The Development and Psychometric Properties of LIWC2015

James W. Pennebaker, Ryan L. Boyd,
Kayla Jordan, and Kate Blackburn

The University of Texas at Austin

E-mail correspondence should be sent to pennebaker@mail.utexas.edu or ryanboyd@utexas.edu. Other correspondence should be sent to James W. Pennebaker, Department of Psychology, The University of Texas at Austin, 108 E. Dean Keeton Stop A8000, Austin, TX 78712-1043.

The official reference to this paper is:

Pennebaker, J.W., Boyd, R.L., Jordan, K., & Blackburn, K. (2015). *The development and psychometric properties of LIWC2015*. Austin, TX: University of Texas at Austin. DOI: 10.15781/T29G6Z

The Development and Psychometric Properties of LIWC2015

The ways people use words in their daily lives can provide rich information about their beliefs, fears, thinking patterns, 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 began amassing increasingly compelling evidence that the words we use have tremendous psychological value (Gottschalk & Glaser, 1969; Stone, Dunphy, Smith, & Ogilvie, 1966; Weintraub, 1989).

Although promising, the early computer methods floundered because of the sheer complexity of the task. Extensive samples of text were not digitized, computers were slow and unwieldy, and there was little agreement about which features of natural language were most related to psychological states. Everything changed in the 1990s with the advent of efficient desktop computers, improved data storage technology, and the explosion of the internet. These factors allowed for the easy collection of large stores of books, conversations, and other digitized text samples.

In order to provide an efficient and effective method for studying the various emotional, cognitive, and structural components present in individuals' verbal and written speech samples, we originally developed a text analysis application called Linguistic Inquiry and Word Count, or LIWC. The first LIWC application was developed as part of an exploratory study of language and disclosure (Francis, 1993; Pennebaker, 1993). The second (LIWC2001) and third (LIWC2007) versions updated the original application with an expanded dictionary and a more modern software design (Pennebaker, Francis, & Booth, 2001; Pennebaker, Booth, & Francis, 2007).

The most recent evolution, LIWC2015 (Pennebaker, Booth, Boyd, & Francis, 2015), has significantly altered both the dictionary and the software options. Importantly, the LIWC2015 software and dictionary are new, rather than a basic update to previous versions of LIWC. As with previous versions, however, 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 LIWC2015 Framework

Both the standard downloadable and web-based versions of the LIWC2015 application rely on an internal default dictionary that defines which words should be counted in the target text files. Note that the LIWC2015 processor is an executable file and cannot be read or opened. To avoid confusion in the subsequent discussion, words contained in texts that are read and analyzed by LIWC2015 are referred to as ***target words***. Words in the LIWC2015 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 LIWC2015 Main Text Processing Module

Because the software application is written in a cross-platform language, it runs identically on PC and Mac computers via the Java Virtual Machine. LIWC2015 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, PDF, RTF, or standard Microsoft Word files (i.e., .doc and .docx). Unlike previous versions, the software can now process text on a line by line basis within and across columns inside of multiple spreadsheet formats, including those saved as .xls, .xlsx, and .csv files.

During operation, LIWC2015 accesses a single text file, a group of files, or texts within a spreadsheet and analyzes each sequentially. For each file, LIWC2015 reads one target word at a time. As each target word is processed, the dictionary file is searched, looking for a dictionary match with the current target word. If the target word is matched with a dictionary word, the appropriate word category scale (or scales) 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.

For each text file, approximately 90 output variables are written as one line of data to an output file. This data record includes the file name and word count, 4 summary language variables (analytical thinking, clout, authenticity, and emotional tone), 3 general descriptor categories (words per sentence, percent of target words captured by the dictionary, and percent of words in the text that are longer than six letters), 21 standard linguistic dimensions (e.g., percentage of words in the text that are pronouns, articles, auxiliary verbs, etc.), 41 word categories tapping psychological constructs (e.g., affect, cognition, biological processes, drives), 6 personal concern categories (e.g., work, home, leisure activities), 5 informal language markers (assents, fillers, swear words, netspeak), and 12 punctuation categories (periods, commas, etc). A complete list of the standard LIWC2015 scales is included in Table 1.

The Default LIWC2015 Dictionary

The LIWC2015 Dictionary is the heart of the text analysis strategy. The default LIWC2015 Dictionary is composed of almost 6,400 words, word stems, and select emoticons. Each dictionary entry additionally defines one or more word categories or subdictionaries. For example, the word *cried* is part of five word categories: sadness, negative emotion, overall affect, verbs, and past focus. Hence, if the word *cried* is found in the target text, each of these five subdictionary scale scores will be incremented. As in this example, many of the LIWC2015 categories are arranged hierarchically. All sadness words, by definition, belong to the broader "negative emotion" category, as well as the "overall affect words" category. Note too that word stems can be captured by the LIWC2015 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 an ingestion word (including hungry, hungrier, hungriest). The asterisk, then, denotes the acceptance of all letters, hyphens, or numbers following its appearance.

Each of the default LIWC2015 categories is composed of a list of dictionary words that define that scale. Table 1 provides a comprehensive list of the default LIWC2015 dictionary categories, scales, sample scale words, and relevant scale word counts.

Table 1. LIWC2015 Output Variable Information

Category	Abbrev	Examples	Words in category	Internal Consistency (Uncorrected α)	Internal Consistency (Corrected α)
Word count	WC	-	-	-	-
Summary Language Variables					
Analytical thinking	Analytic	-	-	-	-
Clout	Clout	-	-	-	-
Authentic	Authentic	-	-	-	-
Emotional tone	Tone	-	-	-	-
Words/sentence	WPS	-	-	-	-
Words > 6 letters	Sixltr	-	-	-	-
Dictionary words	Dic	-	-	-	-
Linguistic Dimensions					
Total function words	funct	it, to, no, very	491	.05	.24
Total pronouns	pronoun	I, them, itself	153	.25	.67
Personal pronouns	ppron	I, them, her	93	.20	.61
1st pers singular	i	I, me, mine	24	.41	.81
1st pers plural	we	we, us, our	12	.43	.82
2nd person	you	you, your, thou	30	.28	.70
3rd pers singular	shehe	she, her, him	17	.49	.85
3rd pers plural	they	they, their, they'd	11	.37	.78
Impersonal pronouns	ipron	it, it's, those	59	.28	.71
Articles	article	a, an, the	3	.05	.23
Prepositions	prep	to, with, above	74	.04	.18
Auxiliary verbs	auxverb	am, will, have	141	.16	.54
Common Adverbs	adverb	very, really	140	.43	.82
Conjunctions	conj	and, but, whereas	43	.14	.50
Negations	negate	no, not, never	62	.29	.71
Other Grammar					
Common verbs	verb	eat, come, carry	1000	.05	.23
Common adjectives	adj	free, happy, long	764	.04	.19
Comparisons	compare	greater, best, after	317	.08	.35
Interrogatives	interrog	how, when, what	48	.18	.57
Numbers	number	second, thousand	36	.45	.83
Quantifiers	quant	few, many, much	77	.23	.64
Psychological Processes					
Affective processes	affect	happy, cried	1393	.18	.57
Positive emotion	posemo	love, nice, sweet	620	.23	.64
Negative emotion	negemo	hurt, ugly, nasty	744	.17	.55
Anxiety	anx	worried, fearful	116	.31	.73
Anger	anger	hate, kill, annoyed	230	.16	.53
Sadness	sad	crying, grief, sad	136	.28	.70
Social processes	social	mate, talk, they	756	.51	.86
Family	family	daughter, dad, aunt	118	.55	.88

Category	Abbrev	Examples	Words in category	Internal Consistency (Uncorrected α)	Internal Consistency (Corrected α)
Friends	friend	buddy, neighbor	95	.20	.60
Female references	female	girl, her, mom	124	.53	.87
Male references	male	boy, his, dad	116	.52	.87
Cognitive processes	cogproc	cause, know, ought	797	.65	.92
Insight	insight	think, know	259	.47	.84
Causation	cause	because, effect	135	.26	.67
Discrepancy	discrep	should, would	83	.34	.76
Tentative	tentat	maybe, perhaps	178	.44	.83
Certainty	certain	always, never	113	.31	.73
Differentiation	differ	hasn't, but, else	81	.38	.78
Perceptual processes	percept	look, heard, feeling	436	.17	.55
See	see	view, saw, seen	126	.46	.84
Hear	hear	listen, hearing	93	.27	.69
Feel	feel	feels, touch	128	.24	.65
Biological processes	bio	eat, blood, pain	748	.29	.71
Body	body	cheek, hands, spit	215	.52	.87
Health	health	clinic, flu, pill	294	.09	.37
Sexual	sexual	horny, love, incest	131	.37	.78
Ingestion	ingest	dish, eat, pizza	184	.67	.92
Drives	drives		1103	.39	.80
Affiliation	affiliation	ally, friend, social	248	.40	.80
Achievement	achieve	win, success, better	213	.41	.81
Power	power	superior, bully	518	.35	.76
Reward	reward	take, prize, benefit	120	.27	.69
Risk	risk	danger, doubt	103	.26	.68
Time orientations	TimeOrient				
Past focus	focuspast	ago, did, talked	341	.23	.64
Present focus	focuspresent	today, is, now	424	.24	.66
Future focus	focusfuture	may, will, soon	97	.26	.68
Relativity	relativ	area, bend, exit	974	.50	.86
Motion	motion	arrive, car, go	325	.36	.77
Space	space	down, in, thin	360	.45	.83
Time	time	end, until, season	310	.39	.79
Personal concerns					
Work	work	job, majors, xerox	444	.69	.93
Leisure	leisure	cook, chat, movie	296	.50	.86
Home	home	kitchen, landlord	100	.46	.83
Money	money	audit, cash, owe	226	.60	.90
Religion	relig	altar, church	174	.64	.91
Death	death	bury, coffin, kill	74	.39	.79
Informal language	informal		380	.46	.84
Swear words	swear	fuck, damn, shit	131	.45	.83
Netspeak	netspeak	btw, lol, thx	209	.42	.82
Assent	assent	agree, OK, yes	36	.10	.39
Nonfluencies	nonflu	er, hm, umm	19	.27	.69
Fillers	filler	Imean, youknow	14	.06	.27

Table 1 Notes: "Words in category" refers to the number of different dictionary words and stems that make up the variable category. All alphas were computed on a sample of ~181,000 text files from several of our language corpora (see Table 2). Uncorrected internal consistency alphas are based on Cronbach estimates; corrected alphas are based on Spearman Brown. See the Reliability and Validity section below. Note that the LIWC2015 dictionary generally arranges categories hierarchically. There are some exceptions to the hierarchy rules. For example, *Social processes* include a large group of words that denote social processes, including all non-first-person-singular personal pronouns as well as verbs that suggest human interaction (talking, sharing) -- many of these words do not belong to any of the *Social processes* subcategories. Another example is *Relativity*, which includes a large number of words that cannot be found in any of its subcategories.

LIWC2015 Dictionary Development

The selection of words defining the LIWC2015 categories involved multiple steps over several years. Originally, the 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.

The most recent version of the dictionary, LIWC2015, is a completely new version compared to earlier ones. Dictionaries can now accommodate numbers, punctuation, and even short phrases. 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 emotion word.

A handful of new categories have been added 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. 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 presented below). We present here a complete overview of the process used to create the LIWC2015 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 LIWC2007 dictionary as a starting point. Within the Psychological Processes category, for example, the emotion 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, Clark, & Tellegen, 1988), Roget's Thesaurus, and standard English dictionaries. Following the creation of preliminary category word lists, 2-6 judges individually generated word lists for each category, then group brain-storming sessions among 4-8 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.

Step 2. Judge Rating Phase. Once the grand list of words was amassed, each word in the dictionary was examined by a group of 4-8 judges and qualitatively rated in terms of "goodness of fit" for each category. In order for a word to remain in a given category, a majority of judges had to agree on its inclusion. In cases of disputes, several corpora and online sources were referenced to determine a word's common use, inflection, and 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, texts from several sources were analyzed using the Meaning Extraction Helper (MEH; Boyd, 2015) to determine how frequently dictionary words were used in various contexts. These sources included blog posts, spoken language studies, Twitter, Facebook, novels, student writings, and several others. Dictionary words that did not occur at least once in multiple corpora were omitted from the dictionary.

Step 4. Candidate Word List Generation. In order to expand the dictionary, we explored several sources of language for high-frequency words that had not been added by judges. Using MEH, high-frequency words were quantified as a percentage of total words for hundreds of thousands of text files from multiple studies and sources. For several linguistic categories (e.g., verbs, adjectives), the Stanford Natural Language Toolkit (NLTK; Toutanova, Klein, Manning, & Singer, 2003) was used in conjunction with MEH to identify common words. All candidate words were then correlated with all dictionary categories in order to detect common words that were not yet included in the dictionary. Words that correlated positively with dictionary categories were added to a list of candidate words for possible inclusion. Following this, 4-8 judges reviewed the candidate list and 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*.

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 for ~181,000 text files hailing from 5 corpora, totalling ~231,000,000 words (see Table 2). All words for each category were treated as a "response" and 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 2-8 judges 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. Several linguistic categories, such as *pronouns* and *adverbs*, constitute established linguistic constructs and were therefore not a part of the omission process. We discuss the psychometric evaluation procedures in extensive detail in the next section.

Step 6. Refinement Phase. After Steps 1 through 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. Note that the psychometrics of each language category changed negligibly during each refinement phase. During the last stage of the final refinement phase, two judges reviewed the dictionary for mistakes.

Step 7. Addition of Summary Variables.

A major change from earlier versions of LIWC is the inclusion of four new summary variables: analytical thinking (Pennebaker et al., 2014), clout (Kacewicz et al., 2012), authenticity (Newman et al., 2003), and emotional tone (Cohn et al., 2004). Each summary variable was derived from previously published findings from our lab and converted to percentiles based on standardized scores from large comparison samples. It must be emphasized that the summary variables are the only non-transparent dimensions in the LIWC2015 output.

A Note about the LIWC2015 Language Categories

For those who are familiar with LIWC2007, some of the LIWC2015 categories and results will be a bit jarring. Some of the original categories have been removed, largely due to their consistently low base rates, low internal reliability, or their infrequent use by researchers:

Past tense verbs	Present tense verbs	Future tense verbs	Human words
Inhibition words	Inclusives	Exclusives	

The following is a list of categories that are either a) new to LIWC2015, or b) substantially different from their counterparts in previous versions. While other LIWC2015 categories may also be slightly different from those in previous versions, categories from previous versions of LIWC that are presented in the list below have undergone substantial revision.

Common verbs	Common adjectives	Common comparison words	Interrogatives
Female references	Male references	Cognitive processes	Differentiation words
Drives	Affiliation words	Achievement words	Power words
Risk words	Reward words	Past focus words	Present focus words
Future focus words	Informal language	Netspeak words	Quantifiers

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

LIWC2015: Internal Reliability and External Validity

Assessing the reliability and validity of text analysis programs is a tricky business. On the surface, one would think that you could determine the internal reliability of a LIWC scale the same way it is done with a questionnaire. With a questionnaire that taps anger or aggression, for example, participants complete a self-report asking a number of questions about their feelings or behaviors related to anger. Reliability coefficients 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 internally consistent.

A similar strategy can be used with words. But be warned: the psychometrics of natural language use are not as straight-forward as with questionnaires. The reason is obvious once you think

about it. Once you say something, you generally don't need to say it again in the same paragraph or essay. The nature of discourse, then, is we usually say something and then move on to the next topic. Repeating the same idea over and over again is generally bad form in language, yet this is a staple 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.

The LIWC Anger scale, for example, is made up of 230 anger-related words and word stems. 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 230 anger words across a select group of text files and then calculate the intercorrelations of the word use. Indeed, in Table 1, we include these internal reliability statistics, including those of Anger where the alpha reliabilities range between .52 (corrected) and .07 (uncorrected) depending on how it is computed. In order to calculate these statistics, each dictionary word was measured as a percentage of total words per text. These scores were then entered as an "item" in a standard Cronbach's alpha calculation, providing raw alpha scores for each word category, separately for each corpora. Uncorrected alphas in Table 1 are averages of each corpora's alpha score. Importantly, the uncorrected method tends to grossly underestimate reliability in language categories due the highly variable base rates of word usage within any given category. Corrected alphas were computed using the Spearman-Brown prediction formula (Brown, 1910; Spearman, 1910), and are generally a more accurate approximation of each category's "true" internal consistency.

Issues of validity are also a bit tricky. We can have people complete a questionnaire that assesses their general moods and then have them write an essay which we then subject to the LIWC program. We can also have judges evaluate the essay for its emotional content. In other words, we can get self-reported, judged, and LIWC numbers that all reflect a participant's anger.

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. The findings suggested that LIWC successfully measures positive and negative emotions, a number of cognitive strategies, several types of thematic content, and various language composition elements. The level of agreement between judges' ratings and LIWC's objective word count strategy provides support for LIWC's external validity.

Since the first version of LIWC, hundreds of studies have found the LIWC categories to be valid across dozens of psychological domains. As a starting point for exploring this body of literature, we recommend a close reading of Tausczik and Pennebaker (2010).

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 dozens of others in the United States, England, Canada, New Zealand, and Australia. For purposes of comparison, text from several dozens of studies have been analyzed using the updated LIWC2015 dictionary. As can be seen in Table 2, these analyses reflect the utterances of over 80,000 writers or speakers totaling over 231 million words. We provide a brief description of each dataset below.

Table 2. Summary Information for LIWC2015 Statistics

	Blogs	Expressive writing	Novels	Natural Speech	NY Times	Twitter
Total files	37,295	6,179	875	3,232	34,929	35,269
Total authors	37,295	2,510	441	2,174	Unknown	35,269
Total words	119,449,058	2,526,709	57,467,183	2,566,446	26,007,632	23,172,994

Note: All texts for all corpora required a minimum of 25 words for inclusion in our analyses. All texts with fewer than 25 words were omitted for all statistics reported in this document.

Blogs. This is an expanded version of the corpus described in Schler, Koppel, Argamon, and Pennebaker (2006). All blog posts were merged by individual prior to analysis, reflecting the entirety of each person's blog.

Expressive writing. This dataset consists of 29 samples from experiments where people were randomly assigned to write either about deeply emotional topics (emotional writing) or about relatively trivial topics such as plans for the day (control writing). Individuals from all walks of life – ranging from college students to psychiatric prisoners to elderly and even elementary-aged individuals – are represented in these studies. Only the emotional writing topics were included in the current analyses.

Novels. This is a sample of novels acquired from Project Gutenberg (http://www.gutenberg.org/) that had been tagged as "literature". All novels were written in the English language by authors who lived between approximately 1660 and 2008. The number of authors presented in Table 2 reflects only known authors of the works analyzed -- works for which the author was unknown were not included in this figure, but included in analyses.

Natural speech. The speech samples included diverse transcripts from multiple contexts, including people wearing audio recorders over days or weeks, strangers interacting in a waiting room, couples talking about problems, and open-air tape recordings of people in public spaces.

New York Times. A collection of articles published online at the New York Times website (http://www.nytimes.com). Articles were collected from the New York Times internet archives and include various types of work, including editorials, features, U.S. and world news, letters to

the editor, and so on. All articles were published between January and July of 2014. Author information was not preserved for this dataset, so the true number of authors is unknown.

Twitter. Individual Twitter posts (i.e., "tweets") were collected from the public profiles of users whose names were entered into the Analyze Words webpage (http://analyzewords.com). Each user's tweets were combined into a single unit of observation for analysis.

As can be seen in Table 3, the LIWC2015 version captures, on average, over 86 percent of the words people use in writing and speech. Note that except for total word count and words per sentence and the four summary variables (Analytic, Clout, Authentic, and Tone), all means in Table 3 are expressed as percentage of total words used in any given language sample. Simple statistical tests indicate that nearly all language categories differ significantly between contexts.

Table 3. LIWC2015 Output Variable Information

Category	Blogs	Expressive writing	Novels	Natural Speech	NY Times	Twitter	Grand Means	Mean SDs
Linguistic Processes								
Word count (mean)	3206.45	408.94	65716.49	794.17	744.62	660.24	11921.82	10274.32
Analytic	49.89	44.88	70.33	18.43	92.57	61.94	56.34	17.58
Clout	47.87	37.02	75.37	56.27	68.17	63.02	57.95	17.51
Authentic	60.93	76.01	21.56	61.32	24.84	50.39	49.17	20.92
Tone	54.50	38.60	37.06	79.29	43.61	72.24	54.22	23.27
Words/sentence*	18.40	18.42	16.13	-	21.94	12.10	17.40	16.38
Words>6 letters	14.38	13.62	16.30	10.42	23.58	15.31	15.60	3.76
Dictionary words	85.79	91.93	84.52	91.60	74.62	82.60	85.18	5.36
Total function words	53.10	58.27	54.51	56.86	42.39	46.08	51.87	5.13
Total pronouns	16.20	18.03	15.15	20.92	7.41	13.62	15.22	3.61
Personal pronouns	10.66	12.74	10.35	13.37	3.56	9.02	9.95	3.02
1st pers singular	6.26	8.66	2.63	7.03	0.63	4.75	4.99	2.46
1st pers plural	0.91	0.81	0.61	0.87	0.38	0.74	0.72	0.83
2nd person	1.32	0.68	1.39	4.04	0.34	2.41	1.70	1.35
3rd pers singular	1.50	2.01	4.80	0.77	1.53	0.64	1.88	1.53
3rd pers plural	0.68	0.57	0.92	0.65	0.68	0.47	0.66	0.60
Impersonal pronouns	5.53	5.28	4.79	7.53	3.84	4.60	5.26	1.62
Articles	6.00	5.70	8.35	4.34	9.08	5.58	6.51	1.79
Prepositions	12.60	14.27	14.27	10.29	14.27	11.88	12.93	2.11
Auxiliary verbs	8.75	9.25	7.77	12.03	5.11	8.27	8.53	2.04
Adverbs	5.88	6.02	4.17	7.67	2.76	5.13	5.27	1.61
Conjunctions	6.43	7.46	6.28	6.21	4.85	4.19	5.90	1.57
Negations	1.81	1.69	1.68	2.42	0.62	1.74	1.66	0.86
Other Grammar								
Common verbs	17.03	18.63	15.42	21.01	10.23	16.33	16.44	2.93
Common adjectives	4.53	4.52	4.36	4.13	4.52	4.89	4.49	1.30
Comparisons	2.17	2.42	2.13	2.35	2.39	1.89	2.23	0.95
Interrogatives	1.51	1.49	1.53	2.44	1.26	1.43	1.61	0.76

Category	Blogs	Expressive writing	Novels	Natural Speech	NY Times	Twitter	Grand Means	Mean SDs
Number	1.89	1.87	1.23	2.19	3.55	1.98	2.12	2.07
Quantifiers	2.27	2.35	1.80	1.93	1.94	1.85	2.02	0.83
Psychological Processes								
Affective processes	5.79	4.77	4.81	6.54	3.82	7.67	5.57	1.99
Positive emotion	3.66	2.57	2.67	5.31	2.32	5.48	3.67	1.63
Negative emotion	2.06	2.12	2.08	1.19	1.45	2.14	1.84	1.09
Anxiety	0.27	0.50	0.44	0.14	0.25	0.24	0.31	0.32
Anger	0.68	0.49	0.51	0.36	0.47	0.75	0.54	0.59
Sadness	0.44	0.50	0.55	0.23	0.29	0.43	0.41	0.40
Social processes[b]	8.95	8.69	12.26	10.42	7.62	10.47	9.74	3.38
Family	0.46	0.77	0.39	0.31	0.33	0.36	0.44	0.63
Friends	0.40	0.55	0.25	0.37	0.18	0.43	0.36	0.40
Female references	0.91	1.37	1.88	0.55	0.62	0.54	0.98	1.26
Male references	1.31	1.47	4.09	0.80	1.38	0.84	1.65	1.34
Cognitive processes	11.58	12.52	9.84	12.27	7.52	9.96	10.61	3.02
Insight	2.28	2.66	2.11	2.46	1.54	1.92	2.16	1.08
Causation	1.46	1.65	1.03	1.45	1.42	1.41	1.40	0.73
Discrepancy	1.56	1.74	1.48	1.45	0.89	1.54	1.44	0.80
Tentative	2.82	2.89	2.27	3.06	1.74	2.35	2.52	1.09
Certainty	1.56	1.51	1.45	1.38	0.76	1.43	1.35	0.70
Differentiation	3.31	3.40	2.82	3.73	2.03	2.62	2.99	1.18
Perceptual processes	2.58	2.38	3.74	2.11	2.42	2.96	2.70	1.20
See	1.04	0.80	1.58	0.78	0.88	1.39	1.08	0.78
Hear	0.75	0.48	1.26	0.63	1.06	0.82	0.83	0.62
Feel	0.64	0.92	0.76	0.61	0.35	0.56	0.64	0.52
Biological processes	2.16	2.59	2.17	1.23	1.44	2.60	2.03	1.39
Body	0.74	0.69	1.24	0.31	0.41	0.77	0.69	0.64
Health	0.61	0.93	0.48	0.38	0.57	0.54	0.59	0.65
Sexual	0.17	0.09	0.08	0.09	0.10	0.24	0.13	0.30
Ingestion	0.54	0.86	0.39	0.35	0.41	0.86	0.57	0.83
Drives	6.87	7.35	5.84	6.39	7.60	7.50	6.93	2.03
Affiliation	2.20	2.45	1.39	2.06	1.69	2.53	2.05	1.28
Achievement	1.27	1.37	0.91	0.99	1.82	1.45	1.30	0.82
Power	2.07	2.02	2.46	1.72	3.62	2.17	2.35	1.12
Reward	1.49	1.56	1.04	1.73	1.07	1.86	1.46	0.81
Risk	0.46	0.54	0.53	0.30	0.56	0.46	0.47	0.41
Time orientations								
Past focus	4.25	5.83	7.06	3.78	4.09	2.81	4.64	2.06
Present focus	10.95	10.45	6.21	15.28	5.14	11.74	9.96	2.80
Future focus	1.60	1.85	1.19	1.45	0.80	1.60	1.42	0.90
Relativity	14.23	16.19	14.56	12.12	14.47	13.99	14.26	3.18
Motion	2.15	2.58	2.34	2.20	1.70	1.94	2.15	1.03
Space	6.43	6.96	7.82	5.86	7.76	6.51	6.89	1.96
Time	5.86	7.01	4.71	4.28	5.17	5.75	5.46	1.81
Personal Concerns								
Work	2.04	2.64	1.20	2.87	4.49	2.16	2.56	1.81
Leisure	1.50	1.17	0.56	1.11	1.67	2.11	1.35	1.08

Category	Blogs	Expressive writing	Novels	Natural Speech	NY Times	Twitter	Grand Means	Mean SDs
Home	0.49	0.99	0.56	0.34	0.47	0.43	0.55	0.63
Money	0.59	0.41	0.45	0.44	1.47	0.74	0.68	0.83
Religion	0.39	0.20	0.34	0.14	0.25	0.35	0.28	0.57
Death	0.15	0.12	0.26	0.04	0.22	0.19	0.16	0.29
Informal Language	2.09	0.45	0.53	7.10	0.29	4.68	2.52	1.65
Swear words	0.35	0.09	0.05	0.25	0.02	0.49	0.21	0.37
Netspeak	0.92	0.05	0.10	1.35	0.16	3.23	0.97	1.17
Assent	0.33	0.10	0.14	3.29	0.05	1.82	0.95	0.72
Nonfluencies	0.42	0.17	0.24	1.96	0.07	0.39	0.54	0.49
Fillers	0.11	0.04	0.01	0.46	0.00	0.04	0.11	0.27
Punctuation*								
Total Punctuation	24.18	12.41	23.68	-	19.02	27.46	21.35	9.01
Periods	10.29	6.17	6.04	-	5.88	9.07	7.49	3.76
Commas	4.15	3.17	7.09	-	6.60	2.76	4.75	1.94
Colons	0.43	0.21	0.12	-	0.27	2.15	0.64	0.85
Semicolons	0.10	0.04	0.53	-	0.17	0.67	0.30	0.53
Question marks	0.59	0.15	0.60	-	0.15	1.40	0.58	1.00
Exclamation marks	1.16	0.12	0.49	-	0.02	3.21	1.00	1.35
Dashes	0.99	0.39	2.14	-	1.23	1.21	1.19	1.38
Quotation marks	0.71	0.22	3.90	-	2.23	1.30	1.67	1.36
Apostrophes	3.85	1.40	2.19	-	1.56	3.32	2.46	4.94
Parentheses	0.90	0.32	0.06	-	0.54	0.81	0.53	0.87
Other punctuation	1.00	0.23	0.52	-	0.36	1.56	0.73	1.70

Notes: Grand Means are the unweighted means of the six genres; Mean SDs refer to the unweighted mean of the standard deviations across the six genre categories.
*In calculating grand means and standard deviations for the words per sentence (WPS) and punctuation categories, the natural speech corpus was excluded due to differing transcription rules across documents.

In many ways, Table 3 points to the important role that context plays in people's use of language. Not surprisingly, the topics of writing – as reflected in the current concerns category – vary substantially as a function of genre. More striking, however, are the large differences in people's use of function words as well as punctuation from genre to genre (cf., Biber, 1988).

Comparing LIWC2015 with LIWC2007

For users of LIWC2007, 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 six corpora.

To get a sense of how much a dictionary has changed from the LIWC2007 to the LIWC2015 versions, look at the LIWC2015/2007 Correlation column. The lower the correlation, the more change across the two versions.

Table 4. Comparisons Between LIWC2015 and LIWC2007: Means, Standard Deviations, and Correlations

LIWC Dimension	Output Label	LIWC2015 mean	LIWC2007 mean	LIWC 2015/2007 Correlation[1]
Word count	WC	11,921.82	11,852.99	1.00
Summary Variables				
Analytical thinking	Analytic	56.34	-	-
Clout	Clout	57.95	-	-
Authentic	Authentic	49.17	-	-
Emotional tone	Tone	54.22	-	-
Language Metrics				
Words per sentence*	WPS	17.40	25.07	0.74
Words>6 letters	Sixltr	15.60	15.89	0.98
Dictionary words	Dic	85.18	83.95	0.94
Function Words	function	51.87	54.29	0.95
Total pronouns	pronoun	15.22	14.99	0.99
Personal pronouns	ppron	9.95	9.83	0.99
1st pers singular	i	4.99	4.97	1.00
1st pers plural	we	0.72	0.72	1.00
2nd person	you	1.70	1.61	0.98
3rd pers singular	shehe	1.88	1.87	1.00
3rd pers plural	they	0.66	0.66	0.99
Impersonal pronouns	ipron	5.26	5.17	0.99
Articles	article	6.51	6.53	0.99
Prepositions	prep	12.93	12.59	0.96
Auxiliary verbs	auxverb	8.53	8.82	0.96
Common adverbs	adverb	5.27	4.83	0.97
Conjunctions	conj	5.90	5.87	0.99
Negations	negate	1.66	1.72	0.96
Other Grammar				
Regular verbs	verb	16.44	15.26	0.72
Adjectives	adj	4.49	-	-
Comparatives	compare	2.23	-	-
Interrogatives	interrog	1.61	-	-
Numbers	number	2.12	1.98	0.98
Quantifiers	quant	2.02	2.48	0.88
Affect Words	affect	5.57	5.63	0.96
Positive emotion	posemo	3.67	3.75	0.96
Negative emotion	negemo	1.84	1.83	0.96
Anxiety	anx	0.31	0.33	0.94
Anger	anger	0.54	0.6	0.97
Sadness	sad	0.41	0.39	0.92
Social Words	social	9.74	9.36	0.96
Family	family	0.44	0.38	0.94
Friends	friend	0.36	0.23	0.78

LIWC Dimension	Output Label	LIWC2015 mean	LIWC2007 mean	LIWC 2015/2007 Correlation[1]
Female referents	female	0.98	-	-
Male referents	male	1.65	-	-
Cognitive Processes[2]	cogproc	10.61	14.99	0.84
Insight	insight	2.16	2.13	0.98
Cause	cause	1.40	1.41	0.97
Discrepancies	discrep	1.44	1.45	0.99
Tentativeness	tentat	2.52	2.42	0.98
Certainty	certain	1.35	1.27	0.92
Differentiation[3]	differ	2.99	2.48	0.85
Perceptual Processes	percept	2.70	2.36	0.92
Seeing	see	1.08	0.87	0.88
Hearing	hear	0.83	0.73	0.94
Feeling	feel	0.64	0.62	0.92
Biological Processes	bio	2.03	1.88	0.94
Body	body	0.69	0.68	0.96
Health/illness	health	0.59	0.53	0.87
Sexuality	sexual	0.13	0.28	0.76
Ingesting	ingest	0.57	0.46	0.94
Drives and Needs	drives	6.93	-	-
Affiliation	affiliation	2.05	-	-
Achievement	achieve	1.30	1.56	0.93
Power	power	2.35	-	-
Reward focus	reward	1.46	-	-
Risk focus	risk	0.47	-	-
Time Orientations[4]				
Past focus	focuspast	4.64	4.14	0.97
Present focus	focuspresent	9.96	8.1	0.92
Future focus	focusfuture	1.42	1.00	0.63
Relativity	relativ	14.26	13.87	0.98
Motion	motion	2.15	2.06	0.93
Space	space	6.89	6.17	0.96
Time	time	5.46	5.79	0.94
Personal Concerns				
Work	work	2.56	2.27	0.97
Leisure	leisure	1.35	1.37	0.95
Home	home	0.55	0.56	0.99
Money	money	0.68	0.70	0.97
Religion	relig	0.28	0.32	0.96
Death	death	0.16	0.16	0.96
Informal Speech	informal	2.52	-	-
Swear words	swear	0.21	0.17	0.89
Netspeak	netspeak	0.97	-	-
Assent	assent	0.95	1.11	0.68

LIWC Dimension	Output Label	LIWC2015 mean	LIWC2007 mean	LIWC 2015/2007 Correlation[1]
Nonfluencies	nonfl	0.54	0.30	0.84
Fillers	filler	0.11	0.40	0.29
All Punctuation*	Allpunc	21.35	21.65	0.98
Periods	Period	7.49	7.56	0.98
Commas	Comma	4.75	4.75	1.00
Colons	Colon	0.64	0.73	0.98
Semicolons	SemiC	0.3	0.29	0.97
Question marks	QMark	0.58	0.58	1.00
Exclamation marks	Exclam	1.00	1.00	1.00
Dashes	Dash	1.19	1.21	0.98
Quotation marks	Quote	1.67	1.64	0.93
Apostrophes	Apostro	2.46	2.52	0.94
Parentheses (pairs)	Parenth	0.53	0.63	0.90
Other punctuation	OtherP	0.73	0.72	0.95

* Due to differences in punctuation rules for transcriptions, the natural language corpus was excluded when computing means and correlations for punctuation categories as well as words per sentence.

[1] Correlation is the average correlation between the 2007 and 2015 dictionaries across six corpora. Low correlations (<.80) are to be expected due to the large category differences between the two versions.

[2] Cognitive processes is conceptually similar to the cognitive mechanisms LIWC2007 category. The newer cognitive process dimension restricts constituent words to true markers of cognitive activity.

[3] Differentiation is conceptually similar to the 2007 exclusive category.

[4] Time Orientation categories are similar to the 2007 categories past, present, and future but are more unified to reflect a general time orientation instead of just verb tense usage.

LIWC Dictionary Translations

The LIWC dictionaries have been translated into several languages, including Spanish, German, Dutch, Norwegian, Italian, Portuguese. Several other language translations are underway, including Arabic, Korean, Turkish, and Chinese. To date, these translations have relied on the LIWC2001 or LIWC2007 dictionaries rather than LIWC2015.

Unlike previous versions of LIWC, the current version is bundled exclusively with the original English dictionary versions. LIWC dictionary translations, as well as other published dictionaries, will be made available at the official LIWC dictionary repository (http://dictionaries.liwc.net/). If you would like to build a non-English LIWC2015 dictionary or if you have built one independently would like to add it to the repository, contact the first author at pennebaker@mail.utexas.edu.

Helpful References

Argamon, S., Koppel, M., Fine, J., & Shimoni, A. R. (2003). Gender, genre, and writing style in formal written texts. *Text*, 23, 32-346.

Argamon, S., Koppel, M., Pennebaker, J. W., & Schler, J. (2009). Automatically profiling the author of an anonymous text. *Communications of the Association for Computing Machinery (CACM)*, 52, 119-123.

Baayen, R. H., Piepenbrock, R., & Bulickers, L. (1995). The CELEX Lexical Database (Release I) [CD ROM]. Philadelphia: Linguistic Data Consortium, University of Pennsylvania.

Back, M. D., Küfner, A. C., & 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*, 837-838.

Baddeley, J. L., Daniel, G. R., & Pennebaker, J. W. (2015). How Henry Hellyer's use of language foretold his suicide. *Crisis, 32*, 288-292.

Bazarova, N. N., Taft, J. G., Choi, Y. H., & Cosley, D. (2012). Managing impressions and relationships on Facebook: Self-presentational and relational concerns revealed through the analysis of language style. *Journal of Language and Social Psychology, 32*, 121-141.

Biber, D. (1988). *Variation across speech and writing.* Cambridge: Cambridge University Press.

Boroditsky, L. (2001). Does language shape thought? Mandarin and English speakers' conception of time. *Cognitive Psychology, 43,* 1-22.

Bosson, J. K., Swann, W. B., Jr., & 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,* 631-643.

Boyd, R. L. (2015). MEH: Meaning Extraction Helper [Software]. Available from http://meh.ryanb.cc

Boyd, R. L., & Pennebaker, J. W. (2015). Did Shakespeare write *Double Falsehood*? Identifying individuals by creating psychological signatures with text analysis. *Psychological science*, *26,* 570-582.

Brewer, M. B., & Gardner, W. (1996). Who is this "We"? Levels of collective identity and self representations. *Journal of Personality & Social Psychology, 71,* 83-93.

Brown, R. (1968). *Words and things: An introduction to language*. NY: Free Press.

Bruner, J. S. (1973). *Beyond the information given: Studies in the psychology of knowing.* London: W. W. Norton.

Bucci, W. (1995). The power of the narrative: a multiple code account. In J. W. Pennebaker (Ed.), *Emotion, Disclosure, and Health* (pp. 93-122). Washington, DC: American Psychological Association.

Buchanan, L., Westbury, C., & Burgess, C. (2001). Characterizing semantic space: Neighborhood effects in word recognition. *Psychonomic Bulletin & Review*, *8*, 531-544.

Carey, A. L., Brucks, M. S., Küfner, A. C. P., Holtzman, N. S., Deters, F. G., Back, M. D., Donnellan, M. B., et al. (2015). Narcissism and the user of personal pronouns revisited. *Journal of Personality and Social Psychology, 109*, 1-15.

Campbell, R. S. & Pennebaker, J. W. (2003). The secret life of pronouns: Flexibility in writing style and physical health. *Psychological science, 14*, 60-65.

Chambers, J. K., Trudgill, P., & Schilling-Estes, N., (2004). *The handbook of language variation and change.* London: Blackwell.

Chung, C. K., & Pennebaker, J. W. (2013). Using computerized text analysis to track social processes. In T. Holtgraves (Ed.), *Handbook of language and social psychology* (pp. 219-23). New York, NY: Oxford.

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). Hershey, PA: IGI Global.

Chung, C. K., & Pennebaker, J. W. (2005). Assessing quality of life through natural language use: Implications of computerized text analysis. In W. R. Lenderking and D. A. Revicki (eds.), *Advancing health outcomes research methods and clinical applications* (pp. 79-94). Washington, DC: Degnon Associates.

Chung, C. K., & Pennebaker, J. W. (2007). The psychological functions of function words. In K. Fiedler (Ed.), *Social communication* (pp. 343-359). New York, NY: Psychology Press.

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,* 96-132.

Cohn, M. A., Mehl, M. R., & Pennebaker, J. W. (2004). Linguistic markers of psychological change surrounding September 11, 2001. *Psychological science, 15,* 687-93.

Crammer, K. & Singer, Y. (2003). Ultraconservative online algorithms for multiclass problems. *Journal of Machine Learning Research*, *3*, 951-991.

Damasio, A. R. (1995). *Descartes' error: Emotion, reason and the human brain.* NY: Harper Collins.

Davison, K. P, & Pennebaker, J. W., & Dickerson, S. S. (2000). Who talks? The social psychology of illness support groups. *American Psychologist*, *55*, 205-217.

De Choudhury, M., Counts, S., & Horvitz, E. (2013, April). Predicting postpartum changes in emotion and behavior via social media. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 3267-3276). ACM.

Feixas, G., Geldschlager, H., & Neimeyer, R. A. (2002). Content analysis of personal constructs. *Journal of Constructivist Psychology, 15,* 1-19.

Fiedler, K., & Semin, G. R. (1992). Attribution and language as a socio-cognitive environment. In G. R. Semin, and K. Fiedler (Eds.), *Language, Interaction, and Social Cognition (*pp. 58-78.) Thousand Oaks, CA: Sage Publications, Inc.

Fitzsimmons, G. M., & Kay, A. C. (2004). Language and interpersonal cognition: Causal effects of variations in pronoun usage on perceptions of closeness. *Personality and Social Psychology Bulletin, 5,* 547-557.

Foltz, P. W. (1996). Latent semantic analysis for text-based research. *Behavior Research Methods, Instruments & Computers, 28,* 197-202.

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

Gazzaniga, M. S. (2005). *The ethical brain.* NY: Dana Press.

Genkin, A., Lewis, D. D., and Madigan, D. (2006). Large-scale Bayesian logistic regression for text categorization. *Technometrics, 49*, 291-304.

Gill, A. (2003). Personality and language. The projection and perception of personality in computer mediated communication. Unpublished doctoral dissertation. University of Edinburgh, Scotland.

Gill, A. J., Oberlander, J., & Austin, E. (2006). The perception of e-mail personality at zero-acquaintace. *Personality and Individual Differences, 40,* 497-507.

Gortner, E. M., & Pennebaker, J. W. (2003). The anatomy of a disaster: Media coverage and community-wide health effects of the Texas A&M Bonfire tragedy. *Journal of Social and Clinical Psychology, 22,* 580-603.

Gottschalk, L. A. (1997). The unobtrusive measurement of psychological states and traits. In C. W. Roberts (Ed.) *Text analysis for the social sciences: Methods for drawing statistical inferences from texts and transcripts (*pp. 117-129*).* Mahwah, NJ: Erlbaum.

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

Graesser, A. C., Gernsbacher, M. A., & Goldman, S. R. (2003). Introduction to the Handbook of Discourse Processes. In A. C. Graesser, M. A. Gernsbacher, and S. R. Goldman, *Handbook of discourse processes* (pp. 1-23). Mahwah, NJ: Lawrence Erlbaum Associates.

Graesser, A. C., Lu, S., Jackson, G. T., Mitchell, H., Ventura, M., Olney, A., & Louwerse, M. M. (2004). AutoTutor: A tutor with dialogue in natural language. *Behavioral Research Methods, Instruments, and Computers, 36,* 180-193.

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,* 193-202.

Graham, L. E., Scherwitz, L., & Brand, R. (1989). Self reference and coronary heart disease incidence n the Western Collaborative Group Study. *Psychosomatic Medicine, 51,* 137-144.

Graybeal, A., Seagal, J. D., & Pennebaker, J. W. (2002). The role of story-making in disclosure writing: The psychometrics of narrative. *Psychology and Health, 17*, 571-581.

Groom, C. J., & Pennebaker, J. W. (2005). The language of love: Sex, sexual orientation, and language use in online personal advertisements. *Sex Roles, 52*, 447-461.

Groom, C. J., & Pennebaker, J. W. (2003). Words. *Journal of Research in Personality, 36*, 615-621.

Hajek, C., & Giles, H. (2003). New directions in intercultural communication competence. In J. O. Greene and B. R. Burleson (Eds.), *Handbook of communication and social interaction skills (*pp.935-957). Mahwah, NJ: Lawrence Erlbaum Associates, Publishers.

Halliday, M. A. K., & Matthiessen, C. (2004). *An introduction to functional grammar* (3rd ed.). London: Arnold.

Hart, R. P., Jarvis, S. E., Jennings, W. P., & Smith-Howell, D. (2005). *Political keywords: Using language that uses us*. NY: Oxford University Press.

Hartley, J., Pennebaker, J. W., & Fox, C. (2003). Using new technology to assess the academic writing styles of male and female pairs and individuals. *Journal of Technical Writing and Communication, 33*, 243-261.

Hartley, J., Sotto E., & Pennebaker, J. W. (2003). Speaking versus typing: A case-study of the effects of using voice-recognition software on academic correspondence. *British Journal of Educational Technology, 34*, 5-16.

Hartley, J., Sotto, E. and Pennebaker, J. W. (2002). Style and substance in psychology: Are influential articles more readable than less influential ones. *Social Studies of Science, 32*, 321-334.

Heberlein, A. S., Adolphs, R., Pennebaker, J. W., & Tranel, D. (2003). Effects of damage to right-hemisphere brain structures on spontaneous emotional and social judgments. *Political Psychology, 24,* 705-726.

Holtgraves, T. (2011). Text messaging, personality, and the social context. *Journal of Research in Personality*, *45*, 92-99.

Holtzman, N. S., Vazire, S., & Mehl, M. R. (2010). Sounds like a narcissist: Behavioral manifestations of narcissism in everyday life. *Journal of Research in Personality*, *44*, 478-484.

Ireland, M. E., & Henderson, M. D. (2014). Language style matching, engagement, and impasse in negotiations. *Negotiation and conflict management research*, *7*, 1-16.

Ireland, M. E., Slatcher, R. B., Eastwick, P. W., Scissors, L. E., Finkel, E. J., & Pennebaker, J. W. (2011). Language style matching predicts relationship initiation and stability. *Psychological science, 22*, 39-44.

Kacewicz, E., Pennebaker, J. W., Davis, M., Jeon, M., & Graesser, A. C. (2013). Pronoun use reflects standings in social hierarchies. *Journal of Language and Social Psychology*, 33, 125-143.

Kanagawa, C., Cross, S. E., & Markus, H. R. (2001). "Who am I?" The cultural psychology of the conceptual self. *Personality and Social Psychology Bulletin, 27,* 90-103.

Kashima, E. S., & Kashima, Y. (1998). Culture and language: The case of cultural dimensions and personal pronoun use. *Journal of Cross-Cultural Psychology, 29,* 461-486.

Kashima, E. S., & Kashima, Y. (2005). Erratum to Kashima and Kashima (1998) and reiteration. *Journal of Cross-Cultural Psychology, 36,* 396-400.

Koppel, M., Schler, J., & Zigdon, K. (2005, August). Determining an author's native language by mining a text for errors. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining* (pp. 624-628). ACM.

Koppel, M., Schler, J., Argamon, S., & Pennebaker, J. W. (2006). Effects of age and gender on blogging. Presented at *AAAI 2006 Spring Symposium on Computational Approaches to Analysing Weblogs*, Stanford, CA, March 2006.

Lee, Chang H., Nam, K., & Pennebaker, J. W. (2004). Is writing as much phonological as speaking? Homophone usage across speaking and writing. *Psychologia: An International Journal of Psychology in the Orient, 47*, 1-9.

Lepore, S. J., & Smyth, J. M. (2002). *The writing cure: How expressive writing promotes health and emotional well-being.* Washington: American Psychological Association.

Li, J., Zheng, R., & Chen, H. (2006). From fingerprint to writeprint. *Communications of the ACM*, *49*, 76-82.

Liehr, P., Mehl, M. R., Summers, L.C., & Pennebaker, J. W. (2004). Connecting with others in the midst of a stressful upheaval on September 11, 2001. *Applied Nursing Research, 17*, 2-9.

Liehr, P., Takahashi, R., Nishimura, C., Frazier, L., Kuwajima, I. & Pennebaker, J. W. (2002). Embodied language: Comparison of the cardiac and stroke health experience for Japanese elders. *Journal of Nursing Scholarship, 34*, 27-32

Lyons, E. J., Mehl, M. R., & Pennebaker, J. W. (2006). Linguistic self-presentation in anorexia: Differences between pro-anorexia and recovering anorexia internet language use. *Journal of Psychosomatic Research, 60,* 253-256.

Markus, H. R., & Kitayama, S. (1991). Culture and the self: Implications for cognition, emotion, and motivation. *Psychological Review, 98,* 224-253.

McAdams, D. P. (2001). The psychology of life stories. *Review of General Psychology, 5,* 100-122.

Mehl, M. R., Pennebaker, J. W. (2003). The social dynamics of a cultural upheaval: Social interactions surrounding September 11, 2001. *Psychological Science, 14,* 579-85.

Mehl, M. R., & Pennebaker, J.W . (2003). The sounds of social life: A psychometric analysis of students' daily social environments and conversations. *Journal of Personality and Social Psychololgy, 84*, 857-870.

Mehl, M. R., Robbins, M. L., & Holleran, S. E. (2012). 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*, 30-50.

Miller, G. A. (1995). *The Science of Words.* NY: Scientific American Library.

Mitchell, T. (1999). *Machine Learning*. NY: McGraw-Hill.

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*, 211-236.

Newman, M. L., Pennebaker, J. W., Berry, D. S., & Richards, J. M. (2003). Lying words: Predicting deception from linguistic style. *Personality and Social Psychology Bulletin, 29,* 665-675.

Niederhoffer, K. G. & Pennebaker, J. W. (2002). Linguistic style matching in social interaction. *Journal of Language and Social Psychology, 21,* 337-360.

Nisbett, R. E. (2003). *The geography of thought: How Asians and Westerners think differently.* NY: Free Press.

Oberlander, J., & Gill, A. J. (2006). Language with character: A stratified corpus comparison of individual differences in e-mail communication. *Discourse Processes, 42,* 239-270.

Peng, K., & Nisbett, R. E. (1999). Culture, dialectics, and reasoning about contradiction. *American Psychologist, 54,* 741-754.

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

Pennebaker, J. W. (2002). What our words can say about us: Towards a broader language psychology. *Psychological Science Agenda, 15*, 8-9.

Pennebaker, J. W., Booth, R. J., & Francis, M. E. (2007). *Linguistic Inquiry and Word Count (LIWC): LIWC2007*. Austin, TX: LIWC.net.

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

Pennebaker, J. W. & Campbell, R. S. (2000). The effects of writing about traumatic experience. *Clinical Quarterly, 9*, 17-21.

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*, 1-10.

Pennebaker, J. W. & Chung, C.K. (2005). Tracking the social dynamics of responses to terrorism: Language, behavior, and the Internet. In S. Wessely and V.N. Krasnov (Eds.), *Psychological responses to the new terrorism: A NATO-Russia dialogue* (pp. 159-170). Holland, Amsterdam: ISO Press.

Pennebaker, J. W. & Graybeal, A. (2001). Patterns of natural language use: Disclosure, personality, and social integration. *Current Directions in Psychological Science, 10,* 90-93.

Pennebaker, J. W. & Lee, Chang H. (2002). The power of words in social, clinical, and personality psychology. *The Korean Journal of Thinking and Problem Solving, 12*, 35-43.

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

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

Pennebaker, J. W., Groom, C. J., Loew, D., & Dabbs, J. M. (2004). Testosterone as a social inhibitor: Two case studies of the effect of testosterone treatment on language. *Journal of Abnormal Psychology, 113*, 172-175.

Pennebaker, J. W., & Ireland, M. (2008). Analyzing words to understand literature. In W. van Peer and J. Auracher (Eds.), *New beginnings for the study of literature* (pp. 24-48). Cambridge, UK: Cambridge Scholars Publishing.

Pennebaker, J. W., & Ireland, M. E. (2011). Using literature to understand authors: The case for computerized text analysis. *Scientific Study of Literature*, *1*, 34-48.

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

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

Pennebaker, J. W., Mehl, M. R., & Niederhoffer, K. (2003). Psychological aspects of natural language use: Our words, our selves. *Annual Review of Psychology, 54,* 547-577.

Pennebaker, J. W., Slatcher, R. B., & Chung, C. K. (2005). Linguistic markers of psychological state through media interviews: John Kerry and John Edwards in 2004, Al Gore in 2000. *Analysis of Social and Public Policy, 5*, 1-9.

Pennebaker, J. W., & Stone, L. D. (2003). Words of wisdom: Language use over the lifespan. *Journal of Personality and Social Psychology, 85,* 291-301.

Ramirez-Esparza, N., & Pennebaker, J. W. (2006). Do good stories produce good health? Exploring words, language, and culture. *Narrative Inquiry, 16*, 211-219.

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.

Robinson, R. L., Navea, R., & Ickes, W. (2013). Predicting final course performance from students' written self-introductions: A LIWC analysis. *Journal of Language and Social Psychology*, 32, 469-479.

Rochon, E., & Saffran, E. M., Berndt, R. S., & Schwartz, M. F. (2000). Quantitative analysis of aphasic sentence production: Further development and new data. *Brain and Language, 72,* 193-218.

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

Rude, S. S., Gortner, E. M., & Pennebaker, J. W. (2004). Language use of depressed and depression-vulnerable college students. *Cognition & Emotion, 18,* 1121-1133.

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*, 261-269.

Scherwitz, L., Berton, K., & Leventhal, H. (1978). Type A behavior, self-involvement, and cardiovascular response. *Psychosomatic Medicine, 40,* 593-609.

Schiller, R., Tellegen, A., & Evens, J. (1995). An idiogrpahic and nomothetic study of personality description. In J. N. Butcher and C. D. Spielberger (Eds.), *Advances in personality assessment* (pp. 1-23). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Schultheiss, O. C., & Brunstein, J. C. (2001). Assessment of implicit motives with a research version of the TAT: Picture profiles, gender differences, and relations to other personality measures. *Journal of Personality Assessment, 77, Special issue: More data on the current Rorschach controversy,* 71-86.

Scott, M. (1996). *WordSmith.* NY: Oxford University Press.

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys*, *34*, 1-47.

Semin, G. R., Rubini, M., & Fiedler, K. (1995). The answer is in the question: The effect of verb causality on the locus of explanation. *Personality & Social Psychology Bulletin, 21,* 834-841.

Skoyen, J. A., Randall, A. K., Mehl, M. R., & Butler, E. A. (2014). "We" overeat, but "I" can stay thin: Pronoun use and body weight in couples who eat to regulate emotion. *Journal of Social and Clinical Psychology*, *33*, 743-766.

Slatcher, R. B. & Pennebaker, J. W. (2006). How do I love thee? Let me count the words: The social effects of expressive writing. *Psychological Science, 17*, 660-664.

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*, 63-75.

Slobin, D. (1996). From "thought" and "language" to "thinking" for "speaking". From J. J. Gumperz and S. J. Levinson (Eds.), *Rethinking linguistic relativity* (pp. 70-96). New York, NY: Cambridge University Press.

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

Stirman, S. W., & Pennebaker, J. W. (2001). Word use in the poetry of suicidal and non-suicidal poets. *Psychosomatic Medicine, 63,* 517-522.

Stone, L. D., & Pennebaker, J. W. (2002). Trauma in real time: Talking and avoiding online conversations about the death of Princess Diana. *Basic & Applied Social Psychology, 24,* 172-182.

Stone, L. D. & Pennebaker, J. W. (2002). Trauma in real time: Talking and avoiding online conversations about the death of Princess Diana. *Basic and Applied Social Psychology, 24*, 172-182.

Stone, P. J., Dunphy, D. C., & Smith, M. S. (1966). *The General Inquirer: A Computer Approach to Content Analysis.* Cambridge: MIT Press.

Tannen, D. (1993). *Framing in discourse.* London: Oxford University Press.

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*, 24-54.

Tausczik, Y., Faasse, K., Pennebaker, J. W., & Petrie, K. J. (2012). Public anxiety and information seeking following the H1N1 outbreak: blogs, newspaper articles, and Wikipedia visits. *Health Communication*, *27*, 179-185.

Toma, C. L., & Hancock, J. T. (2010, February). Reading between the lines: linguistic cues to deception in online dating profiles. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work* (pp. 5-8). ACM.

Tumasjan, A., Sprenger, T. O., Sandner, P. G., & Welpe, I. M. (2010). Predicting elections with twitter: What 140 characters reveal about political sentiment. *ICWSM*, *10*, 178-185.

Van Petten, C., & Kutas, M. (1991). Influences of semantic and syntactic context on open- and closed-class words. *Memory & Cognition, 19,* 95-112.

Van Swol, L. M., & Carlson, C. L. (2015). Language use and influence among minority, majority, and homogeneous group members. *Communication Research, 43, 1-18*.

Väyrynen, J., & Honkela, T. (2005). Comparison of independent component analysis and singular value decomposition in word context analysis. *Proceedings of AKRR*, *5*, 135-140.

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.

Weber-Fox, C., & Neville H. J. (2001). Sensitive periods differentiate processing of open- and closed-class words: An event-related brain potential study of bilinguals. *Journal of Speech, Language, and Hearing Research, 44,* 1338-1353.

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

Williams‐Baucom, K. J., Atkins, D. C., Sevier, M., Eldridge, K. A., & Christensen, A. (2010). "You" and "I" need to talk about "us": Linguistic patterns in marital interactions.

Personal Relationships, *17*, 41-56.

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,* 8-16.

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. *Diagnostica*, *54*, 85-98.

Zijlstra, H., Van Meerveld, T., Van Middendorp, H., Pennebaker, J. W., & Geenen, R. (2004). De Nederlandse versie van de 'linguistic inquiry and word count'(LIWC). *Gedrag & gezondheid*, *32*, 271-281.

Acknowledgements

Portions of the research reported in this manual were made possible by grants from the National Institutes of Health (MH52391), National Science Foundation (IIS-1344257), the Army Research Institute (W5J9CQ12C0043), and the Templeton Foundation. Special thanks go to Cindy Chung. Cindy's mastery of language, thoughtful feedback, and valuable insights have been vital to the ongoing longevity of the LIWC project. We are also deeply indebted to a number of people who have helped with different phases of LIWC, including: Martha Francis, Laura King, Yitai Seah, Jenna Baddelley, Molly Ireland, Yla Tausczik, Matthias Mehl, Richard Slatcher, Jason Ferrell, Sam Gosling, and Gabriella Harari. We are particularly indebted to the LIWC2015 Development Team of Kiki Adams, Jennifer Caplan, Zachary Reese, Courtney Wang, and Nick Abbs.