Operator's Manual

Linguistic Inquiry and Word Count: LIWC2007

James W. Pennebaker, Roger J Booth,
and Martha E. Francis

The University of Texas at Austin
and
The University of Auckland, New Zealand

The LIWC2007 software, Operator's Manual, and the LIWC2007 Language Analysis Manual are published by LIWC.net, Austin, Texas 78703 USA.

Contents

Getting Started 1
Running LIWC2007 on a PC 2
Running LIWC2007 on a Macintosh 2
Reading and Analyzing LIWC2007 Output 3
Customizing LIWC2007 Output 3
Analyzing Text in Segments 3
Handling numerals, abbreviations and emoticons 3
Handling punctuation ..... **Error! Bookmark not defined.**
Creating and Using Custom Dictionaries 4
Conditional Categories 6
Preparing Written Text For LIWC2007 Analysis 6
1. Text file organization. 6
2. Text file computer entry. 6
3. Cleaning the text files. 6
Naming Text Files 7
Typing Conventions: Writing and Interview Samples 7
1. Spelling, abbreviations, contractions. 7
2. End of sentence markers and hyphens. 7
3. Other common problems: 8
Transcribing Oral Transcripts: Special Problems 8
1. Nonfluencies 8
2. Fillers. 9
3. Transcribers' comments. 9
Technical Support 9
Getting Some Practice: Running the Samples 10

Getting Started

The LIWC2007 program comes with the following files:

LIWC2007 the actual application file (LIWC2007.EXE for Windows and LIWC2007 for Macintosh) incorporating the LIWC2007 and LIWC2001 master dictionaries. Note that the LIWC Student Versions only include the internal LIWC2007 dictionary and no other ancillary files.

SAMPLES a directory of sample text files, including 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)

DICTIONARIES The dictionaries included:

LIWC2007.dic is a copy of the internal default dictionary. Note that this is not the actual internal dictionary that the LIWC2007 program runs. Any changes to this dictionary will only take effect if this dictionary is loaded as part of the "Load New Dictionary" command in the "Dictionary" menu.

LIWC2001.dic is a copy of the internal default dictionary used on the original LIWC2001 program

Spanish2001.dic is a Spanish translation of the LIWC2001 dictionary developed by Ramirez-Esparza, N., Pennebaker, J.W., Garcia, F.A., & Suria, R. (2007). La psychología del uso de las palabras: Un programa de comutadora que analiza textos en Español (The psychology of word use: A computer program that analyzes texts in Spanish). *Revista Mexicana de Psicología, 24*, 85-99.

German2001.dic is a German translation of the LIWC2001 dictionary developed by Wolf, M., Horn, A., Mehl, M., Haug, S., Pennebaker, J. W., & Kordy, H. (2008, in press). Computergestützte quantitative Textanalyse: Äquivalenz und Robustheit der deutschen Version des Linguistic Inquiry and Word Count [Computer-aided quantitative text analysis: Equivalence and robustness of the German adaption of the Linguistic Inquiry and Word Count]. *Diagnostica*.

Pronoun.dic is a short sample dictionary of pronouns. It is included as a simple example of the dictionary system.

LIWC2007 word category file (LIWC2007dictionary poster.xls) is an Excel file that lists all the words that are in the LIWC2007 dictionary by category.

Running LIWC2007 on a PC

To run the application, double click on the LIWC2007 icon or LIWC2007.EXE file. Once the LIWC2007 application launches, explore the various options.

To analyze whatever text files you specify, go into the "File" menu and select " Process Text..." (or click on the 'Ask LIWC2007 to process a file(s)' icon). Multiple files can be processed in one of two ways. Either shift-click on all the filenames your require, or alternatively, you can analyse all files in a particular directory by clicking the " Select All" button. The Select All option will analyze all files in that particular directory that are text (.txt) or Word document (.doc) files. If you have directories within the current directory, you can have LIWC2007 process all the text files within these as well by checking the " Include files in enclosed directories" checkbox before clicking the "Select All" button".

Tip: If you have a large number of text files to process, it is generally most efficient to put them all in one directory (or directories) and then use the " Select All" button on that directory. You can also select multiple files within a directory by using shift-clicking or control-clicking. On a PC computer, point to a particular file and hold down the shift key before clicking. You can then point to a later file in the same directory and again depress the shift key before clicking. All files between the two clicked-on files will now be selected. Control-clicking simply requires that you hold down the control key and click on the individual files you wish to select.

You then get the opportunity to specify a name and location for your output file before LIWC2007 begins processing. LIWC2007 processes the files sequentially, showing you its progress, storing the output in the file you specified and then displaying results in a window on the screen. 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 programs such as Excel, SPSS, or SAS.

Running LIWC2007 on a Macintosh

To run the application, double click on the LIWC2007 icon. Once the LIWC2007 application launches, explore the various options.

To analyze whatever text files you specify, go into the "File" menu and select " Process Text...". Select the files that you with to analyze by shift-clicking or command-clicking them in the dialog box.

You then get the opportunity to specify a name and location for your output file before LIWC2007 begins processing. LIWC2007 processes the files sequentially, showing you its progress, storing the output in the file you specified and then displaying results in a window on the screen. 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.

Reading and Analyzing LIWC2007 Output

LIWC2007 stores the output in the file you specified and then displays results in a window on the screen. By default, all LIWC2007 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 any LIWC2007 output file, choose the "Open" command within the "File" Menu (or click on the 'Open an existing document' icon in Windows) and specify an output filename. 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, is recommended.

Customizing LIWC2007 Output

In some cases, you may prefer to analyze only a subset of language dimensions rather than the full set of variables. To do this, open the "Categories" menu. Within each option (e.g., standard information, linguistic dimensions etc.), check boxes are available for each LIWC2007 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 and will be saved when the application is quit. To use all dimensions, choose "Use all categories" from the "Categories" menu.

Analyzing Text in Segments

Each text file analyzed by LIWC2007 can be treated as a whole or broken into segments in one of three ways. This is controlled by the "Analyze in segments…" command on the "Options" menu. You have four choices here: (1) *No text segmentation*; (2) *Define number of segments*, in which case you can choose how many segments you wish to divide your text files into; (3) *Define words per segment*, in which case you choose how many words in each segment; and (4) *Define segment delimiter*, in which case you will have segments of your text separated by a number of blank lines and LIWC2007 will use these to break your text into segments.The active segmentation is displayed in the Windows version of LIWC2007 on the status line at the bottom of the application window, and the the Macintosh version in a floating window called "Analysis Status". Note that in the LIWC output file, the second column refers to the actual segment sequence.

Handling numerals and punctuation

The "Extras…" item of the "Categories" menu allows you to determine how LIWC2007 handles numerals (e.g. 12, 38, 156). In each case you can have LIWC2007 ignore them by

clicking on the "Ignore them" radio button or have them listed as separate categories by clicking on the "List them separately" radio button.

Numerals can also be added to the "numbers" category. The default "numbers" category looks only for words (e.g., seven, thousands). By clicking on the "Add to 'numbers' category" button, numeral sequences are considered word units and are counted in the same way as number words.

The "Punctuation…" item of the "Categories" menu allows you to determine how LIWC2007 counts and reports punctuation characters. By clicking on the item in the dialog box you can switch on or off the following punctuation characters and LIWC2007 will count them and report them as a percentage of total words:

Period	.
Comma	,
Colon	:
SemiC	; (semi-colon)
QMark	? (question mark)
Exclam	! (exclamation mark)
Dash	-
Quote	" (quotation mark)
Apostro	' (apostrophe)
Parenth	() or [] or {} (LIWC2007 will count each pair of parentheses)
OtherP	(other punctuation includes all ASCII characters from 33-47, 58-64, 91-96, 123-126 not included in the list above i.e. all non-alphanumeric and non-control characters)
AllPct	All punctuation

Creating and Using Custom Dictionaries

As well as containing a dictionary and category set integral to the application, LIWC2007 permits use of dictionaries and categories created by the user. This is done by selecting the "Load new dictionary…" option from the "Dictionary" menu. External dictionaries must be plain text files set out in the following format. For example, one could create a specific self-referencing dictionary:

```
%
1       I
2       me
3       my
4       we
5       us
6       our
```

7	singular
8	plural
9	possessive
%	
me	2 7
mine	3 7 9
my	3 7 9
myself	2 7
our	6 8 9
us	5 8
we'*	4 8

If your external dictionary includes category definitions, they must precede the dictionary and be enclosed between % delimiters as shown above. If your dictionary doesn't include category definitions, LIWC2007 will use the default internal categories. There must only be one category definitions per line beginning with the category number and followed by the category name separated from the number by space(s) and/or tab(s). LIWC2007 will accept up to 999 categories.

Each number refers to the category to which each word is assigned. Hence, the word "me" is associated with category 2 (the solo word dictionary of "me") and category 7 (1st person singular), the word "our" is associated with categories 7 (our), 8 (1st person plural), and 9 (possessive).

The dictionary list comprises one word or word-stem per line followed by a list of category numbers with which the word is associated. Again all elements in the line are separated by space(s) and/or tab(s). In the example above, the word "us" is associated with categories 5 and 8. Word-stems are partial words terminated by an asterisk. Thus, in the dictionary, use of an asterisk (*) at the end of the word signals LIWC2007 to ignore all subsequent letters. Consequently, "we'*" will count the words we're, we'll, we'd, etc. in categories 4 and 8.

Helpful tips from users:*

- The first line of the dictionary must be the % symbol, followed by the Category numbers and names. After the last Category name, a second % symbol must be inserted to signal the beginning of the Word entries and numbers.
- It is more efficient to create your own dictionary than to try to integrate your dictionary into the LIWC2007 default dictionary.
- Category names should be ONE word only with no punctuation. Category names do not need to be in alphabetical order.
- Words should be listed in alphabetical order. Single words only but NO numbers or punctuation (exceptions include apostrophe and hyphen).

OK:	Love	500	
	Love-sick	500	503
NOT OK	Lov4sale	521	
	Love sick	500	503

- Word entries should only appear once. Linking words to multiple categories is fine.

  OK: Love 500 504

  NOT OK: Love 500

  Love 504

- Be careful in the use of asterisks and avoid double counting words. For example, the following case is problematic:

  NOT OK: Thank* 27

  Thanksgiving 27 94

In this case, the word "Thanksgiving" would be counted twice in category 27.

*****With special thanks to Nancy Collins at UC-Santa Barbara.

Phrases rather than words

LIWC was originally created to examine words or word stems. LIWC2007 can now search for phrases. This option is currently available on the Mac version but not yet on the PC version.

Preparing Written Text For LIWC2007 Analysis

The accuracy of LIWC2007 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 LIWC2007 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] -- 4568day1E.txt, 4568day2E.txt, and 4568day3E.txt

2. Text file computer entry.

Essays should be entered into the computer using Microsoft Word documents or as standard text or ASCII files. Files prepared with other word processing programs (e.g., Word Perfect) will not work. Also, files in formats such as pdf, html, jpg, etc will not provide valid output. A good rule of thumb is that files ending in .txt or .doc will probably provide accurate results; other filetypes probably won't.

3. Cleaning the text files.

Each 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 LIWC2007 converts all text files to lower case before processing them, grammar, capitalization, and sentence structure do not need to be corrected.

Naming 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**. LIWC2007 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, LIWC2007 will analyze them as a single writing sample.
2. **The file name should be descriptive,** including ID number, condition, and question or day number.
3. **Files must be in TEXT or Word Document format.** LIWC2007 cannot read WordPerfect or other word processing files. Note that virtually all word processing programs allow you to convert your files into ASCII, TEXT, or Word 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. LIWC2007 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. It is best to use standard United States spelling (although the standard default dictionary also contains most British English spellings as well).

   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.**

   The Words per sentence (WPS) category is based on the number of times that end-of-sentence markers are detected. These include all periods (.), question marks, and

exclamation points. One potential problem is that 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 LIWC2007 as part of the word. LIWC2007, for example, lists "self-esteem" as a meaningful word in one of its dictionaries. In cases of hyphenated phrases such as "this-or-that" LIWC2007 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." LIWC2007 will think that "store-I" is one word. Insert blanks on either side of the hyphens so that both words will be counted.

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 LIWC2007 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.

LIWC2007 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. LIWC2007 will count the xxxx as a spoken word but not assign it to a dictionary. For entire passages, don't insert anything.

Technical Support

Technical support for set-up and hardware/software compatibility can be obtained by sending an email to webmaster@liwc.net. 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) or from the second author, Roger Booth, School of Medical Science, The University of Auckland, Auckland, New Zealand (rj.booth@auckland.ac.nz). More extended consultation is available on a fee basis.

Getting Some Practice: Running the Samples

Included with the LIWC2007 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 LIWC2007 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 LIWC2007 "open file" menu.

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

1. Start the LIWC2007 application by clicking on the LIWC2007 icon or LIWC2007.EXE.
2. Within the LIWC2007 application, go into the "File" menu and press "Process text…".
3. Navigate your way to the directory containing the sample files.

4. Click the "Select All" button.
5. LIWC2007 will display a standard dialog box with a default name (LIWC results.dat) and location for the file to contain the results. You can change these if you wish.
6. Press the "Save" button. Voila!

On completion, LIWC2007 will save the results in the specified file and also open it and display the data in a window for you to see. 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 LIWC results.dat file. 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	Sixltr	Dic	funct	pronoun	i	we	posemo	negemo
Lincoln.txt	3639	28.43	23.36	82.77	61.14	10.52	1.59	0.63	3.24	1.70
FDR.txt	1881	22.13	23.34	85.06	56.14	12.39	1.65	3.35	4.36	2.45
Clinton.txt	1584	17.22	20.71	86.36	57.20	14.77	0.88	7.83	4.67	1.64
Huckcln.txt	603	21.54	8.79	88.06	64.84	18.24	2.49	1.16	2.32	1.00
Huckraw.txt	654	21.80	8.10	76.76	55.50	15.90	2.75	1.07	2.29	0.92
Plath.txt	100	33.33	26.00	74.00	39.00	5.00	0	0	4.00	5.00
Sexton.txt	237	14.81	12.24	89.45	60.34	23.21	13.50	0	7.59	4.22
radio.txt	272	5.44	7.72	93.75	58.46	21.32	6.99	1.10	6.25	0.74
talkshow.txt	621	24.84	18.20	93.24	57.33	15.62	0.64	2.42	2.74	1.13
Abstr1.txt	107	17.83	45.79	76.64	34.58	0.93	0	0	7.48	1.87
Abstr2.txt	196	24.50	36.73	66.84	39.29	0.51	0	0	1.53	0

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] and words per sentence [WPS]) 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 with 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 LIWC2007 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.