The Development and Psychometric Properties of LIWC-22

Ryan L. Boyd, Ashwini Ashokkumar,
Sarah Seraj, and James W. Pennebaker

The University of Texas at Austin

General email correspondence should be sent to Ryan L. Boyd or James W. Pennebaker. Specific inquiries about LIWC-22 should be directed to the LIWC team via https://www.liwc.app/contact.

The official reference for this paper is:

Boyd, R. L., Ashokkumar, A., Seraj, S., & Pennebaker, J. W. (2022). *The development and psychometric properties of LIWC-22*. Austin, TX: University of Texas at Austin. https://www.liwc.app

Table of Contents

The Development and Psychometric Properties of LIWC-22

The words that people use in everyday life tell us about their psychological states: their beliefs, emotions, thinking habits, lived experiences, social relationships, and personalities. From the time of Freud's writings about "slips of the tongue" to the early days of computer-based text analysis, researchers across the social sciences have amassed an extensive body of evidence showing that people's words have tremendous psychological value. To appreciate some of the truly great pioneers, check out (Allport, 1942), Gottschalk and Gleser (1969), Stone et al., (1966), and Weintraub (1989).

Although promising, the early computer methods floundered because of the sheer complexity of the task. In order to provide a better method for studying verbal and written speech samples, we originally developed a text analysis application called Linguistic Inquiry and Word Count, or LIWC (pronounced "Luke"). The first LIWC application was developed as part of an exploratory study of language and disclosure (Francis & Pennebaker, 1992). The second (LIWC2001), third (LIWC2007), fourth (2015), and now fifth (LIWC-22) versions updated the original application with increasingly expanded dictionaries and sophisticated software design (Pennebaker et al., 2001, 2007, 2015).

The most recent evolution, LIWC-22 (Pennebaker et al., 2022), has significantly altered both the dictionary and the software options to reflect new directions in text analysis. As with previous versions, the program is designed to analyze individual or multiple language files quickly and efficiently. At the same time, the program attempts to be transparent and flexible in its operation, allowing the user to explore word use in multiple ways.

The LIWC-22 Text Processing Modules

At its core, LIWC-22 consists of software and a "dictionary" — that is, a map that connects important psychosocial constructs and theories with words, phrases, and other linguistic constructions. To avoid confusion, words contained in texts that are read and analyzed by LIWC-22 are referred to as ***target words***. Words in the LIWC-22 dictionary file will be referred to as ***dictionary words***. Groups of dictionary words that tap a particular domain (e.g., negative emotion words) are variously referred to as "subdictionaries" or "word categories" or just "categories."

The Main Text Processing Module: LIWC Analysis

LIWC-22 is designed to accept written or transcribed verbal text which has been stored as a digital, machine-readable file in one of multiple formats, including plain text (.txt), PDF, RTF, or standard Microsoft Word files (.docx). The software can also process texts inside of multiple spreadsheet formats, including those saved as Comma Separated Values (CSV) or Microsoft Excel (.xlsx) format. The default LIWC-22 dictionary can be run in two ways: 1) via the standard Graphical User Interface (GUI), as has been the case with all previous versions of LIWC, and 2) via a Command Line Interface (CLI) from your command line, or from other platforms that can interface with your system's command line (for example, R or Python). LIWC-22 is compatible with both PC and Mac platforms.

During operation, the LIWC-22 processing module accesses each text in your dataset, compares the language within each text against the LIWC-22 dictionary. Like all previous versions of LIWC, the LIWC-22 text processing module works by counting all of the words in a target text, then calculating the percentage of total words that are represented in each of the LIWC subdictionaries. After processing each text sample, the LIWC processing module writes the output to a data table that can be exported in several formats of your choosing, including spreadsheets (e.g., CSV for MS Excel), a newline-delimited JSON, and so on. As described in more detail below, the LIWC output includes the file name, total word count, and the percentage of words that were captured in the text for each language dimension.

Companion Processing Modules

All previous versions of LIWC have been defined largely by their central feature: the LIWC basic processing module. An important distinguishing characteristic of LIWC-22 is that the program includes a group of companion processing modules that provide additional analytic methods for language researchers. These additional features provide new ways to analyze and understand your text samples. These features include the following:

Dictionary workbench

Historically, creating custom dictionaries was a tedious, esoteric process that only a conscientious programmer could love. The dictionary workbench makes it possible for users to build a dictionary using a simple point-and-click interface, accompanied by a built-in error-checking tool. When completed, users can evaluate the psychometric properties of their dictionary — namely, Cronbach's α — on a dataset of their choosing. Note that the underlying format of the LIWC-22 dictionary is considerably more flexible and dynamic than past versions of LIWC. Nevertheless, older LIWC-formatted dictionary files can be imported and seamlessly converted into the new dictionary format.[1]

Word frequencies and word clouds

When analyzing a new data set, it is sometimes helpful to see a word frequency table to determine what words are most commonly used. In addition to word frequencies, LIWC-22 can build a word cloud that can be saved in an image format.

Topic modeling with the Meaning Extraction Method

The traditional LIWC program was built on the assumption that many of the most important social and psychological dimensions of language could be found through the analysis of function words (e.g., pronouns, prepositions) and emotion words. Sometimes, however, it is important to know the content of what people are saying in addition to their linguistic styles. Although there are various types of topic modeling, LIWC-22's topic modeling feature was built in consideration of a psychologically informed method: the Meaning Extraction Method (MEM; Chung & Pennebaker, 2008). The MEM allows users to subject their text analyses to factor analyses (SVD, or other methods) to discover dominant themes and meaning in their dataset.

[1] Note that the older LIWC dictionary format (.dic) can be converted into the newer format (.dicx) but the new cannot be converted into the old format.

Narrative arc

Language researchers have struggled with ways to devise computerized methods to analyze narrative structure. Recent work on the Arc of Narrative points to three underlying processes that are shared in most stories: staging, plot progression, and cognitive tension (Boyd, Blackburn, et al., 2020). The narrative arc module automatically assesses texts for how each narrative structure "unfolds" throughout the story, providing corresponding graphs and metrics that reflect the degree to which each text resembles a normative narrative shape.

Language style matching

When describing the "style" of someone's language, we typically are basing our evaluation on the ways in which authors use function words (e.g., pronouns, articles, auxiliary verbs, prepositions, and other short, common words). The Language Style Matching (LSM) module compares the language style among different texts. LSM analyses can provide a metric of how similarly two people talk with each other in dyads or larger groups, whether they write in similar styles, or more broadly how similarly two or more groups may be writing or thinking alike (Ireland & Pennebaker, 2010).

Contextualizer

When trying to understand what words can tell us about people, context is crucial (Boyd & Schwartz, 2021). Often, we might have a sense that certain types of words are being used a certain way, or that they reflect a particular type of psychosocial process. However, it is important — crucial, even — that we critically examine our assumptions about word use. Often, the best way to do this is to look at *how* words are being used in their broader context; this approach is often referred to as "keywords in context." The Contextualizer module offers a simple way to extract words and their immediate context. For example, if we want to better understand what the word "love" is conveys in a dataset, we can extract a set number of words before and after each appearance of the word "love" in our texts. From there, it is possible to simply inspect the usage of the word or, better yet, come up with a text analysis method (within LIWC-22, perhaps?) to analyze the words surrounding its usage.

Case studies

All previous versions of LIWC were written for researchers who typically analyzed large numbers of text files. Many users, however, have wanted to be able to simply dive into a single text to understand it more deeply through a close analysis. The Case Studies module essentially brings the other modules into a single location to allow users to focus on and explore a single text. While other people are out climbing mountains, dancing at the club with their soul mates, or enjoying expensive designer drugs with movie stars, you can bask in the cold, sterile glow of your computer screen, obsessively analyzing the narrative structure of a coworker's email or LIWCing the dialogue from your favorite episode of *Rick and Morty*.

Prepare transcripts

A new module has been added that helps users to clean conversation transcripts to run different types of LIWC analyses.

The LIWC-22 Dictionary and its Development

The LIWC-22 Dictionary is the heart of the text analysis strategy. The internal dictionary is composed of over 12,000 words, word stems, phrases, and select emoticons. Each dictionary entry is part of one or more categories, or subdictionaries, designed to assess various psychosocial constructs. For example, the word *cried* is part of 10-word categories: affect, tone_pos, emotion, emo_neg, emo_sad, verbs, focuspast, communication, linguistic, and cognition. Hence, if the word *cried* is found in the target text, each of these 10 subdictionary scale scores will be incremented. Most, but not all, of the LIWC-22 categories are arranged hierarchically. All sadness words, by definition, belong to the broader "emo_neg", "emotion", "tone_neg", as well as the overall "affect" category. Note too that word stems can be captured by the LIWC-22 system. For example, the dictionary includes the stem *hungr** which allows for any target word that matches the first five letters to be counted as "food" word (including hungry, hungrier, hungriest). The asterisk, then, denotes the acceptance of all letters, hyphens, or numbers following its appearance.

Each of the default LIWC-22 categories is composed of a list of dictionary words designed to capture that dimension. The selection of words that make up the categories has involved multiple steps. When LIWC was first conceived, the idea was to identify a group of words that tapped into basic emotional and cognitive dimensions often studied in social, health, and personality psychology. As our understanding of the psychology of verbal behavior has matured, the breadth and depth of word categories in the LIWC dictionary has expanded considerably.

For LIWC-22, we have completely rebuilt the text processing engine, including the flexibility of LIWC-formatted dictionaries. Dictionaries can now accommodate numbers, punctuation, short phrases, and even regular expressions. These additions allow the user to read "netspeak" language that is common in Twitter and Facebook posts, as well as SMS (short messaging service, a.k.a. "text messaging") and SMS-like modes of communication (e.g., Snapchat, instant messaging). For example, "b4" is coded as a preposition and ":)" is coded as a positive tone word.

In this latest version of LIWC, several new categories have been added, others overhauled considerably, and a small number have been removed. With the advent of more powerful analytic methods and more diverse language samples, we have been able to build more internally-consistent language dictionaries with enhanced psychometric properties, in general. This means that many of the dictionaries in previous LIWC versions may have the same name, but the words making up the dictionaries have been altered (categories subjected to major changes are described in a later section).

Development of the LIWC-22 Dictionary

The construction of the LIWC dictionaries has significantly evolved over the years. Earlier iterations of the LIWC dictionary relied extensively on large groups of human raters. With increasing computational power, however, more recent versions have depended on establishing a harmony between the domain expertise and knowledge of human raters and sets of increasingly complex algorithms and statistical models. Below, we present an overview of the process used to create the LIWC-22 dictionary.

Step 1. Word Collection. In the design and development of the LIWC category scales, sets of words were first generated for each conceptual dimension, using the LIWC2015 dictionary as a starting point. Within the Psychological Processes category, for example, the "affect" subdictionaries were based on words from several sources, including previous versions of the LIWC dictionary. We drew on common emotion rating scales, such as the PANAS (Watson et al., 1988), Roget's Thesaurus, and standard English dictionaries. Following the creation of preliminary category word lists, 3-4 judges individually generated word lists for each category, then held group brainstorming sessions in which additional words relevant to the various dictionaries were generated and added to the initial lists. Similar schemes were used for the other subjective dictionary categories.

Step 2. Judge Rating Phase. Once the grand list of words was amassed, each word in the dictionary was independently examined by 3-4 judges and qualitatively rated in terms of "goodness of fit" for each category. In order for a word to be retained in a given category, a majority of judges had to agree on its inclusion. In cases of disputes, judges individually and jointly inspected several corpora and online sources to help determine a word's most common use, inflection, and psychological meaning. Words for which judges could not decide on appropriate category placement were removed from the dictionary.

Step 3. Base Rate Analyses. Once a working version of the dictionary was constructed from judges' ratings, the Meaning Extraction Helper (MEH; Boyd, 2018) was used to determine how frequently dictionary words were used in various contexts across a large, diverse corpus of texts: we refer to this as the "Test Kitchen" corpus. The Test Kitchen corpus contains 15,000 texts from a diverse set of 15 corpora, including blog posts, spoken language studies, social media, novels, student writing, and several others; we discuss this corpus in much greater detail in a later section. Most relevant to this section, however, is that these analyses were used to root out dictionary words that did not occur at least once across multiple corpora.

Step 4. Candidate Word List Generation. The 5,000 most frequently-used words in the Test Kitchen corpus were identified. Of these 5,000 words, those that were not already in the LIWC dictionary were considered candidates for inclusion. For several linguistic categories (e.g., verbs, adjectives), Stanford's CoreNLP and custom-made analytic software was used to identify high base rate exemplars that were treated as candidates for inclusion (see: Boyd, 2020; Manning et al., 2014). All candidate words were then correlated with all dictionary categories in order to identify common words that were 1) not yet included in the dictionary, and 2) showed acceptable conceptual and statistical fit with existing categories. Words that correlated positively with dictionary categories were added to a list of candidate words for possible inclusion. All candidate words were reviewed by teams of 3-4 judges who voted on 1) whether words should be included in the dictionary and 2) whether words were a sound conceptual fit for specific dictionary categories. Judges' rating procedures were parallel to those outlined in *Step 2*. Finally, the four

authors (RLB, AA, SS, and JWP) jointly worked on evaluating each word in randomly-assigned teams of two to determine whether they should be cross-categorized into other LIWC-22 categories.

Step 5. Psychometric Evaluation. Following all previously-described steps, each language category was separated into its constituent words. Each word was then quantified as a percentage of total words. All words for each category were used to compute internal consistency statistics for each language category as a whole. Words that were detrimental to the internal consistency of their overarching language category were added to a candidate list of words for omission from the final dictionary. A group of 4 judges (the authors of this document) then reviewed the list of candidate words and voted on whether words should be retained. Words for which no majority could be established were omitted.[2] Several linguistic categories, such as *pronouns* and *prepositions*, constitute established linguistic constructs and were therefore not a part of the omission process.

Step 6. Refinement Phase. After Steps 1-5 were complete, they were repeated in their entirety. This was done to catch any possible mistakes/oversights that might have occurred throughout the dictionary creation process. The psychometrics of each language category changed negligibly during each refinement phase. During the last stage of the final refinement phase, all four judges reviewed the dictionary in its entirety for mistakes.

Step 7. Addition of Summary Variables. In addition to standard LIWC dimensions based on percentage of total words, four summary variables were calculated: analytical thinking (Pennebaker et al., 2014), clout (Kacewicz et al., 2014), authenticity (M. L. Newman et al., 2003), and emotional tone (Cohn et al., 2004). Each summary variable builds upon previously-published research from our lab; measures are calculated, then converted to percentiles based on standardized scores from large comparison corpora. The summary variables are the only non-transparent dimensions in the LIWC-22 output. The summary measures have been adjusted against new norms but are conceptually consistent with the scores calculated in LIWC2015.

[2] Worry not: the authors remain very good friends to this day.

LIWC-22: Establishing the Psychometrics

From the beginning, the top priority of creating LIWC has been to build a scientifically sound system that is both reliable and valid. For each iteration of LIWC, the dictionaries have been modernized to try to keep up with subtle (and not-so-subtle) shifts in language. At the same time, the world of text-based data science has grown exponentially, providing new methods and data that facilitate increasingly well-validated versions of the dictionaries. For LIWC-22, we have been able to build a large text corpus that includes traditional and contemporary English language samples across multiple contexts. This "Test Kitchen" corpus was used for multiple purposes in the creation and testing of the LIWC-22 dictionary, ranging from word selection to the assessment of the dictionaries' reliability and validity.

The Test Kitchen Corpus

The assessment of any text analysis system requires a large set of text samples drawn across multiple authors and contexts. Several impressive corpora exist, including archives from Twitter, Facebook, Reddit, movie transcripts, Wikipedia, the British National Corpus, Project Gutenberg, and, for researchers associated with Google, just about anything ever posted on the web. The challenge for psychological researchers, however, is to assemble a large array of texts that broadly represent the ways words are used by everyday people in everyday life.

In previous versions of LIWC development, we relied on whatever datasets we had collected or could find. For LIWC-22, we sought to build a curated corpus that would broadly represent the many ways in which language is used. Once built, we could rely on the corpus as a "test kitchen" to both quantify and qualify our LIWC-22 dictionary and, at the same time, obtain estimates about the context-dependence of verbal behavior (insofar as word use reflects a certain class of verbal behavior as well as social behavior and psychologically meaningful behavior more broadly defined).

The Test Kitchen corpus was constructed from randomly selected subsets of text from across 15 different types of English language sets. The original datasets included thousands, sometimes millions of writings or transcribed speech samples, including blogs, emails, movie dialog, social media posts, natural conversations, etc. Some of the data repositories were collected by our or other labs, others came from public archives. From each of the 15 data sets, we randomly selected 1,000 text samples with a minimum of 100 words. For any texts with more than 10,000 words, an algorithm was written to select 10,000 continuous words from a random starting point in the document. As can be seen in Table 1, the Test Kitchen corpus includes 1,000 texts from each of the 15 different sources, for a total of 15,000 texts, each averaging over 2,000 words. The overall word count of the entire corpus is over 31 million words. To the degree possible, all personally identifying information was stripped.

Note that for most corpora, single texts reflected the writings from a single person. For example, each text from the Blog or Email corpus included multiple blog entries or emails from the same person. For additional information on the Test Kitchen Corpus, see Appendix A. Due to the nature of several of the data sources, the Test Kitchen corpus is not readily available for research use and cannot be made publicly available.

Table 1. The Test Kitchen Corpus of 31 Million Words

Corpus	Description	Word Count *M* (SD)
Applications	Technical college admissions essays	1506 (501)
Blogs	Personal blogs from blogger.com	2144 (1920)
Conversations	Natural conversations	614 (495)
Enron Emails	Internal emails from Enron	316 (376)
Facebook	Facebook posts from mypersonality.com	2195 (2034)
Movies	Transcribed movie dialogue	6633 (2459)
Novels	Novels from Project Gutenberg	5703 (189)
NYT	New York Times articles	744 (494)
Reddit	Individuals' Reddit comments	1751 (1945)
Short Stories	Short stories	2977 (2211)
SOC	Stream of consciousness essays	656 (256)
Speeches	U.S. Congressional speeches	950 (1241)
TAT	Thematic Apperception Test, online website	326 (63)
Tweets	Collected tweets from individual accounts	4442 (2858)
Yelp	Restaurant reviews posted to Yelp	99 (1)
Overall mean		**2070 (2466)**

Note: Each corpus is composed of 1,000 texts, each originally between approximately 100 to 10,000 words. After data collection, some texts became slightly smaller or larger because of data cleaning procedures. For a more detailed description of the Test Kitchen Corpus, see Appendix A.

Quantifying the Reliability and Validity of LIWC-22

Assessing the reliability and validity of text analysis programs is a tricky business. One might reasonably assume that it is psychometrically acceptable to calculate the internal consistency of a LIWC category in the same way as one might do with a self-report questionnaire; this assumption would be wrong. Consider the fact that a questionnaire that taps into the construct of anger or aggression, for example, typically asks participants to respond to a number of questions about their feelings or behaviors related to anger, in general. Reliability coefficients, then, are computed by correlating people's responses to the various questions. The more highly they correlate, the reasoning goes, the more the questionnaire items all measure the same thing. Voila! The scale is deemed to be internally consistent, and therefore reliable.

A similar strategy can be used with words. But be warned: the psychometrics of natural language are not as straight-forward as with questionnaires. The reason becomes deceptively obvious to those who dedicate their lives to the study of verbal behavior. Once you say something, you generally do not need to say it again in the same social media post, essay, or conversation. The nature of discourse, then, is we usually say something and then move on to the next thought or topic. Repeating the same idea over and over again is not the rule in verbal behavior, but the exception. Yet, this repeated, concurrent sampling of a construct is the foundation of self-report questionnaire design. It is important, then, to understand that acceptable boundaries for natural language reliability coefficients are lower than those commonly seen elsewhere in psychological tests. Put simply: behavior — particularly verbal behavior — has very different psychometric properties than more "broad strokes" assessments of human psychology commonly used elsewhere.

The LIWC-22 Anger scale, for example, is made up of 181 anger-related words, word stems, and phrases. In theory, the more that people use one type of anger word in a given text, the more they should use other anger words in the same text. To test this idea, we can determine the degree to which people use each of the 181 anger words across a select group of text files and then calculate the intercorrelations of the word use. In order to calculate these statistics, each dictionary word was measured as a percentage of total words per text (Cronbach's α) or, alternatively, in a binary "present versus absent" manner (Kuder–Richardson Formula 20; Kuder & Richardson, 1937). The scores were then entered as an "item" in a standard internal consistency calculation, providing internal consistency metrics across all corpora.

Both Kuder–Richardson Formula 20 and the raw, unstandardized Cronbach's α metric are presented in Table 2; both metrics were derived from the entire Test Kitchen corpus. Importantly, the traditional Cronbach's α method, calculated from relative word frequencies, tends to sorely underestimate reliability in language categories due the highly variable base rates of word usage within any given category. The Kuder–Richardson Formula 20 should generally be considered to be a more "true" approximation of each category's true internal consistency.

Table 2. LIWC-22 Language Dimensions and Reliability

Category	Abbrev.	Description/Most frequently used exemplars	Words/ Entries in category*	Internal Consistency: Cronbach's α	Internal Consistency: KR-20
Summary Variables					
Word count	WC	Total word count			
Analytical thinking	Analytic	Metric of logical, formal thinking	-	-	-
Clout	Clout	Language of leadership, status	-	-	-
Authentic	Authentic	Perceived honesty, genuineness	-	-	-
Emotional tone	Tone	Degree or positive (negative) tone	-	-	-
Words per sentence	WPS	Average words per sentence	-	-	-
Big words	BigWords	Percent words 7 letters or longer	-	-	-
Dictionary words	Dic	Percent words captured by LIWC	-	-	-
Linguistic Dimensions	Linguistic		4933	0.36	1.00
Total function words	function	the, to, and, I	499/1443	0.28	0.99
Total pronouns	pronoun	I, you, that, it	74/286	0.43	0.97
Personal pronouns	ppron	I, you, my, me	42/221	0.24	0.95
1st person singular	i	I, me, my, myself	6/74	0.49	0.85
1st person plural	we	we, our, us, lets	7/17	0.43	0.78
2nd person	you	you, your, u, yourself	14/59	0.37	0.82
3rd person singular	shehe	he, she, her, his	8/30	0.58	0.83
3rd person plural	they	they, their, them, themsel*	7/20	0.36	0.69
Impersonal pronouns	ipron	that, it, this, what	32/68	0.43	0.91
Determiners	det	the, at, that, my	97/293	-0.19	0.95
Articles	article	a, an, the, alot	3/103	0.12	0.61
Numbers	number	one, two, first, once	44/61	0.57	0.87
Prepositions	prep	to, of, in, for	83/302	0.16	0.95
Auxiliary verbs	auxverb	is, was, be, have	25/282	0.44	0.97
Adverbs	adverb	so, just, about, there	159/514	0.63	0.97
Conjunctions	conj	and, but, so, as	49/65	0.11	0.89
Negations	negate	not, no, never, nothing	8/247	0.49	0.92
Common verbs	verb	is, was, be, have	1560	0.60	0.99
Common adjectives	adj	more, very, other, new	1507	0.26	0.99
Quantities	quantity	all, one, more, some	422	0.45	0.96
Psychological Processes					
Drives	Drives	we, our, work, us	1477	0.58	0.98
Affiliation	affiliation	we, our, us, help	384	0.43	0.94
Achievement	achieve	work, better, best, working	277	0.53	0.92
Power	power	own, order, allow, power	856	0.67	0.96
Cognition	Cognition	is, was, but, are	1403	0.68	0.99
All-or-none	allnone	all, no, never, always	35	0.37	0.88
Cognitive processes	cogproc	but, not, if, or, know	1365	0.67	0.99
Insight	insight	know, how, think, feel	383	0.43	0.96
Causation	cause	how, because, make, why	169	0.21	0.90
Discrepancy	discrep	would, can, want, could	108	0.29	0.91
Tentative	tentat	if, or, any, something	230	0.52	0.94
Certitude	certitude	really, actually, of course, real	131	0.22	0.88
Differentiation	differ	but, not, if, or	325	0.38	0.94
Memory	memory	remember, forget, remind, forgot	26	0.23	0.64
Affect	Affect	good, well, new, love	2999	0.64	0.99
Positive tone	tone_pos	good, well, new, love	1020	0.61	0.98
Negative tone	tone_neg	bad, wrong, too much, hate	1530	0.62	0.98
Emotion	emotion	good, love, happy, hope	1030	0.61	0.97
Positive emotion	emo_pos	good, love, happy, hope	337	0.52	0.93
Negative emotion	emo_neg	bad, hate, hurt, tired	618	0.52	0.95
Anxiety	emo_anx	worry, fear, afraid, nervous	120	0.37	0.80
Anger	emo_anger	hate, mad, angry, frustr*	181	0.30	0.82
Sadness	emo_sad	:(, sad, disappoint*, cry	134	0.25	0.80
Swear words	swear	shit, fuckin*, fuck, damn	462	0.79	0.93
Social processes	Social	you, we, he, she	2760	0.43	0.99
Social behavior	socbehav	said, love, say, care	1632	0.49	0.98
Prosocial behavior	prosocial	care, help, thank, please	242	0.49	0.89
Politeness	polite	thank, please, thanks, good morning	142	0.58	0.87
Interpersonal conflict	conflict	fight, kill, killed, attack	305	0.43	0.88
Moralization	moral	wrong, honor*, deserv*, judge	356	0.37	0.90
Communication	comm	said, say, tell, thank*	350	0.42	0.95
Social referents	socrefs	you, we, he, she	1232	0.35	0.97
Family	family	parent*, mother*, father*, baby	194	0.48	0.89
Friends	friend	friend*, boyfriend*, girlfriend*, dude	102	0.27	0.75
Female references	female	she, her, girl, woman	254	0.56	0.89
Male references	male	he, his, him, man	230	0.62	0.91

Table 2. LIWC-22 Language Dimensions and Reliability (Cont'd)

Category	Abbrev.	Description/Most frequently used exemplars	Words/ Entries in category*	Internal Consistency: Cronbach's α	Internal Consistency: KR-20
Expanded Dictionary					
Culture	Culture	car, united states, govern*, phone	772	0.67	0.92
Politics	politic	united states, govern*, congress*, senat*	339	0.75	0.91
Ethnicity	ethnicity	american, french, chinese, indian	239	0.39	0.79
Technology	tech	car, phone, comput*, email*	202	0.41	0.82
Lifestyle	lifestyle	work, home, school, working	1437	0.67	0.97
Leisure	leisure	game*, fun, play, party*	295	0.57	0.91
Home	home	home, house, room, bed	122	0.35	0.83
Work	work	work, school, working, class	547	0.74	0.95
Money	money	business*, pay*, price*, market*	281	0.69	0.91
Religion	relig	god, hell, christmas*, church	241	0.60	0.90
Physical	physical	medic*, food*, patients, eye*	1993	0.74	0.98
Health	health	medic*, patients, physician*, health	715	0.79	0.92
Illness	illness	hospital*, cancer*, sick, pain	259	0.52	0.79
Wellness	wellness	healthy, gym*, supported, diet	118	0.31	0.57
Mental health	mental	mental health, depressed, suicid*, trauma*	126	0.40	0.63
Substances	substances	beer*, wine, drunk, cigar*	154	0.31	0.72
Sexual	sexual	sex, gay, pregnan*, dick	357	0.53	0.86
Food	food	food*, drink*, eat, dinner*	379	0.76	0.93
Death	death	death*, dead, die, kill	109	0.46	0.83
States					
Need	need	have to, need, had to, must	55	0.11	0.76
Want	want	want, hope, wanted, wish	56	0.19	0.76
Acquire	acquire	get, got, take, getting	74	0.15	0.85
Lack	lack	don't have, didn't have, *less, hungry	89	0.03	0.65
Fulfilled	fulfill	enough, full, complete, extra	49	0.04	0.68
Fatigue	fatigue	tired, bored, don't care, boring	66	0.27	0.63
Motives					
Reward	reward	opportun*, win, gain*, benefit*	62	0.37	0.74
Risk	risk	secur*, protect*, pain, risk*	128	0.28	0.86
Curiosity	curiosity	scien*, look* for, research*, wonder	76	0.26	0.79
Allure	allure	have, like, out, know	105	0.68	0.98
Perception	Perception	in, out, up, there	1834	0.59	0.99
Attention	attention	look, look* for, watch, check	130	0.16	0.86
Motion	motion	go, come, went, came	485	0.42	0.97
Space	space	in, out, up, there	617	0.41	0.98
Visual	visual	see, look, eye*, saw	226	0.49	0.94
Auditory	auditory	sound*, heard, hear, music	255	0.49	0.91
Feeling	feeling	feel, hard, cool, felt	157	0.32	0.90
Time orientation					
Time	time	when, now, then, day	464	0.50	0.97
Past focus	focuspast	was, had, were, been	699	0.71	0.98
Present focus	focuspresent	is, are, I'm, can	373	0.60	0.96
Future focus	focusfuture	will, going to, have to, may	138	0.32	0.92
Conversational	Conversation	yeah, oh, yes, okay	500	0.73	0.96
Netspeak	netspeak	:), u, lol, haha*	439	0.73	0.96
Assent	assent	yeah, yes, okay, ok	50	0.41	0.72
Nonfluencies	nonflu	oh, um, uh, i i	21	0.49	0.74
Fillers	filler	rr*, wow, sooo*, youknow	24	0.23	0.61

***Notes:** "Words/Entries in category" refers to the number of different words and/or entries that make up the variable category. For function words, two numbers are included: number of different key words in the category and the total number of entries. For example, there are only about 8 primary negation words (e.g., wasn't, isn't) which are parts of 247 different entries (e.g., "wasn't happy"). All alphas were computed on a sample of 15,000 texts from the Test Kitchen corpus (see Table 1). For the purpose of "translating" LIWC category internal consistencies with relation to what one might expect to see in other assessment methods, such as self-report questionnaires, we recommend that readers focus primarily on the Kuder–Richardson Formula 20 (KR-20) values.

Establishing the validity of the various LIWC dimensions is an outstandingly large and difficult topic. Almost by definition, the various LIWC content categories are face valid. The more challenging question concerns how inter- and intra-personal psychological processes are reflected in language use. For example, do people who use a high rate of "affiliation" words actually feel a high need for affiliation? Are they already well-connected with other people, or are they experiencing a high need due to a *lack* of meaningful social connections? What are the interpersonal effects of a person's high (versus low) use of affiliation words? Does the use of affiliation language correlate with or predict other objective measures of social connection, interpersonal needs, and spatiotemporal proximity to people who make for good affiliative prospects?

It is beyond the scope of this manual to attempt to summarize the outstanding number of studies conducted at the intersection of text analysis and psychosocial processes since 1992. Using the search query "LIWC text analysis" on Google Scholar, over 2,400 studies and/or papers are retrieved from the year 2021 alone. In the dozens of studies from our own labs, correlations between LIWC affect or emotion categories from texts people write and their self-reports of the relevant affective feelings typically range from .05 to .40, averaging around ~.15 to ~.20. The correlations between judges' ratings of people's writing samples and the LIWC scores of the authors' writing samples are typically a bit higher, in the .15 to .30 range (a range similar to the correlation between people's self-reported and judges' ratings). We find slightly higher correlations among self-reports, judges' ratings, and LIWC for cognitive and social processes. Note that the correlations are highly dependent on the context and what the instructions or topics of the writing samples (for general reviews, see: Boyd & Schwartz, 2021; Pennebaker, 2011; Tausczik & Pennebaker, 2010).

While the various dimensions of LIWC have been extensively validated over the years, across thousands of studies, and by hundreds of independent research labs, it is critical to appreciate the fact that human psychology is (perhaps ironically) complex beyond words. For scholars both within and outside of the psychological sciences, we emphasize the importance of approaching language data with an understanding that verbal behavior — and thus, LIWC measures derived from such behavior — are best suited to capturing some aspects of human psychology better than others. One should not necessarily expect self-report questionnaires and LIWC scores of the same constructs to correlate strongly or, in some cases, even at all. Indeed, different approaches to measuring a construct are often not correlated, but still validly reflect different (but equally valid) aspects of human psychology (Ganellen, 2007; Mauss & Robinson, 2009)**.** In other cases, different approaches to measurement of the same construct often are tapping into truly different constructs altogether, as is often the case for self-reports versus more objective behavioral measures (Boyd, Pasca, et al., 2020; Boyd & Pennebaker, 2017).

Variability and Context in the Measurement of Verbal Behavior

Across analyses of various types of verbal behavior (speech, social media posts, personal writing, formal writing, etc.) the LIWC-22 dictionary captures, on average, between 80% and 90% of all words people use. However, the ways in which context influences the words that people use — and the properties of those words — are many (van Dijk, 1999, 2009). The base rate of word use varies considerably depending on what people are talking about and the context of their communication. While it may be appropriate to use high rates of words about sex and death in your therapist's office, a tactful conversationalist may be less inclined to use the same words at high frequency while delivering a toast at their deeply religious grandmother's 105th birthday party.

In Table 3, we present descriptive statistics for all LIWC-22 measures, highlighting distinctions between different "types" of language data, ranging from highly informal (e.g., social media posts, conversations) to highly formal (e.g., newspaper articles).[3] With the exception of total word count, words per sentence, and the four summary variables (Analytic, Clout, Authentic, and Tone), all means in the table below are expressed as percentage of total words used in any given language sample.

Importantly, Table 3 provides a glimpse into the ways in which word use shifts as a function context. Transcribed daily conversations, for example, typically have much shorter and common words, more positive emotion words, and far more I-words (e.g., I. me, my) than New York Times articles. In fact, every corpus in Table 3 has its own unique linguistic fingerprint. This is particularly apparent when looking at the four summary variables and function word categories.

[3] A comprehensive report of the means and standard deviations across the entire Test Kitchen corpus are available for download at: https://www.liwc.app/static/documents/LIWC-22.Descriptive.Statistics-Test.Kitchen.xlsx

Table 3. Selected Descriptive Statistics from the Test Kitchen Corpus

Category	Twitter		Conv		Blogs		NYT		Others	
Texts (*N*)	1,000		1,000		1,000		1,000		11,000	
Summary Variables	**mean**	SD	**mean**	SD	**mean**	SD	**mean**	SD	**mean**	SD
Word count	**4442.1**	2858.2	**613.7**	495.2	**2144.4**	1919.6	**744.1**	493.6	**2101.1**	2511.7
Analytic	**42.86**	27.48	**11.03**	9.27	**38.70**	23.20	**87.62**	12.41	**51.27**	28.08
Clout	**49.10**	28.36	**61.01**	21.01	**29.99**	23.45	**53.90**	17.59	**49.99**	29.27
Authentic	**52.33**	25.58	**55.20**	22.44	**68.08**	23.50	**28.90**	19.61	**49.52**	28.39
Tone	**68.00**	26.36	**63.88**	24.51	**44.99**	20.48	**37.08**	22.30	**46.24**	25.95
Words/sentence	**30.79**	100.62	**5.77**	2.20	**14.27**	6.03	**19.88**	6.21	**17.01**	26.22
Big words	**15.98**	5.63	**11.34**	2.80	**14.19**	4.21	**24.77**	4.54	**17.48**	6.70
Dictionary words	**83.53**	7.54	**93.60**	2.55	**88.30**	6.36	**80.34**	5.61	**88.50**	4.98
Linguistic variables	**63.69**	9.59	**79.10**	4.28	**71.51**	7.28	**58.79**	6.41	**70.10**	7.13
function	**47.45**	8.39	**58.68**	4.16	**56.62**	6.52	**46.75**	5.33	**55.46**	6.45
pronoun	**15.38**	5.17	**17.54**	2.77	**16.28**	4.03	**7.65**	3.20	**15.02**	4.80
ppron	**11.10**	4.16	**9.77**	2.39	**11.04**	3.35	**4.15**	2.65	**10.29**	3.93
i	**5.49**	3.57	**3.02**	1.53	**6.27**	2.98	**0.67**	1.22	**4.42**	3.41
we	**0.97**	1.02	**0.84**	0.84	**0.91**	0.83	**0.38**	0.57	**0.91**	1.23
you	**3.06**	1.62	**4.24**	1.70	**1.34**	1.32	**0.36**	0.64	**1.49**	1.82
shehe	**0.84**	0.71	**0.76**	1.04	**1.56**	1.47	**1.64**	1.62	**2.28**	2.79
they	**0.54**	0.39	**0.79**	0.69	**0.71**	0.58	**0.70**	0.61	**0.96**	1.12
ipron	**4.29**	1.58	**7.77**	1.87	**5.24**	1.56	**3.50**	1.46	**4.73**	1.90
determiners	**11.51**	2.63	**10.80**	2.20	**13.83**	2.37	**15.53**	2.12	**14.79**	2.76
article	**4.62**	1.73	**2.80**	1.19	**6.06**	1.89	**9.26**	1.83	**7.11**	2.32
number	**2.49**	2.00	**2.27**	2.64	**2.03**	1.72	**3.24**	2.91	**2.01**	2.63
preposition	**10.78**	2.87	**11.91**	2.31	**12.84**	2.04	**14.63**	1.80	**13.41**	2.65
auxiliar verb	**7.97**	2.51	**13.14**	2.05	**8.99**	1.92	**5.72**	1.78	**8.59**	2.75
adverb	**5.10**	1.83	**8.97**	2.16	**6.34**	1.68	**3.23**	1.44	**5.08**	2.17
conjunction	**4.24**	1.44	**7.02**	2.00	**6.74**	1.46	**5.45**	1.34	**6.30**	1.85
negations	**1.92**	1.02	**2.50**	1.22	**1.80**	0.79	**0.77**	0.55	**1.59**	1.06
verb	**16.33**	3.80	**22.48**	3.13	**17.78**	3.17	**11.51**	2.96	**17.00**	4.18
adjective	**6.48**	2.11	**6.71**	1.83	**5.95**	1.15	**5.93**	1.61	**5.93**	1.74
quantity	**3.44**	1.13	**4.16**	1.50	**4.28**	1.14	**4.59**	2.11	**4.01**	1.73
Drives	**4.74**	2.32	**2.96**	1.62	**3.81**	1.55	**5.41**	2.63	**4.48**	2.35
affiliation	**2.31**	1.47	**1.53**	1.08	**1.93**	1.12	**1.41**	1.05	**2.00**	1.51
achieve	**1.29**	1.19	**0.97**	0.71	**1.00**	0.55	**1.34**	1.02	**1.27**	1.07
power	**1.22**	1.15	**0.50**	0.63	**0.93**	0.82	**2.85**	2.30	**1.27**	1.22
Cognition	**10.25**	3.24	**15.46**	3.21	**12.84**	3.11	**8.63**	3.15	**11.89**	3.56
allnone	**1.62**	0.83	**1.97**	1.10	**1.45**	0.68	**0.58**	0.51	**1.32**	0.88
cogproc	**8.57**	2.84	**13.36**	3.12	**11.30**	2.84	**7.95**	2.96	**10.49**	3.32
insight	**1.85**	0.84	**3.30**	1.36	**2.55**	1.03	**1.54**	0.99	**2.40**	1.19
cause	**1.18**	0.60	**1.54**	0.79	**1.37**	0.56	**1.26**	0.73	**1.38**	0.81
discrep	**1.81**	0.79	**1.82**	0.96	**1.83**	0.71	**0.94**	0.68	**1.77**	0.97
tentat	**1.51**	0.69	**3.25**	1.44	**2.43**	0.94	**1.43**	0.92	**2.07**	1.32
certitude	**0.59**	0.41	**1.23**	0.81	**0.73**	0.42	**0.30**	0.36	**0.60**	0.59
differ	**2.46**	1.05	**3.93**	1.25	**3.48**	0.99	**2.81**	1.20	**3.13**	1.39
memory	**0.12**	0.12	**0.08**	0.18	**0.11**	0.14	**0.04**	0.12	**0.09**	0.17
Affect	**8.96**	4.48	**5.50**	2.23	**5.54**	1.64	**3.79**	1.67	**5.13**	2.25
tone_pos	**6.05**	4.52	**4.18**	2.05	**3.39**	1.16	**2.33**	1.25	**3.32**	1.84
tone_neg	**1.85**	1.07	**1.01**	0.81	**1.76**	0.93	**1.38**	1.25	**1.55**	1.10
emotion	**3.02**	2.13	**2.06**	1.52	**2.12**	1.08	**0.80**	0.68	**1.82**	1.43
emo_pos	**2.03**	1.96	**1.44**	1.34	**1.17**	0.71	**0.35**	0.42	**1.03**	1.07
emo_neg	**0.76**	0.56	**0.44**	0.46	**0.81**	0.59	**0.38**	0.44	**0.67**	0.65
emo_anx	**0.10**	0.11	**0.06**	0.14	**0.14**	0.19	**0.08**	0.15	**0.14**	0.22
emo_anger	**0.18**	0.20	**0.09**	0.20	**0.18**	0.26	**0.11**	0.24	**0.13**	0.23
emo_sad	**0.17**	0.23	**0.05**	0.15	**0.15**	0.23	**0.06**	0.15	**0.13**	0.26
swear	**1.08**	1.42	**0.18**	0.44	**0.33**	0.53	**0.02**	0.09	**0.20**	0.46
Social processes	**13.76**	4.83	**11.45**	2.97	**10.50**	3.27	**10.09**	3.69	**12.32**	4.21
socbehav	**5.15**	3.99	**2.88**	1.31	**3.48**	1.21	**3.90**	1.93	**3.88**	1.71
prosocial	**1.17**	0.97	**0.40**	0.47	**0.44**	0.35	**0.54**	0.54	**0.78**	0.84
polite	**1.17**	3.74	**0.31**	0.37	**0.21**	0.24	**0.08**	0.22	**0.36**	0.52
conflict	**0.27**	0.25	**0.10**	0.19	**0.23**	0.26	**0.41**	0.67	**0.22**	0.32
moralization	**0.40**	0.40	**0.14**	0.25	**0.28**	0.26	**0.30**	0.47	**0.25**	0.33
communication	**1.78**	0.97	**1.48**	0.94	**1.58**	0.75	**1.81**	1.18	**1.57**	1.14
social referents	**8.51**	2.66	**8.43**	2.35	**6.82**	2.50	**5.95**	2.75	**8.33**	3.70
family	**0.52**	0.87	**0.21**	0.39	**0.43**	0.51	**0.33**	0.72	**0.43**	0.62
friend	**0.30**	0.29	**0.20**	0.34	**0.20**	0.26	**0.06**	0.16	**0.16**	0.30
female	**0.79**	0.63	**0.55**	0.81	**0.92**	1.10	**0.71**	1.28	**1.56**	2.44
male	**1.23**	0.96	**0.72**	0.85	**1.38**	1.28	**1.48**	1.54	**1.58**	1.84

Table 3. Selected Descriptive Statistics from the Test Kitchen Corpus (Cont'd)

Category	Twitter		Conv		Blogs		NYT		Others	
Expanded Dictionary	**mean**	SD	**mean**	SD	**mean**	SD	**mean**	SD	**mean**	SD
Culture	**0.89**	1.25	**0.54**	0.65	**0.93**	1.00	**2.18**	2.24	**0.74**	1.16
politic	**0.42**	0.92	**0.10**	0.24	**0.29**	0.72	**1.29**	1.90	**0.35**	0.94
ethnicity	**0.16**	0.32	**0.16**	0.38	**0.13**	0.28	**0.43**	0.70	**0.12**	0.33
technology	**0.32**	0.69	**0.29**	0.44	**0.52**	0.57	**0.46**	0.91	**0.27**	0.52
Lifestyle	**3.88**	2.60	**4.34**	2.29	**3.37**	1.71	**5.94**	3.16	**4.25**	2.72
leisure	**0.98**	0.78	**0.82**	0.95	**0.81**	0.67	**1.07**	1.41	**0.54**	0.73
home	**0.26**	0.43	**0.28**	0.44	**0.38**	0.39	**0.33**	0.57	**0.41**	0.53
work	**1.70**	2.30	**2.85**	1.97	**1.48**	1.27	**3.39**	2.45	**2.55**	2.44
money	**0.54**	0.74	**0.40**	0.49	**0.47**	0.63	**1.36**	2.01	**0.73**	1.27
religion	**0.53**	1.16	**0.11**	0.24	**0.34**	0.65	**0.23**	0.65	**0.20**	0.42
Physical	**2.17**	1.23	**1.01**	1.05	**1.93**	1.21	**1.58**	2.03	**2.67**	2.34
health	**0.45**	0.58	**0.21**	0.40	**0.40**	0.45	**0.51**	1.23	**0.85**	1.50
illness	**0.12**	0.21	**0.05**	0.17	**0.12**	0.17	**0.16**	0.58	**0.23**	0.47
wellness	**0.05**	0.10	**0.02**	0.10	**0.04**	0.17	**0.04**	0.18	**0.05**	0.18
mental	**0.04**	0.08	**0.01**	0.07	**0.04**	0.09	**0.03**	0.14	**0.04**	0.14
substances	**0.08**	0.23	**0.03**	0.12	**0.06**	0.18	**0.05**	0.29	**0.06**	0.22
sexual	**0.13**	0.23	**0.04**	0.20	**0.11**	0.29	**0.08**	0.29	**0.08**	0.25
food	**0.61**	0.68	**0.33**	0.67	**0.47**	0.67	**0.42**	1.28	**0.74**	1.73
death	**0.17**	0.26	**0.04**	0.14	**0.11**	0.17	**0.18**	0.43	**0.15**	0.28
States										
need	**0.48**	0.32	**0.53**	0.60	**0.52**	0.36	**0.27**	0.35	**0.52**	0.55
want	**0.49**	0.39	**0.48**	0.45	**0.45**	0.35	**0.14**	0.20	**0.39**	0.41
acquire	**0.87**	0.49	**0.96**	0.62	**0.92**	0.45	**0.52**	0.43	**0.84**	0.62
lack	**0.08**	0.11	**0.14**	0.30	**0.14**	0.17	**0.08**	0.18	**0.14**	0.29
fulfill	**0.11**	0.29	**0.07**	0.15	**0.14**	0.14	**0.12**	0.18	**0.16**	0.24
fatigue	**0.08**	0.12	**0.07**	0.16	**0.13**	0.18	**0.01**	0.05	**0.07**	0.21
Motives										
reward	**0.26**	0.54	**0.05**	0.13	**0.11**	0.16	**0.34**	0.57	**0.18**	0.31
risk	**0.22**	0.25	**0.10**	0.20	**0.20**	0.20	**0.33**	0.44	**0.25**	0.35
curiosity	**0.26**	0.32	**0.44**	0.46	**0.34**	0.27	**0.29**	0.41	**0.38**	0.51
allure	**8.54**	2.25	**11.59**	2.91	**7.61**	1.90	**3.58**	1.50	**6.66**	2.75
Perception	**8.30**	2.32	**9.01**	2.45	**9.20**	2.13	**8.96**	2.41	**9.47**	3.00
attention	**0.57**	0.59	**0.32**	0.40	**0.41**	0.27	**0.37**	0.39	**0.50**	0.48
motion	**1.53**	0.73	**1.65**	0.92	**1.80**	0.78	**1.33**	0.84	**1.72**	0.98
space	**4.95**	1.89	**6.09**	2.10	**5.74**	1.46	**6.34**	1.89	**6.07**	1.99
visual	**1.14**	0.64	**0.74**	0.68	**0.96**	0.54	**0.77**	0.72	**1.01**	0.81
auditory	**0.36**	0.51	**0.30**	0.45	**0.33**	0.33	**0.22**	0.52	**0.31**	0.41
feeling	**0.44**	0.31	**0.60**	0.58	**0.54**	0.43	**0.23**	0.32	**0.46**	0.49
time	**4.96**	3.68	**3.65**	1.42	**4.97**	1.42	**4.38**	1.59	**4.36**	1.77
focuspast	**2.79**	1.47	**4.14**	1.83	**4.41**	1.96	**4.76**	2.29	**4.99**	3.06
focuspresent	**5.35**	1.79	**7.66**	1.94	**5.06**	1.60	**2.98**	1.49	**4.45**	2.55
focusfuture	**1.53**	0.69	**1.89**	1.19	**1.77**	0.86	**0.93**	0.74	**1.51**	1.09
Conversation	**3.07**	2.29	**6.40**	3.39	**1.25**	1.65	**0.11**	0.23	**0.81**	1.47
netspeak	**2.48**	2.08	**1.22**	2.22	**0.89**	1.46	**0.09**	0.20	**0.51**	1.24
assent	**0.40**	0.46	**3.21**	2.06	**0.24**	0.30	**0.03**	0.09	**0.19**	0.33
nonfluencies	**0.20**	0.32	**1.60**	1.30	**0.16**	0.22	**0.01**	0.04	**0.12**	0.28
filler words	**0.13**	0.18	**0.61**	1.11	**0.05**	0.11	**0.00**	0.01	**0.03**	0.12
All punctuation	**19.07**	10.97	**35.81**	10.94	**22.72**	10.16	**15.67**	3.96	**20.99**	18.18
Periods	**5.73**	5.04	**18.57**	6.20	**10.11**	7.93	**5.97**	1.87	**8.08**	7.03
Comma	**1.89**	1.51	**6.02**	4.10	**4.14**	2.06	**6.70**	1.79	**3.84**	2.87
Question Mark	**1.16**	0.96	**3.02**	1.70	**0.56**	0.66	**0.17**	0.51	**1.20**	11.32
Exclamation points	**3.22**	4.63	**0.51**	1.64	**1.06**	1.75	**0.03**	0.11	**1.12**	3.50
Apostrophes	**1.45**	1.91	**4.48**	2.60	**2.77**	1.85	**0.00**	0.04	**2.09**	2.06
Other punctuation	**5.63**	7.10	**3.20**	3.55	**4.09**	4.16	**2.80**	2.24	**4.66**	6.87

Notes: The corpora include aggregated Twitter posts (Twitter), natural conversations (Conv), blog entries (Blogs), New York Times articles NYT), and the 11 remaining Test Kitchen corpora (Other). The Twitter and Blog samples reflect up to 10,000 words selected from 1,000 different individuals' Twitter or blogging histories. The Summary Variables from Analytic Thinking to Emotional Tone are standardized composite variables transformed to a scale from 1 to 100. Big words refer to the percentage of total words that are 7 letters or longer. Dictionary words refer to the percentage of words that were counted by LIWC within each text. All categories listed under "Select LIWC variables" are based on the percentage of total words.

Significant Changes in LIWC-22 from Previous Versions of LIWC

Virtually every aspect of LIWC-22 has undergone moderate-to-large changes and updates from previous versions. Certainly, the current iteration of LIWC has been updated to a far greater extent than any previous version. At the broadest level, we have redesigned the overall structure of the dictionary by dividing the categories into "Basic" and "Expanded" super-categories. The Basic Dictionary includes most of the dimensions from earlier LIWC versions, at least, at a conceptual level. The Expanded Dictionary includes majorly updated versions of traditional LIWC categories and, additionally, introduces a host of new categories and variables.

New and updated LIWC-22 language categories

A substantial group of new variables have been added to LIWC. Some are based on recent research findings, psychological domains that have been overlooked in the past, and our own interests. Other variables have been included because of theoretical shifts in the social sciences or, more broadly, in culture. For returning LIWC users, Table 4 lists the means and standard deviations of all LIWC-22 variables alongside their LIWC2015 equivalents, based on the Test Kitchen corpus. Simple correlations between the new and old LIWC variables are also included.

Within the Basic Dictionary, the following additions and significant changes include:

- *Determiners* are a standard part of speech used by linguists that refer to words that precede nouns that specify a quantity (the *first* rule, *three* toys) or clarify the noun's meaning (*this* table, *our* table, *the* table, *any* table). We have now added determiners to the list of function word categories.

- *Cognition* has been added as a general category that reflects different ways people think or refer to their thinking. Cognition serves as the overarching dimension that includes the subcategories of all-or-none thinking, cognitive processes, and memory. Notable changes include:

    - *All-or-none* or absolutist language (e.g., all, none, never). All-or-none thinking — more formally known as "dichotomous thinking" — refers to a style of thinking that tends to be over-generalized and more extreme. All-or-none thinking has been theorized, researched, and explored consistently throughout the history of psychology (Jonason et al., 2018; Metfessel, 1940; Neuringer, 1961). Consistent with recent research on absolutist language and depression (Al-Mosaiwi & Johnstone, 2018), we have split our previous "certainty" category into two separate, weakly-correlated constructs: "all-or-none thinking" and "certitude."

    - *Certitude* has replaced the original cognitive processing dimension of certainty. Unlike all-or-none thinking, *certitude* appears to reflect a degree of bravado, boasting of certainty that often reveals an insecurity or lack of truly verifiable, concrete information, which we've labeled "certitude." Examples: "I love you, really" or "I'm positive that I've studied enough."

    - *Memory* words (e.g., remember, forget) reflect people's references and attention to their memories, beliefs about memory, and the processes of recall and forgetting.

- *Affect* dimensions have changed considerably. Since 2015, a number of important studies have been published that highlight some clear shortcomings of past affect/emotion dictionaries (e.g., Jaidka et al., 2020; Sun et al., 2020). Two important shortcomings of past approaches were a) the nature of emotion language has changed considerably since the first versions of LIWC (particularly the way in which swear words are used), and b) traditionally, we have not distinguished between the constructs of "emotion words" and "sentiment." These issues have been corrected with the following revised/new categories:

    - We now conceptualize the positive tone *(tone_pos)* and negative tone *(tone_neg)* dictionaries as reflecting sentiment rather than emotion *per se*. Although quite similar to the old posemo and negemo categories, the two "tone" dictionaries include words related to positive and negative emotion (e.g., happy, joy, sad, angry) and also words related to those emotions (e.g., birthday, beautiful, kill, funeral). Note that swear words are <u>not</u> included as a subordinate category to either the tone_positive or tone_negative categories.

    - *Emotion*, positive emotion (*emo_pos*), negative emotion (*emo_neg*), anxiety (*emo_anx*), anger (*emo_ang*), and sadness (*emo_sad*) variables are now restricted to true emotion labels, as well as words that strongly imply emotions. For example, the word "laughter" strongly suggests a behavior that, in most cases, flows from a positive affective state. Consequently, the LIWC-22 emotion words are much more "pure" and have lower base rates than previous LIWC editions. Note that, like all superordinate LIWC categories, the superordinate emotion categories include more words than their subordinate categories. For example, the *emo_neg* category contains many undifferentiated negative emotion words that extend beyond the mere sum total of *emo_anx*, *emo_ang*, and *emo_sad* categories.

    - *Swear* words have changed dramatically since LIWC was first developed. In the 1990s, swear words were overwhelmingly used to express anger. They now are as likely or, particularly in informal contexts, *more* likely to reflect positive sentiment. Although the swear word category has been expanded, it now is only part of the overall affect and tone dictionaries rather than positive or negative tone/emotion categories.

- *Social behaviors* category. In the past, we simply included an all-inclusive "social processes" variable. The social behaviors dimension generally seeks to reflect a broad set of social behaviors, or references to them. The "social behaviors" variable is the overarching category that is made up of words associated with the following subordinate categories:

- *Prosocial* behaviors or referents that signal helping or caring about others, particularly at the interpersonal level (e.g., Penner. et al., 2005).

- *Politeness* markers include words such as “please”, “thank you”, and similar words suggesting adherence to social norms and manners (Brown & Levinson, 1978; Holtgraves & Joong-nam, 1990).

- *Interpersonal conflict* words, such as fight, kill, argue, reflecting referents to concepts indicative of or reflecting conflict, broadly defined (e.g., Barki & Hartwick, 2004).

- *Moralization* or words reflecting judgemental language, often where the speaker makes a moral evaluation (either good or bad) about another’s behavior or character (Brady et al., 2020).

- *Communication* words are terms describing an act of communication, such as talk, explain, or disagreement.

The Expanded Dictionary includes the following new dimensions and variables:

- The *Culture* category is an overarching dimension of three cultural domains:

    - *Politics* is a broadly-construed variable that includes words commonly used in political (e.g., congress, parliament, president, democratic) or legal (court, law) discourse.

    - *Ethnicity* refers to words that identify national, regional, linguistic, ethnic, or racial identities. Words reflecting racial or ethnic slurs are generally excluded and are instead included as part of the swear category.

    - *Technology* words refer to scientific and technological devices and inventions (e.g., wifi, nuclear, computer). This category ranges from common devices that fall under the broadest conceptualizations of technology use in “cybernetics” (Kline, 2009) to common innovations that have had observable impact on human culture and society.

- *Health* dimension additions. Previous LIWC versions that included physical and health terms have now been broken into more defined categories that include:
    - *Illness* which includes disease names and illness related physical symptoms.
    - *Wellness* terms include words such as yoga, diet, exercise, fasting.
    - *Mental health* terms typically refer to diagnoses (e.g., bipolar, neurosis) or behaviors (suicide, addiction).
    - *Substances* refer to drugs (including alcohol) often associated with abuse or psychological escape.
- *States* refer to many short-term or transient internal states that can drive or inhibit behaviors (see, e.g., Kenrick et al., 2010; Schaller et al., 2017). The six states dimensions include:
    - *Need* states are associated with language that necessitates specific actions or behaviors to ensure the person's well-being and survival. Implicit to the concept of "need" is that important outcomes a requisite upon a limited set of specific actions.
    - *Want* words signal the authors' desires or preferences. Often, "wants" are philosophically distinguished from "needs" in that needs are conceptualized as innate and requisite for survival, whereas wants are learned and generally more associated with additional satisfaction above and beyond basic needs (e.g., Buttle, 1989; Mari, 2008; Oliver, 2014).
    - *Acquire* words reflect the searching for, finding, or obtaining of objects, states, or goals that serve one's needs or wants.
    - *Lack* is an expression of a missing physical object or abstract state associated with one's needs, wants, or goals. That is, a discrepancy between a current state and a desired or more complete alternate state.
    - *Fulfill* refers to the language of a biological or psychological state of completion, satisfaction of a goal, satiation, or "having enough."
    - *Fatigue* words often reflect exhaustion, boredom, or expended effort (see, e.g., Hockey, 2013).

- *Motives* refer to underlying states that drive, guide, or pull a person to behave. In addition to the *reward* and *risk* (which were part of LIWC2015), two additional variables have been added:

    - *Curiosity* words reflect the authors' search for or interest in new knowledge or experiences. This is thought to be a correlate of openness to new experiences.

    - *Allure* is a dimension derived from the world of advertising (Kannan & Tyagi, 2013) made up of words commonly used in successful ads and persuasive communications.

Lastly, the four "summary measures" of *Analytic*, *Clout*, *Authenticity*, and *Tone* have been re-normed to better reflect their base rates across a wider set of texts.[4] The "raw" means are noticeably lower for the summary measures than what would be seen with the LIWC2015 versions – while the correlations will generally be very high (close to +1.00), you may see that the scores themselves are different in LIWC-22. Consequently, if you are comparing summary variables that were originally run with LIWC2015, the LIWC-22 variables will have different values. See Table 4 for the relevant statistics.

[4] Note that the underlying algorithms for all four summary measures are fundamentally the same as in previous versions of LIWC. Conceptually, however, the emotion-related words that serve as the components of the *Tone* variable have been updated and made cleaner as described earlier in the discussion of the affect variables.

LIWC categories no longer included

We have removed a small number of variables due to their consistently low base rates, low internal reliability, or their infrequent use by researchers. These include:

- Comparison words (greater, best, after)
- Interrogatives (who, what, where)
- Relativity (sum of time, space, motion words)
- Certain low base-rate punctuation (colons, semicolons, dashes, quotation marks, parentheses)

Note that the LIWC-22 application comes with the original internal dictionaries for both LIWC2001, LIWC2007, and LIWC2015 for those who want to rely on older versions of the dictionary as well as to compare LIWC-22 analyses with those provided by older versions of the software.

For users of LIWC2015, a new edition of LIWC that uses a different dictionary can be an unsettling experience. Most of the older dictionaries have been slightly changed, some have been substantially reworked (e.g., social words, cognitive process words), and several others have been removed or added. To assist in the transition to the new version of LIWC, we include Table 4 which lists the means, standard deviations, and correlations between the two dictionary versions. These analyses are based on the corpora detailed in Tables 2 and 3. All numbers presented in Table 4 are the average results from all 15 Test Kitchen corpora.

To get a sense of how much a dictionary has changed from the LIWC2015 to the LIWC-22 versions, look at the “LIWC-22/2015 Correlation” column. The lower the correlation, the more change across the two versions.

Table 4. Comparisons Between LIWC-22 and LIWC2015

LIWC-22 (LIWC2015) Variable	LIWC-22		LIWC2015		LIWC-22/ LIWC2015 Correlation
	Mean	SD	Mean	SD	
WC	2070.5	2466.4	2070.1	2466.1	1.00
Analytic	49.6	29.8	59.5	28.0	0.99
Clout	49.6	28.4	62.5	22.1	0.95
Authentic	50.0	28.0	42.7	26.9	0.94
Tone	48.2	26.4	55.9	28.3	0.84
Words/sentence	17.2	34.7	17.2	34.7	1.00
Big words	17.2	6.7	17.2	6.7	1.00
Dictionary words	88.0	5.9	86.6	6.2	0.96
Linguistic	69.6	8.2			
function	54.6	7.1	51.8	6.8	0.90
pronoun	14.8	5.0	14.9	5.0	1.00
ppron	9.9	4.1	9.7	4.0	1.00
i	4.3	3.4	4.2	3.4	1.00
we	0.9	1.1	0.9	1.1	1.00
you	1.7	1.9	1.7	1.9	1.00
shehe	2.0	2.5	2.0	2.5	1.00
they	0.9	1.0	0.9	1.0	1.00
ipron	4.9	2.0	5.2	2.0	0.98
det	14.3	2.9			
article	6.7	2.6	6.7	2.6	1.00
number	2.1	2.6	2.2	2.6	1.00
prep	13.2	2.7	13.1	2.7	1.00
auxverb	8.7	2.9	8.6	2.8	0.99
adverb	5.3	2.4	4.9	2.3	0.97
conj	6.2	1.9	6.0	1.8	0.97
negate	1.6	1.1	1.7	1.1	0.99
verb	17.0	4.4	16.4	4.4	0.99
adj	6.0	1.7	4.5	1.5	0.74
quantity	4.0	1.7	2.0	0.9	0.56
Drives	4.4	2.3	7.5	2.5	0.79
affiliation	1.9	1.4	2.2	1.5	0.90
achieve	1.2	1.0	1.5	1.1	0.90
power	1.3	1.3	2.5	1.4	0.71
Cognition	22.2	4.9			
allnone	1.3	0.9			0.65[a]
cogproc	10.4	3.4	10.7	3.5	0.95
insight	2.4	1.2	2.2	1.1	0.91
cause	1.4	0.8	1.5	0.8	0.91
discrep	1.7	1.0	1.6	0.9	0.82
tentat	2.1	1.3	2.4	1.4	0.93
certitude (certainty)	0.6	0.6	1.5	0.8	0.33[a]
differ	3.1	1.4	2.9	1.4	0.91
memory	0.1	0.2			
Affect	5.4	2.6	5.6	2.4	0.87
tone_pos (posemo)	3.5	2.2	3.7	2.1	0.86
tone_neg (negemo)	1.5	1.1	1.8	1.2	0.84
emotion	1.9	1.5			0.75[b]
emo_pos	1.1	1.2			0.74[b]
emo_neg	0.7	0.6			0.68[b]
emo_anx (anx)	0.1	0.2	0.3	0.3	0.69[b]
emo_anger (anger)	0.1	0.2	0.6	0.7	0.51[b]
emo_sad (sad)	0.1	0.2	0.4	0.4	0.48[b]
swear	0.3	0.6	0.3	0.6	0.98
Social	12.1	4.2	10.2	3.9	0.89[c]
socbehav	3.9	2.0			
prosocial	0.7	0.8			
polite	0.4	1.1			
conflict	0.2	0.3			
moral	0.3	0.3			
comm	1.6	1.1			
socrefs	8.1	3.5			0.91[c]
family	0.4	0.6	0.5	0.7	0.93
friend	0.2	0.3	0.4	0.4	0.67
female	1.3	2.2	1.3	2.2	1.00
male	1.5	1.7	1.5	1.7	0.99

Table 4. Comparisons Between LIWC-22 and LIWC2015 (Cont'd)

LIWC-22 (LIWC2015) Variable	LIWC-22		LIWC2015		LIWC-22/ LIWC2015 Correlation
	Mean	SD	Mean	SD	
Culture	0.8	1.3			
politic	0.4	1.0			
ethnicity	0.1	0.4			
tech	0.3	0.6			
lifestyle	4.3	2.7			
leisure	0.6	0.8	1.2	1.1	0.84
home	0.4	0.5	0.5	0.6	0.89
work	2.5	2.4	3.1	2.7	0.95
money	0.7	1.3	0.8	1.3	0.97
relig	0.2	0.5	0.3	0.6	0.97
Physical (bio)	2.4	2.2	2.4	1.9	0.92
health	0.7	1.4	0.8	1.2	0.94[d]
illness	0.2	0.4			0.69[d]
wellness	0.0	0.2			
mental	0.0	0.1			
substances	0.1	0.2			
sexual	0.1	0.3	0.2	0.3	0.76
food (ingest)	0.7	1.6	0.7	1.3	0.95
death	0.1	0.3	0.2	0.3	0.91
need	0.5	0.5			
want	0.4	0.4			
acquire	0.8	0.6			
lack	0.1	0.3			
fulfill	0.1	0.2			
fatigue	0.1	0.2			
reward	0.2	0.3	1.5	1.0	0.29
risk	0.2	0.3	0.5	0.5	0.63
curiosity	0.4	0.5			
allure	7.0	3.1			
Perception	9.3	2.9	2.5	1.4	0.53
attention	0.5	0.5			
motion	1.7	0.9	2.1	1.0	0.79
space	6.0	2.0	6.8	2.1	0.78
visual (see)	1.0	0.8	1.0	0.8	0.88
auditory (hear)	0.3	0.4	0.7	0.7	0.65
feeling (feel)	0.5	0.5	0.6	0.5	0.86
time	4.4	1.9	5.2	2.0	0.90
focuspast	4.7	2.9	4.4	2.7	0.98
focuspresent	4.7	2.5	10.1	4.4	0.89
focusfuture	1.5	1.1	1.4	0.9	0.86
Conversation	1.3	2.3			
netspeak	0.7	1.5	0.6	1.3	0.95
assent	0.4	1.0	0.5	1.1	0.98
nonflu	0.2	0.6	0.4	0.7	0.92
filler	0.1	0.3	0.1	0.3	0.88
AllPunc	21.6	16.8	21.7	16.9	1.00
Period	8.6	7.3	8.6	7.3	1.00
Comma	4.1	3.0	4.1	3.0	1.00
QMark	1.2	9.7	1.2	9.7	1.00
Exclam	1.1	3.3	1.1	3.4	1.00
Apostro	2.1	2.2	2.1	2.2	1.00
OtherP	4.5	6.4	4.5	6.4	1.00

Notes: Analyses are based on the full Test Kitchen Corpus of 15,000 files. Because several significant changes were made in the LIWC-22 dictionaries, the following notes help explain the apparent discrepancies in the means and/or correlations between the same variables.

[a] The LIWC2015 variable "certainty" proved to be measuring two overlapping constructs, all-or-none thinking (or "allnone") and a form of grandiose talking, which we now call "certitude". Correlations with LIWC-22 allnone and certitude variables are both correlated with the 2015 variable of certainty.

[b] Whereas earlier versions of LIWC labeled all affect dictionaries as "affect" or "emotion" dimensions, LIWC-22 makes a finer distinction between broad sentiment and more targeted emotions. For LIWC-22, the variables "affect", "tone_pos," and "tone_neg" correspond to the LIWC2015 affect, posemo, and negemo dimensions (which we now consider sentiments). The new dimensions of "emotion," "emo_pos," "emo_neg," "emo_anx," "emo_anger," and "emo_sad" are based on specific emotion words. Note that the old anxiety, anger, and sadness dimensions were sentiment-based and no longer exist. However, the correlations with the current emotion versions suggest that the emotion dimensions still overlap with previous sentiments.

[c] The LIWC-22 social dimensions are greatly expanded. In theory, the overarching "social" dimension is quite similar to the LIWC2015 "social" category. In many ways, however, the old social dimension was based primarily on social referents, as is apparent in the table. Note that both LIWC-22 "social" and "social referents" are correlated with LIWC2015 "social."

[d] The LIWC2015 "health" dimension included both health and illness words. These dimensions have been broken into separate categories in LIWC-22. Correlations for "health" and "illness" are both correlated with LIWC2015 "health."

LIWC Dictionary Translations

Over the years, the LIWC dictionaries have been translated into several languages in collaboration with researchers all over the world, including:

- Brazilian Portuguese (Carvalho et al., 2019; Filho et al., 2013)
- Chinese (Huang et al., 2012)
- Dutch (Boot et al., 2017; van Wissen & Boot, 2017)
- French (Piolat et al., 2011)
- German (Meier et al., 2018; Wolf et al., 2008)
- Italian (Agosti & Rellini, 2007)
- Japanese (Igarashi et al., 2021)
- Norwegian (Goksøyr, 2019)
- Romanian (Dudău & Sava, 2020)
- Russian (Kailer & Chung, 2011)
- Serbian (Bjekić et al., 2014)
- Spanish (Ramírez-Esparza et al., 2007)
- Turkish (Müderrisoğlu, 2012)
- Ukrainian (Zasiekin et al., 2018)

To date, these translations have relied primarily on the LIWC2001, LIWC2007, or LIWC2015 dictionaries.

The various LIWC dictionary translations, as well as other published dictionaries, are available to academic users at the LIWC dictionary repository (https://www.liwc.app/dictionaries).

WARNING: Translating LIWC into another language *sounds* deceptively simple. It's not. We can provide multiple emails from people who have tried and ultimately gave up after years of effort. Orchestrating a scientific, psychologically, and culturally valid translation of the LIWC dictionary often requires a team of full-time researchers working exclusively for a year or more on the translation process. The translation of LIWC is not for the faint of heart.

We and others have begun to encourage a much simpler approach: machine translation. Rather than translating the dictionary, we strongly recommend translating your text samples into English and then running the English LIWC on them. We know, we know, Google Translate doesn't do a perfect job. However, LIWC is heavily based on the analysis of common function, emotion, and other everyday words which all translate with great reliability. In general, automated translations of text data provide results that are quite trustworthy (Barbosa et al., 2021; Meier et al., 2021; Windsor et al., 2019).

Seriously — we have worked with many very, very serious scholars who were dedicated to creating a high-quality dictionary translation, yet ultimately never finished. If, after reading this warning, you would still like to build a non-English LIWC-22 dictionary or if you have built one independently and would like to add it to the repository, please contact Ryan Boyd and/or James Pennebaker.

Acknowledgements

Portions of the research reported in this manual were made possible by funds from the University of Texas at Austin, the Templeton Foundation (48503, 62256), the National Science Foundation (SES1758835, IIS-1344257, DGE-161040, IIS2107524), the National Institutes of Health (MH117172, GM112697), the Department of Justice (DJF-BAA15F06718R0006603, NIJ-2021-60007), Microsoft Research, and Receptiviti, Inc. None of the above-mentioned institutions/organizations played any role, direct or otherwise, in the development of LIWC-22, collection or interpretation of data, or in the writing of this report.

We are also deeply indebted to a number of people who have helped with different phases of building and evaluating LIWC-22, including current and former graduate students:

> Kate Blackburn, Serena Brandler, Cindy Chung, Pelin Cunningham-Erdogdu, Martha Francis, Molly Ireland, Kayla Jordan, Remy Mallett, Alexi Martel, Matthias Mehl, Kate Niederhoffer, Shruti Padke, Nikkita Sarna, Miti Shah, Richard Slatcher, Mohini Tellakat, and Katie Williams.

We are particularly indebted to the LIWC-22 Dictionary Development Team, including:

> Yaretzi Campos, Isabel Webb Carey, John Henry Cruz, Oliver Davidson, Darby Edmonson, Abby Evans, Katherine Ewend, Jana Fakhreddine, Rebecca Hernandez, Jonnie Rose Hontanosas, Grace Jumonville, Gabriela Leyva-Montiel, Jade Marion, Steven Mesquiti, Omar Olivarez, Emily Pencis, Laura Sowin, Leela Srinivasan, Molly Sun, Rachel Thompson, Juo-Lin Tsai, Nishi Uppuluri, Weixi Wang, Andreas Weyland, and Aliah Zewail.

Finally, many of the ideas behind this project were influenced by the work of our fellow text analysis experts, including (but in no way limited to):

- H. Andrew Schwartz, Lyle Unger, and the core WWBP team
- Eric Horvitz and his research group at Microsoft Research
- Morteza Dehghani and the Computational Social Sciences Laboratory
- Rada Mihalcea and her students at the University of Michigan

The programming and development of LIWC-22 was expertly done by our dear colleagues at SVAPS Systems, headed by Alexander Chingarev and his remarkable team. The layout and design of LIWC-22 was greatly aided by the Houston-based firm Pennebaker Designs.

References

Agosti, A., & Rellini, A. (2007). *The Italian LIWC dictionary*. LIWC.app.

Allport, G. W. (1942). *The use of personal documents in psychological science*. Social Science Research Council.

Al-Mosaiwi, M., & Johnstone, T. (2018). In an absolute state: Elevated use of absolutist words is a marker specific to anxiety, depression, and suicidal ideation. *Clinical Psychological Science*, *6*(4), 529–542. https://doi.org/10.1177/2167702617747074

Auxier, B., & Anderson, M. (2021, April 7). Social media use in 2021. *Pew Research Center: Internet, Science & Tech*. https://www.pewresearch.org/internet/2021/04/07/social-media-use-in-2021/

Barbosa, A., Ferreira, M., Ferreira Mello, R., Dueire Lins, R., & Gasevic, D. (2021). The impact of automatic text translation on classification of online discussions for social and cognitive presences. *LAK21: 11th International Learning Analytics and Knowledge Conference*, 77–87. https://doi.org/10.1145/3448139.3448147

Barki, H., & Hartwick, J. (2004). Conceptualizing the construct of interpersonal conflict. *International Journal of Conflict Management*, *15*(3), 216–244. https://doi.org/10.1108/eb022913

Bjekić, J., Lazarević, L. B., Živanović, M., & Knežević, G. (2014). Psychometric evaluation of the Serbian dictionary for automatic text analysis—LIWCser. *Psihologija*, *47*(1), 5–32. https://doi.org/10.2298/PSI1401005B

Boot, P., Zijlstra, H., & Geenen, R. (2017). The Dutch translation of the Linguistic Inquiry and Word Count (LIWC) 2007 dictionary. *Dutch Journal of Applied Linguistics*, *6*(1), 65–76. https://doi.org/10.1075/dujal.6.1.04boo

Boyd, R. L. (2018). *MEH: Meaning Extraction Helper [Software]* (2.1.06) [Computer software]. https://meh.ryanb.cc

Boyd, R. L. (2020). *BUTTER: Basic Unit-Transposable Text Experimentation Resource*. https://www.butter.tools

Boyd, R. L., Blackburn, K. G., & Pennebaker, J. W. (2020). The narrative arc: Revealing core narrative structures through text analysis. *Science Advances*, *6*(32), 1–9. https://doi.org/10.1126/sciadv.aba2196

Boyd, R. L., Pasca, P., & Lanning, K. (2020). The personality panorama: Conceptualizing personality through big behavioural data. *European Journal of Personality*, *34*(5), 599–612. https://doi.org/10.1002/per.2254

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

Boyd, R. L., & Schwartz, H. A. (2021). Natural language analysis and the psychology of verbal behavior: The past, present, and future states of the field. *Journal of Language and Social Psychology*, *40*(1), 21–41. https://doi.org/10.1177/0261927X20967028

Boyd, R. L., Wilson, S. R., Pennebaker, J. W., Kosinski, M., Stillwell, D. J., & Mihalcea, R. (2015). Values in words: Using language to evaluate and understand personal values. *Proceedings of the Ninth International AAAI Conference on Web and Social Media*, 31–40. http://www.aaai.org/ocs/index.php/ICWSM/ICWSM15/paper/view/10482

Brady, W. J., Crockett, M. J., & Van Bavel, J. J. (2020). The MAD model of moral contagion: The role of motivation, attention, and design in the spread of moralized content online. *Perspectives on Psychological Science*, *15*(4), 978–1010. https://doi.org/10.1177/1745691620917336

Brown, P., & Levinson, S. C. (1978). Universals in language usage: Politeness phenomena. *Questions and Politeness: Strategies in Social Interaction*, 56–311.

Buttle, F. (1989). The social construction of needs. *Psychology & Marketing*, *6*(3), 197–210. https://doi.org/10.1002/mar.4220060304

Carvalho, F., Rodrigues, R. G., Santos, G., Cruz, P., Ferrari, L., & Guedes, G. P. (2019). Evaluating the Brazilian Portuguese version of the 2015 LIWC Lexicon with sentiment analysis in social networks. *Anais Do Brazilian Workshop on Social Network Analysis and Mining (BraSNAM)*, 24–34. https://doi.org/10.5753/brasnam.2019.6545

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

Cohn, M. A., Mehl, M. R., & Pennebaker, J. W. (2004). Linguistic markers of psychological change surrounding September 11, 2001. *Psychological Science*, *15*(10), 687–693. https://doi.org/10.1111/j.0956-7976.2004.00741.x

Dudău, D. P., & Sava, F. A. (2020). The development and validation of the Romanian version of Linguistic Inquiry and Word Count 2015 (Ro-LIWC2015). *Current Psychology*. https://doi.org/10.1007/s12144-020-00872-4

Filho, P. P. B., Pardo, T. A. S., & Aluísio, S. M. (2013). An evaluation of the Brazilian Portuguese LIWC dictionary for sentiment analysis. *Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology*, 215–219.

Francis, M. E., & Pennebaker, J. W. (1992). Putting stress into words: The impact of writing on physiological, absentee, and self-reported emotional well-being measures. *American Journal of Health Promotion*, *6*(4), 280–287. https://doi.org/10.4278/0890-1171-6.4.280

Ganellen, R. J. (2007). Assessing normal and abnormal personality functioning: Strengths and weaknesses of self-report, observer, and performance-based methods. *Journal of Personality Assessment*, *89*(1), 30–40. https://doi.org/10.1080/00223890701356987

Goksøyr, A. (2019). *Norsk versjon av Language Inquiry and Word Count 2007 (LIWC2007no). Oversettelse og psykometriske egenskaper*. https://app.cristin.no/results/show.jsf?id=1678137

Gottschalk, L. A., & Gleser, G. C. (1969). The measurement of psychological states through the content analysis of verbal behavior. University of California Press.

Hockey, R. (2013). The psychology of fatigue: Work, effort and control. Cambridge University Press.

Holtgraves, T., & Joong-nam, Y. (1990). Politeness as universal: Cross-cultural perceptions of request strategies and inferences based on their use. *Journal of Personality and Social Psychology*, *59*(4), 719–729. https://doi.org/10.1037/0022-3514.59.4.719

Huang, C.-L., Chung, C. K., Hui, N., Lin, Y.-C., Seih, Y.-T., Lam, B. C. P., Chen, W.-C., Bond, M. H., & Pennebaker, J. W. (2012). The development of the Chinese Linguistic Inquiry and Word Count dictionary. *Chinese Journal of Psychology*, *54*(2), 185–201.

Igarashi, T., Okuda, S., & Sasahara, K. (2021). *Development of the Japanese Version of the Linguistic Inquiry and Word Count Dictionary 2015 (J-LIWC2015)*. https://doi.org/10.31234/osf.io/5hq7d

Ireland, M. E., & Pennebaker, J. W. (2010). Language style matching in writing: Synchrony in essays, correspondence, and poetry. *Journal of Personality and Social Psychology*, *99*(3), 549–571. https://doi.org/10.1037/a0020386

Jaidka, K., Giorgi, S., Schwartz, H. A., Kern, M. L., Ungar, L. H., & Eichstaedt, J. C. (2020). Estimating geographic subjective well-being from Twitter: A comparison of dictionary and data-driven language methods. *Proceedings of the National Academy of Sciences*, *117*(19), 10165–10171. https://doi.org/10.1073/pnas.1906364117

Jonason, P. K., Oshio, A., Shimotsukasa, T., Mieda, T., Csathó, Á., & Sitnikova, M. (2018). Seeing the world in black or white: The Dark Triad traits and dichotomous thinking. *Personality and Individual Differences*, *120*, 102–106. https://doi.org/10.1016/j.paid.2017.08.030

Jordan, K. N., Sterling, J., Pennebaker, J. W., & Boyd, R. L. (2019). Examining long-term trends in politics and culture through language of political leaders and cultural institutions. *Proceedings of the National Academy of Sciences*, *116*(9), 3476–3481. https://doi.org/10/ggdb37

Jurafsky, D., Ranganath, R., & McFarland, D. (2009). Extracting social meaning: Identifying interactional style in spoken conversation. *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 638–646. https://aclanthology.org/N09-1072

Kacewicz, E., Pennebaker, J. W., Davis, M., Jeon, M., & Graesser, A. C. (2014). Pronoun use reflects standings in social hierarchies. *Journal of Language and Social Psychology*, *33*(2), 125–143. https://doi.org/10.1177/0261927X13502654

Kailer, A., & Chung, C. K. (2011). *The Russian LIWC2007 dictionary*. Pennebaker Conglomerates. https://www.liwc.app

Kannan, R., & Tyagi, S. (2013). Use of language in advertisements. *English for Specific Purposes World*, *37*(13), 1–10.

Kenrick, D. T., Neuberg, S. L., Griskevicius, V., Becker, D. V., & Schaller, M. (2010). Goal-driven cognition and functional behavior: The fundamental-motives framework. *Current Directions in Psychological Science*, *19*(1), 63–67. https://doi.org/10.1177/0963721409359281

Kline, R. (2009). Where are the cyborgs in cybernetics? *Social Studies of Science*, *39*(3), 331–362. https://doi.org/10.1177/0306312708101046

Kosinski, M., Matz, S. C., Gosling, S. D., Popov, V., & Stillwell, D. (2015). Facebook as a research tool for the social sciences: Opportunities, challenges, ethical considerations, and practical guidelines. *American Psychologist*, *70*(6), 543–556. https://doi.org/10.1037/a0039210

Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika*, *2*(3), 151–160. https://doi.org/10.1007/BF02288391

Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S., & McClosky, D. (2014). *The Stanford CoreNLP natural language processing toolkit*. https://doi.org/10.3115/v1/P14-5010

Mari, C. (2008). *Consumer motivation: Foundations for a theory of consumption* (Edizioni Scientifiche Italiene). Social Science Research Network.

Mauss, I. B., & Robinson, M. D. (2009). Measures of emotion: A review. *Cognition & Emotion*, *23*(2), 209–237. https://doi.org/10.1080/02699930802204677

Mehl, M. R., Pennebaker, J. W., Crow, D. M., Dabbs, J., & Price, J. H. (2001). The Electronically Activated Recorder (EAR): A device for sampling naturalistic daily activities and conversations. *Behavior Research Methods, Instruments, & Computers*, *33*(4), 517–523. https://doi.org/10.3758/BF03195410

Meier, T., Boyd, R. L., Mehl, M. R., Milek, A., Pennebaker, J. W., Martin, M., Wolf, M., & Horn, A. B. (2021). (Not) Lost in translation: Psychological adaptation occurs during speech translation. *Social Psychological and Personality Science*, *12*(1), 131–142. https://doi.org/10.1177/1948550619899258

Meier, T., Boyd, R. L., Pennebaker, J. W., Mehl, M. R., Martin, M., Wolf, M., & Horn, A. B. (2018). *"LIWC auf Deutsch": The development, psychometrics, and introduction of DE-LIWC2015* (pp. 1–42). University of Zurich. https://dx.doi.org/10.17605/OSF.IO/TFQZC

Metfessel, M. (1940). The all-or-none nature of emotional thinking. *The Journal of Psychology*, *9*, 323–326. https://doi.org/10.1080/00223980.1940.9917698

Minkov, E., Balasubramanyan, R., & Cohen, W. W. (2008). Activity-centred search in email. *Proceedings of the Fifth Conference on Email and Anti-Spam*, 58–63. https://www.ceas.cc/2008/papers/ceas2008-paper-54.pdf

Morgan, W. G. (1995). Origin and history of the Thematic Apperception Test images. *Journal of Personality Assessment*, *65*(2), 237–254. https://doi.org/10.1207/s15327752jpa6502_2

Müderrisoğlu, S. (2012, April). Türkçe psikolojik metin analizi programı: LIWC Türkçe [Turkish psychological text analysis program: Turkish LIWC] [Poster]. 17th National Congress of Psychology, Istanbul, Turkey.

Murray, H. A. (1943). *Thematic apperception test*. Harvard University Press.

Neuringer, C. (1961). Dichotomous evaluations in suicidal individuals. *Journal of Consulting Psychology*, *25*(5), 445–449. https://doi.org/10.1037/h0046460

Newman, M. L., Pennebaker, J. W., Berry, D. S., & Richards, J. M. (2003). Lying words: Predicting deception from linguistic styles. *Personality and Social Psychology Bulletin*, *29*(5), 665–675. https://doi.org/10.1177/0146167203029005010

Oliver, R. L. (2014). Satisfaction: A behavioral perspective on the consumer. Routledge.

Pennebaker, J. W. (2011). The secret life of pronouns: What our words say about us. Bloomsbury.

Pennebaker, J. W., Booth, R. J., Boyd, R. L., & Francis, M. E. (2015). *Linguistic Inquiry and Word Count: LIWC2015*. Pennebaker Conglomerates. www.liwc.net

Pennebaker, J. W., Booth, R. J., & Francis, M. E. (2007). *LIWC2007: Linguistic Inquiry and Word Count*. Pennebaker Conglomerates, Inc. www.liwc.net

Pennebaker, J. W., Boyd, R. L., Booth, R. J., Ashokkumar, A., & Francis, M. E. (2022). *Linguistic Inquiry and Word Count: LIWC-22*. Pennebaker Conglomerates. https://www.liwc.app

Pennebaker, J. W., Chung, C. K., Frazee, J., Lavergne, G. M., & Beaver, D. I. (2014). When small words foretell academic success: The case of college admissions essays. *PLOS ONE*, *9*(12), e115844. https://doi.org/10.1371/journal.pone.0115844

Pennebaker, J. W., Francis, M. E., & Booth, R. J. (2001). *Linguistic Inquiry and Word Count (LIWC): LIWC2001* (pp. 1–21). Lawrence Erlbaum.

Penner., L. A., Dovidio., J. F., Piliavin., J. A., & Schroeder., D. A. (2005). Prosocial behavior: Multilevel perspectives. *Annual Review of Psychology*, *56*(1), 365–392. https://doi.org/10.1146/annurev.psych.56.091103.070141

Piolat, A., Booth, R. J., Chung, C. K., Davids, M., & Pennebaker, J. W. (2011). La version française du dictionnaire pour le LIWC: Modalités de construction et exemples d'utilisation. *Psychologie Française*, *56*(3), 145–159. https://doi.org/10.1016/j.psfr.2011.07.002

Ramírez-Esparza, N., Pennebaker, J. W., García, F. A., & Suriá, R. (2007). La psicología del uso de las palabras: Un programa de computadora que analiza textos en español. *Revista Mexicana de Psicología*, *24*(1), 85–99.

Schaller, M., Kenrick, D. T., Neel, R., & Neuberg, S. L. (2017). Evolution and human motivation: A fundamental motives framework. *Social and Personality Psychology Compass*, *11*(6), e12319. https://doi.org/10.1111/spc3.12319

Schler, J., Koppel, M., Argamon, S., & Pennebaker, J. W. (2006). Effects of age and gender on blogging. *Proceedings of the 2006 AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs*, *86*, 191–197. https://doi.org/10.1155/2015/862427

Stillwell, D., & Kosinski, M. (2012). myPersonality project: Example of successful utilization of online social networks for large-scale social research. *Proceedings of the 1st ACM Workshop on Mobile Systems for Computational Social Science (MobiSys)*, 1–2.

Stone, P. J., Dunphy, D. C., Smith, M. S., & Ogilvie, D. M. (1966). *The General Inquirer: A computer approach to content analysis*. M.I.T. Press.

Sun, J., Schwartz, H. A., Son, Y., Kern, M. L., & Vazire, S. (2020). The language of well-being: Tracking fluctuations in emotion experience through everyday speech. *Journal of Personality and Social Psychology*, *118*(2), 364–387. https://doi.org/10.1037/pspp0000244

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. https://doi.org/10.1177/0261927X09351676

van Dijk, T. A. (1999). Context models in discourse processing. In H. van Oostendorp & S. R. Goldman (Eds.), *The construction of mental representations during reading* (pp. 123–148). Lawrence Erlbaum Associates.

van Dijk, T. A. (2009). *Society and discourse: How social contexts influence text and talk*. Cambridge University Press. https://doi.org/10.1017/CBO9780511575273

van Wissen, L., & Boot, P. (2017). An electronic translation of the LIWC dictionary into Dutch. In I. Kosem, C. Tiberius, M. Jakubíček, J. Kallas, S. Krek, & V. Baisa (Eds.), *Electronic lexicography in the 21st century: Proceedings of eLex 2017 conference* (pp. 703–715). Lexical Computing CZ s.r.o., Brno, Czech Republic. https://elex.link/elex2017/proceedings-download/

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology*, *54*(6), 1063–1070. https://doi.org/10.1037/0022-3514.54.6.1063

Weintraub, W. (1989). *Verbal behavior in everyday life*. Springer.

Windsor, L. C., Cupit, J. G., & Windsor, A. J. (2019). Automated content analysis across six languages. *PLOS ONE*, *14*(11), e0224425. https://doi.org/10.1371/journal.pone.0224425

Wolf, M., Horn, A. B., Mehl, M. R., Haug, S., Pennebaker, J. W., & Kordy, H. (2008). Computergestützte quantitative textanalyse: Äquivalenz und robustheit der deutschen Version des Linguistic Inquiry and Word Count. [Computer-aided quantitative textanalysis: Equivalence and reliability of the German adaptation of the Linguistic Inquiry and Word Count.]. *Diagnostica*, *54*(2), 85–98. https://doi.org/10.1026/0012-1924.54.2.85

Zasiekin, S., Bezuglova, N., Hapon, A., Matiushenko, V., Podolska, O., & Zubchuk, D. (2018). Psycholinguistic aspects of translating LIWC dictionary. *East European Journal of Psycholinguistics*, *5*(1), 111–118. https://doi.org/10.5281/zenodo.1436335

Appendix A: The Test Kitchen Corpus

The underlying logic of creating the Test Kitchen was to bring together a broad set of English language samples that generally reflect the ways people write and speak across a wide array of contexts for the purpose of testing and validating the LIWC-22 program. Over the years, the Pennebaker Lab has acquired hundreds of data sets from our own surveys and experiments, as well as digitized public text archives, and text data from labs of colleagues. The purpose of this report is to describe the different sets of data that make up the Test Kitchen corpus, how they were collected and cleaned, and strengths and weaknesses of each set.

Note that several of these data sets are proprietary and/or may contain personal information. Consequently, the Test Kitchen corpus is not public and cannot be shared in raw form. Randomized, bag-of-words versions of the dataset that cannot be reversed engineered may, in special cases, be available to bonafide university researchers. The conditions under which the Test Kitchen corpus will be shared is at the sole discretion of the first and last authors of this document.

Overview

The Test Kitchen comprises 15 different sets of data, each of which includes 1,000 individual text files with between 100 and 10,000 words. The entire 15,000 file data set is 31 million words. As seen in Table 1, all the files in the Test Kitchen corpus were randomly selected from larger data archives.

Table A1. The Test Kitchen Corpus

Corpus	Description	Test Kitchen *N*	Years Written	Population *N*
Applications	Technical college admissions essays	1,000	2018-2019	2500
Blogs	Individual blogs from blogger.com	1,000	1999-2004	37,296
Conversations	Natural conversations	1,000	1996-2019	4,000
Enron Emails	Internal emails from Enron	1,000	1995-2001	5,367
Facebook	Facebook user timelines from mypersonality.com	1,000	2004-2012	141,000
Movies	Transcribed movie dialogue	1,000	1912-2014	19,970
Novels	Novels from Project Gutenberg	1,000	1789-1970	2,523
NYT	New York Times articles	1,000	1989-2017	18,312
Reddit	Individuals' Reddit comments	1,000	2019-2020	50,000
Short Stories	Short stories	1,000	1819-2016	2092
SOC	Stream of consciousness essays	1,000	2015-2016	1574
Speeches	U.S. Congressional speeches	1,000	1994-2016	357,080
TAT	Thematic Apperception Test, online website	1,000	2011-2019	14,000
Tweets	Collected tweets from individual accounts	1,000	2016-2020	> 1.5 million
Yelp	Published Yelp reviews	1,000	2010-2019	1,048,366

The Test Kitchen corpus files were randomly selected from a larger population of writing samples, labeled "Population N". The "Years written" column refers to the range of years that the texts in the population sample were written.

Corpus and Text File Selection

From the beginning, we sought to collect a diverse set of text files that would be relevant to the needs of a wide array of researchers who might use LIWC-22. It was important to include samples that reflected both informal and formal contexts, spoken and written formats, texts that spoke to narrow and broad audiences, and reflected authors who were in different social and psychological states when their texts were generated.

The final set over-represents texts written in the 21st century (13 of the 15 corpora), recent social media platforms (5 sets), authors under the age of 30 (likely the majority of cases in 7 sets), and the language of college students (3). For most data sets, age, sex, ethnicity, and education were not available and/or linked to the Test Kitchen files.

All Test Kitchen files were randomly selected from much larger data sets. To be eligible for selection, files had to have a minimum of 100 words, written in English, with a minimum LIWC "Dic" score of 65 (meaning that at least 65% of the words in the file were captured by all LIWC2015 dictionaries. Additional screening was conducted on Reddit and Twitter samples to exclude files populated by bots. For the email sample, automated informational emails were deleted.

The randomly selected texts that had more than 10,000 words were trimmed so that the file included 10,000 contiguous words using the following procedure: 1) total word count (WC) within each text was computed and subtracted by 100, yielding a corrected WC (cWC); 2) a random number, *N*, between 0 and cWC was selected; 3) the program identified the *N*th word in the text and selected all words from *N* through *N* plus 9,999 words or to the final word in the text, whichever comes first. This procedure assured that different parts of very long texts were equally likely to be sampled.

Once the final samples were identified, standard cleaning methods were applied, including removal of email addresses, urls, high numbers (>5) letters or punctuation marks in a row, html or related tags or code, etc. It should be noted that on rare occasions, this reduced the word count of a file to fewer than 100 words (less than 0.1 percent of files had fewer than 95 words).

Individual Corpus Characteristics

Applications. The admission applications were written by students applying to two-year masters programs in one of several health practitioner roles. In the essays, students wrote about their backgrounds and why they sought admission to the program. The original corpus from which the 1,000 Test Kitchen were drawn involved 2,500 essays from over 4 years.

Blogs. The complete entries from 37,296 blogs were collected from blogger.com in August 2004. For complete information on corpus design, see (Schler et al., 2006). Blogs contained everything written from their inception to the day they were collected. The final corpus from which the 1,000 Test Kitchen was drawn contained the full content of blogs by 35,385 individuals, ranging in total length from 107 to 481,983 words. Age and gender data were available for 27.4% of bloggers (N = 9688). Among these, approximately half of bloggers were female; ages ranged from 13 to 88 (M = 22.41).

Conversations. The conversation corpus is a collection of text files from several laboratories that have collected a variety of natural conversations among people in lab settings as well as interactions in the real world. Of the 1,000 text files, approximately 280 come from one of about five experiments where people wore digital recording devices for as many as 4 days. The devices, called electronically-activated recorders, or EARs, typically recorded for 30 seconds once every 12 minutes as people went about their lives (Mehl et al., 2001). Only the words uttered by the persons wearing the EAR were included in the data set. About 350 recordings came from a large study by Jurafsky et al. (Jurafsky et al., 2009) that involved heterosexual speed dating interactions. Each file included only a single person's language. A third project of approximately 220 interactions were conversations between two strangers in a get-to-know-you exercise (Kacewicz et al., 2014). Two sets of studies involved business students (N = 70) and psychology students (N = 50) participating in groups tasks. Most of the studies involved college students between 18 and 30 years old. Finally, about 30 files were gathered from microphones placed in public places on a university campus (e.g., dining hall, student post office, hallways) and were transcribed without knowledge of anyone speaking (see Pennebaker, 2011, chapter 9).

Emails. In December, 2001, one of the world's most powerful energy and financial institutions, Enron Corporation, collapsed and declared bankruptcy. The ensuing scandal resulted in multiple trials of the senior management. As part of the legal battles, most of Enron's corporate emails were initially made public. Soon thereafter, a smaller email set of about 520,000 emails were made public involving about 150 of the senior management of the company (see Minkov et al., 2008). The emails to and from this group involved 5,367 individuals, from which the 1,000 email text corpus was derived.

Facebook. Between 2004 and 2012, David Stillwell and his colleagues created one of the most impressive data sets in the history of psychology (Kosinski et al., 2015; Stillwell & Kosinski, 2012). Mypersonality.com offered users the opportunity to complete a wide set of questionnaires that provided feedback about their traits, desires, emotions, etc. They were also given the option to link their Facebook account so that the researchers could analyze their friendship networks, likes, the words they wrote, and the pictures they uploaded. The mypersonality.com website stopped data collection in 2012 and eventually stopped sharing its data in 2018. Our lab worked with Stillwell and Michal Kosinski and have been able to rely on 141,000 individuals'

anonymized Facebook posts (Boyd et al., 2015). One thousand were randomly selected for the Test Kitchen corpus.

Movies. A corpus of 19,970 movie subtitles (*M* word count = 7,231) was provided by OpenSubtitles.org, which is created, maintained, and updated by a massive volunteer-based crowdsourcing effort. The corpus consists only of subtitles in English corresponding to movies that were either originally released in English or, for international movies, where the dialogue has been translated to English. Data cleaning involved removing all text unrelated to the dialog such as timestamps. The original data was collected for a separate project (see Boyd, Blackburn, et al., 2020 for more details).

Novels. A large corpus of texts was originally extracted from the 2010 Project Gutenberg DVD (http://www.gutenberg.org), then winnowed down to 2,523 novels based on their official Library of Congress classifications, which served as the basis for a dataset for Boyd et al (2020). All novels collected were written by authors living between 1789-1970, with most publishing in the 1800s (median publication date = 1886). The 1,000 texts for the Test Kitchen were randomly selected from the Boyd et al. (2020) archive.

New York Times. The corpus (*N* = ~2.7M) contains multiple types of articles, including editorials, features, opinions articles, world, U.S., and local news, letters to the editor, etc. which were published in the New York Times between 1989 and 2017. The articles were collected from the publication's website for the Boyd et al. (2020) project. The overall mean word count for this corpus was 827.

Reddit. All comments that were made on the *r/askreddit* subreddit between December 1, 2019 and February 15, 2020 were harvested and aggregated by user. Only those users with a minimum of 1,000 words and who met the standard data cleaning thresholds were retained, resulting in approximately 50,000 text files. Although Reddit does not collect demographic information, estimates for 2019 suggest that 63 percent of users are male with the median age somewhere around 29 (Auxier & Anderson, 2021).

Short Stories. Short Stories were collected from 36 different online sources not restricted by copyright laws and freely available to the public. All short stories were written between 1819-2016, the majority of which were self-published online after 1995. Approximately, 10 percent were initially released through publishing houses after copyrights expired. The original sample of 2,092 short stories were collected as part of the Boyd et al. (2020) arc of narrative project.

Speeches. Each year, members of the U.S. senate and house deliver a number of speeches, most of which are published in the *Congressional Record*. The speech corpus includes 357,080 public speeches, debates, tributes, etc. between 1994 and 2016 delivered by 2,035 members of the senate and house. For more information on the corpus, see Jordan et al. (2019).

Stream of consciousness. Students in an online introductory psychology class were asked to write for 20 minutes in a stream of consciousness way. The students who wrote the 1,574 essays averaged 18.8 years old, with 60.7 percent identifying as female. More detailed information about the procedures is available from Vine et al. (2020).

Thematic Apperception Test (TAT). The TAT corpus was collected from online users visiting the website www.secretlifeofpronouns.com or www.utpsyc.org which gave people the option to complete several online quizzes and tests. The TAT was originally developed by Henry D. Murray and Christiana Morgan to catalog people's underlying motivations and personality through their use of words and themes (Morgan, 1995; Murray, 1943). The test asks participants to view an ambiguous picture and to then compose a story based on that drawing. All TAT narratives were based on a single drawing that depicts two people in a laboratory. For more detail on the data set and methods, see Boyd et al. (2020).

Twitter. Since 2016, the first author has been collecting the content of user timelines posted to Twitter. Using an English-language filter, a custom pipeline is used to collect tweets from random users anywhere in the world; up to their ~3,200 most recent tweets are collected, then warehoused. For high-activity users, 3,200 tweets often does not extend back more than 1-2 years. For low-activity users, 3,200 tweets may constitute their entire history of Twitter use. To date, this corpus consists of > 1.5 million user timelines. For the purpose of the Test Kitchen, 1,000 user random timelines were randomly sampled and aggregated during the summer of 2020. Hashtags, URLs, and usernames were removed from all tweet texts prior to aggregation.

Yelp reviews. Yelp makes its reviews public, including ways to easily download the reviews with metadata (https://www.yelp.com/dataset). We worked with a corpus of over one million reviews posted between 2010 and 2019. For the Test Kitchen corpus, we limited the 1,000 entries to reviews that were exactly 100 words. Note that, with cleaning, the mean number of words dropped to 99.

Appendix B: Recommended Further Reading

Alvero, A. J., Giebel, S., Gebre-Medhin, B., Antonio, A. lising, Stevens, M. L., & Domingue, B. W. (2021). Essay content and style are strongly related to household income and SAT scores: Evidence from 60,000 undergraduate applications. *Science Advances*. https://doi.org/10.1126/sciadv.abi9031

Argamon, S., Koppel, M., Pennebaker, J. W., & Schler, J. (2007). Mining the blogosphere: Age, gender and the varieties of self-expression. *First Monday*, *12*(9). https://firstmonday.org/ojs/index.php/fm/article/view/2003

Ashokkumar, A., & Pennebaker, J. W. (2021). Social media conversations reveal large psychological shifts caused by COVID-19's onset across U.S. cities. *Science Advances*, *7*(39), eabg7843. https://doi.org/10.1126/sciadv.abg7843

Back, M. D., Küfner, A. C. P., & Egloff, B. (2011). "Automatic or the people?" Anger on September 11, 2001, and lessons learned for the analysis of large digital data sets. *Psychological Science*, *22*(6), 837–838. https://doi.org/10.1177/0956797611409592

Baddeley, J. L., Daniel, G. R., & Pennebaker, J. W. (2011). How Henry Hellyer's use of language foretold his suicide. *Crisis*, *32*(5), 288–292. https://doi.org/10.1027/0227-5910/a000092

Bantum, E. O., & Owen, J. E. (2009). Evaluating the validity of computerized content analysis programs for identification of emotional expression in cancer narratives. *Psychological Assessment, 21*(1), 79–88. https://doi.org/10.1037/a0014643

Barrett, L. F., Adolphs, R., Marsella, S., Martinez, A. M., & Pollak, S. D. (2019). Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements: *Psychological Science in the Public Interest*. https://doi.org/10.1177/1529100619832930

Bayram, A. B., & Ta, V. P. (2019). Diplomatic chameleons: Language style matching and agreement in international diplomatic negotiations. *Negotiation and Conflict Management Research*, *12*(1), 23–40. https://doi.org/10.1111/ncmr.12142

Berger, J., & Packard, G. (2021). Using natural language processing to understand people and culture. *American Psychologist*, No Pagination Specified-No Pagination Specified. https://doi.org/10.1037/amp0000882

Biber, D. (1988). *Variation across speech and writing*. Cambridge University Press. https://doi.org/10.1017/CBO9780511621024

Bierstetel, S. J., Farrell, A. K., Briskin, J. L., Harvey, M. W., Gable, S. L., Ha, T., Ickes, W., Lin, W.-F., Orina, M. M., Saxbe, D., Simpson, J. A., Ta, V. P., & Slatcher, R. B. (2020). Associations between language style matching and relationship commitment and satisfaction: An integrative data analysis. *Journal of Social and Personal Relationships*, *37*(8–9), 2459–2481. https://doi.org/10.1177/0265407520923754

Blackburn, K. G., Wang, W., Pedler, R., Thompson, R., & Gonzales, D. (2020). Linguistic markers in women's discussions on miscarriage and abortion illustrate psychological responses to their experiences. *Journal of Language and Social Psychology*, 0261927X20965643. https://doi.org/10.1177/0261927X20965643

Block, J. H., Fisch, C. O., Obschonka, M., & Sandner, P. G. (2019). A personality perspective on business angel syndication. *Journal of Banking & Finance*, *100*, 306–327. https://doi.org/10.1016/j.jbankfin.2018.10.006

Boals, A., & Klein, K. (2005). Word use in emotional narratives about failed romantic relationships and subsequent mental health. *Journal of Language and Social Psychology*, *24*(3), 252–268. https://doi.org/10.1177/0261927X05278386

Boals, A., & Perez, A. S. (2009). Language use predicts phenomenological properties of Holocaust memories and health. *Applied Cognitive Psychology*, *23*(9), 1318–1332. https://doi.org/10.1002/acp.1538

Bosson, J. K., Swann, W. B., & Pennebaker, J. W. (2000). Stalking the perfect measure of implicit self-esteem: The blind men and the elephant revisited? *Journal of Personality and Social Psychology*, *79*(4), 631–643. https://doi.org/10.1037/0022-3514.79.4.631

Boyd, R. L. (2017). Psychological text analysis in the digital humanities. In S. Hai-Jew (Ed.), *Data Analytics in Digital Humanities* (pp. 161–189). Springer International Publishing. https://doi.org/10.1007/978-3-319-54499-1_7

Boyd, R. L. (2018). Mental profile mapping: A psychological single-candidate authorship attribution method. *PLOS ONE*, *13*(7), e0200588. https://doi.org/10.1371/journal.pone.0200588

Boyd, R. L., & Pennebaker, J. W. (2015). Did Shakespeare write Double Falsehood? Identifying individuals by creating psychological signatures with text analysis. *Psychological Science*, *26*(5), 570–582. https://doi.org/10.1177/0956797614566658

Boyd, R. L., & Pennebaker, J. W. (2018). Building a personalized college major selection web page. *PsyArXiv*. https://doi.org/10.31234/osf.io/grf9x

Brewer, M. B., & Gardner, W. (1996). Who is this "We"? Levels of collective identity and self representations. *Journal of Personality and Social Psychology*, *71*(1), 83–93. https://doi.org/10.1037/0022-3514.71.1.83

Brown, R. (1968). *Words and things*. Free Press.

Bucci, W. (1995). The power of the narrative: A multiple code account. In *Emotion, disclosure, & health* (pp. 93–122). American Psychological Association. https://doi.org/10.1037/10182-005

Buchanan, L., Westbury, C., & Burgess, C. (2001). Characterizing semantic space: Neighborhood effects in word recognition. *Psychonomic Bulletin & Review*, *8*(3), 531–544. https://doi.org/10.3758/bf03196189

Campbell, R. S., & Pennebaker, J. W. (2003). The secret life of pronouns: Flexibility in writing style and physical health. *Psychological Science*, *14*(1), 60–65. https://doi.org/10.1111/1467-9280.01419

Carey, A. L., Brucks, M. S., Küfner, A. C. P., Holtzman, N. S., gro\s se Deters, F., Back, M. D., Donnellan, M. B., Pennebaker, J. W., & Mehl, M. R. (2015). *Narcissism and the use of personal pronouns revisited* (Vol. 109). American Psychological Association. https://doi.org/10.1037/pspp0000029

Chung, C. K., & Pennebaker, J. W. (2012). Linguistic Inquiry and Word Count (LIWC): Pronounced "Luke," ... And other useful facts. In P. M. McCarthy & C. Boonthum-Denecke (Eds.), *Applied natural language processing: Identification, investigation and resolution* (pp. 206–229). IGI GLobal. http://doi:10.4018/978-1-60960-741-8.ch012

Chung, C. K., & Pennebaker, J. W. (2018a). Textual analysis. In *Measurement in Social Psychology* (pp. 153–173). Routledge.

Chung, C. K., & Pennebaker, J. W. (2018b). What do we know when we LIWC a person? Text analysis as an assessment tool for traits, personal concerns and life stories. In *The SAGE handbook of personality and individual differences: The science of personality and individual differences* (pp. 341–360). Sage. https://doi.org/10.4135/9781526451163.n16

Chung, C., & Pennebaker, J. W. (2007). The psychological functions of function words. *Social Communication*, *1*, 343–359.

Clarkson, P. M., Ponn, J., Richardson, G. D., Rudzicz, F., Tsang, A., & Wang, J. (2020). A textual analysis of US corporate social responsibility reports. *Abacus*, *56*(1), 3–34. https://doi.org/10.1111/abac.12182

Conway, L. G., Conway, K. R., Gornick, L. J., & Houck, S. C. (2014). Automated integrative complexity. *Political Psychology*, *35*(5), 603–624. https://doi.org/10.1111/pops.12021

Conway, L. G., Conway, K. R., & Houck, S. C. (2020). Validating Automated Integrative Complexity: Natural language processing and the Donald Trump test. *Journal of Social and Political Psychology*, *8*(2), 504–524. https://doi.org/10.5964/jspp.v8i2.1307

Coppersmith, G. (2022). Digital life data in the clinical whitespace. *Current Directions in Psychological Science*, 09637214211068839. https://doi.org/10.1177/09637214211068839

De Choudhury, M., Counts, S., & Horvitz, E. (2013). Predicting postpartum changes in emotion and behavior via social media. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 3267–3276. https://doi.org/10.1145/2470654.2466447

Dean, H. J., & Boyd, R. L. (2020). Deep into that darkness peering: A computational analysis of the role of depression in Edgar Allan Poe's life and death. *Journal of Affective Disorders*, *266*, 482–491. https://doi.org/10.1016/j.jad.2020.01.098

Dehghani, M., & Boyd, R. L. (Eds.). (2022). *Handbook of language analysis in psychology*. The Guilford Press.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186. https://doi.org/10.18653/v1/N19-1423

Dowell, N. M. M., McKay, T. A., & Perrett, G. (2021). It's not that you said it, it's how you said it: Exploring the linguistic mechanisms underlying values affirmation interventions at scale. *AERA Open*, *7*, 23328584211011612. https://doi.org/10.1177/23328584211011611

Drouin, M., Boyd, R. L., Hancock, J. T., & James, A. (2017). Linguistic analysis of chat transcripts from child predator undercover sex stings. *The Journal of Forensic Psychiatry & Psychology*, *28*(4), 437–457. https://doi.org/10.1080/14789949.2017.1291707

Drouin, M., Boyd, R. L., & Romaneli, M. G. (2018). Predicting recidivism among internet child sex sting offenders using psychological language analysis. *Cyberpsychology, Behavior, and Social Networking*, *21*(2), 78–83. https://doi.org/10.1089/cyber.2016.0617

Dzogang, F., Lightman, S., & Cristianini, N. (2018). Diurnal variations of psychometric indicators in Twitter content. *PLOS ONE*, *13*(6), e0197002. https://doi.org/10.1371/journal.pone.0197002

Edwards, T., & Holtzman, N. S. (2017). A meta-analysis of correlations between depression and first person singular pronoun use. *Journal of Research in Personality*, *68*, 63–68. https://doi.org/10.1016/j.jrp.2017.02.005

Eichstaedt, J. C., Kern, M. L., Yaden, D. B., Schwartz, H. A., Giorgi, S., Park, G., Hagan, C. A., Tobolsky, V. A., Smith, L. K., Buffone, A., Iwry, J., Seligman, M. E. P., & Ungar, L. H. (2021). Closed- and open-vocabulary approaches to text analysis: A review, quantitative comparison, and recommendations. *Psychological Methods*, *26*(4), 398–427. https://doi.org/10.1037/met0000349

Eichstaedt, J. C., Schwartz, H. A., Kern, M. L., Park, G., Labarthe, D. R., Merchant, R. M., Jha, S., Agrawal, M., Dziurzynski, L. A., Sap, M., Weeg, C., Larson, E. E., Ungar, L. H., & Seligman, M. E. P. (2015). Psychological language on Twitter predicts county-level heart disease mortality. *Psychological Science*, *26*(2), 159–169. https://doi.org/10.1177/0956797614557867

Eichstaedt, J. C., Smith, R. J., Merchant, R. M., Ungar, L. H., Crutchley, P., Preoţiuc-Pietro, D., Asch, D. A., & Schwartz, H. A. (2018). Facebook language predicts depression in medical records. *Proceedings of the National Academy of Sciences*, 201802331. https://doi.org/10.1073/pnas.1802331115

Entwistle, C., Horn, A. B., Meier, T., & Boyd, R. L. (2021). Dirty laundry: The nature and substance of seeking relationship help from strangers online. *Journal of Social and Personal Relationships*, *38*(12), 3472–3496. https://doi.org/10.1177/02654075211046635

Fitzsimons, G. M., & Kay, A. C. (2004). Language and interpersonal cognition: Causal effects of variations in pronoun usage on perceptions of closeness. *Personality & Social Psychology Bulletin*, *30*(5), 547–557. https://doi.org/10.1177/0146167203262852

Garcia, D., Pellert, M., Lasser, J., & Metzler, H. (2021). Social media emotion macroscopes reflect emotional experiences in society at large. *ArXiv:2107.13236 [Cs]*. http://arxiv.org/abs/2107.13236

Garcia, D., & Rimé, B. (2019). Collective emotions and social resilience in the digital traces after a terrorist attack. *Psychological Science*, *30*(4), 617–628. https://doi.org/10.1177/0956797619831964

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. https://doi.org/10.3758/s13428-017-0875-9

Giles, H. (Ed.). (2012). *The handbook of intergroup communication* (1st edition). Routledge.

Giles, H., & Ogay, T. (2007). Communication Accommodation Theory. In B. B. Whaley & W. Samter (Eds.), *Explaining communication: Contemporary theories and exemplars* (1st Edition, pp. 293–310). Routledge.

Golbeck, J. (2018). Predicting alcoholism recovery from Twitter. In R. Thomson, C. Dancy, A. Hyder, & H. Bisgin (Eds.), *Social, Cultural, and Behavioral Modeling* (pp. 243–252). Springer International Publishing.

Golbeck, J. A. (2016). Predicting personality with social media. *AIS Transactions on Replication Research*, *2*(2), 1–10.

Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers*, *36*(2), 193–202. https://doi.org/10.3758/BF03195564

Graybeal, A., Sexton, J. D., & Pennebaker, J. W. (2002). The role of story-making in disclosure writing: The psychometrics of narrative. *Psychology & Health*, *17*(5), 571–581. https://doi.org/10.1080/08870440290025786

Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, *21*(3), 267–297. https://doi.org/10.1093/pan/mps028

Groom, C. J., & Pennebaker, J. W. (2005). The language of love: Sex, sexual orientation, and language use in online personal advertisements. *Sex Roles*, *52*(7), 447–461. https://doi.org/10.1007/s11199-005-3711-0

Guntuku, S. C., Klinger, E. V., McCalpin, H. J., Ungar, L. H., Asch, D. A., & Merchant, R. M. (2021). Social media language of healthcare super-utilizers. *Npj Digital Medicine*, *4*(1), 1–6. https://doi.org/10.1038/s41746-021-00419-2

Hall, M., & Caton, S. (2017). Am I who I say I am? Unobtrusive self-representation and personality recognition on Facebook. *PLOS ONE*, *12*(9), e0184417. https://doi.org/10.1371/journal.pone.0184417

Hart, R. P., Daughton, S., & LaVally, R. (2017). *Modern rhetorical criticism* (4th ed.). Routledge. https://doi.org/10.4324/9781315203584

Hoemann, K., Wu, R., LoBue, V., Oakes, L. M., Xu, F., & Barrett, L. F. (2020). Developing an understanding of emotion categories: Lessons from objects. *Trends in Cognitive Sciences*, *24*(1), 39–51. https://doi.org/10.1016/j.tics.2019.10.010

Hogenraad, R. (2020). The way of visionaries: Foresight and imagination, computed. *Quality & Quantity*. https://doi.org/10.1007/s11135-020-01071-w

Hogenraad, R., McKenzie, D. P., & Péladeau, N. (2003). Force and influence in content analysis: The production of new social knowledge. *Quality & Quantity: International Journal of Methodology*, *37*(3), 221–238. https://doi.org/10.1023/A:1024401325472

Holtgraves, T. M. (Ed.). (2014). *The Oxford handbook of language and social psychology*. Oxford University Press.

Horn, A. B., & Maercker, A. (2016). I and We- ruminative self-focus and we-ness in couples and wellbeing. *European Health Psychologist*, *18*(S), 594.

Horn, A. B., & Meier, T. (in press). Language in close relationships. In M. Dehghani & R. L. Boyd (Eds.), *The handbook of language analysis in psychology*. Guilford Press.

Ireland, M. E., Schwartz, H. A., Chen, Q., Ungar, L. H., & Albarracín, D. (2015). Future-oriented tweets predict lower county-level HIV prevalence in the United States. *Health Psychology*, *34*(Suppl), 1252–1260. https://doi.org/10.1037/hea0000279

Jordan, K. N., Pennebaker, J. W., & Ehrig, C. (2018). The 2016 U.S. Presidential Candidates and How People Tweeted About Them. *SAGE Open*, *8*(3), 215824401879121. https://doi.org/10/gf5266

Karan, A., Rosenthal, R., & Robbins, M. L. (2019). Meta-analytic evidence that we-talk predicts relationship and personal functioning in romantic couples. *Journal of Social and Personal Relationships*, *36*(9), 2624–2651. https://doi.org/10.1177/0265407518795336

Kennedy, B., Ashokkumar, A., Boyd, R. L., & Dehghani, M. (2022). Text analysis for Psychology: Methods, principles, and practices. In M. Dehghani & R. L. Boyd (Eds.), *The handbook of language analysis in psychology*. Guilford Press. https://doi.org/10.31234/osf.io/h2b8t

Kennedy, B., Atari, M., Mostafazadeh Davani, A., Hoover, J., Omrani, A., Graham, J., & Dehghani, M. (2021). Moral concerns are differentially observable in language. *Cognition*, *212*, 104696. https://doi.org/10/gk495q

Kern, M. L., Park, G., Eichstaedt, J. C., Schwartz, H. A., Sap, M., Smith, L. K., & Ungar, L. H. (2016). Gaining insights from social media language: Methodologies and challenges. *Psychological Methods*, *21*(4), 507–525. https://doi.org/10.1037/met0000091

Lanning, K., Pauletti, R. E., King, L. A., & McAdams, D. P. (2018). Personality development through natural language. *Nature Human Behaviour*, *2*(5), 327–334. https://doi.org/10.1038/s41562-018-0329-0

Lanning, K., Wetherell, G., Warfel, E. A., & Boyd, R. L. (2021). Changing channels? A comparison of Fox and MSNBC in 2012, 2016, and 2020. *Analyses of Social Issues and Public Policy*. https://doi.org/10.1111/asap.12265

Lin, Y., Yu, R., & Dowell, N. (2020). LIWCs the same, not the same: Gendered linguistic signals of performance and experience in online STEM courses. In I. I. Bittencourt, M. Cukurova, K. Muldner, R. Luckin, & E. Millán (Eds.), *Artificial Intelligence in Education* (pp. 333–345). Springer International Publishing. https://doi.org/10.1007/978-3-030-52237-7_27

Loveys, K., Torrez, J., Fine, A., Moriarty, G., & Coppersmith, G. (2018). Cross-cultural differences in language markers of depression online. *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, 78–87. https://doi.org/10.18653/v1/W18-0608

Markowitz, D. M. (2021). The meaning extraction method: An approach to evaluate content patterns from large-scale language data. *Frontiers in Communication*, *6*. https://doi.org/10.3389/fcomm.2021.588823

Markowitz, D. M. (2022). Psychological trauma and emotional upheaval as revealed in academic writing: The case of COVID-19. *Cognition and Emotion*, *36*(1), 9–22. https://doi.org/10.1080/02699931.2021.2022602

Markowitz, D. M., & Shulman, H. C. (2021). The predictive utility of word familiarity for online engagements and funding. *Proceedings of the National Academy of Sciences*, *118*(18). https://doi.org/10.1073/pnas.2026045118

Mehl, M. R. (2006). Quantitative text analysis. In M. Eid & E. Diener (Eds.), *Handbook of multimethod measurement in psychology* (pp. 141–156). American Psychological Association. https://doi.org/10.1037/11383-011

Mehl, M. R., Raison, C. L., Pace, T. W. W., Arevalo, J. M. G., & Cole, S. W. (2017). Natural language indicators of differential gene regulation in the human immune system. *Proceedings of the National Academy of Sciences*, *114*(47), 12554–12559. https://doi.org/10.1073/pnas.1707373114

Mehl, M. R., Robbins, M. L., & Holleran, S. E. (2013). How taking a word for a word can be problematic: Context-dependent linguistic markers of extraversion and neuroticism. *Journal of Methods and Measurement in the Social Sciences*, *3*(2), 30–50. https://doi.org/10.2458/v3i2.16477

Meier, T., Boyd, R. L., Mehl, M. R., Milek, A., Pennebaker, J. W., Martin, M., Wolf, M., & Horn, A. B. (2020). Stereotyping in the digital age: Male language is "ingenious", female language is "beautiful" – and popular. *PLOS ONE*, *15*(12), e0243637. https://doi.org/10.1371/journal.pone.0243637

Mergenthaler, E. (1996). Emotion-abstraction patterns in verbatim protocols: A new way of describing psychotherapeutic processes. *Journal of Consulting and Clinical Psychology, 64*(6), 1306–1315. https://doi.org/10.1037//0022-006x.64.6.1306

Mergenthaler, E., & Bucci, W. (1999). Linking verbal and non-verbal representations: Computer analysis of referential activity. *British Journal of Medical Psychology*, *72*(3), 339–354. https://doi.org/10.1348/000711299160040

Metzler, H., Rimé, B., Pellert, M., Niederkrotenthaler, T., Natale, A. D., & Garcia, D. (2021). *Collective emotions during the COVID-19 outbreak*. PsyArXiv. https://doi.org/10.31234/osf.io/qejxv

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, & K. Q. Weinberger (Eds.), *Advances in Neural Information Processing Systems 26* (pp. 3111–3119). Curran Associates. http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf

Miller, G. A. (1996). *The science of words*. Scientific American Library. https://books.google.com/books?id=5SxvQgAACAAJ

Newman, M. L., Groom, C. J., Handelman, L. D., & Pennebaker, J. W. (2008). Gender differences in language use: An analysis of 14,000 text samples. *Discourse Processes*, *45*(3), 211–236. https://doi.org/10.1080/01638530802073712

Niculae, V., Kumar, S., Boyd-Graber, J., & Danescu-Niculescu-Mizil, C. (2015). Linguistic harbingers of betrayal: A case study on an online strategy game. *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 1650–1659. http://www.aclweb.org/anthology/P15-1159

Obschonka, M., & Fisch, C. (2018). Entrepreneurial personalities in political leadership. *Small Business Economics*, *50*(4), 851–869. https://doi.org/10.1007/s11187-017-9901-7

Orvell, A., Kross, E., & Gelman, S. A. (2017). How "you" makes meaning. *Science*, *355*(6331), 1299–1302. https://doi.org/10.1126/science.aaj2014

Pellert, M., Metzler, H., Matzenberger, M., & Garcia, D. (2021). Validating daily social media macroscopes of emotions. *ArXiv:2108.07646 [Cs]*. http://arxiv.org/abs/2108.07646

Pennebaker, J. W. (1997). Writing about emotional experiences as a therapeutic process. *Psychological Science*, *8*(3), 162–166. https://doi.org/10.1111/j.1467-9280.1997.tb00403.x

Pennebaker, J. W., & Chung, C. K. (2013). Counting little words in big data: The psychology of individuals, communities, culture, and history. In *Social Cognition and Communication* (pp. 25–42). Taylor and Francis. https://doi.org/10.4324/9780203744628

Pennebaker, J. W., & Ireland, M. E. (2011). Using literature to understand authors: The case for computerized text analysis. *Scientific Study of Literature*, *1*(1), 34–48. https://doi.org/10.1075/ssol.1.1.04pen

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

Pennebaker, J. W., Mehl, M. R., & Niederhoffer, K. G. (2003). Psychological aspects of natural language use: Our words, our selves. *Annual Review of Psychology*, *54*(1), 547–577. https://doi.org/10.1146/annurev.psych.54.101601.145041

Pennebaker, J. W., & Stone, L. D. (2003). Words of wisdom: Language use over the life span. *Journal of Personality and Social Psychology*, *85*(2), 291–301. https://doi.org/10.1037/0022-3514.85.2.291

Pérez-Rosas, V., Wu, X., Resnicow, K., & Mihalcea, R. (2019). What makes a good counselor? Learning to distinguish between high-quality and low-quality counseling conversations. *Proceedings of the 57th Conference of the Association for Computational Linguistics*, 926–935. https://www.aclweb.org/anthology/P19-1088

Raffaelli, Q., Mills, C., de Stefano, N.-A., Mehl, M. R., Chambers, K., Fitzgerald, S. A., Wilcox, R., Christoff, K., Andrews, E. S., Grilli, M. D., O'Connor, M.-F., & Andrews-Hanna, J. R. (2021). The think aloud paradigm reveals differences in the content, dynamics and conceptual scope of resting state thought in trait brooding. *Scientific Reports*, *11*(1), 19362. https://doi.org/10.1038/s41598-021-98138-x

Ramírez-Esparza, N., Chung, C. K., Kacewicz, E., & Pennebaker, J. W. (2008). The psychology of word use in depression forums in English and in Spanish: Testing two text analytic approaches. *Proceedings of the Second International Conference on Weblogs and Social Media*, 102–108.

Robbins, M. L., Mehl, M. R., Smith, H. L., & Weihs, K. L. (2013). Linguistic indicators of patient, couple, and family adjustment following breast cancer. *Psycho-Oncology*, *22*(7), 1501–1508. https://doi.org/10.1002/pon.3161

Robinson, R. L., Navea, R., & Ickes, W. (2013). Predicting final course performance from students' written self-introductions. *Journal of Language and Social Psychology*, *32*(4), 469–479. https://doi.org/10.1177/0261927X13476869

Saha, K., Yousuf, A., Boyd, R. L., Pennebaker, J. W., & De Choudhury, M. (2022). Social media discussions predict mental health consultations on college campuses. *Scientific Reports*, *12*(1), 123. https://doi.org/10.1038/s41598-021-03423-4

Sbarra, D. A., Smith, H. L., & Mehl, M. R. (2012). When leaving your ex, love yourself: Observational ratings of self-compassion predict the course of emotional recovery following marital separation. *Psychological Science*, *23*(3), 261–269. https://doi.org/10.1177/0956797611429466

Scheffer, M., Leemput, I. van de, Weinans, E., & Bollen, J. (2021). The rise and fall of rationality in language. *Proceedings of the National Academy of Sciences*, *118*(51). https://doi.org/10.1073/pnas.2107848118

Schoonvelde, M., Schumacher, G., & Bakker, B. N. (2019). Friends with text as data benefits: Assessing and extending the use of automated text analysis in political science and political psychology. *Journal of Social and Political Psychology*, *7*(1), 124-143–143. https://doi.org/10.5964/jspp.v7i1.964

Schultheiss, O. (2013). Are implicit motives revealed in mere words? Testing the marker-word hypothesis with computer-based text analysis. *Frontiers in Psychology*, *4*, 748. https://doi.org/10.3389/fpsyg.2013.00748

Semin, G. R., & Fiedler, K. (1988). The cognitive functions of linguistic categories in describing persons: Social cognition and language. *Journal of Personality and Social Psychology*, *54*(4), 558–568. https://doi.org/10.1037/0022-3514.54.4.558

Seraj, S., Blackburn, K. G., & Pennebaker, J. W. (2021). Language left behind on social media exposes the emotional and cognitive costs of a romantic breakup. *Proceedings of the National Academy of Sciences*, *118*(7). https://doi.org/10.1073/pnas.2017154118

Simchon, A., Guntuku, S. C., Simhon, R., Ungar, L. H., Hassin, R. R., & Gilead, M. (2020). Political depression? A big-data, multimethod investigation of Americans' emotional response to the Trump presidency. *Journal of Experimental Psychology. General*, *149*(11), 2154–2168. https://doi.org/10.1037/xge0000767

Slatcher, R. B., Chung, C. K., Pennebaker, J. W., & Stone, L. D. (2007). Winning words: Individual differences in linguistic style among U.S. presidential and vice presidential candidates. *Journal of Research in Personality*, *41*(1), 63–75. http://dx.doi.org/10.1016/j.jrp.2006.01.006

Srivastava, S. B., Goldberg, A., Manian, V. G., & Potts, C. (2018). Enculturation trajectories: Language, cultural adaptation, and individual outcomes in organizations. *Management Science*, *64*(3), 1348–1364. https://doi.org/10.1287/mnsc.2016.2671

Stark, L. (2018). Algorithmic psychometrics and the scalable subject. *Social Studies of Science*, *48*(2), 204–231. https://doi.org/10.1177/0306312718772094

Sterling, J., Jost, J. T., & Bonneau, R. (2020). Political psycholinguistics: A comprehensive analysis of the language habits of liberal and conservative social media users. *Journal of Personality and Social Psychology*, *118*(4), 805–834. https://doi.org/10.1037/pspp0000275

Stirman, S. W., & Pennebaker, J. W. (2001). Word use in the poetry of suicidal and nonsuicidal poets. *Psychosomatic Medicine*, *63*(4), 517–522. https://doi.org/10.1097/00006842-200107000-00001

Ta, V. P., Boyd, R. L., Seraj, S., Keller, A., Griffith, C., Loggarakis, A., & Medema, L. (2021). An inclusive, real-world investigation of persuasion in language and verbal behavior. *Journal of Computational Social Science*. https://doi.org/10.1007/s42001-021-00153-5

Van Der Zee, S., Poppe, R., Havrileck, A., & Baillon, A. (2022). A personal model of Trumpery: Linguistic deception detection in a real-world high-stakes setting. *Psychological Science*, *33*(1), 3–17. https://doi.org/10.1177/09567976211015941

Wilson, S. R., Shen, Y., & Mihalcea, R. (2018). Building and validating hierarchical lexicons with a case study on personal values. In S. Staab, O. Koltsova, & D. I. Ignatov (Eds.), *Social Informatics* (pp. 455–470). Springer International Publishing.

Winter, D. G., & McClelland, D. C. (1978). Thematic analysis: An empirically derived measure of the effects of liberal arts education. *Journal of Educational Psychology*, *70*(1), 8–16. https://doi.org/10.1037/0022-0663.70.1.8

Wojcik, S. P., Hovasapian, A., Graham, J., Motyl, M., & Ditto, P. H. (2015). Conservatives report, but liberals display, greater happiness. *Science*, *347*(6227), 1243–1246. https://doi.org/10.1126/science.1260817

Zhang, J., Chang, J., Danescu-Niculescu-Mizil, C., Dixon, L., Hua, Y., Taraborelli, D., & Thain, N. (2018). Conversations gone awry: Detecting early signs of conversational failure. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1350–1361. https://doi.org/10.18653/v1/P18-1125

Zhang, J., Pennebaker, J., Dumais, S., & Horvitz, E. (2020). Configuring audiences: A case study of email communication. *Proceedings of the ACM on Human-Computer Interaction*, *4*(CSCW1), 1–26. https://doi.org/10.1145/3392871

Changelog

2022-05-16
Included note about summary measure norms. Addition of minor details around new affective category construction.

2022-04-18
Fixed grammatical typo on Page 4.