**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](#_bookmark0) 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 | 586 (510) |
| 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** |  | **2128 (2778)** |

**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.

LIWC-22 Development Manual Page 11

# 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 of 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 |

LIWC-22 Development Manual Page 13

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).