that identified this topic in Wordnet. Subsequently, we went down the Wordnet semantic hierarchy for this noun, obtaining all its elements (i.e., for nationality, different nations). Then, for each element, we obtained associated terms (we used derivationally related forms to transform nouns into adjectives, for example from warmth to warm, and obtain additional relevant terms). For example, the cluster analysis suggested many words related to nationalities. A glance at the Wordnet hierarchy indicates that all nationalities are grouped under the word "inhabitant," so we used this single word to complete the nationality dictionary. A few specific unaccounted-for responses were added manually to a catch-all dictionary with words that denoted lack of knowledge (e.g., "I don't know" or "?"). Finally, we added some occupation-related words from the Bureau of Labor Statistics (2016) for completion, as an existing complementary dictionary (only 6 responses, < .001% of total, were accounted by this dictionary, and all those responses were also in Wordnet)[6].

Having these expanded dictionaries, we reviewed all the words that appeared in more than one dictionary. Then, for words that clearly belonged to a subset of those dictionaries, based on proximity to the seed word in Wordnet, we manually deleted them from the additional dictionaries. Note that this step is not necessary, and a version of the dictionary with no such deletions performed similarly.

In order to shorten the dictionaries, making them more manageable for validation studies, we used Word2Vec to exclude very rare words. Specifically, if a word did not appear in the

[6] We note that R's wordnet does not retrieve all the hyponyms for words in the format that Wordnet Online (http://wordnetweb.princeton.edu/perl/webwn) does. In particular, we identified that "instances" of hypernyms are not retrieved by R wordnet. At first, we dealt with this by manually adding 1660 hyponyms of seed words through Wordnet Online. This process may have missed some hyponyms. However, since then we added helper functions to obtain "instances" and other relations that might be missed by R wordnet's hyponyms for users' ease and thoroughness.

Word2Vec database we used for internal reliability (see next section; trained on a very large Google News corpora), it was deleted from the dictionaries. The inclusion of rare words is less detrimental than the deletion of words participants may use, so this approach aimed to be conservative in deletions. An even shorter version of the dictionaries (not used here) can be obtained by removing words that appear in multiple dictionaries (12% of preprocessed words).

The final dictionary data contained 14,449 words (13,930 preprocessed words) across 28 dictionaries. Final dictionaries varied in length from 7 (lack of knowledge) to 2,402 (morality) preprocessed words (see Table 2 and online repository for full dictionaries). For each sense, we also obtained their SentiWordnet (Baccianella, Esuli, & Sebastiani, 2010) valence. For words with multiple senses, the direction and valence were averaged across senses.

Table 2. Dictionary characteristics

Dictionary	Words	High	Low	Pos	Neg	Preprocessed	High	Low	Pos	Neg
Sociability	1210	505	430	0.21	0.24	1148	479	421	0.21	0.24
Morality	2523	477	1865	0.14	0.19	2404	458	1791	0.14	0.19
Ability	999	611	303	0.2	0.14	950	590	298	0.2	0.14
Assertiveness	774	453	269	0.16	0.16	731	423	255	0.17	0.17
Health	1477	39	1432	0.07	0.22	1427	35	1384	0.07	0.22
Status	595	291	193	0.17	0.14	560	279	183	0.17	0.14
Work	2051	NA	NA	0.03	0.02	1957	NA	NA	0.02	0.02
Politics	400	87	109	0.08	0.09	391	86	107	0.08	0.09
Religion	818	784	30	0.06	0.05	804	771	30	0.06	0.05
Beliefs - other	119	NA	NA	0.09	0.07	117	NA	NA	0.09	0.07
Inhabitant	664	NA	NA	0	0.01	657	NA	NA	0	0.01
Country	312	NA	NA	0	0.01	306	NA	NA	0	0.01
Feeling	1164	NA	NA	0.2	0.3	1088	NA	NA	0.2	0.31
Relative	215	NA	NA	0.04	0.02	214	NA	NA	0.04	0.02
Clothing	602	NA	NA	0.01	0.02	567	NA	NA	0.01	0.02
Ordinariness	147	52	88	0.17	0.23	146	52	88	0.17	0.23
Body part	390	NA	NA	0.03	0.03	353	NA	NA	0.02	0.03
Body properties	349	NA	NA	0.12	0.13	329	NA	NA	0.12	0.13
Skin	59	NA	NA	0.1	0.16	55	NA	NA	0.09	0.17
Body covering	216	NA	NA	0.01	0.04	208	NA	NA	0.01	0.04
Beauty	223	168	47	0.3	0.13	208	155	46	0.31	0.13
Insults	40	NA	NA	0.04	0.34	39	NA	NA	0.04	0.35
STEM	781	NA	NA	0.02	0.01	726	NA	NA	0.02	0.01
Humanities	83	NA	NA	0.09	0.02	80	NA	NA	0.08	0.02
Art	404	NA	NA	0.03	0.02	371	NA	NA	0.03	0.02
Social groups	31	NA	NA	0.06	0.06	31	NA	NA	0.06	0.06
Lacks knowledge	7	NA	NA	0.06	0.05	7	NA	NA	0.06	0.05
Fortune	28	NA	NA	0.25	0.19	27	NA	NA	0.24	0.2

Note. Words are the original words obtained, including different forms of a word (e.g., plural and singular) while preprocessed words collapses across these by lemmatizing, deleting symbols, among others (see development data section for preprocessing procedures). High and Low refers to the number of words for each direction of the dictionary, when available. Valence (Pos: Positive, Neg: Negative) was obtained from Sentiwordnet.

The final dictionaries accounted for 77% of the development data responses (47% of unique responses), indicating a significant improvement in coverage from the Wordnet expansion. Given the potential for overfitting to these specific data, in a confirmatory study (described in the first validation study), the dictionaries accounted for 84% of the responses (58% of unique responses). Thus, the dictionaries can account for the vast majority of stereotype-relevant responses in traditional groups explored in stereotyping research.

Having created extensive dictionaries that were able to capture over 3/4 of the content of social group stereotypes, the next steps were to ascertain the reliability and validity of the dictionaries. In order to explore the internal reliability of the dictionaries, we used similarity metrics from the semantic content in large text corpora (Word2Vec and Glove).

Reliability

A simple[7] inferential test for the consistency of dictionaries is to check whether the average pairwise similarity between words from the same dictionary is larger than the average pairwise similarity between words from different dictionaries. Indeed, using Word2vec similarities such a test reveals that words within a dictionary are more co-similar (M = .177) than words between dictionaries (M = .097), $t(29.14) = -8.69$, $p < .001$, $d = 1.67$. The same is the case when using Glove similarities, with within-dictionary words being more similar (M = .183) than between-dictionaries (M = .089), $t(29.25) = -8.46$, $p < .001$. $d = 1.63$. (Lacks knowledge dictionary not included due to its large number of abbreviations and non-words.) This large effect size denotes the semantic consistency of the Wordnet network, and therefore, our dictionaries.

[7] Although the test is "simple," combinations for such a large number of words quickly add up: using words found in Word2vec this resulted in 104,394,025 pairwise similarities.

Previous reports on the development of dictionaries (e.g., Pennebaker et al., 2015) have made use of more traditional metrics of internal reliability, such as alpha. Although the field's standard, alpha has some limitations as a measure of internal consistencies of dictionaries. Alpha measures a test's (i.e., a collection of items) correlational consistency, and usually for instruments that present all items to all participants. For dictionaries, on the other hand, not all items (i.e., words) are presented to all participants, so alpha, a measure of a test's consistency, given all the items presented, might be less appropriate. Additionally, alpha depends on the number of items, so our dictionaries, which tended to have very large numbers of words, will have high alphas regardless of the inter-item correlations.

Nonetheless, given the lack of standards in the field for measures of reliability of dictionaries, we present here the alphas obtained for our dictionaries. In order to obtain these alphas, we used the previously described measures of similarities as the inter-item correlations and applied the formula for alpha (as described in Chakrabartty, 2018). The results (see Table 3) indicated high internal reliabilities for most dictionaries (with the exception of the "lacks knowledge" dictionary, which had very few items found in word2vec and glove).

Table 3. Dictionary Cronbach Alphas based on Word2Vec and Glove.

Dictionary	Word2Vec	Glove	Dictionary	Word2Vec	Glove
Sociability	0.995	0.996	**Body property**	0.985	0.985
Morality	0.997	0.998	**Skin**	0.944	0.953
Ability	0.993	0.994	**Body covering**	0.978	0.975
Assertiveness	0.991	0.993	**Beauty**	0.981	0.981
Health	0.995	0.995	**Insults**	0.936	0.945
Status	0.985	0.989	**STEM**	0.994	0.993
Work	0.995	0.995	**Humanities**	0.963	0.959
Politics	0.987	0.987	**Art**	0.983	0.979
Religion	0.994	0.992	**Social Groups**	0.91	0.933
Beliefs - other	0.978	0.979	**Lacks knowledge**	0.374	0.481
Inhabitant	0.989	0.987	**Fortune**	0.858	0.898
Country	0.975	0.985	**Warmth**	0.998	0.998
Feeling	0.995	0.997	**Competence**	0.996	0.997
Relative	0.981	0.98	**Beliefs**	0.996	0.995
Clothing	0.993	0.992	**Geography**	0.991	0.989
Ordinariness	0.971	0.975	**Appearance**	0.996	0.996
Body part	0.986	0.988			

Validity as Related to Rating Scales

Given that scales are the traditional and most commonly used way of gathering information in psychology, we first tested how our dictionaries related to warmth, competence, and beliefs scales. For each group, in addition to seven open-ended responses, we collected participants Likert-type ratings on stereotype content dimensions[8]. We planned to test how well the scale ratings were predicted by our sociability, morality, ability, assertiveness, beliefs, and status dictionaries, all of which have been linked to these dimensions in the literature.

Method

Participants (N = 251) were recruited through Amazon Mechanical Turk (Mage = 33.3, 52% female; 76% White, 10% Black, 5% Hispanic, 4% Asian).

[8] In fact, the development data also included warmth and competence scales that largely replicate the results from this confirmatory analysis (see Supplement).

In an initial block, participants saw a sample of 4 social groups, from the same social groups as the development data. The instructions read: “Please indicate how the following people are viewed by society. Please note that we are not interested in your personal beliefs, but in how you think these people are viewed by others”. They were also told to use one word per box, two maximum, and then saw the prompt “As viewed by society, what are the characteristics of a person who is…” followed by the social group and seven boxes for responses. These responses were preprocessed in the same way as those from the development data.

In a second block, they saw the same groups, but rated them on scales. The prompt read “Please indicate how the following people would be viewed by society. Please note that we are not interested in your personal beliefs, but in how you think these people are viewed by others.” This was followed by “To what extent would most individuals in our society view a person who is **(social group)** as…” and a 1 (not at all) to 5 (extremely) scale for the items “Friendly/Sociable,” “Trustworthy/Moral,” “Self-confident/Assertive,” “Competent/Skilled,” “Wealthy/High-status,” “Politically conservative,” “Religious.” These items corresponded to the facets of sociability, morality, assertiveness, ability, status, politics, and religion. To form indexes, “Friendly/Sociable” and “Trustworthy/Moral” were combined for warmth (alpha = .76), “Self-confident/Assertive” and “Competent/Skilled” were combined for competence (alpha = .86), and “Politically conservative” and “Religious” were combined for beliefs (alpha = .7). After these blocks, participants completed demographic questions.

Analyses were mixed-effects models with participants and social groups as random factors, and observations were each participants’ responses to a group (i.e., aggregating across the seven responses for the text data). In terms of relevant variables, note that while scales measure only direction (e.g., low to high competence in a 5-point scale), our theory-driven

dictionaries measure as separate variables both prevalence and direction. *Prevalence* refers to the number of words related to the dimension (e.g., out of a participant's seven responses, more competence-related words indicate higher prevalence of competence). Thus, in this study, prevalence of each dictionary dimension was a count variable ranging from 0 to 7. On the other hand, *direction* refers to the antonymy dimensional end of the word. Words high on a dimension (e.g., *friendly* for Warmth) were coded as 1 for that dimension's direction, and words low on the dimension (e.g., *unfriendly* for Warmth) were coded as -1. If direction was unknown it was coded as 0, and if the response was not in the dictionary, it was coded as missing. To aggregate across the seven responses, we averaged these numerical codes (e.g., if a participant had five warmth-related words, two low and three high, their warmth direction score would be (3–2)/5 = 0.2). Thus, dictionary direction variables ranged from -1 to 1.

As a result of dictionaries partitioning information into prevalence and direction, different analytical strategies arise. In particular, we focus first on the more straightforward test of using the dictionary direction to predict the scales. If the dictionary is valid, dictionary direction should predict scale ratings: the higher the direction score, the higher the scale score. Therefore, we used linear models to predict scales from corresponding dictionary direction indicators.

Two additional models use dictionaries' partitioning of data to gain additional insights. Similar to incremental validity, if the prevalence of a dimension also predicts the scales, it would be evidence of how the dictionaries in these tasks can provide additional information (note this is a sufficient but not necessary test, as prevalence is from the start not necessarily expected to correlate with scales). We incorporate prevalence in two ways. First, models predicting dictionary prevalence from their corresponding scales use both linear and quadratic effects in Poisson models. A quadratic effect is likely, such that a higher number of dimensional words

relates to both high and low scale ratings for that dimension. That is, groups that are extreme in their scale score for a dimension (e.g., are very competent or very incompetent) may also elicit more responses related to that dimension (e.g., because it is more salient; Fiske, 1980). Conversely, participants may rate groups in the middle of the scale for a dimension either because they feel neutral, because they do not know how the group is stereotyped along that content, or because it is socially undesirable to rate the group on the dimension (see Nadler, Weston & Voyles, 2015); this might produce fewer open-ended responses for that dimension.

The second way we test for a possible role of prevalence is through its interaction with direction. These models are probably superior to the previously described models, as they incorporate the direction information encoded by the scales and are thus more likely to be predictive. We use linear models with the direction by prevalence interaction as the relevant predictor of the corresponding scale.

Results

Direction. We found the expected patterns of results, with all dictionary direction indicators providing predictive significance. We note however that the religion direction indicator was not significant for predicting the religion item. This was probably due to the low rate of religion-related open-ended responses, which greatly lowered the useable data for models with this variable. See Table 4 for all direction results[9].

[9] As expected, informal observations of cross-dictionary models (e.g., competence direction predicting Warmth scales) showed smaller and/or non-significant results compared to models for congruent dimension dictionaries. In order to conduct formal model comparisons we had to make additional assumptions and recode the direction variable such that missing values in the aggregate model were treated as zeroes (i.e., neutral direction). Using this recoding, all models presented in Table 4 (congruent dimensions) were significantly a better fit than cross-dictionary models, *p*s < .001.

Table 4. Prediction of scales by dictionaries direction and direction by prevalence interaction.

Outcome	Predictor	b	Beta	t	df	P	Marg. R^2
Warmth	Warmth direction	0.449	0.36	11.99	756.99	<.001	0.14
Warmth	Morality direction	0.503	0.422	11.5	591.88	<.001	0.182
Warmth	Sociability direction	0.39	0.337	9.37	471.24	<.001	0.12
Competence	Competence direction	0.424	0.302	10.15	832.41	<.001	0.115
Competence	Ability direction	0.32	0.242	6.46	618.71	<.001	0.068
Competence	Assertiveness direction	0.399	0.294	8.52	593.62	<.001	0.1
Morality	Morality direction	0.536	0.391	11.1	593.9	<.001	0.161
Sociability	Sociability direction	0.41	0.329	8.25	477.87	<.001	0.116
Ability	Ability direction	0.389	0.274	7.14	611.61	<.001	0.091
Assertiveness	Assertiveness direction	0.38	0.265	6.94	591.31	<.001	0.081
Beliefs	Beliefs direction	0.189	0.155	2.3	224.3	0.022	0.027
Beliefs	Politics direction	0.232	0.191	2.62	181.51	0.01	0.04
Beliefs	Religion direction	0.609	0.285	2.12	43.95	0.039	0.091
Politics	Politics direction	0.33	0.23	3.05	185.79	0.003	0.056
Religion	Religion direction	0.41	0.175	1.41	42.12	0.167	0.038
Warmth	Warmth direction*count	0.165	0.315	6.15	737.39	<.001	0.189
Warmth	Morality direction*count	0.17	0.278	4.64	578.57	<.001	0.231
Warmth	Sociability direction*count	0.204	0.292	3.95	440.57	<.001	0.143
Competence	Competence direction*count	0.097	0.187	3.9	823.62	<.001	0.15
Competence	Ability direction*count	0.031	0.046	0.83	606.38	0.409	0.073
Competence	Assertiveness direction*count	0.105	0.125	2.15	574.01	0.032	0.111
Beliefs	Beliefs direction*count	0.087	0.099	0.96	220.16	0.34	0.035
Beliefs	Politics direction*count	0.098	0.098	0.8	173.41	0.425	0.048
Morality	Morality direction*count	0.19	0.273	4.7	582.68	<.001	0.206
Sociability	Sociability direction*count	0.207	0.272	3.34	441.46	0.001	0.138
Ability	Ability direction*count	0.056	0.077	1.37	607.48	0.17	0.101
Assertiveness	Assertiveness direction*count	0.142	0.163	2.48	576.97	0.013	0.094
Politics	Politics direction*count	0.135	0.122	0.9	178.73	0.369	0.067

Note. Outcomes are scales and predictors are either the dictionary direction or the interaction between direction and prevalence (count). Models with religion as variable either needed further simplification (e.g., deletion of random intercepts for group), or were not computable, due to the low number of responses related to religion, such as in the case of interaction effects. Marginal R^2 are provided for the models.

Prevalence. In terms of predicting dictionary prevalence from the scales, we only found effects for warmth and its facets. In particular, for sociability we found linear effects, such that more positive sociability ratings predicted more sociability-related words. On the other hand, we

found quadratic effects for morality and combined warmth, such that more extreme scores on the scale, either high or low, predicted more morality- or warmth- related words. See Table 5 for all results and Figure 2 for an illustration of the quadratic effect of scale warmth on warmth prevalence.

Interaction. Finally, we tested for interaction effects between dictionary direction and prevalence. We found that dictionary prevalence interacted with direction for both warmth and competence, as well as a number of the facets (see Table 4 and Figure 3 for an example visualization). This suggests that dictionaries provide separable pieces of information in the form of prevalence and direction that are both predictive of their scale counterparts.

Table 5. Prediction of dictionary prevalence by scales.

Outcome	Predictor	Effect	b	B	Z	p	Marg. R^2
Morality	Morality	linear	-1.26	-0.033	-0.99	0.323	0.03
Morality	Morality	quadratic	5.46	0.142	5.44	<.001	0.03
Sociability	Sociability	linear	3.06	0.101	2.13	0.033	0.01
Sociability	Sociability	quadratic	2.47	0.082	1.88	0.06	0.01
Warmth	Warmth	linear	-2.02	-0.052	-1.65	0.099	0.02
Warmth	Warmth	quadratic	3.99	0.104	4.1	<.001	0.02
Morality	Warmth	linear	-1.04	-0.023	-1.03	0.3	0.01
Morality	Warmth	quadratic	2.78	0.061	3.37	<.001	0.01
Sociability	Warmth	linear	4.18	0.139	2.63	0.009	0.01
Sociability	Warmth	quadratic	1.04	0.035	0.78	0.432	0.01
Ability	Ability	linear	0.3	0.01	0.2	0.84	0
Ability	Ability	quadratic	1.04	0.03	0.96	0.336	0
Assertiveness	Assertiveness	linear	1.03	0.03	0.72	0.47	0
Assertiveness	Assertiveness	quadratic	0.56	0.017	0.49	0.624	0
Competence	Competence	linear	1.28	0.026	1.13	0.258	0
Competence	Competence	quadratic	0.48	0.01	0.58	0.563	0
Ability	Competence	linear	-0.32	-0.01	-0.21	0.974	0
Ability	Competence	quadratic	0	0.04	0.03	0.974	0
Assertiveness	Competence	linear	2.72	0.08	1.8	0.07	0.01
Assertiveness	Competence	quadratic	1.42	0.04	1.22	0.224	0.01
Politics	Politics	linear	2.03	0.114	0.708	0.479	0
Politics	Politics	quadratic	1.59	0.089	0.745	0.456	0
Religion	Religion	linear	14.73	1.412	1.74	0.081	0.03
Religion	Religion	quadratic	2.51	0.241	0.446	0.655	0.03
Beliefs	Beliefs	linear	1.89	0.089	0.72	0.475	0
Beliefs	Beliefs	quadratic	0.53	0.025	0.28	0.778	0
Politics	Beliefs	linear	2.22	0.125	0.771	0.44	0
Politics	Beliefs	quadratic	1.67	0.094	0.786	0.432	0
Religion	Beliefs	linear	16.24	1.56	1.65	0.099	0.04
Religion	Beliefs	quadratic	-4.81	-0.461	-0.81	0.417	0.04

Note. Outcomes are dictionary prevalence and predictors are the corresponding scale score. Marginal R^2 are provided for the models.

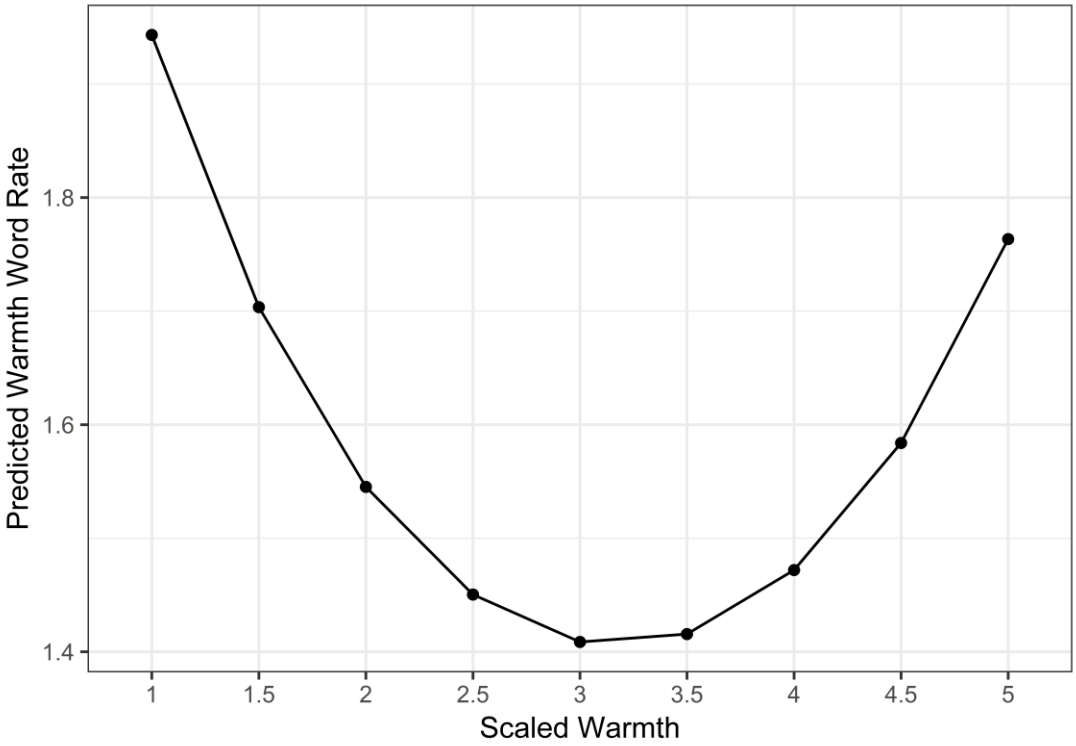

Figure 2. Quadratic prediction of scale warmth on dictionary warmth prevalence.

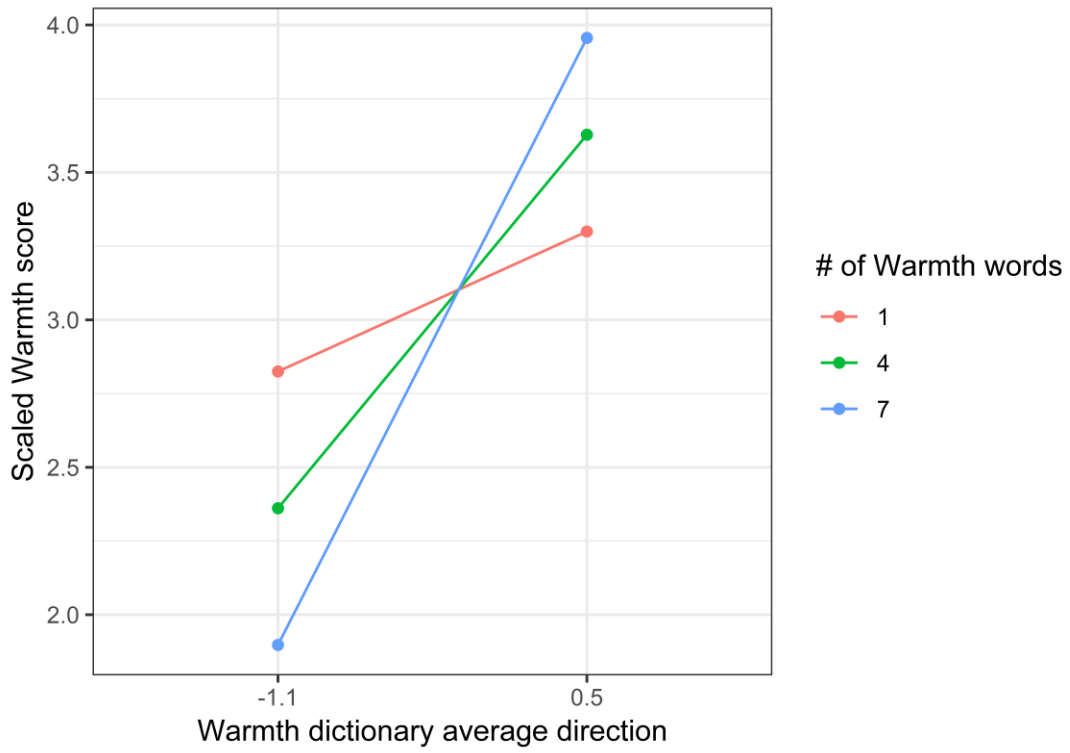

Figure 3. Interaction prediction between warmth direction and prevalence (i.e., # of warmth-related words) on scale warmth.

Validity as Related to Human Judgment

A second test of validity used human ratings of thematic identification to study the extent to which the dictionary words reflect human semantic judgments. Specifically, we expected to show that coders appropriately identify words from a dictionary as belonging to it.

Method

Participants (N = 245) were recruited through Amazon Mechanical Turk (*M*age = 33.3; 61% male; 78% White, 9% Black, 6% Asian, 2% Hispanic). Participants saw 13 questions, each of which presented a random sample of 6 of the words of a dictionary. Instructions asked participants to identify the common theme of the six words and to rate on a scale from 1 (*Not at all*) to 6 (*Extremely*) how well they fit into a condensed list of our dictionaries, in lay terms: sociability/friendliness, morality/trustworthiness, confidence/autonomy, ability/skill,

socioeconomic status, political or religious beliefs, health, work/professions, body properties/parts/appearance, familiarity/family, feelings/emotions, and geography. Participants were told to consider words from both directions (e.g., both morality and immorality) to refer to the same theme and were asked to base their responses on the objective meaning of the words rather than personal opinion.

Analyses consisted of a series of mixed models (participants as random intercepts), one for each dictionary. If human judgments agree with the dictionary, we should expect a linear relation between participants' responses and the dictionary words. For example, if the block presented a subset of morality words, data from this block was used for the morality model. Because each block had 12 questions, one for each dictionary label (e.g., morality/trustworthiness, confidence/autonomy), each of these questions (or labels) became a level in a contrast-coded factor predictor. The response to the questions was the outcome variable, allowing us to statistically compare the mean scale response for each question, for each dictionary. Thus, we explored whether participants scored highest on the congruent response option for each dictionary. For example, we expected the morality score (vs. scores for other dimensions) to be higher for the morality dictionary.

Results

Analyses largely supported all the expected patterns. Given the large number of tests, results are summarized in Table 6. In general, the congruent score for each dimension was significantly higher than the score for all other incongruent dimensions. Exceptions were only for the sociability and assertiveness dictionaries, where scores on the emotion response option were not significantly lower than scores on the congruent dimensions. Note that this was only the case for sociability when correcting for multiple testing. Possibly, sociability and assertiveness are

simply more highly correlated to emotional words; for example, the Stereotype Content Model posits that stereotypes trigger emotions (Fiske et al., 2002) or because others' positive or negative emotions tell us about their (at least state) friendliness, or because approach and avoidance emotions (see Elliot, Eder, & Harmon-Jones, 2013) relate to assertiveness. In any case, participants were able to correctly identify the dictionary dimensions by using only random subsets of 6 words.

Table 6. Estimated values and pairwise comparisons between congruent and incongruent response options.

	sociability	morality	assertiveness	ability	status	beliefs	health	work	body	family	emotions	geography
Sociability	**3.69**	2.92***	2.8***	2.7***	2.61***	2.58***	2.56***	2.56***	2.49***	2.73***	3.46^	2.47***
Morality	3.11**	**3.42**	2.84***	2.77***	2.78***	2.72***	2.52***	2.75***	2.47***	2.62***	3.04***	2.48***
Assertiveness	2.94***	2.84***	**3.4**	3.14**	2.66***	2.6***	2.65***	2.87***	2.54***	2.6***	3.27	2.47***
Ability	2.87***	2.73***	3.15***	**3.77**	2.74***	2.57***	2.55***	2.82***	2.57***	2.63***	2.87***	2.47***
Status	2.86***	2.73***	3.1***	3.14***	**3.52**	2.6***	2.63***	2.88***	2.5***	2.7***	2.88***	2.51***
Beliefs	2.71***	2.9***	2.74***	2.68***	2.77***	**3.81**	2.47***	2.71***	2.58***	2.68***	2.82***	2.79***
Health	2.4***	2.54***	2.45***	2.47***	2.42***	2.32***	**4.14**	2.48***	3.18***	2.52***	2.72***	2.46***
Work	2.63***	2.55***	2.69***	3.24***	2.86***	2.54***	2.62***	**3.86**	2.56***	2.54***	2.62***	2.62***
Body	2.69***	2.59***	2.65***	2.67***	2.7***	2.57***	2.7***	2.6***	**3.74**	2.56***	2.67***	2.56***
Family	2.84***	2.71***	2.65***	2.64***	2.76***	2.52***	2.53***	2.62***	2.59***	**3.89**	2.86***	2.55***
Emotions	3.09***	2.78***	3.03***	2.69***	2.65***	2.5***	2.69***	2.52***	2.56***	2.62***	**4.01**	2.43***
Geography	2.6***	2.58***	2.54***	2.49***	2.76***	2.76***	2.44***	2.56***	2.58***	2.71***	2.55***	**4.25**

Each row is a dictionary and each column a response option. Values are coefficient for the response in the specified dictionary. P-values refer to the pairwise comparison between that baseline (the congruent response score) and the column response (e.g., the morality response for the sociability dictionary, 2.92, is significantly smaller than the sociability response for the sociability dictionary, 3.69). P-values for each outcome control for family-wise multiple comparisons by using the Dunnet method for 11 tests. For sociability dictionary, emotion and sociability are different at p = .009 when not adjusting multiple comparisons. For assertiveness dictionary, emotion and assertiveness are at p = .116 when not adjusting for multiple comparisons.

* p < .05, ** p < .01, *** p < .001, ^ NS only when controlling for multiple testing. **Bold** (diagonal), congruent scores, no pairwise comparisons with themselves, so no significance testing is provided, all are different from zero.

Discussion

This paper reviewed a way to create text analysis instruments. Our approach is more automated than existing human-coded approaches and based on standardized sources such as Wordnet and word embeddings such as Word2Vec's and Glove's trained models. Furthermore, we provided guidelines for the evaluation of text analysis instruments, including coverage, reliability, and validity. Finally, we illustrated how our approach can develop dictionaries that meet the criteria for these three indicators, using the stereotype content literature as a context.

Summary of Proposed Approach

The approach presented here consists of a series of steps that can be modified to the question of interest. Our steps to develop dictionaries follow:

1. Identify seed words related to the construct of interest from the literature or through a data-driven approach using cluster analyses or open-ended questions. Test coverage if desired for a benchmark.
2. Expand these seed words by using Wordnet to obtain synonyms, antonyms, hyponyms, and other associated terms.
3. Test coverage using a model task that measures the construct in a theoretically-relevant context. In this case, a measure of open-ended stereotypes of multiple social groups seemed an appropriate prototypical task for the dictionaries and was thus used to measure coverage. If coverage is acceptable (which could vary by researcher expectations), move to step 5.
4. If coverage seems inadequate, the researcher could look at unaccounted-for words. If the number of unaccounted words is large, a cluster analysis using word embeddings may facilitate identification of relevant constructs neglected in the initial seed word

selection. These seed words could be expanded by repeating step 2 but could also be simplified by finding construct-relevant "hypernym" word-senses in Wordnet that would facilitate expansion by getting all its hyponyms. For example, if many unaccounted words seem related to emotional states (e.g., sadness, anger), the hypernym word-sense for *feeling* (sense 1) could be expanded into all its hyponyms (and their synonyms, antonyms, etc.) to cover for all these unaccounted words. Once coverage is acceptable, move to the next step.

5. Use word embeddings as an independent and standard source of semantic cosimilarity in order to determine dictionary reliability. Reliability could be evaluated by looking at average cosimilarities within and between dictionaries (as reported here for consistency with current practices, alpha could also be calculated). Further research could explore developing more sophisticated metrics of reliability.
6. If reliability is acceptable, validate the dictionaries by using human-coding on subsets of the words and by correlating dictionary word-frequency with theoretically-related measures, among other possible methods.

Summary of Current Studies

The current studies used the above steps to develop dictionaries for the measurement of stereotype content. The field of stereotype content is ever-growing but suffers from a deficit of discovery-driven studies exploring open-ended stereotype responses, and a lack of access to online text data due to the limitations of current instruments. Thus, besides illustrating the steps in this tutorial, we expect these dictionaries to be broadly applicable to stimulate the field of stereotype content. The previous steps, as applied to stereotype content, follow:

1. We identified 341 words (455 senses) for the literature-relevant constructs of sociability, morality/trustworthiness, ability, assertiveness/dominance, status, and political and religious beliefs, as well as indicators of their direction (i.e., high or low on the dimensions). In our case, this was a fully theory-driven step, but it was complemented by data-driven dictionaries in the following steps. These words accounted for only about 20% of participants' stereotypes.
2. We used Wordnet to expand the seed words into fuller dictionaries. We also identified additional seed words from unaccounted-for responses and expanded those as well. After some additional manual revisions and deletion of rare words (non-necessary steps but which made further analyses easier), we had the final version of the instrument with 28 dictionaries and 14,449 words (13,930 preprocessed words). We also obtained the valence for all these words, in addition to direction for most of the dictionaries.
3. The expansion of words resulted in over 80% coverage in a confirmatory study. We considered this coverage to be acceptable, and it was a considerable improvement from the existing words in the literature.
4. Given that after seed expansion and identification of additional seed words in previous steps brought coverage to a satisfactory level for such a diverse response set, we moved on to the next step.
5. We obtained the pairwise similarities between all words in our dictionaries using Word2Vec and Glove word embeddings. These metrics indicated that words within dictionaries were more semantically similar than words between dictionaries, suggesting the expected internal consistency.

6. We validated our dictionaries in two separate studies. First, for the theory-driven dictionaries, we used open-ended data to predict scaled warmth, competence, and beliefs ratings. These results supported that the dictionaries we created predicted social group's predicted warmth, competence, and beliefs. But beyond this, the dictionaries also tapped into information about stereotype content not extractable from the scales, namely prevalence of the different dimensions. This finding highlights an example of the benefits of using open-ended text data with the aid of dictionaries. In a second study, we presented human coders with subset of words from each dictionary and asked them to rate how much the subsets referred to different contents. Participants were able to accurately place the words into the expected dictionary with great concordance. Thus, we successfully created and validated high-coverage dictionaries in an area that lacked such instruments.

Text data is vital for new and renewing fields of psychology. Developments in machine learning fields such as natural language processing have opened the door for psychologists to tap into these so-far underused sources of information. As evidenced from theoretical developments and revisions based on discovery-oriented open-ended measures in fields such as emotion and social categorization research, embracing text data in experimental contexts is of great value to psychologists. In addition, archival text data are now ubiquitous (e.g., from social media or internet archives) and provides out-of-lab insights into human behavior. Being able to identify constructs of interest in these sources and create instruments for their measurement will generate opportunities to expand the science by allowing us to ask new questions, or extend previous findings to a wider variety of contexts of potential higher ecological validity.

In this paper we have provided a tutorial on how to create mostly automated dictionaries for the measurement of psychological constructs in text data. Our approach provides several advantages. For example, existing approaches heavily rely on multiple human judges to create the totality of the dictionaries over multiple iterations of group discussion and subjective decisions. This process is resource-intensive and time-consuming, and risks introducing selection biases, based on a specific group of judges who might differ from other judges.

On the other hand, the approach provided here largely automates the process, greatly reducing the time and resources necessary for the creation of the dictionaries. (We provide all the R code used here for readers to be able to implement the procedures for creating dictionaries of their own.) To be sure, the initial steps of our approach necessitate two judges to make decisions about seed words/senses. In our illustration the initial seed words were obtained from the literature and were thus independent from us and theory-driven. However, the selection of senses for Wordnet expansion can potentially introduce some subjectivity for multi-senses words[10]. Admittedly, also, Wordnet itself was created by human judges, but it has been extensively tested, validated, and used, and it provides a standardized source for expansion.

As an alternative to theory-driven seeds that we also utilized, researchers could opt to select a single hypernym noun seed word for expansion, when available in Wordnet. For novel areas where seed words might not be found in the literature, these hypernym seeds can be obtained in a data-driven fashion by exploring development text data through clustering algorithms (via word embedding similarity) or topic models. Using this alternative generative approach could result in even less subjective decisions from researchers.

[10] Further development of this approach could introduce automated sense-disambiguating techniques. For example, given a set of seed words it would be possible to obtain their word embeddings using Wordnet vector similarity metrics (Patwardhan & Pedersen, 2006) in order to select the senses most similar to the average vector representation of all seed words.

The use of Wordnet in the development of dictionaries has multiple other advantages. Given Wordnet's multi-sense network it is possible to obtain valence scores for specific senses using Sentiwordnet. While most sentiment analyses rely on the words, Sentiwordnet provides different valence scores for each sense. This is important for many psychologically-relevant words that share meaning with less psychologically relevant words, such as *warmth*, a central concept in the field of stereotype content illustrated here, which has multiple other meaning including of course for physical warmth. This advantage reduces noise in sentiment analyses when the context of interest is known. In addition, knowing the sense of a word allows for superior translation to other languages using tools such as Babelnet. Babelnet allows for translation from Wordnet senses into their corresponding sense in other-language wordnets. Translation can depend on context, and this is facilitated by Wordnet's structure, otherwise requiring manual translation by fluent speakers of the target language. Given the neglect of cross-cultural research, using Wordnet and Babelnet to study text in multiple languages can provide a fruitful avenue for the generalizability of psychological findings.

Complementarity with Existing Methods

The current paper also provides more general guidance on the evaluation of the reliability and validity of dictionaries for psychological research. Existing reports do not always follow extensive, standardized protocols to validate the instruments. A benchmark exception protocol is the LIWC authors' development of their updated 2015 dictionaries (Pennebaker, Boyd, Jordan, & Blackburn, 2015). Their paper details the process of judge selection of words for the dictionaries. Then, they report on the internal consistency by using word percentages in large corpora. Thus, for each word in a dictionary (a variable), they obtained its occurrence proportion across thousands of documents, each of which was an observation. Then they calculated alpha

for each dictionary based on these metrics. Although transparent and familiar, this is (as the LIWC authors admit), a noisy method based on co-occurrences of words from the same dictionary within a topic. Arguably, if a document is about achievement, the writer might use multiple words related to achievement, such as success and failure in the same document. However, the writer of the document might well use the same words throughout, or the document might only briefly mention the topic using a single word from the dictionary.

The method we introduce here for the evaluation of internal consistency completely sidesteps this issue and provides a complementary measure of semantic similarity. In fact, word embeddings such as Word2vec and Glove are semantic models that represent words as vectors and facilitate multiple operations on them. Furthermore, trained models are freely accessible and easily implemented through available packages in statistical software such as R.

The LIWC authors provided the conventional Cronbach's alphas as a measure of internal reliability, which we also do here. However, we note that Cronbach's alpha may not always be ideal for these kinds of data. Dictionaries differ from traditional scales in multiple respects, including that (a) many more items are often included in dictionaries, which is known to result in larger alphas, and (b) traditional metrics such as correlation might be inappropriate for many semantic indicators, particularly when words are treated as observations. Instead, perhaps average similarities within the dictionary can be tested against a benchmark (e.g., words between dictionaries, as here). However, the ideal solution would be the development of more specialized metrics for this purpose. We suggest that future research could tackle this issue.

We explored convergent validity by using open-ended data, obtaining word-frequencies of the dictionaries on that data, and correlating it to rating data. In addition to this approach, we successfully presented subsets of words from each dictionary to measure the degree to which

they were able to match the label of the concordant dictionary. Others may use additional methods to provide evidence for the validity of their dictionaries, in a similar fashion to more traditional scales.

Limitations and Future Directions

As research on natural language evolves, so will the opportunities for text analysis in psychology. The use of dictionaries has been growing in some areas for some time, and we provide here a guide on how to develop novel dictionaries in a more automated and standardized manner for constructs currently lacking them. However, dictionaries, as any instrument, have limitations.

Limitations of dictionaries occur when used in longer text data routinely found from most online sources such as social media. When dealing with sentences and paragraphs, a simple word-counting approach misses some of the sentence-level structure, potentially resulting in more noisy estimates. For example, negations, modifiers, and sarcasm are missed by dictionaries in a word-counting approach. Some potential fixes incorporate more advanced techniques such as dependency parsing for negations and modifiers. These methods identify relationships between words in a sentence, but highly complicate the analyses. For these reasons, particular care is required when using dictionaries in long-text data. For controlled experiments, we recommend designing studies with this limitation in mind, asking for example for single-word responses which are adequate for numerous cases, such as the one illustrated here with stereotyping.

Among other limitations, dictionaries provide categorical measures of a word's semantic association to an overarching topic. However, words differ on their prototypicality for the specified construct (e.g., in a dictionary for sociability, the word sociable may be more

prototypical than the word extroverted) and may belong to multiple constructs simultaneously to different extents (e.g., the word extroverted may be classified as related both to sociability and to assertiveness to different extents). In addition, even if dictionaries have good coverage for the task of interest, as here, they may not provide usable coding for short texts that may be common in some sources (e.g., some social media). Finally, language is dynamic, and even though most words in a dictionary will continue to be used in the same way over many years, some may change meaning, or the dictionary may miss new words or senses in the future.

In response to some of the previously mentioned limitations, a recent paper (Garten et al., 2018) proposes using a mix of dictionaries and word embeddings to code text data. In this manner, an average word embedding vector would be obtained for all the words in a dictionary, representing the prototypical vector for the construct. Then, similarities could be obtained between each word response and the prototypical vector, providing a continuous measure of semantic similarity to the construct. In addition to providing a continuous measure rather than an all-or-nothing metric, this approach can be used with non-exhaustive dictionaries. We believe this approach is promising and complementary to the one presented here.

We explored our development and validation Study 1 data, predicting warmth and competence scales with both word-counting (i.e., word-counting plus the directional indicator) and word embedding coding (with three variables, one for high and one for low senses plus their interaction) and using either the full dictionaries or the seed dictionaries. Based on marginal R^2 values (see Supplement), in three of the four tests, word counting + direction and full dictionary word embeddings performed similarly. In three of the four tests, full dictionary word embeddings outperformed word embeddings based on just seed words. In one of the tests, word counting + direction notably underperformed both word embedding models. Thus, although

variable, in general we found evidence for better performance when using Wordnet-expanded dictionaries, particularly in combination with word embeddings. This suggests that our procedure to expand dictionaries can be used independently in a word counting approach or combined with novel word embeddings approaches (Garten & colleagues, 2018), often outperforming either method based on smaller dictionaries. Additionally, if a smaller dictionary is preferred, in cases where even a small number of words is difficult to obtain from the literature, our method can be used for more limited expansion.

We note that although the full dictionary word embedding tended to outperform the word counting + direction model, the latter partitions the data in a qualitatively different way that may be uniquely useful for some questions. For example, in questions about the content of responses, word counting approaches allow for the creation of taxonomies and the measurement of coverage. This is a useful discovery-oriented step when aiming to study a process for which constructs of interest are unknown. In our illustration, by exploring word counting coverage we were able to find relevant constructs that were not previously included in our seed words. Additionally, an experimental paradigm like the validation Study 1 provides both a measure of accessibility of the dimension (i.e., how many times is the dimension is mentioned) and of the direction. Using word embeddings makes it more difficult to estimate the accessibility of the dimension, because no response is clearly classified into one dimension. Thus, each method has distinct strengths.

Beyond these advantages, Wordnet expansions allow for a linguistic examination of the properties of the dimension (e.g., are there more positive or negative words, more high or low dimension dictionaries). Finally, as mentioned previously, Wordnet is available in several

languages (e.g., see Babelnet; https://babelnet.org/), allowing for sense-specific translation.

Conclusion

In this paper, we provide guidance and examples on how to import natural language processing methods, specifically Wordnet and word embeddings, into the automation, creation, and evaluation of psychological text instruments. Text data open the possibilities to ask novel questions about behavior in the laboratory and beyond and may provide a way to improve both psychological theory and practice. We hope that the procedures outlined here greatly facilitate the use of text data to complement traditional approaches in psychology.

Open Practices Statement

The data, code, and materials for all studies are available at https://osf.io/yx45f/?view_only=570a9017944d4ecfa35a88e690f081d2. None of the studies were preregistered.

References

Abele, A. E., & Bruckmüller, S. (2011). The bigger one of the "Big Two"? Preferential processing of communal information. *Journal of Experimental Social Psychology, 47*(5), 935-948. Doi: 10.1016/j.jesp.2011.03.028

Abele, A. E., & Wojciszke, B. (2014). Communal and agentic content in social cognition: A dual perspective model. *Advances in Experimental Social Psychology, 50,* 195–255. Doi: 0.1016/B978-0-12-800284-1.00004-7

Abele, A. E., Uchronski, M., Suitner, C., & Wojciszke, B. (2008). Towards and operationalization of fundamental dimensions of agency and communion: Trait content ratings in five countries considering valence and frequency of word occurrence. *European Journal of Social Psychology, 38*(7), 1202-1217. Doi: 10.1002/ejsp.575

Abele, A. E., Hauke, N., Peters, K., Louvet, E., Szymkow, A., & Duan, Y. (2016). Facets of the fundamental content dimensions: Agency with competence and assertiveness—Communion with warmth and morality. *Frontiers in psychology, 7,* 1810. Doi: 10.3389/fpsyg.2016.01810

Baccianella, S., Esuli, A., & Sebastiani, F. (2010). *Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining*. In Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC'10), pages 2200–2204, Valletta.

Bergsieker, H. B., Leslie, L. M., Constantine, V. S., & Fiske, S. T. (2012). Stereotyping by omission: Eliminate the negative, accentuate the positive. *Journal of Personality and Social Psychology, 102*(6), 1214-1238. Doi: 10.1037/a0027717

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM, 55*(4), 77-84. Doi: 10.1145/2133806.2133826

Chakrabartty, S. N. (2018). Cosine Similarity Approaches to Reliability of Likert Scale and Items. *Romanian Journal of Psychological Studies, 1*(6), February-2018. Available at SSRN: https://ssrn.com/abstract=3202379

Decter-Frain, A., & Frimer, J. A. (2016). Impressive words: linguistic predictors of public approval of the US congress. *Frontiers in psychology, 7*, 240. Doi: 10.3389/fpsyg.2016.00240

Diehl, M., Owen, S., & Youngblade, L. (2004). Agency and communion attributes in adults' spontaneous self-representations. *International journal of behavioral development, 28*(1), 1-15. Doi: 10.1080/01650250344000226

Dupree, C. H., & Fiske, S. T. (in press.) Self-presentation in interracial settings: The competence downshift by White liberals. *Journal of Personality and Social Psychology*.

Ellemers, N. (2017). *Morality and the regulation of social behavior: Groups as moral anchors*. London: Routledge

Elliot, A. J., Eder, A. B., & Harmon-Jones, E. (2013). Approach–avoidance motivation and emotion: Convergence and divergence. *Emotion Review*, *5*(3), 308-311. Doi: 10.1177/1754073913477517

Feinerer, I. & Hornik, K. (2017). *wordnet: WordNet Interface*. R package version 0.1-14, https://CRAN.R-project.org/package=wordnet.

Fellbaum, C. (1998, ed.) *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.

Fiske, S. T. (1980). Attention and weight in person perception: The impact of negative and extreme behavior. *Journal of Personality and Social Psychology, 38*(6), 889-906. Doi: 10.1037/0022-3514.38.6.889

Fiske, S. T. (2018). Stereotype content: Warmth and competence endure. *Current Directions in Psychological Science, 27*(2) 67–73. Doi: 10.1177/0963721417738825

Fiske, S. T., Cuddy, A. J. C., Glick, P., & Xu, J. (2002). A model of (often mixed) stereotype content: Competence and warmth respectively follow from perceived status and competition. *Journal of Personality and Social Psychology, 82*, 878–902. Doi: 10.1037/0022-3514.82.6.878

Garten, J., Hoover, J., Johnson, K. M., Boghrati, R., Iskiwitch, C., & Dehghani, M. (2018). Dictionaries and distributions: Combining expert knowledge and large scale textual data content analysis. *Behavior research methods, 50*(1), 344-361. doi: 10.3758/s13428-017-0875-9

Gendron, M., Roberson, D., & Barrett, L. F. (2015). Cultural variation in emotion perception is real: A response to Sauter, Eisner, Ekman, and Scott (2015). *Psychological science, 26*(3), 357-359. doi: 10.1177/0956797614566659

Goodwin, G. P. (2015). Moral character in person perception. *Current Directions in Psychological Science, 24*(1), 38-44. doi: 10.1177/0963721414550709

Iliev, R., Dehghani, M., & Sagi, E. (2015). Automated text analysis in psychology: Methods, applications, and future developments. *Language and Cognition, 7*(2), 265-290. doi: 10.1017/langcog.2014.30

Katz, D., & Braly, K. (1933). Racial stereotypes of one hundred college students. *The Journal of Abnormal and Social Psychology, 28*(3), 280. doi: 10.1037/h0074049

Koch, A., Imhoff, R., Dotsch, R., Unkelbach, C., & Alves, H. (2016). The ABC of stereotypes about groups: Agency/socioeconomic success, conservative–progressive beliefs, and communion. *Journal of Personality and Social Psychology, 110*(5), 675.

Koch et al., under review. Judging Warmth is a Personal Matter: Less Consensus in Rating Groups' Communion than Other Stereotype Dimensions Reconciles the Agency-Beliefs-Communion (ABC) Model with the Stereotype Content Model.

Michalke, M. (2017). *koRpus: An R Package for Text Analysis (Version 0.10-2)*. Available from https://reaktanz.de/?c=hacking&s=koRpus

Mikolov, T., Chen, K., Corrado, G., Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint*, arXiv:1301.3781.

Miller, G. A. (1995). WordNet: a lexical database for English. *Communications of the ACM, 38*(11), 39-41. Doi: 10.1145/219717.219748

Miller, G. A., & Charles, W. G. (1991). Contextual correlates of semantic similarity. *Language and cognitive processes, 6*(1), 1-28. Doi: 10.1080/01690969108406936

Nadler, J. T., Weston, R., & Voyles, E. C. (2015). Stuck in the middle: the use and interpretation of mid-points in items on questionnaires. *The Journal of General Psychology, 142*(2), 71-89. Doi: 10.1080/00221309.2014.994590

Navigli, R., & Ponzetto, S. P. (2012). BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence, 193*, 217-250. Doi: 10.1016/j.artint.2012.07.001

Nicolas, G., Bai, X., & Fiske, S. T. (in press). Exploring research methods blogs in psychology: Who posts what about whom, with what effect. *Perspectives on Psychological Science*.

Nicolas, G., Bai, X., Fiske, S. T., Terache, J., Carrier, A., & Yzerbyt, V., Koch, A., Imhoff, R., & Unkelbach, C. (under review). Warmth is central, among social cognitive dimensions: Natural language analysis of intergroup information seeking.

Nicolas, G., Skinner, A. L., & Dickter, C. L. (2018). Other than the sum: Hispanic and Middle Eastern categorizations of Black-White mixed-race faces. Social and Personality Psychological Science. doi: 10.1177/1948550618769591

Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). *The development and psychometric properties of LIWC2015*. Available at http://liwc.net/howliwcworks.php.

Pennington, J., Socher, R., & Manning, C. (2014). *Glove: Global vectors for word representation*. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP) (pp. 1532-1543).

Pietraszkiewicz, A., Formanowicz, M., Gustafsson Sendén, M., Boyd, R. L., Sikstrom, S., & Sczesny, S. (2018). The big two dictionaries: Capturing agency and communion in natural language. *European Journal of Social Psychology*. doi: 10.1002/ejsp.2561

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*(1), 24-54. Doi: 10.1177/0261927X09351676

Oosterhof, N. N., & Todorov, A. (2008). The functional basis of face evaluation. *Proceedings of the National Academy of Sciences, 105*(32), 11087-11092. Doi: 10.1073/pnas.0805664105

Uchronski, M. (2008). Agency and communion in spontaneous self-descriptions: Occurrence and situational malleability. *European Journal of Social Psychology, 38*(7), 1093-1102. Doi: 10.1002/ejsp.563

Wojciszke, B. (1994). Multiple meanings of behavior: Construing actions in terms of competence or morality. *Journal of Personality and Social Psychology, 67*(2), 222. Doi: 10.1037/0022-3514.67.2.222

Wojciszke, B., Bazinska, R., & Jaworski, M. (1998). On the dominance of moral categories in impression formation. *Personality and Social Psychology Bulletin, 24*(12), 1251-1263. Doi: 10.1177/01461672982412001

Wojciszke, B., Baryla, W., Parzuchowski, M., Szymkow, A., & Abele, A. E. (2011). Self-esteem is dominated by agentic over communal information. *European Journal of Social Psychology, 41*(5), 617-627. Doi: 10.1002/ejsp.791