Software and this manual are available through,
Erlbaum Publishers, Mahwah, NJ
(for more info, contact www.erlbaum.com)

Linguistic Inquiry and Word Count (LIWC)

James W. Pennebaker and Martha E. Francis

1999

The University of Texas at Austin

Contents

Getting Started

The LIWC program comes with the following files:

- LIWC.EXE — the actual program file
- DIC.AID — the master dictionary
- SETUP.TXT — a readable file with instructions for setup on older machines
- INSTALL.BAT — a simple DOS program for setting up LIWC on your C drive
- \SAMPLES — a subdirectory of sample text files, including
  - 3 inauguration speeches by Lincoln, Franklin Roosevelt, and Clinton (Lincoln.txt, FDR.txt, Clinton.txt)
  - 2 poems by Sylvia Plath and Anne Sexton (Plath.txt, Sexton.txt)
  - 2 talk show segments: Howard Stern (radio), Donna Shelala (TV) (Radio.txt, talkshow.txt)
  - 2 files of a passage from Huckleberry Finn -- one original, one "cleaned" (Huckraw.txt, Huckcln.txt)
  - 2 psychology journal abstracts (Abstr1.txt, Abstr2.txt)

Installing LIWC on Your PC

If you are using Windows95 or more recent version, follow these instructions:

➔ Insert the LIWC disk into Drive A

➔ From the Start button, choose INSTALL and type A:INSTALL and press ENTER. A directory, named WLIWC, will be created with the required files in it.

You are ready for action.

Running LIWC

To run the program, go to the WLIWC directory (which can be found on the C: drive via the My Computer icon on the desktop). Double click on the LIWC icon or LIWC.EXE file. Once the LIWC program starts, explore the various options.

To analyze whatever text files you specify, go into the FILE menu and select PROCESS. To find the input files you want to run, press the BROWSE button. Assuming you want to run one or more files in the directory supplied with this program called \SAMPLE, go into that directory and select the file to be analyzed. Note that only files ending in ".TXT" will initially be listed unless you change the "List files of type" box to "All Files (*.*)". Once you find the file you want to run, select it and press OK.

IMPORTANT: You can analyze as many files as you like from the same directory at once. To do so, you must select one file and then change the input file-name in the Files to Process box according to DOS conventions. Assume, for example, you have a series of files named R01.a, R01.b, R02.a, etc. If you have clicked on "R01.a" to be processed, change it to

"R??.?" (or *.?) to analyze all files beginning with R followed by two characters and with one character after the dot.

Specify an output file name that will **not** be included in your input file. For example, if you specify the files to be analyzed as "*.*" (which would indicate all possible files in the directory), any output file that was named would, by definition be analyzed as well. The philosophical implications of the computer's analyzing the output as input are too existential for even the fastest machines to ponder without crashing.

Finally, the LIWC output file automatically goes in the directory of the input files. Trying to force the output to another directory will result in an error. Note that you do not have to specify the complete path for the output file -- only the file name.

Reading and Analyzing LIWC Output

By default, all 74 LIWC output variables are listed consecutively in the output file. The output file is saved in tab-delimited text that includes the variable names on the first line. This allows it to be read directly into SPSS or Excel programs.

To view the output file, choose the Open command within the File Menu (or click on the open folder icon) and specify the output file-name. Alternatively, the output file can be opened with any word processing program (e.g., Word, Word Perfect). For the best view of the output file, however, a spreadsheet program, such as Excel or SPSS, is recommended.

Customizing LIWC Output

In some cases, the LIWC user may prefer to analyze only a subset of language dimensions rather than the full 74 variables. To do this, open the Preferences menu. Within each option (e.g., statistical information, standard linguistic categories), check boxes are available for each LIWC dimension. By clicking on each dimension and removing the check mark, the output category can be omitted from the analyses. Note that the category preferences will remain in effect until they are re-checked.

Creating and Using Custom Dictionaries

If you are planning to create your own user-defined categories for analysis, it is best to make a copy of the file DIC.AID (the master dictionary) and give it another file name, such as DICNEW.AID. You can create up to 10 user-defined categories. To do this, first examine the structure of the master dictionary file. Note that the words within it are in alphabetical order and are followed by a series of numbers, ranging from 01 to 68.

Each number refers to the category to which each word is assigned. Hence, the word "a" is associated with category 9 (articles), the word "abandon" is associated with categories 12 (affect), 16 (negative emotion), 19 (sad), 20 (cognitive mechanism), and 24 (inhibition). The highest category number in the DIC.AID file is 68 which refers to filler words and phrases.

To create your own categories, you must use category numbers 69, 70, 71, ... up to 78. If, for example, you wanted to create a category of common fruits (apples, bananas), you would insert all of the fruit words you wanted to count and put the number 69 adjacent to them. If you then added a category of vegetables, you would assign the next available number, in this case 70.

To name the new categories for the LIWC output, click the Define Categories entry in the Options menu. Simply replace the "not defined" with the category name you prefer. Note that the name should be less than 8 letters and should not overlap with any of the existing category names.

Finally, to run the custom dictionary, specify its filename and location by choosing "Select custom dictionary" in the Options menu. Note that once the new dictionary has been named, it will remain the default dictionary.

Before you begin creating multiple new categories, a few caveats are in order:

➔ It is not recommended that you change any of the existing categories. The ability to compare results from study to study and sample to sample depends on the use of the same words that to up the general categories.

➔ Words can be in multiple categories. If you considered yourself to be a free-thinking vegetarian, you might put the word "tomato" into both the fruit and vegetable categories.

➔ In the dictionary, use of an asterisk (*) at the end of the word signals the computer to ignore all subsequent letters. Hence, "plum*" will count the words plum, plums, plumbing, plump, etc.

➔ If new categories are added, it is highly recommended that two or more judges agree about the assignment of the words to the particular categories. For a description of the methods for getting reliable content categories, see the section "The Development and Psychometric Properties of LIWC."

Preparing Written Text For LIWC Analysis

The accuracy of LIWC output data is determined by the quality of the text files that are analyzed. In order to insure best results, it is necessary to properly prepare text essays for LIWC analysis. The essential steps for essay text organization, entry, and editing are as follows:

1) **Text file organization**. Each language sample should be put in its own file and named in a systematic and meaningful way. For example, data from a study with two conditions and three days of writing might be saved in files using this naming strategy:

[PARTICIPANT#][DAY#].[CONDITION] -- 45681.E, 45682.E, and 45683.E

2) **Text file computer entry**. Essays should be entered into the computer using a standard word processing package (e.g., Word for Windows, WordPerfect) that allows conversion into an ASCII or text file.

3) **Cleaning the text files**. Each text file to be analyzed should be examined and adjusted for misspellings and inappropriate word use (e.g., "its" rather than "it's"). It is always wise to run all files through standard spell-check programs. Because LIWC only examines words, grammar, capitalization, and sentence structure do not need to be corrected.

Naming Text Files

Because the file names are part of the output file, certain conventions should be adopted in the preparation of the files and file names:

1. Separate files for separate text samples. LIWC analyzes data one file at a time. If participants write responses to two questions or perhaps write on two separate days, each question or day should be a separate file. If responses to both questions (or both days' writing) are within the same file, LIWC will analyze them as a single writing sample.

2. The file name should be descriptive, including ID number, condition, and question or day number. Filenames should not be longer than 8 letters and/or digits. You may also include up to three numbers/letters to the right of the period in the filename in place of another file extension. Examples: S456F001.DAT, 6011, R12.3

3. Files must be in TEXT or ASCII format. LIWC cannot read WordPerfect, Word, or other word processing files. Note that virtually all word processing programs allow you to convert your files into ASCII or TEXT format.

Typing Conventions: Writing and Interview Samples

In making corrections or cleaning text files, keep in mind what your goals are in analyzing the data. LIWC does not discriminate between upper- and lower-case letters. It can only count words that are in its dictionaries. Misspellings, colloquialisms, foreign words, and abbreviations are usually not in the dictionaries. The following items should be checked before any files are analyzed:

1. Spelling, abbreviations, contractions.

Correct all spelling errors. Use standard United States spelling.

Meaningful abbreviations should be spelled out. "Jan" should be January. More obscure abbreviations or acronyms, such as "AT&T", can remain as such unless you have reason to want the term to be expanded and counted as four separate words: "American Telephone and Telegraph".

Common verb contractions are in the dictionary and do not need to be changed. These include: don't, won't, isn't, shouldn't, can't, couldn't, I'm, I'll, I'd, we're, we'd, you're, he's, it's, etc. Most others will be simply counted as possessive nouns: "Sally's shoes" will be counted the same way as "Sally's going to the store." In the second case, change "Sally's" to "Sally is."

2. End of sentence markers and hyphens.

Two of the 74 output variables are based on end-of-sentence markers: words per sentence (WPS) and percentage of sentences ending in question marks (Qmarks). All periods (.), question marks, and exclamation points are counted as end of sentence marks. Common abbreviations (such as Dr., Ms., U.S.A., D.O.A.) will be counted as multiple sentences unless the periods are removed. Be careful that the removal of the periods doesn't make a new word. For example, the United States, or U.S., becomes "US" (1st person plural pronoun) when the periods are removed. In this case, change it to "USA".

Time markers (e.g., 6 a.m. or 7:30 p.m.) can also be a problem. Because "a.m." without the periods is a verb, "am", change time to 6am or 7:30pm.

When words start or end with hyphens, they are read by LIWC as part of the word. LIWC, for example, lists "self-esteem" as a meaningful word in one of its dictionaries. In cases of hyphenated phrases such as "this-or-that" LIWC will search for a single word and won't find it. To correct, change "this-or-that" to "this - or - that".

Watch out for hyphens between phrases, as in "we went to the store--I don't know why." LIWC will think that "store--I" is one word. Insert blanks on either side of the hyphens so that both words will be counted.

Note that the program ignores all other punctuation marks: commas, colons, semi-colons, quotation marks, asterisks, pound signs, as well as @, #, $, %, ^, &, (,), +, =, etc.

3. Other common problems:

Typed entry	**Change to:**
w/	with
b/	between
&	and
‘cause	because
gotta	got to
lotta	lot of
and/or	and - or
‘an or ‘n	and
mos	months
sec	second
@	at

Transcribing Oral Transcripts: Special Problems

Although not designed for spoken language, we have found LIWC to be useful in analyzing conversations and interviews. To accommodate certain dimensions of spoken language, we have adopted the following conventions:

1. Nonfluencies

Hm, hmm, uh, uhh, uhm, um, umm, and er are part of the nonfluency dictionary. Other forms will not be caught (e.g., ooooh should be changed to um if used as a nonfluency).

Stuttering can be accommodated by altering the stuttering part of a phrase to a nonfluency marker. For example, “The, the bo-, the boat went into the water” could be changed to “Uh, the boat went into the water.” The transcriber will have to decide how many uh’s would be appropriate.

Uh-uh and uh-huh should be changed to “no” and “yes”. Huh? should be changed to “what?” Or, if you are very, very proper, to “Excuse me madam, I didn’t quite catch what you said.”

2. Fillers.

Everyday speech is littered with “meaningless” fillers. Unfortunately, these fillers use some of the most important words in our dictionaries. Watch out for the following:

You know. As in, “we went, you know, to the store and, you know, bought gum.” Change to one word: youknow. “We went, youknow, to the store...”

I mean. As in, “we went, I mean, to the store...” Change to one word: Imean.

I don’t know. As in, “we went, I don’t know, to the store...” Change to: Idontknow.

Like. “We went, like, to like the store and like we like bought like gum.” Be careful with like because sometimes it is used appropriately. As a nonfluency, change it to: rrlike. Note that all words starting with “rr” will be coded as a nonfluency. Hence, if you are transcribing audiotapes made in the 1950’s, the word “well” would likely be used the way “like” is today. Hence, you would enter it as “rrwell.”

3. Transcribers’ comments.

LIWC is designed only for spoken language. Transcribers often insert remarks, such as [subject laughs], [shaky voice], [whispers]. We recommend removing these.

Occasionally, the transcriber cannot understand a word or passage. Rather than writing [can't understand word] or [?], the transcriber should put a nonsense word, such as "xxxx" in its place. LIWC will count the xxxx as a spoken word but not assign it to a dictionary. For entire passages, don't insert anything.

Common Problems in Running LIWC

Symptom	Likely problem	How to fix it
Floating point error: invalid	You have specified all files *.* in your input statement, which caused your output file to be run as a text file	Don't use *.* for your input files
Floating point error: invalid	One or more of the files you are analyzing is empty, contains only one word, or is in a format other than text	Don't try to analyze empty files. Either remove the files, or put something in them
Floating point error: invalid	You are specifying your output file in a different directory from your input files	Your output file must be in the same directory as your input files
Done! And the screen is frozen	You may not have specified an output file	Hit control-alt-del once and exit the program. Start over and specify an output file
Dictionary not found! And the screen is frozen	You specified an alternative dictionary that LIWC couldn't find	In the Select Custom Dictionary menu, enter the correct new path/name of your new dictionary, for example: C:\wliwc\newdic.aid
The HELP menu doesn't work	The HELP facility has not been implemented in this LIWC version	Consult this document for assistance

Technical Support

Technical support for set-up and hardware/software compatibility can be obtained by contacting the publisher, LEA Software and Alternative Media (201-236-9500 ext. 122). Further assistance is available from the first author, James W. Pennebaker, Department of Psychology, The University of Texas, Austin, Texas 78712 (Pennebaker@psy.utexas.edu). More extended consultation is available on a fee basis.

Getting Some Practice: Running the Samples

Included with the LIWC program is a subdirectory called SAMPLES. It is composed of 11 text files of varying lengths. These include:

Inaugural addresses of Lincoln, Franklin Roosevelt, and Clinton at the beginning of the first term of office:

LINCOLN.TXT
FDR.TXT
CLINTON.TXT

Two poems from Anne Sexton and Sylvia Plath:

SEXTON.TXT
PLATH.TXT

Two rather dry abstracts from esoteric social psychology journals by esteemed social psychologists:

ABSTR1.TXT
ABSTR2.TXT

Two transcripts from the media -- one from the Howard Stern Show; the other from a morning program interview with Donna Shelala:

RADIO.TXT
TALKSHOW.TXT

A passage from Mark Twain's Huckleberry Finn which is presented in its original, unedited form as well as in a form translated into "proper" American English. The purpose of these two forms is to give the researcher a sense of how extensive editing can change the output (not as much as you might think):

HUCKRAW.TXT
HUCKCLN.TXT

This group of files is intended to give the LIWC user a sense of the diversity of text samples that can be analyzed and the similarities and differences among them. To appreciate that nature of the samples, simply open any of them in WordPad, Word, WordPerfect, or even the LIWC "open file" menu.

Here is a step-by-step procedure for LIWCing the 11 files:

1. Start the LIWC program by clicking on LIWC.EXE or its icon.
2. Within the LIWC program, go into the FILE menu and press PROCESS.
3. Within the Files to Process box, press the BROWSE button adjacent to the Input Files line. Find the "samples" folder and double click on it.
4. Click on one of the files, such as ABSTR1.TXT and then press OK.
5. In the Files to Process box, replace ABSTR1.TXT with *.TXT (this tells the program to analyze all files within this folder that end with .TXT).
6. Specify an output file name, such as Samples.dat.
7. Press OK. Voila!

To see if the data were analyzed properly, use the LIWC open file option to open "SAMPLES.DAT" file. Beautiful, isn't it? You can scroll the file to the right and see that all 74 variables are there as are the file names.

To see the data more completely, however, use either Excel or SPSS to open SAMPLES.DAT. If you use SPSS, open the file as a tab-delimited file and be sure to check the box "Read variable names." The first part of the output file should look something like this in Excel:

FILENAME	WC	WPS	Qmarks	Unique	Dic	SixItr	Pronoun	I	We	Self
LINCOLN.TXT	3624	27.45	16.7	27.7	69	23.6	6.5	1.6	0.6	2.2
FDR1932.TXT	1875	22.06	0	37.8	70.2	23.4	8.3	1.7	3.4	5

CLINTN92.TXT	1580	16.99	0	37.7	71.3	21	11.2	0.9	7.8	8.7
SEXTON.TXT	237	13.17	16.7	43.9	85.2	12.2	19.8	13.5	0	13.5
PLATH.TXT	98	32.67	0	78.6	58.2	25.5	5.1	0	0	0
ABSTR1.TXT	102	17	0	62.7	65.7	47.1	0	0	0	0
ABSTR2.TXT	185	23.13	0	53.5	54.1	38.4	0	0	0	0
RADIO.TXT	272	5.44	18	51.5	83.1	7.7	16.9	7	1.1	8.1
TALKSHOW.TXT	620	24.8	0	39	74	18.1	9.2	0.6	2.4	3.1
HUCKRAW.TXT	655	14.24	23.9	45.3	66.3	8.5	13.9	2.7	1.1	3.8
HUCKCLN.TXT	608	14.48	23.8	44.1	74.3	8.7	14.8	2.5	1.2	3.6

OK, your file doesn't look exactly like this. And yes, there are another 65 variables in your output file. However, even this small sample of verbal material yields some intriguing findings.

Important: All variables (except raw word count [WC], words per sentence [WPS], and percentage of sentences ending in question marks [Qmarks]) reflect percentage of total words. So, for example, 1.6% of Lincoln's inaugural address was comprised of 1st person singular "I" words (I, me, my) compared to 7% of the speech sample from Howard Stern. Clinton, more than any president, used a tremendously high rate of 1st person plural words (e.g., we us) in his speech (7.8%). Natural spoken text generally has a lower percentage of long words (i.e., words greater than six letters [sixltrs]) than formal text. Other striking differences (e.g., use of emotion words) can be seen in the actual LIWC analysis.

By looking at this table, it is easy to see how language use can differ from person to person and from context to context. Obviously, when attempting to get a reliable picture of language use within a given person or situation, the more and lengthier the text samples, the better.

The Development and Psychometric Properties of LIWC

The ways that individuals talk and write provide windows into their emotional and cognitive worlds. Over the last three decades, researchers have provided evidence to suggest that people's physical and mental health can be predicted by the words they use (Gottschalk & Glaser, 1969; Rosenberg & Tucker, 1978; Stiles, 1992). More recently, a large number of studies have found that having individuals write or talk about deeply emotional experiences is associated with improvements in mental and physical health (e.g., Pennebaker, 1997; Smyth, 1997). Text analyses based on these studies indicate that those individuals who benefit the most from writing tend to use relatively high rates of positive emotion words, a moderate number of negative emotion words, and, most importantly, an increasing number of cognitive or thinking words from the first to last days of writing (e.g., Pennebaker & Francis, 1996; Pennebaker, Mayne, & Francis, 1997).

In order to provide an efficient and effective method for studying the various emotional, cognitive, structural, and process components present in individuals' verbal and written speech samples, we developed a text analysis program called Linguistic Inquiry and Word Count, or LIWC. The first LIWC program was developed as part of an exploratory study of language and disclosure (Francis, 1993; Pennebaker, 1993). As described below, the second version of LIWC is an updated revision of the original program. Both LIWC programs have been written in Turbo

C++ (a Borland software product), which can operate on PCs. The current LIWC is best suited for Windows-based platforms. LIWC programs are designed to analyze written text on a word by word basis, calculate the percentage words in the text that match each of up to 82 language dimensions, and generate output as a tab-delimited text file that can be directly read into application programs, such as SPSS for Windows, Excel, etc.

The LIWC Framework

The LIWC program is made up of two basic files. The first, named "LIWC.EXE," is the text processing program that reads the files to be analyzed, calculates the text dimensions, and creates the output file. The second, named "DIC.AID," is the master dictionary file that defines which words should be counted in the target text files. Note that the LIWC.EXE file is an executable file and cannot be read or opened. The master dictionary file, however, is in text format and can be read and modified to suit the user's needs.

To avoid confusion in the subsequent discussion, text words that are read and analyzed by the LIWC program are referred to as ***target words***. Words in the LIWC 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.

The LIWC Main Text Processing Module

LIWC is designed to accept written or transcribed verbal text which has been stored as a text or ASCII file using any of the popular PC word processing software packages (e.g., WordPerfect, Word for Windows). LIWC accesses a single file or group of files and analyses each sequentially, writing the output to a single file. Processing time for a page of single-spaced text is typically 0.5 to 10 seconds depending on the computer. LIWC reads each designated text file, one target word at a time. As each target word is processed, the LIWC dictionary file is searched, looking for a dictionary match with the current target word. If the target word matches the dictionary word, the appropriate word category scale for that word is incremented. As the target text file is being processed, counts for various structural composition elements (e.g., word count and sentence punctuation) are also incremented.

With each text file, up to 84 output variables are written as one line of data to a designated output file. This data record includes the file name, 17 standard linguistic dimensions (e.g., word count, percentage of pronouns, articles), 25 word categories tapping psychological constructs (e.g., affect, cognition), 10 dimensions related to "relativity" (time, space, motion), and 19 personal concern categories (e.g., work, home, leisure activities). The program also has the flexibility to analyze up to 10 categories defined by the user. A complete list of the standard LIWC scales is included in Table 1.

The LIWC Dictionary

The LIWC Dictionary is the heart of the text analysis strategy. The LIWC Dictionary is composed of 2,290 words and word stems. Each word or word stem defines one or more word categories or subdictionaries. For example, the word "cried" is part of four word categories: sadness, negative emotion, overall affect, and a past tense verb. Hence, if it is found in the target text, each of these four subdictionary scale scores will be incremented. As in this example, many of the LIWC categories are arranged hierarchically. All anger words, by definition, will be categorized as negative emotion and overall emotion words. Note too that word stems can be captured by the LIWC system. For example, the LIWC Dictionary includes the stem "hungr*" which allows for any target word that matches the first five letters to be counted as an eating word (this would include hungry, hungrier, hungriest). The asterisk, then, denotes the acceptance of all letters, hyphens, or numbers following its appearance.

Each of the 74 preset LIWC categories is composed of a list of dictionary words that define that scale. Table 1 provides a comprehensive list of the LIWC dictionary categories, scales, sample scale words, and relevant scale word counts.

LIWC Dictionary Development. The selection of words defining the LIWC categories involved multiple steps over several years. The initial idea was to identify a group of words that tapped basic emotional and cognitive dimensions often studied in social, health, and personality psychology. With time, the domain of word categories expanded considerably.

In the design and development of the LIWC category scales, sets of words were first generated for each category scale. Within the Psychological Processes category, for example, the emotion or affective subdictionaries were based on words from several sources. We drew on common emotion rating scales, such as the PANAS (Watson, Clark, & Tellegen, 1988), Roget's Thesaurus, and standard English dictionaries. Following the creation of preliminary category word lists, brain-storming sessions among 3-6 judges were held in which words relevant to the various scales were generated and added to the initial scale lists. Similar schemes were used for the other subjective dictionary categories.

Once the broad word lists were amassed, those words in the Psychological Processes and Personal Concerns and most in the Relativity (excluding verb tense) categories were then rated by three independent judges. In this phase of development, the judges were instructed to focus on both the inclusion and exclusion of words in each LIWC Dictionary scale list. First, the judges indicated whether each word in the scale list should or should not be included on the particular scale in question. Second, they were instructed to include additional words they felt should be included in the scale. After the completion of the first judging phase, all category scale word lists were updated by the following set of rules: 1) a word remained on the scale list if two out of three judges agreed, 2) a word was deleted from the scale list if at least two of the three judges agreed it should be excluded, and 3) a word was added to the scale list if two out of three judges agreed. Due to the objective nature of elements in the Standard Language Dimensions category (e.g., articles, pronouns, prepositions), judges' ratings were not collected for the various scale lists in that category.

The second rating phase involved the discrimination of LIWC category word elements. Judges were given category level alphabetized word lists (e.g., all Cognitive Process words) and asked first to indicate whether each word in the list should or should not be included in the high-level category in question. Second, judges were instructed to indicate in which, if any, of the mid-level scale lists the word should be included (e.g., Insight, Causation). Percentages of agreement for judges' ratings were acceptable for all LIWC Category and scale lists (ranging from a low of 86% agreement for Optimism to 100% agreement for Relatives).

After completion of the second judging phase, all category scale word lists were updated by the following rules: 1) a word remained on the scale list if two out of three judges agreed and 2) a word was deleted from the scale list if at least two of the three judges agreed. The final percentages of judges' agreement for this second pass ranged from 93% agreement for Insight to 100% agreement for Eating, Metaphysical, Friends, Relatives, and Humans.

The initial LIWC judging took place in 1992-1994. A significant LIWC revision was undertaken in 1997 to streamline the original program and dictionaries. Text files from several dozen studies, totaling over 8 million words were analyzed using LIWC as well as WordSmith, a powerful word count program used in discourse analysis. Original LIWC categories that were used at very low rates (less than 0.3 percent of words made up the category) or that suffered from consistently poor reliability or validity were omitted. Several new categories, including social processes, several personal concern categories, and the relativity dimensions, were added

following the same stringent judge-based procedures described above (including both passes). Finally, once the entire new LIWC dictionary was assembled, any words that were not used at least 0.005 percent of the time in our previous text files or were not listed in Francis and Kucera's (1982) Frequency Analysis of English Usage were excluded.

LIWC's External Validity

One of the first tests of the validity of the LIWC scales was undertaken by Pennebaker and Francis (1996) as part of an experiment in which first year college students wrote about the experience of coming to college. During the writing phase of the study, 72 Introductory Psychology students met as a group on three consecutive days to write on their assigned topics. Participants in the experimental condition (n = 35) were instructed to write about their deepest thoughts and feelings concerning the experience of coming to college. Those in the control condition (n = 37) were asked to describe any particular object or event of their choosing in an unemotional way. After the writing phase of the study was completed, four judges rated the participants' essays on various emotional, cognitive, content, and composition dimensions designed to correspond to selected LIWC Dictionary scales.

Using LIWC output and judges' ratings, Pearson correlational analyses were performed to test LIWC's external validity. Results, presented in Table 1, reveal that the LIWC scales and judges' ratings are highly correlated. These findings suggest that LIWC successfully measures positive and negative emotions, a number of cognitive strategies, several types of thematic content, and various language composition elements. As can be seen in Table 1, two LIWC-judge correlations are presented. The first, Judge 1, is based on overall ratings of the entire essay set (210 total essays across conditions). The second correlation, Judge 2, refers to the mean within-condition correlation – a much more stringent test of reliability. The level of agreement between judges' ratings and LIWC's objective word count strategy provides support for LIWC's external validity.

Base Rates of Word Usage

In evaluating any text analysis program, it is helpful to get a sense of the degree to which language varies across settings. Since 1986, we have been collecting text samples from a variety of studies – both from our own lab as well as from others in the United States, Canada, and New Zealand. For purposes of comparison, four classes of text from 43 separate studies were analyzed and compared. As can be seen in Table 2, these analyses reflect the utterances of at least 1,695 writers or speakers totaling over 1.6 million words. Twenty of the samples are based on individuals from all walks of life – ranging from college students to psychiatric prisoners to elderly and even elementary-aged individuals – who were asked to write about deeply emotional topics. Fifteen samples, which were generally the control groups of the emotion writing groups, wrote about relatively trivial topics, such as plans for the day or descriptions of ordinary events or objects. A third class of text was based on a random sampling of pages from the 30 top-selling paperback fiction books of 1995. Finally, we analyzed data from seven observational studies in which participants were tape recorded while engaged in conversations with others. The speech samples ranged from strangers interacting in a waiting room, to couples talking about problems, to televised interviews, to open-air tape recordings of people in public spaces.

As can be seen in Table 3, the LIWC version captures, on average, 80 percent of the words people used in writing and speech. Note that except for total word count and words per sentence, all means in Table 3 are expressed as percentage of total word use in any given speech/text sample. Across all of the studies, for example, 15.2 percent of words used were pronouns, 5.8 percent articles, and 4.0 percent were emotional words. Simple oneway ANOVAs

indicated that word usage was significantly different across the four settings for all but one of the word categories (religion words).

References

Francis, W.N., & Kucera, H. (1982). *Frequency analyses of English usage: Lexicon and grammar*. Boston: Houghton Mifflin.

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

Pennebaker, J. W. (1997). Writing about emotional experiences as a therapeutic process. *Psychological Science, 8*, 162-166.

Pennebaker, J.W., Colder, M., & Sharp, L.K. (1990). Accelerating the coping process. *Journal of Personality and Social Psychology, 58,* 528-537.

Pennebaker, J.W., & Francis, M.E. (1996). Cognitive, emotional, and language processes in disclosure. *Cognition and Emotion, 10*, 601-626.

Pennebaker, J. W., Mayne, T., & Francis, M. E. (1997). Linguistic predictors of adaptive bereavement. *Journal of Personality and Social Psychology, 72*, 863-871.

Rosenberg, S.D. & Tucker, G.J. (1978). Verbal behavior and schizophrenia: The semantic dimension. *Archives of General Psychiatry, 36*, 1331-1337.

Stiles, W.B. (1992). *Describing talk: A taxonomy of verbal response modes*. Newbury Park, CA: Sage.

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*, 1063-1070.

The research reported in this paper was made possible by a grant from the National Institutes of Health (MH52391). We are deeply indebted to a number of people who helped with different phases of this project: Laura King, Cheryl Hughes, Becky Smith, Kathy Davison, Janie Keller, Mary Sue Hayward, Brooke Novales, Anne Vano, Michael Crow, Sally Dickerson, and Bernard Rimé.

Table 1. LIWC Output Variable Information

Dimension	Abbrev	Examples	# Words	Judge 1	Judge 2
I. STANDARD LINGUISTIC DIMENSIONS					
Word Count	WC				
Words per sentence	WPS				
Sentences ending with ?	Qmarks				
Unique words (type/token ratio)	Unique				
% words captured, dictionary words	Dic				
% words longer than 6 letters	Sixltr				
Total pronouns	Pronoun	I, our, they, you're	70		
1st person singular	I	I, my, me	9		
1st person plural	We	we, our, us	11		
Total first person	Self	I, we, me	20	.78	.47
Total second person	You	you, you'll	14		
Total third person	Other	she, their, them	22		
Negations	Negate	no, never, not	31		
Assents	Assent	yes, OK, mmhmm	18		
Articles	Article	a, an, the	3		
Prepositions	Preps	on, to, from	43		
Numbers	Number	one, thirty, million	29		
II. PSYCHOLOGICAL PROCESSES					
Affective or Emotional Processes	Affect	happy, ugly, bitter	615		
Positive Emotions	Posemo	happy, pretty, good	261	.63	.33
Positive feelings	Posfeel	happy, joy, love	43		
Optimism and energy	Optim	certainty, pride, win	69	.37	.22
Negative Emotions	Negemo	hate, worthless,	345	.75	.38
Anxiety or fear	Anx	nervous, afraid, tense	62	.57	.40
Anger	Anger	hate, kill, pissed	121	.57	.41
Sadness or depression	Sad	grief, cry, sad	72	.66	.29
Cognitive Processes	Cogmech	cause, know, ought	312		
Causation	Cause	because, effect,	49	.39	.31
Insight	Insight	think, know,	116	.73	.23
Discrepancy	Discrep	should, would, could	32	.53	.20
Inhibition	Inhib	block, constrain	64		
Tentative	Tentat	maybe, perhaps,	79	.49	.21
Certainty	Certain	always, never	30		
Sensory and Perceptual Processes	Senses	see, touch, listen	111		
Seeing	See	view, saw, look	31		
Hearing	Hear	heard, listen, sound	36		
Feeling	Feel	touch, hold, felt	30		
Social Processes	Social	talk, us, friend	314		

Communication	Comm	talk, share, converse	124		
Other references to people	Othref	1st pl, 2nd, 3rd per	54		
Friends	Friends	pal, buddy, coworker	28	.74	.69
Family	Family	mom, brother, cousin	43	.81	.80
Humans	Humans	boy, woman, group	43		
III. RELATIVITY					
Time	Time	hour, day, oclock	113		
Past tense verb	Past	walked, were, had	144	.75	.75
Present tense verb	Present	walk, is, be	256		
Future tense verb	Future	will, might, shall	14		
Space	Space	around, over, up	71		
Up	Up	up, above, over	12		
Down	Down	down, below, under	7		
Inclusive	Incl	with, and, include	16		
Exclusive	Excl	but, except, without	19		
Motion	Motion	walk, move, go	73		
IV. PERSONAL CONCERNS					
Occupation	Occup	work, class, boss	213		
School	School	class, student, college	100	.27	.25
Job or work	Job	employ, boss, career	62		
Achievement	Achieve	try, goal, win	60		
Leisure activity	Leisure	house, TV, music	102		
Home	Home	house, kitchen, lawn	26		
Sports	Sports	football, game, play	28		
Television and movies	TV	TV, sitcom, cinema	19		
Music	Music	tunes, song, cd	31		
Money and financial issues	Money	cash, taxes, income	75		
Metaphysical issues	Metaph	God, heaven, coffin	85		
Religion	Relig	God, church, rabbi	56		
Death and dying	Death	dead, burial, coffin	29		
Physical states and functions	Physcal	ache, breast, sleep	285		
Body states, symptoms	Body	ache, heart, cough	200	.45	.61
Sex and sexuality	Sexual	lust, penis, fuck	49		
Eating, drinking, dieting	Eating	eat, swallow, taste	52		
Sleeping, dreaming	Sleep	asleep, bed, dreams	21		
Grooming	Groom	wash, bath, clean	15		
APPENDIX: EXPERIMENTAL DIMENSIONS					
Swear words	Swear	damn, fuck, piss	29		
Nonfluencies	Nonfl	uh, rr*	6		
Fillers	Fillers	youknow, Imean	6		

Table 2 Summary Information for LIWC Statistics

	Text Samples				
	Emotion Writing	Control Writing	Books	Talking	Totals
Number of files	2,028	1,473	300	777	4,578
Number of writers/speakers	768	469	30	428	1,695
Number of words	665,184	443,668	200,016	306,439	1,615,307
Number of studies	20	15	1	7	43

Emotion writing studies require participants to write about their emotions and thoughts about personally relevant topics; Control Writing involves writing about non-emotional topics, such as plans for the day or descriptions of ordinary objects or events; Books refers to a semi-random sample of pages from the 30 best-selling fiction books of 1995; Talking files come from transcripts collected from individuals who are talking in non-experimental settings (i.e., correlational studies).

Table 3. LIWC Means Across 43 Studies

Dimension	Emotion Writing	Control Writing	Books	Talking	Mean (sd)
I. LINGUISTIC DIMENSIONS					
Word Count	327	301	667	394	353 (278)
Words per sentence	20.9	19.4	13.0	10.9	18.2 (29.9)
Sentences ending with ?	2.1	0.4	9.8	21.6	5.4 (13.1)
Unique words (type/token ratio)	51.5	50.1	48.6	50.8	50.8 (9.8)
% words captured, dictionary words	83.3	78.9	74.0	75.4	79.9 (7.8)
% words longer than 6 letters	13.1	14.1	16.4	10.1	13.1 (5.1)
Total pronouns	17.2	12.4	13.6	15.8	15.2 (5.2)
1st person singular	10.6	8.2	2.7	5.6	8.5 (4.7)
1st person plural	0.8	1.5	0.5	1.0	1.1 (1.6)
Total first person	11.4	9.7	3.3	6.6	9.5 (4.7)
Total second person	0.4	0.3	1.5	4.0	1.0 (1.9)
Total third person	3.3	1.2	7.0	2.5	2.7 (2.8)
Negations	2.3	0.8	1.9	2.8	1.8 (1.6)
Assents	0.1	0.0	0.2	1.6	0.4 (1.7)
Articles	5.0	7.2	7.5	4.3	5.8 (2.8)
Prepositions	12.6	15.2	13.2	9.2	12.9 (3.5)
Numbers	1.0	1.0	1.0	1.5	1.1 (1.2)
II. PSYCHOLOGICAL PROCESSES					
Affective or Emotional Processes	5.3	2.3	3.9	4.0	4.0 (2.4)
Positive Emotions	2.7	1.7	2.2	2.7	2.4 (1.6)
Positive feelings	0.9	0.3	0.7	0.9	0.7 (0.8)
Optimism and energy	0.5	0.4	0.5	0.3	0.5 (0.6)
Negative Emotions	2.6	0.6	1.6	1.3	1.6 (1.7)
Anxiety or fear	0.6	0.2	0.4	0.3	0.4 (0.6)
Anger	0.7	0.2	0.6	0.5	0.5 (0.8)
Sadness or depression	0.7	0.2	0.4	0.2	0.4 (0.7)
Cognitive Processes	7.8	4.1	6.1	7.3	6.4 (3.2)
Causation	1.1	0.6	0.6	1.1	0.9 (0.9)
Insight	2.5	1.1	1.9	2.4	2.0 (1.5)
Discrepancy	2.7	1.1	2.1	1.7	2.0 (1.5)
Inhibition	0.3	0.3	0.4	0.2	0.3 (0.4)
Tentative	2.5	1.6	1.8	2.2	2.1 (1.5)
Certainty	1.4	0.7	0.9	0.9	1.1 (0.9)
Sensory and Perceptual Processes	2.5	2.2	3.2	2.6	2.4 (1.6)
Seeing	0.5	0.8	0.9	1.0	0.7 (0.8)
Hearing	1.1	0.8	1.7	1.3	1.1 (1.1)
Feeling	0.8	0.3	0.4	0.2	0.5 (0.7)

Social Processes	9.5	6.0	13.1	10.9	8.8 (4.9)
Communication	1.7	1.3	2.3	1.9	1.7 (1.4)
Other references to people	4.8	3.0	9.2	7.6	5.0 (3.6)
Friends	0.6	0.3	0.1	0.1	0.4 (0.7)
Family	1.2	0.4	0.4	0.2	0.7 (1.3)
Humans	0.8	0.4	0.7	0.7	0.6 (0.8)
III. RELATIVITY					
Time	5.0	6.2	3.5	3.4	5.0 (2.8)
Past tense verb	6.9	5.8	7.5	4.5	6.2 (4.4)
Present tense verb	10.2	8.7	6.2	13.7	10.1 (5.1)
Future tense verb	1.0	1.7	0.9	0.9	1.2 (1.6)
Space	2.4	3.3	3.0	2.6	2.8 (1.7)
Up	1.2	2.0	1.3	1.2	1.4 (1.1)
Down	0.2	0.5	0.4	0.2	0.3 (0.5)
Inclusive	6.3	7.1	5.9	4.5	6.2 (2.4)
Exclusive	4.0	2.6	3.1	3.8	3.5 (1.7)
Motion	1.3	2.6	1.1	1.6	1.8 (1.6)
IV. PERSONAL CONCERNS					
Occupation	2.5	3.9	1.3	1.7	2.7 (2.3)
School	1.2	2.6	0.3	0.8	1.5 (1.9)
Job or work	0.5	0.8	0.6	0.5	0.6 (1.0)
Achievement	0.9	0.8	0.6	0.5	0.8 (0.8)
Leisure activity	1.1	2.6	0.9	0.8	1.5 (1.9)
Home	0.8	1.5	0.5	0.3	0.9 (1.3)
Sports	0.3	0.7	0.1	0.2	0.4 (1.0)
Television and movies	0.1	0.3	0.1	0.1	0.1 (0.4)
Music	0.1	0.3	0.2	0.1	0.2 (0.4)
Money and financial issues	0.3	0.3	0.3	0.5	0.3 (0.6)
Metaphysical issues	0.4	0.2	0.4	0.2	0.3 (0.7)
Religion	0.2	0.2	0.2	0.1	0.2 (0.5)
Death and dying	0.2	0.0	0.2	0.0	0.1 (0.5)
Physical states and functions	1.2	2.5	2.0	0.9	1.6 (1.8)
Body states, symptoms	0.6	0.8	1.5	0.5	0.7 (1.1)
Sex and sexuality	0.3	0.1	0.2	0.2	0.2 (0.5)
Eating, drinking, dieting	0.2	1.1	0.2	0.2	0.5 (1.0)
Sleeping, dreaming	0.1	0.6	0.2	0.1	0.3 (0.6)
Grooming	0.1	0.6	0.1	0.1	0.2 (0.7)
APPENDIX: EXPERIMENTAL DIMENSIONS					
Swear words	0.1	0.0	0.1	0.2	0.1 (0.4)
Nonfluencies	0.0	0.0	0.0	0.5	0.1 (0.4)
Fillers	0.0	0.0	0.0	0.2	0.04 (0.3)