"LIWC auf Deutsch": The Development, Psychometrics, and Introduction of DE-LIWC2015

Preprint · February 2019

DOI: 10.31234/osf.io/uq8zt

13 authors, including:

Tabea Meier
Northwestern University
14 PUBLICATIONS **53** CITATIONS
SEE PROFILE

Ryan L. Boyd
Lancaster University
111 PUBLICATIONS **3,143** CITATIONS
SEE PROFILE

James W. Pennebaker
University of Texas at Austin
385 PUBLICATIONS **57,419** CITATIONS
SEE PROFILE

Matthias R. Mehl
The University of Arizona
187 PUBLICATIONS **10,281** CITATIONS
SEE PROFILE

“LIWC auf Deutsch”: The Development, Psychometrics, and Introduction of DE-LIWC2015

Tabea Meier[1,2], Ryan L. Boyd[3], James W. Pennebaker[3], Matthias R. Mehl[4], Mike Martin[1,2,5], Markus Wolf[1], and Andrea B. Horn[1,2]

[1]University of Zurich, Dept. of Psychology, Switzerland
[2]University of Zurich, URPP “Dynamics of Healthy Aging”, Switzerland
[3]The University of Texas at Austin, Dept. of Psychology, USA
[4]The University of Arizona, Dept. of Psychology, Tucson, USA
[5]Collegium Helveticum, Zurich, Switzerland

Correspondence regarding the German adaptation of LIWC2015 (DE-LIWC2015) should be sent to Tabea Meier: t.meier@psychologie.uzh.ch and Andrea B. Horn: a.horn@psychologie.uzh.ch

The DE-LIWC2015 is part of the LIWC2015 system, and is directly available within the most recent version of LIWC2015.

Frequently Asked Questions (FAQs) for the “LIWC auf Deutsch” can be found here: https://osf.io/tfqzc/

The LIWC2015 program is a commercial product distributed by Pennebaker Conglomerates for research purposes and by Receptiviti, Inc for commercial purposes.

The official citation for the DE-LIWC2015 is:

Meier, T., Boyd, R.L., Pennebaker, J.W., Mehl, M.R., Martin, M., Wolf, M., & Horn, A.B. (2018). *“LIWC auf Deutsch”: The Development, Psychometrics, and Introduction of DE-LIWC2015.* Retrieved from https://osf.io/tfqzc/.

Table of Contents

“LIWC auf Deutsch”: The Development, Psychometrics, and Introduction of DE-LIWC2015

Introduction

The notion that the words we use in our daily lives are psychologically meaningful has existed since before psychology was formalized as a scientific field. In fact, early discoveries in expressive writing studies demonstrated that writing about emotional upheavals may promote positive psychological outcomes. In combination with findings of such improvements being tied to changes in the use of particular characteristics of language use, James Pennebaker and colleagues developed the Linguistic Inquiry and Word Count (LIWC) (Pennebaker & Francis, 1999; Pennebaker, Francis, & Booth, 2001).

While the modern study of natural language use in the psychological sciences has historically evolved from expressive writing studies (e.g., Horn & Mehl, 2004; Pennebaker & Beall, 1986), research on the “Psychology of Language” has been extended into a variety of topics in the social sciences, including emotional states, health, social interactions, personality, age, and gender (Tausczik & Pennebaker, 2010). At the same time, the digital revolution has led to an increasing availability of language data at a large scale (Boyd & Pennebaker, 2017), and language research has expanded into more modern communication contexts such as online social media (Caton, Hall, & Weinhardt, 2015; Ritter, Preston, & Hernandez, 2014; Schwartz et al., 2013; Stanton, Meston, & Boyd, 2017).

Computerized word count approaches are now more popular than ever, and their application goes beyond psychological research. The accumulation of evidence has led to new insights, and it has been recognized that function words (e.g., pronouns, articles, prepositions) are often particularly important for revealing some of the most interesting aspects of ourselves. This has led to several updates and expansions on the LIWC dictionary since its first introduction, with the most recent version being LIWC2015 (Pennebaker, Boyd, Jordan, & Blackburn, 2015).

In this manual, we introduce a new version of the German adaptation of the Linguistic Inquiry and Word Count (LIWC), called the DE-LIWC2015. The aim of the present work was to develop an update to the previous version of the German LIWC adaptation (Wolf et al., 2008) that corresponds to the LIWC2015 properties. The overall goal was to enable automated word count analysis that accurately captures the most frequently used words in the German language context, converting them into psychologically meaningful metrics for use in research. For this reason, special emphasis was placed on frequently used words across several different communication contexts (e.g., ranging from informal to formal language samples) during the dictionary development process. As with previous versions, the program is designed to efficiently and quickly analyze individual or multiple language files. Along with the DE-LIWC2015, an optional feature to automatically preprocess texts in order to correctly analyze formal second person pronouns (e.g. “Sie”, “Ihr”, “Ihnen”) is provided.

The present work aims at describing the development and psychometric properties of the DE-LIWC2015 dictionary. Further, this manual provides important information

about preprocessing steps that are needed before running LIWC analysis on German text files, such as to ensure the text files are in the right encoding (UTF-8).

For general information on the LIWC2015 framework and text analysis software, please refer to Pennebaker, Boyd, Jordan & Blackburn (2015).

Some Brief Notes on LIWC

Essentially, the LIWC program consists of two components: The processing machine and the dictionary. The dictionary represents the heart of the program as it specifies exactly what words to count. For the remainder of this manual, words that are part of the DE-LIWC2015 dictionary file will be referred to as “dictionary words”, and words in the analyzed text files will be referred to as “target words”. Collections of dictionary words that cover a particular domain (e.g., positive emotions) will be referred to as “word categories".

For every analyzed text file in LIWC2015 about 90 output variables are generated as one row of data into an output file (Pennebaker, Boyd, et al., 2015). For the DE-LIWC2015, this data includes the file name and word count, four summary language variables (analytical thinking, clout, authenticity, and emotional tone), three general descriptor categories (words per sentence, percent of target words captured by the dictionary, and percentage of words in the text that with more than six letters), 25 standard linguistic dimensions (e.g., percentage of pronouns, articles, prepositions, etc. in the texts), 40 word categories targeting at psychological dimensions (e.g., affect, cognition, biological processes, drives), six personal concern categories (e.g., work, home, leisure activities), five informal language categories (assents, fillers, swear words, netspeak), and 12 punctuation categories (periods, commas, etc.).

The Default DE-LIWC2015 Dictionary

The new DE-LIWC2015 Dictionary is comprised of 18,711 words, word stems, and a few emoticons, forming more than 80 dictionary categories. Each entry in the dictionary is mapped onto one or more categories.

es. For example, the word *“lachen”* is part of three word categories: affective processes, positive emotion and social processes. Therefore, if the word *“lachen”* is found within a target text, each of these word categories will be counted. As demonstrated in this example, the word categories in the dictionary are hierarchically arranged. As an example, all “anxiety” words are also a part of the broader “negative emotion” word category, which in turn is a part of the overall “affective processes” category as well.

Additionally, the LIWC software captures so-called word stems. Within the LIWC framework we do not mean word stems in a linguistic sense, but rather we refer to fragments of words that have carefully been selected in order to capture variants of words with the same beginning. For instance, the dictionary includes the word stem *“schönste*”*, which allows counting of any target words that start with this stem as a “comparison word" (including *"schönste"*, *"schönster"*, *"schönstes"*, etc.). The asterisk denotes the acceptance of all letters, numbers or hyphens following its position.

The DE-LIWC2015 contains all original LIWC2015 categories (Pennebaker, Boyd, et al., 2015), along with a few additional categories that are unique to the German dictionary. A comprehensive overview of all default DE-LIWC2015 dictionary categories is given in Table 1, including the corresponding output label, example words, word counts and internal consistencies.

Table 1. DE-LIWC2015 Output Variable Information [1]

Category	Output Label	Examples	Words in Category	Internal Consistency (Uncorrected α)	Internal Consistency (Corrected α)
Word Count	WC	-	-	-	-
Summary Variables					
Analytic 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	es, zu, nicht, sehr	810	0.12	0.44
Total pronouns	pronoun	ich, sie, man	174	0.23	0.64
Personal pronouns	ppron	ich, sie, ihm	68	0.27	0.69
1st pers singular	i	ich, mir mein	12	0.41	0.81
1st pers plural	we	wir, uns, unsere	14	0.46	0.84
2nd person	you_total	du, dein, dich, rrsie, rrihr, euch	30	0.36	0.77
2nd pers singular	you_sing	du, dein, dich	14	0.51	0.86
2nd pers plural	you_plur	euch, euer, ihr,	8	0.24	0.65
2nd pers formal	you_formal	rrsie, rrihr, rrihnen	8	0.28	0.70
3rd person	other	sie, ihr, ihm, deren, ihrem	24	0.31	0.73
3rd pers singular	shehe	sie, ihr, ihm	22	0.30	0.72
3rd pers plural	they	sie, deren, ihrem	11	0.31	0.73
Impersonal pronouns	ipron	man, all, manche	109	0.28	0.70
Articles	article	ein, der, die, nen	22	0.17	0.54
Prepositions	prep	ab, auf, danach	186	0.16	0.54
Auxiliary verbs	auxverb	bin, habt, geht's	161	0.17	0.55
Common Adverbs	adverb	außerdem, dabei, gar	279	0.35	0.76
Conjunctions	conj	anstatt, auch, und	87	0.18	0.57
Negations	negate	kein, nein, nichts	39	0.19	0.59
Other Grammar					
Common verbs	verb	abreist, besuchen, esse	5405	0.29	0.72
Common adjectives	adj	lange, frei, schön	5343	0.34	0.76
Comparisons	compare	ähnlich, älter, wichtiger	1910	0.24	0.66
Interrogatives	interrog	inwiefern, wann, warum	50	0.26	0.68
Numbers	number	acht, eins, halb	92	0.30	0.72
Quantifiers	quant	viel, wenig, ziemlich	259	0.24	0.65
Psychological Processes					
Affective processes	affect	glücklich, weinen	4773	0.43	0.82
Positive emotion	posemo	glücklich, liebe, schön	2243	0.42	0.81
Negative emotion	negemo	beleidigt, bösartig, heulen	2739	0.44	0.83
Anxiety	anx	ängstlich, besorgt	430	0.24	0.65
Anger	anger	hass, sauer, zorn	1014	0.45	0.83
Sadness	sad	schluchzen, träne, trauer,	691	0.31	0.73
Social processes	social	gesellig, kumpel reden	3071	0.43	0.82
Family	family	papa, tochter, tante	166	0.39	0.80
Friends	friend	bro, kumpel	124	0.14	0.49
Female references	female	frau, mädchen, weiblich,	142	0.15	0.51
Male references	male	bruder, mann, onkel	156	0.15	0.51
Cognitive processes	cogproc	denken, weil, wissen	3711	0.53	0.87
Insight	insight	denken, realisieren	960	0.26	0.67
Causation	cause	deswegen, grund	448	0.15	0.52
Discrepancy	discrep	sollte, wollte	365	0.36	0.77
Tentative	tentat	eventuell, vielleicht	463	0.42	0.81
Certainty	certain	immer, sicher	690	0.35	0.77
Differentiation	differ	aber, sonst	227	0.38	0.78
Perceptual processes	percept	fühle, höre, schauen	1447	0.24	0.65
See	see	angeschaut, sehe, sicht	354	0.19	0.59
Hear	hear	höre, klang, zuhören	308	0.28	0.70
Feel	feel	fühle, fühlt, glatt	455	0.16	0.53
Biological processes	bio	essen, blut, schmerz	1912	0.35	0.76
Body	body	arm, kopf, muskel	729	0.39	0.79
Health	health	erkältet, klinik, medikament	663	0.37	0.78
Sexual	sexual	geil, heiß, nackt	381	0.32	0.74
Ingestion	ingest	hunger, mahlzeit, pizza	444	0.43	0.82
Drives	drives	Freund, erfolg, gemobbt,	3076	0.33	0.74
Affiliation	affiliation	allianz, freund, sozial	492	0.38	0.78
Achievement	achieve	besser, erfolg, sieg	1193	0.32	0.74
Power	power	gemobbt, herrscher,	1297	0.32	0.73
Reward	reward	jubel, medaille	360	0.21	0.61
Risk	risk	gefahr, kritisch	492	0.12	0.45
Time orientations					
Past focus	focuspast	früher, gestern, war	2061	0.54	0.87
Present focus	focuspresent	aktuell, bin, heute	2037	0.28	0.70
Future focus	focusfuture	bald, später, wird	100	0.19	0.58
Relativity	relative	gegend, region, plötzlich	2991	0.44	0.82
Motion	motion	ankunft, auto, gehen	1400	0.26	0.68
Space	space	unten, über, klein	915	0.22	0.63
Time	time	ab, bisher, dauerhaft	1033	0.49	0.85
Personal concerns					
Work	work	beruf, job, hochschule	1825	0.53	0.87
Leisure	leisure	aktivität, kino, reise	715	0.32	0.74
Home	home	sofa, wohnzimmer	157	0.28	0.70
Money	money	rechnung, schuld, teuer	689	0.41	0.80
Religion	relig	fromm, kirche	338	0.31	0.73
Death	death	begräbnis, tod	272	0.33	0.74
Informal language	informal	aufm, lol, cool	775	0.36	0.77
Swear words	swear	depp, drecksack, motherfucker	244	0.30	0.72
Netspeak	netspeak	likes, lol, ok	288	0.26	0.68
Assent	assent	gell, genau, ja	45	0.18	0.56
Nonfluencies	nonflu	äh, oh, hm	33	0.14	0.50
Fillers	filler	naja, wasweißich, sozusagen	27	0.09	0.38

Note: "Words in category" refers to the number of different words and word stems that make up the dictionary category. All alphas were computed on a sample of ~38,000 texts from several different language corpora (see Table 5). Uncorrected internal consistency alphas are based on Cronbach estimates; corrected alphas are based on Spearman-Brown adjustments (see *Psychometric Properties* section below). Note that some categories are hierarchically arranged in DE-LIWC2015, but there are some exceptions to these hierarchy rules. For example, *Social processes* include a variety of words that stand for social processes, but they also include salutations as well as other words that suggest human interaction (reden, gesellig). Many of these words do not belong to any of the *Social processes* subcategories. Another example is *Informal*, which includes words (e.g. cool) that cannot be found in any of its subcategories.

[1] Please note that, for clarity of presentation, the Table is not in DINA4 format. When printing the pdf version page, scaling options can be used to adjust the oversized pages to paper size.

DE-LIWC2015: Overview of Changes

The following section describes the development of the DE-LIWC2015 dictionary. The primary goal of this process was to provide significant updates from the German LIWC2001 dictionary (Wolf et al., 2008) to better align it with the most recent English LIWC2015 dictionary. Changes in the word categories are thus largely in line with revisions reported in the English 2015 dictionary (Pennebaker, Boyd, et al., 2015).

Compared to the original German LIWC2001 translation, the DE-LIWC2015 dictionary has not only expanded in size, but changes to the categories and structure of the dictionary have also been made. The DE-LIWC2015 dictionary counts around 18,000 words arranged in 77 categories (compared to the German LIWC2001 dictionary, which counted 7,598 words across 68 categories). Furthermore, 4 summary variables that are part of the LIWC2015 system have been included. While some categories have been removed entirely (see Table 2), new word categories have been added (e.g. drives, certain function word categories), and other categories have undergone major revision (e.g. cognitive processes; see Table 3).

Table 2. Removed Categories Since German LIWC2001 (Alphabetical Order)

Communication	Other References
Down	Past tense verbs
Exclusion	Positive Feelings
Future tense verbs	Present tense verbs
Grooming	School
Humans	Self
Inclusion	Sleeping
Inhibition	Sports
Metaphysical	Television
Music	Up
Optimism	

Based on a substantial body of research suggesting that function words are particularly relevant to psychological processes (e.g. Chung & Pennebaker, 2007; Pennebaker, Chung, Frazee, Lavergne, & Beaver, 2015; Tausczik & Pennebaker, 2010), the LIWC dictionaries now include the categories of “conjunctions”, “adverbs”, “quantifiers”, “auxiliary verbs”, “verbs”, “impersonal pronouns” and “total function words”. Furthermore, “total relativity words” were added.

Additionally, some categories have been revised under new names (e.g. senses-> perception, physical-> body), or have, to a great extent, been integrated into new, broader categories (e.g. job, occupation -> work, eating-> ingestion).

For the creation of the DE-LIWC2015, unique characteristics of the German language were taken into account and changes went beyond the procedure reported

in the English version. For this reason, the DE-LIWC2015 now contains a set of pronoun subcategories that is unique to the German dictionary (see “DE-LIWC2015: Special Adaptations to German Automated Word Count Analysis” for more information). Table 3 presents an alphabetical list of categories that are either new to DE-LIWC2015, or were subject to substantial revision from their corresponding categories in the previous German dictionary.

With the advent of more powerful analytic methods and a collection of diverse language samples, we have been able to build a dictionary with more internally consistent word categories. The DE-LIWC2015 can thus be seen as an entirely new entity rather than a simple revision or update to the previous German LIWC2001 translation.[2]

Table 3. *New* and Substantially Revised Categories in the DE-LIWC2015 (Alphabetical Order)

Achievement	*Interrogatives*
Adjectives	*Male references*
Adverbs	*Netspeak*
Affiliation	Past focus words
Auxiliary verbs	*Personal pronouns* (including various German-specific subcategories)
Cognitive processes	*Power*
Comparisons	Present focus words
Conjunctions	*Quantifiers*
Differentiation	*Reward*
Drives	*Risk*
Female references	Time
Future focus words	*Total function words*
Impersonal pronouns	*Total relativity*
Informal language	*Verbs*

Note. Categories in *italics* are newly added categories since the last version, and categories in non-italics have undergone substantial revision. While some of the other LIWC categories may have been altered as well, the here presented categories are the ones that have undergone major revision compared to the previous version of LIWC.

[2] Please note that the LIWC2015 application comes with the original dictionaries for both the German 2001 and 2015 version. It is possible to choose the old dictionary for those who would like to rely on some of the old categories, or to compare analyses of the LIWC2015 dictionary with those provided by older versions.

DE-LIWC2015: The Dictionary Development Process

We here present a complete overview of the development process of the DE-LIWC2015 dictionary. While the process had many stages and was, in part, recursive in nature, the overall process can roughly be divided into 5 broad stages:

1. Initial translation / Development
2. Dictionary expansion
3. Psychometric evaluation
4. Refinement / Revision
5. Addition of summary variables

1. Initial Translation / Development

For the development of the new DE-LIWC2015, we used the previous German LIWC2001 version as starting point. In order to add new words, we relied on a combined approach that included both translation from the English LIWC2015 dictionary, as well as an inductive approach focusing on word base rates. Beyond these inductive approaches, for many of the steps, expert judge ratings decided upon the conceptual fit and ultimate inclusion of new words. All judges involved in this process were native German speakers with a research background in Psychology.

All words from the *content word* categories that had been newly added into the English LIWC 2015 dictionary since the 2001 version were directly translated into German. To make this first step as efficient as possible, we translated words using Google Translate (https://translate.google.com). We then added these machine-translated words into the existing German LIWC 2001 dictionary. The goal of this initial step was to obtain a crude working version of a new German dictionary in the most efficient way.

For the new *function word* dimensions (e.g. conjunctions, prepositions, adverbs, impersonal pronouns, personal pronouns, interrogatives and auxiliary verbs), we added new words after consulting available lists of closed word forms, such as the Stuttgart-Tübingen tagset (Schiller, Teufel, Stöckert, & Thielen, 1999).

We then performed an extensive manual revision on this updated working dictionary. Spelling mistakes and obvious conceptual mismatches resulting from machine translation were omitted in this prior expert rating phase.

Unlike the English dictionary, we established the personal pronoun categories according to the characteristics of the German language and included subcategories that are not part of the English LIWC2015 (e.g. “formal_you”,“other”, “you_singular”, “you_plural”). This procedure was necessary to accomodate the way personal pronouns are structured in German and aimed at minimizing misclassifications, i.e. homographs that are counted in more than one pronoun category: Since LIWC is not sensitive to capitalization, it can inherently not distinguish between *“sie”* as a third

person pronoun and “Sie” as a second person pronoun. To address this issue, we provide an optional, automatic preprocessing possibility in order to analyze formal *you-talk* (e.g. *“Sie”*) in German text samples (see section “DE-LIWC2015: Special Adaptations to German Automated Word Count Analysis”). The implementation of different pronoun subcategories allows for a clearer distinction and a more reliable analysis of pronoun use in German samples.

The preprocessing feature and the new pronoun subcategories of the DE-LIWC2015 are described in detail in one of the sections below (“DE-LIWC2015: Special Adaptations to German Automated Word Count Analysis”).

2. Dictionary Expansion

Step 1: Revision of existing dictionary. Once we obtained a working template of the new German dictionary, we conducted major changes to the internal composition of the dictionary. This involved multiple revision cycles on word stems. Our goal was to remove any overlap between word stems and other dictionary entries. For example, we replaced the previous entry *"Abend*"* with entries such as *"Abend"*, *"abends"*, *"Abendessen*"* and *"Abendmahl"*. While the word *"abends"* is counted as an adverb and a time reference, the word *"Abendmahl"* is counted as a religious word. The removal of the very general word stem *"Abend*"* in this example allows for a more precise tracking of words by preventing an overlap between different entries. For every existing overlap in the dictionary, we replaced the word stem (*) with different inflections of the corresponding word.

Verbs and adjectives were greatly affected. For this reason, we obtained a list of the 200 most frequent adjectives and verbs from two available German text corpora: the deTenTen corpus (Jakubíček, Kilgarriff, Kovář, Rychlý, & Suchomel, 2013) and the Europarl corpus (Koehn, 2005). We searched for verbs and adjectives that had not yet been part of the LIWC dictionary, inflected them in all of their forms and included them into the dictionary.

This revision on the asterisks/word stems was a necessary step due to the way the LIWC processing component works internally, marking a notable improvement over the previous German LIWC dictionary. After all of the overlaps in word stems had been removed, we continued the word collection process.

Step 2: Base rate analysis and candidate word list generation. In this step, we paid particular attention to word base rates, with the aim to include the most common words of the German language into the dictionary. We explored several sources of language for high-frequency words that our judges had not yet added. For this purpose, we collected a large German corpus of different language modalities (formal, informal, spoken and written) from different sources. An overview of the text samples used for the dictionary development process is depicted in Table 5.

Using the Meaning Extraction Helper (Boyd, 2017a), we identified high-frequency German words from these text sources. The 2,000 most frequent words from each of these sources were compared to the working version of the DE-LIWC2015 dictionary in order to detect common words that were not yet included in the dictionary. We placed words that were not yet captured on a list of candidate words for possible inclusion. A judge subsequently reviewed the list, and rated whether 1) the words should be included in the dictionary and 2) whether words made a sound conceptual

fit for each dictionary category. Subsequently, the judge thoroughly reviewed the whole dictionary again, checking for every dictionary entry whether any additional word categories would apply.

Step 3: Word similarity analysis. In order to complete the dictionary expansion phase, we searched for common words that may have been overlooked. Based on the collected German corpus covering different communication contexts (Table 5), we analyzed word similarity of existing words in the dictionary using latent semantic analysis (LSA) (Landauer & Dumais, 1997) to identify additional candidate words for inclusion into the dictionary. We did this for every category we wanted to further expand. For example for the category “risk”, we searched for words that are similar to those that made up this category at that point but were not yet part of the dictionary, and placed them on a candidate list for inclusion. Following a judge rating identical to that described above, we made a decision for each candidate word regarding its inclusion in the dictionary. After having established the preliminary dictionary, it was scanned multiple times for spelling errors.

3. Psychometric Evaluation

Once we completed all previous steps, we separated each word category into its constituent words (see Pennebaker, Boyd, et al., 2015). Then, we quantified each word as a percentage of total words for 38,900 text files stemming from six corpora (see Table 5). For each word category, we treated all words as a "response" and used them to compute internal consistency measures for each language category as a whole. The psychometric evaluation procedures are discussed in detail in the next section “DE-LIWC2015: Psychometric properties”.

Words that undermined the internal consistency of their overarching word category were placed on a candidate list of words for omission from the final dictionary. We did not consider words for omission if they represented a completion of a word group of which the rest was not reducing internal consistency (i.e. inflections of a certain verb or adjective). For example, if the counted frequency of the word "albernen" correlated negatively with the determined total frequency of its assigned word categories, but other variants of this word (e.g. albern, alberner, albernsten) made an acceptable fit with the same category, we still retained the word "albernen" in order to have a complete representation of "albern" and its inflected forms. We then reviewed the list of candidate words and, again, individually judged each word for potential exclusion.

The internal consistencies of most word categories were satisfactory in the initial psychometric evaluation and, as such, a conservative approach in the judge rating / word exclusion phase was adopted. We excluded words that constituted an obvious conceptual mismatch or with very low base rates (not occurring in any of the six corpora) from the dictionary in this process. Changes made in this phase were relatively minor.

4. Refinement / Revision

After phases 1 through 3 were complete, we partially repeated them in an iterative fashion to catch any possible mistakes or oversights that may have occurred in the dictionary development process. Furthermore, in an expert discussion round, some decisions had to be made regarding how to deal with particular challenges unique to

automated word count analyses in German. This included, for example, conventions on how to deal with German homographs (e.g. verbs that can also be adjectives or nouns). All of these challenges and decisions are described in detail in the section below “DE-LIWC2015: Special Adaptations to German Automated Word Count Analysis”.

Note that the psychometrics of word categories changed negligibly over the course of the refinement phases. In a final refinement phase, an individual judge who had never seen the dictionary before reviewed the whole dictionary. Changes suggested by this judge were put on a candidate list for adaptation. In a final expert rating, three judges decided upon these final adaptations. We conducted psychometric evaluation again to achieve the psychometric properties of the final dictionary.

5. Addition of Summary Variables

Upon completion of the final DE-LIWC2015 dictionary, the 4 summary variables corresponding to those in the English LIWC2015 version (Pennebaker, Boyd, et al., 2015) were integrated into the LIWC system: Analytical thinking (Pennebaker, Chung, et al., 2015), Clout (Kacewicz, Pennebaker, Davis, Jeon, & Graesser, 2013), Authenticity (Newman, Pennebaker, Berry, & Richards, 2003), and Emotional Tone (Cohn, Mehl, & Pennebaker, 2004). We computed each summary variable on the basis of previously published research and converted them to percentiles based on standardized scores from large comparison samples.[3]

DE-LIWC2015: Special Adaptations to German Automated Word Count Analysis

The following section provides an overview of certain aspects of the German language that required special attention during the DE-LIWC2015 dictionary development process. Some of these characteristics resulted in the introduction of categories that are unique to the DE-LIWC2015 (a set of pronoun subcategories). Moreover, we here present an optional, automatic preprocessing feature to track formal second person pronouns (e.g. “*Sie”, “Ihr”, “Ihnen”*).

Personal pronouns: Subcategories and polite form (optional preprocessing)

Pronouns in the German language are often homographs: Words in the same spelling are used with different meanings, which provokes several challenges to a word count program such as LIWC. It is important to keep in mind that LIWC does not distinguish between uppercase and lowercase spelling. The word *“sie”* then can refer to either “she” but also “they” or even “you”, when used in the polite (capitalized) form.

For this reason, the DE-LIWC2015 contains several you-subcategories that are illustrated in Table 4 and that are not part of the English LIWC. Please note that with these partitionings of “you” into several subcategories, a more reliable tracking of *you-talk* has been enabled. However, a few ambiguities in certain personal pronoun

[3] We note that the summary variables are the only nontransparent dimensions in the LIWC2015 output. Whereas the output for all other LIWC categories are either percentages or raw counts, the summary variable algorithms are more complex.

categories remain inevitable due to the way German pronouns are structured. In the following paragraphs, we provide an overview of the different pronoun subcategories unique to DE-LIWC2015 and point out a few things you should be aware of.

Table 4. Overview of new DE-LIWC2015 Personal Pronoun Categories

You_Total	You_Sing	You_Plur	You_Formal	Other	SheHe	They
dein	dein	euch	rrihnen	deren	deren	deren
deine	deine	euer	rrihr	derer	dessen	derer
deinem	deinem	eure	rrihrem	ihr	einem	ihnen
deinen	deinen	eurem	rrihren	ihnen	er	ihr
deiner	deiner	euren	rrihrer	ihre	ihm	ihre
deines	deines	eures	rrihres	ihrem	ihn	ihrem
deines-gleichen	deines-gleichen	eures-gleichen	rrihres-gleichen	ihren	ihr	ihren
deinig*	deinig*	ihr	rrsie	ihrer	ihre	ihrer
denkste	denkste			ihres	ihrem	ihres
dich	dich			ihres-gleichen	ihren	ihres-gleichen
dir	dir			sie	ihrer	sie
du	du			dessen	ihrerseits	
euch	haste			einem	ihres	
euer	siehste			er	ihres-gleichen	
eure				ihm	sein	
eurem				ihn	seine	
euren				ihrerseits	seinem	
eures				sein	seinen	
eures-gleichen				seine	seiner	
haste				seinem	seinerseits	
ihr				seinen	seines	
rrihnen				seiner	sie	
rrihr				seinerseits		
rrihrem				seines		
rrihren						
rrihrer						
rrihres						
rrihres-gleichen						
rrsie						
siehste						

Note. The word «ihr» remains part of both the «You_Plur», «SheHe» and «They» category.

Total_You. The “total_you” contains all entries of the you-subcategories. It is what is most comparable to the English “you”-category, since in English “you-words” may refer to either singular, plural, informal or formal addressing of people.

You_Sing. Splitting up “you” into different subcategories enabled us a precise capturing of “singular *you-talk*” - the direct (informal) addressing of a single person –

which is typically the case in couples' conversations for example. The words in the "you sing"-category do not overlap with any of the other ppron-subcategories.

You_Plur. In a similar way, the "you plural"-category, counts "you-talk" when talking to more than one speaker. One important thing to note here is that the word *"ihr"* is part of this category, but at the same time is also part of the *"she/he"* and *"they"* category. For this reason, the "total_you"-category is also slightly affected by this ambiguity. For a word count program, it is not possible to distinguish these different forms, as this only becomes clear in context. Every time DE-LIWC2015 encounters the word *"ihr"*, it will always get counted in all of these categories. Depending on the language context, the use of "you-plural", "she/he" or "they" is more likely to occur in your samples.

You_Formal (preprocessing required). As a novelty, the DE-LIWC2015 comes with an extra category for the polite form ("you_formal"). However, since LIWC does not discriminate between upper- and lower-case spelling (and would inherently not distinguish between *"Sie"* = formal "you") and *"sie"* = "she", "they"), a special solution is required.

For this reason, we provide an optional, preprocessing possibility for the formal you-forms, in which all entries will automatically be recoded into "*rrSie*", "*rrIhnen*" etc. Note that this is an optional feature and may only be of interest to you *if* you have samples containing formal conversation (e.g. psychotherapy transcripts, books) *and if* you are interested in the use of personal pronouns.

An automated and user-friendly solution for the preprocessing is provided by the program *TextEmend* (Boyd, 2018b). The *TextEmend* software can be freely downloaded from the following location:

https://toolbox.ryanb.cc

The accompanying pre-processing script for DE-LIWC2015 can be downloaded from the following location:

https://dx.doi.org/10.17605/OSF.IO/G9M5N [4]

The program and associated pre-processing script follow a conservative approach for the recoding, in which the capitalized forms are only recoded if they are *not* at the beginning of a sentence. This procedure was chosen to reduce ambiguity between capitalized formal *you-talk* and other pronouns such as *"sie"* at the beginning of sentences. Note that in some occasions, people capitalize the word *"ihr"* even when informally referring to a group of people. Depending on the goals of your analyses, this may or may not be relevant. Should you be merely interested in "you-talk" generally in your analyses, this will not represent an issue, since the capitalized *"ihr"* will be counted in "you_total" anyway. Should you have a specific interest in *only* formal *you-talk*, the word "ihr" as an informal reference to *a group* of people should be transcribed in lower-case spelling, since TextEmend would otherwise recode it into "rrIhr" which would then be detected as "you_formal" by LIWC.

She/he and they. What remains the trickiest in German pronoun analyses are the "she/he" and "they" categories. As you can see in Table 4, a lot of words overlap between these two categories. However, splitting *you-talk* in its subcategories and

[4] Note that it is not necessary to use the *TextEmend* software to use the DE-LIWC2015 pre-processing script. Replaces are made using regular expressions, which can be performed using other mainstream scripting languages such as R and Python if you prefer.

preprocessing of the transcripts to separately track “you_formal”, has improved these two categories, since overlaps between them and “you_formal” were omitted.

Other. Since overlaps between the “she/he” and “they”-categories could not fully be ommitted, we include an overall category “other”, which contains all words of its two subcategories “she/he” and “they”. Thus, this overall category contains pronouns people use when they talk about other people, which are better distinguishable from other ppron-subcategories.

Verbs, adjectives, and time focus

As a change to the 2001-version of the DE-LIWC, the new version now contains categories such as “verbs” and “adjectives”. However, one issue about German verbs is that some of them are not clearly distinguishable from other grammatical categories (e.g. adjectives or nouns). For example the word *“glauben”* can be used both as a verb and a noun (remember that LIWC is not sensitive to capitalization), and *“reserviert”* can either be an adjective or a verb (third person singular present, or past participle).

During the dictionary development process, arbitrary decisions had to be taken about these ambivalent words regarding their assignments to the linguistic categories in the DE-LIWC2015. The different issues that occurred and the chosen solutions are exemplified in the following.

Verb and noun overlap. In LIWC2015, there is no separate category for nouns. However, since some verbs are homographs with commonly used nouns (e.g. *“Vertrauen”*), decisions on these cases had to be taken. In cases, in which the verbs were homographic with the mainly used nouns of the same word, these words were *not* included in the verb category. This was true for words such as *“vertrauen”, “vermissen”, “betrug”, “glauben”, “leben”,* to name a few. Only the conjugated forms of these verbs, (e.g*. “lebe”, “lebte”*), for which the spelling is distinct from the commonly used nouns were included in the verb category.

However, many verbs *could* be used as a noun, but do not represent the most commonly-used form of the noun. These cases remained part of the verb category. For example the word “*trauern*” is part of the verb category, as it refers to an action, and it has a more commonly used noun form “*Trauer*”. An analogous procedure was chosen for many other infinitives of verbs, for which there exists an alternative noun version (often marked by a suffix, e.g. –ung). As an example the words “*beleidigen*” or “*überraschen*” remained part of the verb category, since they refer to the action and their alternative substantive “*Beleidigung*” and “*Überraschung*” represent the more typically used noun forms.

This can be seen as a conservative approach to count verbs. By omitting the most ambivalent words in the verb category (those that are also frequently used as nouns), we aimed at a more reliable count of German verbs.

Verb and adjective overlap. In addition to the verb/noun overlaps, some verb forms are furthermore not clearly distinguishable from adjectives. For example, the word “*organisiert*” (organized/organizes) is a verb (third person singular *or* past participle), but it can also be used as an adjective. Words that can be both adjectives and verbs were *only* included in the *verb* category, due to their predominant role as verbs and to prevent redundant double-codings between adjectives and verbs.

For the declined forms of these words (i.e *“organisierte”, “organisierter”, “organisiertes”, “organisierten”, “organisiertem”*), however; these words were included into the adjectives category, as long as they did not represent homographs of a verb. For the declined forms of our example “*organisiert”*, this means that *“organiserter”*, *“organisertes”* and *“organisiertem”* are part of the adjectives category, while *“organisierte”* and *“organisierten”* are only part of the verbs category, due to their predominant use as verbs (past simple third person singular, and first person plural).

Focuspast and focuspresent overlap. As a general rule, the focus past category contains conjugated past tense verb forms and perfect participles in addition to other words referring to the past (e.g. “*gestern”*). As an additional issue, however, words such as “*organisiert*” or “*vertraut*” may represent both present tense form (third person singular) and past tense (as perfect participle). To reduce ambiguity in the time focus categories, such words were included neither in the focuspast nor in the focuspresent category. Other ambiguous words such as “o*rganisierte*”, “*organisierten*” or “*verärgerte*”, which may be used as adjectives in addition to verbs, were also excluded from the focus past category. Perfect participles (which usually have the prefix –ge) were included in the focus past category, but only their base forms and not their declined forms that are sometimes used as adjectives: “*gelebt”* is part of the focus past category, but not its inflected forms “*gelebte”, “gelebtes”* etc. Perfect participles that are synonymous with commonly used nouns (e.g. “*gefallen”*) were not included in the past focus category either.

Focuspresent and focusfuture overlap. The focuspresent category contains conjugated verb forms in the present tense, as well as some other words referring to the present (e.g. “*heute”, “jetzt”*).

In order to reduce ambiguity, only the conjugated verb forms that do not overlap with any other verb tense were included. As a consequence, first person plural forms are not part of the focuspresent category, as they are usually synonymous with the infinitive, which is used in other tenses as well, such as future tense constructions. For this reason, infinitives are not part of the focusfuture category either. The focusfuture category contains words conceptually referring to the future (e.g. “*künftig”*, “*später”*, “*bald”*, “*vorausblicken”*), as well as conjugations of verbs commonly used in future tense construction (e.g. “*werden”*).

Comparison words. The comparison word categories contain the comparatives of adjectives as well as some words generally referring to comparison (e.g*. “ähnlich”*). Regarding the comparatives, only the less ambiguous inflected forms of the comparatives were included into the comparative category. The uninflected forms of comparison words, i.e. masculine adjectives (e.g. “*schöner”, “kleiner”*) are not part of the comparison category due to their ambiguity: “*Schöner”* can refer to a comparison (“Jenes Bild ist *schöner* als das hier”), but it is also frequently used as an adjective preceding a masculine noun (ein *schöner* Mann). Only “*schönere”, “schöneres”, “schönerem”, “schöneren”* and “*schönste*”* are counted as comparison words in this example.

DE-LIWC2015: Psychometric properties

As has previously been described by Pennebaker, Boyd, et al. (2015), assessing the reliability and validity of a text analysis program is not trivial. One might think that the internal reliability of a LIWC scale can be determined the same way as it is done with a questionnaire. However, in a questionnaire designed to capture anger or loneliness, for example, participants complete a self-report asking a number of questions about their construct-relevant feelings and behavior. Reliability coefficients can then be computed by simply correlating peoples' answers to the various questions. The higher they correlate, the higher is the probability that the questionnaire items all measure the same thing- thus, leading us to the conclusion that we can consider the scale as internally consistent.

Although we can use a similar strategy with words, be warned: the psychometric properties of natural language use are much less straight-forward than with questionnaires and this is for a very obvious reason. Once we say something, we usually do not need to say it over and over again in the same paragraph or essay, and particularly not using the same words. Instead, the nature of discourse is that we say something and then move on to using new words, resulting in something of a "sparse" distribution of single-word measures. Repeating the same idea over and over again may be at the core of a self-report questionnaire design, but in language, too many repetitions generally count as bad form. It is thus important to understand that acceptable boundaries for natural language reliability coefficients are lower than those typically used elsewhere in psychological tests.

As an example, the DE-LIWC2015 "anger"-category consists of 1,014 anger-related words and word stems. In theory, we expect that the more a person uses one type of anger word in a given text, the more she should use other anger words in the same text. To test this assumption, we can determine the degree to which people use each of the 1,014 anger words across a selected group of text files and then calculate the intercorrelations of the anger-word use. In Table 1, we indeed present these internal reliability statistics, including those of the "anger"-category, where the alpha reliabilities range between .45 (uncorrected) and .83 (corrected) depending on the way it is computed.

In line with the procedure suggested by Pennebaker, Boyd, et al. (2015), we presented each dictionary word as a percentage of total words per text in order to calculate internal consistency statistics for the DE-LIWC2015. We then treated these scores as "items" in a standard Cronbach's alpha calculation, providing raw alpha scores for each word category, separately for each text corpora used. Uncorrected alphas in Table 1 represent averages of each corpora's alpha score. What is important to note, however, is that the uncorrected method tends to largely underestimate reliability in language categories due to the highly variable base rates of word usage within any given category. The Spearman-Brown prediction formula (Brown, 1910; Spearman, 1910) was used to compute the corrected alphas, which generally represent a more accurate approximation of the "true" internal consistency of each category.

Throughout the last decades, LIWC has increasingly been used in research across a variety of disciplines and languages. Correlates of LIWC-variables go beyond psychological self-report measures and extend onto human coded social interaction and even biomarkers such as gene expression. Some recent examples of topics

studied using LIWC include depression / negative emotionality (Tackman et al., 2018), dyadic processes and coping (Karan, Wright, & Robbins, 2017; Neysari et al., 2016; Robbins, Karan, Lopez, & Weihs, 2018), individual differences and personality (Carey et al., 2015), gender and translation (Meier et al., in prep), psychopathology / psychotherapy research (Friederich et al., 2017; Sonnenschein, Hofmann, Ziegelmayer, & Lutz, 2018; Wolf, Chung, & Kordy, 2010; Wolf, Sedway, Bulik, & Kordy, 2007; Wolf, Theis, & Kordy, 2013), pregnancy / prenatal stress (Schoch-Ruppen, Ehlert, Uggowitzer, Weymerskirch, & La Marca-Ghaemmaghami, 2018) and gene expression (Mehl, Raison, Pace, Arevalo, & Cole, 2017). A nice introduction to LIWC research is given by the following book chapters: Boyd and Pennebaker (2015); Boyd (2017b). We also provide a list of selected LIWC contributions, particularly from the German language context, in the reference list at the end of this document. The vast amount of LIWC-based research implies the validity of the approach to measure psychological properties of an individual from their language.

Base Rates of Word Usage

When analyzing people's word use, it can be extremely helpful (if not essential) to know the degree to which language measures vary across settings and contexts (Pennebaker, Boyd, et al., 2015). It has for example been found that psychological correlates of word use can vary between public or interpersonal language (conversations with other) versus private language (stream-of-consciousness essays) (Mehl, Robbins, & Holleran, 2012; Nowson & Gill, 2014).

For comparison purposes, we analyzed text from several of our studies and other available sources covering different language contexts (written vs. spoken, formal vs. informal) using the updated DE-LIWC2015 dictionary. Below, we provide a brief description of the used text datasets. As presented in Table 5, these text samples reflect the utterances of over 34,000 writers or speakers, with a total language output of over 294 million words.

Table 5. Summary Information for DE-LIWC2015 Statistics

	Formal natural speech	Informal natural speech	TED talks	Expressive writing	Informal writing (Reddit)	Informal writing (Twitter)
Total *N*	3,397	239	2,147	965	17,040	15,112
Total authors	unknown	100	1,733	647	17,040	15,112
Total words	48,361,001	214,481	4,039,559	204,706	84,046,562	157,420,344

Note. For all corpora, texts required a minimum of 100 words to be included in our analyses. We ommitted all texts with fewer than 100 words for all statistics reported in this document.

Formal natural speech. Texts in this dataset came from the German samples of the German—English European Parliament Proceedings Parallel Corpus (Europarl; Koehn, 2005), which consists of the ongoing proceedings of the European Parliament since 1996. While all text files in this corpus used for statistical analysis contained multiple speakers, speaker information was not preserved for this dataset and the true number of speakers is unknown.

Informal natural speech. These speech samples came from the German transcripts of the CALLHOME corpus (Karins, MacIntyre, Brandmair, Lauscher, & McLemore, 1997), which comprises recordings of unscripted phone call conversations of native German speakers.

TED Talks. We obtained transcripts of English TED talks from the official TED talk website (https://www.ted.com), along with their corresponding German translation as well as speaker and transcriber information (where available). Some speakers gave more than one talk, while, at the same time, some translators provided multiple translations.

Modern internet platforms build new opportunities to share insights to a broad audience. With its slogan "ideas worth spreading", the TED talks conference is a popular example of a knowledge spreading opportunity in the digital age to share insights to a broad audience. In this format, academics, entrepreneurs, artists and a variety of other individuals give short talks about their area of expertise, with the videos subsequently being hosted and publicly available on the TED website. Language registered in this corpus can best be considered as somewhere between formal and informal, but uniquely non-spontaneous.

Expressive writing. This dataset consisted of four German samples from both cross-sectional and longitudinal studies (Horn, Holzgang, & Rosenberger, 2017; Horn, Kneisler, Schuster, & Traue, 2010; Horn & Maercker, 2016; Horn, Pössel, & Hautzinger, 2011), in which individuals wrote about deeply personal topics in a stream-of-consciousness manner. Writings ranged from young adolescents who took part in a depression prevention intervention, older adults who wrote about their recent transition to retirement, to depressed patients and individuals who wrote about their romantic relationships. In the study about depression, patients were assigned to write either about time management or an emotional topic. We included texts from both conditions in the current analyses.

Reddit. Reddit is a widely used US-American based social media forum that is gaining popularity in the German-speaking context as well. On this website (https://www.reddit.com/), users can share their posts, which are organized by topic into “subreddits”.

We collected user-made posts from the largest German-speaking subreddit (/r/de), which is primarily oriented toward nation-related discourse for individuals in Germany. Users discuss and exchange ideas about various topics (e.g. sports, politics, leisure) in a threaded, forum-style manner. All posts from this subreddit (approximately 147,000 submissions and ~1.3 million comments) were collected in October, 2017. We then aggregated all posts by user and automatically screened them using automated techniques (“langdetect,” n.d.) to ensure that the majority of a person’s posts were in German. This resulted in the collected posts of over 17,000 German-speaking Reddit users made to the /r/de subreddit.

Twitter. Much like the Reddit corpus, Twitter reflects informal, netspeak language, but due to the nature of twitter post length limits, individual posts are usually somewhat shorter than in Reddit. A script was created to listen to the Twitter “garden hose” API for public Twitter posts (i.e., "tweets") made in the German language.[5] Once a tweet was detected that fit the language criteria, the Twitter Search API was

[5] Tweets were automatically tagged for language by Twitter using their own internal rather than local models.

used to collect other German-language tweets from the same user, going as far back into their public timeline as the API allowed (approximately a maximum of 3,200 tweets). We conducted basic pre-processing of each tweet, removing hashtags, urls, and user mentions (i.e., @'s). Similar to the Reddit corpus, we then aggregated all tweets, by user, into single units for psychometric analysis.

As reported in Table 6, the DE-LIWC2015 dictionary captures, on average 83 percent of the words people use in written and spoken language. Note that, except for total word count, words per sentence, and the four summary variables (Analytic, Clout, Authentic, and Tone), all means in Table 6 are expressed as percentage of total words used in any given language sample.

Scores reported in Table 6 highlight once more what has been emphasized previously – namely the importance of considering context when studying people's language use (e.g. Biber, 1988; Mehl et al., 2012; Nowson & Gill, 2014; Pennebaker, Boyd, et al., 2015). Variability in frequency as a function of language context can be observed not only for certain content word categories, but also for function words and even punctuation markers.

Table 6. DE-LIWC2015 Output Variable Information [6]

Category	Informal natural speech	Formal natural speech	TED Talks	Expressive writing	Reddit	Twitter	Grand Mean	Grand Mean SD
Word Count (mean)	897.41	14,236.39	1,881.49	212.13	4,932.31	10,416.91	5,429.44	9,245.24
Summary Variables								
Analytic Thinking	7.55	94.91	54.52	16.69	52.20	71.28	49.53	20.62
Clout	55.64	76.14	74.36	43.53	51.16	62.96	60.63	14.86
Authentic	39.64	40.42	46.42	76.73	35.09	51.75	48.34	24.41
Emotional tone	60.58	57.40	66.94	59.66	44.21	78.43	61.20	27.69
Words/sentence	7.32	19.96	14.52	17.56	15.55	46.18	20.18	119.94
Words > 6 letters	13.37	34.98	25.58	17.95	21.44	24.10	22.90	4.15
Dictionary words	92.19	85.00	85.97	88.20	73.23	71.73	82.72	6.93
Linguistic Dimensions								
Total function words	58.20	51.87	55.64	59.11	47.57	42.04	52.41	6.11
Total pronouns	16.42	11.10	17.71	21.08	12.56	10.38	14.87	3.35
Personal pronouns	9.53	5.91	9.97	15.55	6.32	5.77	8.84	2.78
1st pers singular	4.03	1.74	2.99	9.51	2.95	2.64	3.98	2.17
1st pers plural	1.03	1.76	2.32	1.93	0.40	0.77	1.37	1.12
2nd person (you_total)	2.16	0.43	1.18	0.34	1.10	1.29	1.08	0.97
2nd pers singular	1.78	0.00	0.22	0.07	0.89	0.79	0.62	0.71
2nd pers plural	0.37	0.05	0.16	0.24	0.16	0.33	0.22	0.36
2nd pers formal	0.01	0.38	0.81	0.04	0.05	0.17	0.24	0.39
3rd person (other)	2.26	1.59	3.03	3.60	1.48	0.94	2.15	1.27
3rd pers singular	2.25	1.54	2.93	3.56	1.45	0.94	2.11	1.25
3rd pers plural	1.14	0.88	1.84	1.32	0.63	0.44	1.04	0.92
Impersonal pronouns	4.14	4.20	6.18	4.41	4.80	3.47	4.53	1.44
Articles	9.46	13.94	10.83	6.38	8.82	7.60	9.51	2.13
Prepositions	6.06	12.62	10.95	9.36	8.55	10.72	9.71	2.20
Auxiliary verbs	11.21	7.19	8.49	10.71	8.58	7.04	8.87	2.11
Common Adverbs	6.11	3.71	4.49	5.75	4.99	4.25	4.88	1.57
Conjunctions	17.43	10.15	12.46	14.21	12.25	9.10	12.60	2.49
Negations	3.18	1.06	1.34	2.08	2.28	1.53	1.91	0.94
Other Grammar								
Common verbs	17.12	13.72	16.18	17.66	13.91	12.55	15.19	2.91
Common adjectives	5.96	6.01	7.18	7.66	6.33	6.69	6.64	1.79
Comparisons	1.87	1.63	2.52	2.28	2.25	1.63	2.03	0.86
Interrogatives	1.87	0.78	1.63	1.30	1.47	1.29	1.39	0.73
Numbers	1.44	1.76	1.80	1.62	1.77	3.60	2.00	2.06
Quantifiers	2.65	2.13	3.10	4.10	3.05	2.32	2.89	1.08
Psychological Processes								
Affective processes	4.45	5.00	5.07	6.75	4.80	7.00	5.51	1.95
Positive emotion	3.14	3.33	3.60	4.47	2.94	5.25	3.79	1.64
Negative emotion	1.32	1.64	1.44	2.27	1.83	1.64	1.69	1.03
Anxiety	0.14	0.23	0.23	0.34	0.19	0.20	0.22	0.30
Anger	0.24	0.47	0.36	0.57	0.57	0.51	0.46	0.58
Sadness	0.80	0.54	0.51	0.93	0.67	0.61	0.68	0.53
Social processes	11.50	12.65	13.82	14.54	9.75	9.40	11.94	3.25
Family	0.42	0.38	0.54	1.24	0.30	0.38	0.54	0.70
Friends	0.17	0.39	0.26	1.04	0.30	0.30	0.41	0.46
Female references	0.18	0.46	0.23	0.67	0.15	0.19	0.31	0.44
Male references	0.82	1.06	0.69	1.71	0.48	0.41	0.86	0.81
Cognitive processes	20.33	16.61	17.38	19.18	16.97	12.80	17.21	3.78
Insight	2.07	2.73	2.92	3.49	2.17	1.88	2.54	1.07
Causation	3.91	2.33	2.87	2.05	2.25	1.66	2.51	0.96
Discrepancy	2.82	2.03	2.53	3.05	2.53	1.96	2.48	1.09
Tentative	4.25	2.59	3.72	4.21	3.83	2.77	3.56	1.39
Certainty	4.48	3.64	3.21	4.50	3.79	2.83	3.74	1.35
Differentiation	4.97	3.49	4.34	5.23	5.26	3.19	4.41	1.51
Perceptual processes	1.89	1.71	2.49	2.04	1.59	2.42	2.02	1.09
See	0.69	0.49	1.22	0.76	0.71	1.29	0.86	0.76
Hear	0.83	0.90	0.70	0.65	0.50	0.58	0.69	0.61
Feel	0.22	0.18	0.35	0.40	0.18	0.30	0.27	0.31
Biological processes	3.41	2.15	3.81	3.70	2.85	3.33	3.21	1.48
Body	0.36	0.24	0.78	0.76	0.39	0.60	0.52	0.59
Health	0.38	0.50	0.90	0.91	0.48	0.52	0.61	0.65
Sexual	0.17	0.16	0.27	0.50	0.27	0.42	0.30	0.44
Ingestion	0.32	0.16	0.32	0.47	0.33	0.72	0.39	0.71
Drives	4.79	14.63	9.99	9.46	5.83	7.00	8.62	2.37
Affiliation	1.52	3.96	3.29	4.66	1.25	2.04	2.79	1.62
Achievement	2.15	4.29	4.20	2.91	2.56	2.83	3.16	1.15
Power	0.87	5.65	2.03	1.35	1.62	1.61	2.19	1.00
Reward	0.22	0.82	0.66	0.60	0.43	0.79	0.59	0.47
Risk	0.33	0.96	0.63	0.62	0.66	0.60	0.63	0.58
Time orientations								
Past focus	3.09	2.40	3.72	3.63	2.22	1.78	2.81	1.63
Present focus	9.77	4.47	5.60	7.80	6.60	6.81	6.84	2.00
Future focus	0.62	1.62	1.00	0.88	0.83	0.93	0.98	0.65
Relativity	14.78	16.77	17.22	17.13	13.79	16.93	16.10	3.32
Motion	1.67	1.81	2.09	1.92	1.57	2.03	1.85	0.92
Space	7.25	10.81	10.29	8.07	7.78	8.61	8.80	2.11
Time	6.63	5.11	5.74	8.03	5.10	7.23	6.31	2.09
Personal concerns								
Work	1.47	7.88	3.71	1.96	3.13	2.96	3.52	1.63
Leisure	0.83	0.47	1.36	1.00	1.37	2.82	1.31	1.06
Home	0.33	0.24	0.25	0.54	0.22	0.36	0.32	0.43
Money	0.69	1.24	0.86	0.39	1.01	1.01	0.87	0.76
Religion	0.45	0.81	0.35	0.23	0.37	0.38	0.43	0.51
Death	0.08	0.16	0.21	0.15	0.15	0.13	0.15	0.32
Informal language	15.63	0.59	1.54	2.58	4.29	5.06	4.95	2.44
Swear words	0.10	0.00	0.03	0.18	0.19	0.20	0.12	0.25
Netspeak	1.59	0.04	0.20	0.39	1.66	2.14	1.00	1.12
Assent	8.32	0.11	0.26	0.55	0.95	0.62	1.80	1.19
Nonfluencies	3.72	0.00	0.04	0.02	0.08	0.16	0.67	0.48
Fillers	1.45	0.00	0.03	0.16	0.28	0.22	0.36	0.36

[6] Please note that, for clarity of presentation, the Table is not in DINA4 format. When printing the pdf version page, scaling options can be used to adjust the oversized pages to paper size.

Table 6 (continued). DE-LIWC2015 Output Variable Information [7]

Category	Informal natural speech	Formal natural speech	TED Talks	Expressive writing	Reddit	Twitter	Grand Mean	Grand Mean SD
Punctuation								
Total Punctuation	30.96	15.80	19.61	14.77	21.76	26.96	21.64	7.30
Periods	14.56	4.76	6.88	6.79	7.27	8.89	8.19	3.15
Commas	11.34	7.44	8.75	5.52	4.68	3.52	6.87	2.53
Colons	0.27	0.31	0.64	0.22	0.51	1.94	0.65	1.23
Semicolons	0.00	0.08	0.07	0.02	0.42	0.15	0.12	0.33
Question Marks	1.70	0.12	0.54	0.12	0.97	1.31	0.79	0.75
Exclamation Marks	0.41	0.33	0.08	0.65	0.58	2.56	0.77	1.87
Dashes	0.79	1.21	1.19	0.29	1.38	2.72	1.26	1.75
Quotation Marks	0.00	0.37	1.01	0.21	1.04	1.26	0.65	0.95
Apostrophes	0.01	0.03	0.08	0.06	0.19	0.62	0.16	0.50
Parentheses	0.13	1.00	0.23	0.51	1.32	1.48	0.78	1.31
Other Punctuation	1.76	0.16	0.14	0.38	3.41	2.51	1.39	3.07

Note. Means of relative word count (% of total words) in the six different corpora used. Grand Means are the unweighted means of the six corpora; mean SDs are the unweighted mean of the standard deviations across the six corpora.

[7] Please note that, for clarity of presentation, the Table is not in DINA4 format. When printing the pdf version page, scaling options can be used to adjust the oversized pages to paper size.

Comparing the DE-LIWC2015 Dictionary with the Previous German LIWC2001 Dictionary

For those who are familiar with the previous German LIWC2001 dictionary, transitioning to the new 2015 German LIWC may cause confusion at first. While the majority of the older word categories have been at least slightly changed, some have been substantially revised (e.g. cognitive processes, perceptual processes), and others have either been removed or newly added. Table 7 aims at providing an overview on how the dictionary has changed by reporting means and correlations between the two German dictionary versions. All these analyses are based on the four corpora Callhome, Europarl, expressive writing and TED talks specified in Tables 5 and 6. All numbers reported in Table 7 are the average results from these four corpora. The Twitter and Reddit corpora were not considered for this analysis due to the lack of netspeak categories in the 2001 version.

The LIWC2015/2001 correlation column quantifies the degree to which the new dictionary has changed from the 2001 version. The lower the correlation, the more change across the two versions with respect to the according category.

What can be highlighted is the meaningful improvement in dictionary coverage. In average, the DE-LIWC2015 now captures 83 percent of the total words (Table 6), compared to the 63 percent coverage of the DE-LIWC2001 (Wolf et al., 2008). Depending on the corpora used, dictionary coverage ranged between 72 and 92 percent in our samples. Improvement in the dictionary coverage is likely the result of the increased size of the new dictionary as well as a focus on words frequency for the inclusion of new words throughout the dictionary development process.

Table 7. Comparisons Between the DE-LIWC2015 and 2001: Means (SD) and Correlations [8]

Category	2015 Output Label	2001 Output Label	DE-LIWC2015 mean (SD)[1]	DE-LIWC2001 mean (SD)[1]	DE-LIWC2015/2001 correlation[1]
Word Count	WC	WC	4,306.85 (5,850.47)	4,306.76 (5,850.08)	1.00
Summary Variables					
Analytic Thinking	Analytic		43.42 (17.18)	-	-
Clout	Clout		62.42 (15.16)	-	-
Authentic	Authentic		50.80 (23.44)	-	-
Emotional tone	Tone		61.15 (27.86)	-	-
Words/sentence	WPS	WPS	14.84 (6.62)	14.85 (6.62)	1.00
Words > 6 letters	Sixltr	Sixltr	22.97 (3.57)	22.89 (3.58)	1.00
Dictionary words	Dic	Dic	87.84 (4.08)	70.00 (4.28)	0.72
Linguistic Dimensions					
Total function words	function.	function.	56.21(4.27)	-	-
Total pronouns	pronoun	pronoun	16.58 (3.03)	10.96 (2.86)	0.87
Personal pronouns	ppron	ppron	10.24 (2.74)	-	-
1st pers singular	i	i	4.57 (2.18)	4.54 (2.17)	1.00
1st pers plural	we	we	1.76 (1.26)	1.69 (1.23)	0.99
2nd person (you_total)	you_total	you	1.03 (0.87)	1.08 (0.85)	0.79
2nd pers singular	you_sing		0.52 (0.55)	-	-
2nd pers plural	you_plur		0.20 (0.34)	-	-
2nd pers formal	you_formal		0.31 (0.40)	-	-
3rd person (other)	other	other	2.62 (1.50)	3.89 (1.69)	0.81
3rd pers singular	shehe		2.57 (1.48)	-	-
3rd pers plural	they		1.30 (1.12)	-	-
Impersonal pronouns	ipron		4.73 (1.36)	-	-
Articles	article	article	10.15 (1.94)	9.83 (1.88)	0.97
Prepositions	prep	preps	9.75 (1.84)	8.32 (1.61)	0.91
Auxiliary verbs	auxverb		9.40 (1.89)	-	-
Common Adverbs	adverb		5.02 (1.46)	-	-
Conjunctions	conj		13.56 (2.13)	-	-
Negations	negate	negate	1.92 (0.93)	1.74 (0.89)	0.97
Other Grammar					
Common verbs	verb		16.17 (2.48)	-	-
Common adjectives	adj		6.70 (1.59)	-	-
Comparisons	compare		2.07 (0.84)	-	-
Interrogatives	interrog		1.40 (0.71)	-	-
Numbers	number	number	1.65 (1.40)	1.47 (1.35)	0.98
Quantifiers	quant		2.99 (1.06)	-	-
Psychological Processes					
Affective processes	affect	affect	5.31 (1.70)	4.36 (1.50)	0.82
Positive emotion	posemo	posemo	3.63 (1.39)	3.12 (1.25)	0.81
Negative emotion	negemo	negemo	1.67 (1.05)	1.24 (0.89)	0.81
Anxiety	anx	anx	0.23 (0.33)	0.20 (0.30)	0.86
Anger	anger	anger	0.41 (0.59)	0.27 (0.46)	0.84
Sadness	sad	sad	0.69 (0.55)	0.29 (0.32)	0.57
Social processes	social	social	13.13 (3.23)	9.66 (2.73)	0.89
Family	family	family	0.64 (0.79)	0.60 (0.69)	0.90
Friends	friend	friend	0.47 (0.49)	0.40 (0.46)	0.86
Female references	female		0.39 (0.51)	-	-
Male references	male		1.07 (0.97)	-	-
Cognitive processes	cogproc	cogmech	18.37 (3.43)	9.84 (2.26)	0.77
Insight	insight	insight	2.80 (1.09)	2.90 (1.11)	0.68
Causation	cause	cause	2.79 (0.98)	1.76 (0.72)	0.73
Discrepancy	discrep	discrep	2.60 (1.09)	1.75 (0.87)	0.85
Tentative	tentat	tentat	3.69 (1.35)	1.28 (0.74)	0.70
Certainty	certain	certain	3.96 (1.33)	2.13 (0.91)	0.43
Differentiation	differ	excl	4.51 (1.44)	2.28 (0.85)	0.52
Perceptual processes	percept	senses	2.03 (1.05)	0.13 (0.21)	0.32
See	see	see	0.79 (0.68)	0.03 (0.10)	0.21
Hear	hear	hear	0.77 (0.65)	0.05 (0.14)	0.33
Feel	feel	feel	0.29 (0.33)	0.01 (0.05)	0.05
Biological processes	bio	physcal	3.27 (1.42)	0.84 (0.78)	0.58
Body	body	body	0.53 (0.61)	0.49 (0.60)	0.76
Health	health		0.67 (0.69)	-	-
Sexual	sexual	sexual	0.28 (0.40)	0.16 (0.33)	0.67
Ingestion	ingest	eating	0.32 (0.58)	0.13 (0.28)	0.66
Drives	drives		9.72 (2.37)	-	-
Affiliation	affiliation		3.36 (1.72)	-	-
Achievement	achiev	achiev	3.39 (1.12)	2.35 (0.91)	0.66
Power	power		2.48 (1.01)	-	-
Reward	reward		0.58 (0.43)	-	-
Risk	risk		0.64 (0.47)	-	-
Time orientations[2]					
Past focus	focuspast	past	3.21 (1.90)	3.27 (1.70)	0.93
Present focus	focuspresent	present	6.91 (1.84)	6.52 (1.63)	0.84
Future focus	focusfuture	future	1.03 (0.64)	0.76 (0.55)	0.90
Relativity	relativ		16.47 (3.00)		
Motion	motion	motion	1.87 (0.88)	1.06 (0.65)	0.72
Space	space	space	9.10 (1.95)	7.17 (1.60)	0.90
Time	time	time	6.38 (1.91)	4.41 (1.55)	0.90
Personal concerns					
Work	work	job	3.76 (1.54)	2.23 (1.03)	0.60
Leisure	leisure	leisure	0.92 (0.77)	1.29 (0.84)	0.50
Home	home	home	0.34 (0.43)	0.53 (0.52)	0.63
Money	money	money	0.80 (0.68)	0.89 (0.68)	0.78
Religion	relig	relig	0.46 (0.48)	0.45 (0.44)	0.54
Death	death	death	0.15 (0.35)	0.13 (0.29)	0.68
Informal language	informal		5.08 (2.34)	-	-
Swear words	swear	swear	0.08 (0.20)	0.08 (0.23)	0.76
Netspeak	netspeak		0.55 (0.71)	-	-
Assent	assent	assent	2.31 (1.49)	1.68 (1.07)	0.74
Nonfluencies	nonflu	nonfl	0.94 (0.62)	0.40 (0.53)	0.29
Fillers	filler	fillers	0.41 (0.37)	0.02 (0.07)	0.35

[1] Means are the unweighted means of relative word count (% of total words) of the four corpora Callhome, Europarl, expressive writing and TED talks specified in the section above, and mean SDs are the unweighted mean of the standard deviations across the same corpora.
Correlations are the average Pearson correlation between the 2001 and 2015 dictionaries across these four corpora. Low correlations in this table are to be expected due to the large category differences between the two versions. Reddit and Twitter were not considered for means and correlations, due to the lack of netspeak categories in the 2001 version.
[2] Time orientation categories are similar to the 2001 categories past, present, and future but are more unified to reflect a general time orientation instead of just verb tense usage.

[8] Please note that, for clarity of presentation, the Table is not in DINA4 format. When printing the pdf version page, scaling options can be used to adjust the oversized pages to paper size.

Comparing DE-LIWC2015 with the English LIWC2015

We used the DE-ENG parallel corpora Europarl and translated TED talk transcripts described above in order to test the DE-LIWC2015 for its equivalence with the English LIWC2015. Means, effect sizes and correlations are reported in Table 8. Following recommendations for unequal sample sizes, we computed effect sizes as Hedge's *g*, although it needs to be noted that in large samples, Hedges' *g* is mostly equivalent to Cohen's *d* (Durlak, 2009; Lakens, 2013; Lenhard & Lenhard, 2016). In addition, we calculated Pearson correlations, which should be understood as metrics of measure equivalence between the two dictionaries.

The sample sizes of the corpora were as follows:

	Europarl	**TED Talks**
German	*N* = 3,397	*N* = 2,147
English	*N* = 3,808	*N* = 2,655

Within these two corpora, the DE-LIWC2015 captures, on average, 85 percent of the total words, which is comparable to the English LIWC2015, which also captures 85 percent of the words in the English equivalent of these corpora.

Equivalence between DE-LIWC2015 and English LIWC2015
Similar to other LIWC dictionary translators (Boot, Zijlstra, & Geenen, 2017), we aimed for high ($r > .50$) correlations between corresponding German and English categories, which we obtained for 66 (~87%) out of the 76 reported correlations for the DE-LIWC2015 dictionary categories. High correlations between German and English word categories can be thought to reflect a general equivalence between the two dictionaries, implying that the way the DE-LIWC2015 represents the according category is comparable to the English LIWC2015 dictionary.

While lower correlations may be the product of differences between the dictionaries, such as translation biases and/or particular structural differences in the languages themselves, it needs to be stated that the correlations furthermore depend on the reference corpora used. The two chosen bilingual corpora in our example represent a highly formal context (Europarl) on the one hand, and a more informal or semi-formal context (TED Talks) on the other hand. As can be seen in Table 8, the correlation coefficients differ between the two corpora, which demonstrates again the importance of context in language analysis.

As previously shown in Table 6, mean base rates differ between the different contexts. This is true both for English and German. Between-corpora differences in correlation coefficients may partially be caused by language-specific differences in the ways certain word categories are used in different corpora. For example, both in the English and in the German samples, means of social processes words are higher in the TED talks than in the Europarl, meaning that in the TED talks, social words were more commonly used than in the very formal context of the Europarl corpus. The same seems to be the case for certain grammar categories, such as comparison words and quantifiers, which were more typically used in TED talks as a rhetorical mean.

Given that the direction of the base rate differences between contexts is the same in both languages, the DE-ENG correlation for the “social processes” category is lower in the Europarl samples (*r* =.76) than in the TED Talks sample (*r* =.85). This could be taken to suggest that between the two languages, there are subtle differences in the degree to which the use of “social processes”-words increases from one corpora to another. Similarly, DE-ENG correlations for the “she/he” categories and “total_you” categories are higher in the more informal TED Talks sample compared to the Europarl. At the same time, “you_total” and “she/he” show higher means in the TED Talks than in the Europarl in both languages, suggesting that the proceedings of the European Parliament are not the most typical context to talk about other people, or to use directive speech reflected in *you-talk*. However, these differences between contexts might be more pronounced for one language than the other.

In the case of “she/he”, one has to keep in mind that – unlike in English – this category can be confounded by the use of formal *you-talk* (“Sie”, “Ihr”). Although we used the preprocessing procedure (described in one of the sections above: “DE-LIWC2015: Special Adaptations to German Automated Word Count Analysis) to track “formal_you”, remember that the conservative preprocessing approach tends to overestimate “she/he” and underestimate “you_formal”, if these pronouns often occur at the beginning of sentences. This may have been more the case in the formal than in the other context.

In addition to a possible slight underestimation of *you-talk* in the German compared to the English text samples (caused by overlaps between “formal_you” and “she/he”), one other potential confounder on “total_you” might be the use of *generic you*, i.e. referring to an unspecified person. The *generic you* is commonly used as rhetorical mean in TED talks (e.g. “you should never stop dreaming”). While the *generic you* can be translated with either a second person pronoun (“you”/”du”) or with an impersonal pronoun (“one”/”man”), this may have created another language-specific bias in these corpora regarding this word category.

These are just a few examples demonstrating how one context may exert a bias on both dictionaries, leading to different degrees of correspondence across different reference corpora.

Most important, however, is that the DE-LIWC2015 seems to show very strong equivalence with the original English LIWC2015 dictionary, in general.

Cross-language differences in word base rates

The LIWC metrics allow analysis across different languages by providing relative category values that are robust against language-specific biases. Still, it has been shown in other LIWC translations that the *base rates* in the word use of certain categories may differ due to specific characteristics of a language (Ramírez-Esparza, Pennebaker, García, & Suriá Martínez, 2007). An obvious example borrowed from the Spanish LIWC adaptation is the frequency of personal pronoun use in English versus Spanish. In Spanish, which is a “pronoun-drop” language, the use of personal pronouns is grammatically redundant and they are typically omitted, whereas in English, a conjugated verb is normally accompanied by a personal pronoun. It is thus not surprising that fewer first person pronouns are counted in Spanish than in English samples in cross-language LIWC-analyses. Such cross-language differences, however, can go beyond pure grammar: It has further been found that Spanish texts

contain more positive emotion words than English comparison samples (Ramirez-Esparza, Chung, Kacewicz, & Pennebaker, 2008; Ramírez-Esparza et al., 2007).

As a resource for researchers who aim to perform multilingual LIWC analysis, we calculated the effect sizes (Hedges' *g*) between the mean LIWC output variables between DE-LIWC2015 and English LIWC2015 (Table 8). For many categories (26 categories out of 76 total dictionary categories), large differences were found (|*g*| > .8) [9]. Given the high DE-ENG correlations for most of the word categories, these cross-language differences in mean base rates are likely to reflect differences in the structure of a language, or possibly not 100% identical hit rates across different dictionaries, which may be due to characteristics of their dictionary development process, rather than meaningful psychological differences.

For example, we see in Table 8 that six-letter words were more common in the German samples compared to the English samples. This is fairly unsurprising, given that the German language generally contains a lot of long words such as compound nouns. However, the high DE-ENG correlation (*r* = .77) for this word category suggests that the ways these words differ throughout the German sample corresponds to the way they differ in the English sample, meaning that if an English text is relatively high on six-letter words compared to the rest of the sample, it is also relatively high in the German sample using the DE-LIWC2015.

Furthermore, articles are more frequently used in German than in English in our corpora. Differences in the use of articles between English and German has previously been described by Wolf et al. (2008). This cross-language difference may be driven by different grammar rules in the languages. While uncountable nouns are mostly used without articles in English, these same nouns are used in conjunction with an article in some cases in German (e.g. "What lovely weather!" // "Was für *ein* schönes Wetter!").

As mentioned earlier, the German "She/He"-category contains homographs with the "they"- and "formal_you" categories, which may have contributed to mean DE-ENG base rate differences in these categories resulting in a large effect size. We applied a preprocessing procedure aiming at minimizing these overlaps, however; an absolute separation of these subcategories remains challenging for the German language (see section "DE-LIWC2015: Special Adaptations to German Automated Word Count Analysis"). In general, homographs tend to be different across languages, which might lead to differences in hit and base rates between LIWC dictionaires across different languages (see also "DE-LIWC2015: Special Adaptations to German Automated Word Count Analysis").

> ***Recommendations for multilingual analyses:*** Should you aim to look at associations of LIWC scores with variables of interest in a multilingual sample in your studies, we recommend standardization or language group-mean (of your sample, or other big samples like the ones reported in this manual) centering of the DE-LIWC2015 variables. Directly comparing the raw scores to each other might be difficult due to large absolute base rate differences across languages. However, standardization is not necessary in monolingual

[9] Effect size thresholds according to Cohen (1988):
|*g*| = .2 : small effect size, |*g*| = .5 : medium effect size, and |*g*| = .8 : large effect size

analyses or in multilingual analyses that rely on methods analyzing linear associations, which is probably the case for most research using LIWC.

We conclude that although there seem to be mean base rate differences across the two languages, which is important to know for multilingual analyses, the way these words are used in different samples covaries between German and English and that thus the DE-LIWC2015 word categories are largely performing equivalently to the English LIWC2015 dictionary.

Table 8. Comparisons Between the DE-LIWC2015 and the English LIWC2015: Means and Equivalence Measures [10]

Category	DE-LIWC2015 Output Label	Europarl				TED Talks				Overall (Europarl and TED Talks)			
		Means		Equivalence		Means		Equivalence		Grand Means		Equivalence	
		DE-LIWC 2015	ENG-LIWC 2015	*g*	Pearson's *r*	DE-LIWC 2015	ENG-LIWC 2015	*g*	Pearson's *r*	DE-LIWC 2015	ENG-LIWC 2015	*g*	Pearson's *r*
Word Count	WC	14,236.39	14,539.14	0.01	1.00	1,881.49	2,045.28	0.18	0.99	8,058.94	8,292.21	0.02	0.99
Summary Variables													
Analytic Thinking	Analytic	94.91	89.19	-0.70	0.79	54.52	54.44	0.00	0.88	74.72	71.82	-0.20	0.83
Clout	Clout	76.14	69.59	-0.81	0.80	74.36	74.64	0.02	0.90	75.25	72.12	-0.30	0.85
Authentic	Authentic	40.42	22.54	-1.44	0.54	46.42	43.08	-0.16	0.81	43.42	32.81	-0.63	0.67
Emotional tone	Tone	57.40	60.94	0.15	0.81	66.94	54.04	-0.57	0.77	62.17	57.49	-0.20	0.79
Words/sentence	WPS	19.96	23.95	1.05	0.71	14.52	16.55	0.31	0.68	17.24	20.25	0.58	0.70
Words > 6 letters	Sixltr	34.98	26.64	-2.86	0.70	25.58	17.89	-2.16	0.83	30.28	22.26	-2.47	0.77
Dictionary words	Dic	85.00	82.64	-0.72	0.72	85.97	86.92	0.26	0.78	85.48	84.78	-0.20	0.75
Linguistic Dimensions													
Total function words	function.	51.87	51.59	-0.11	0.72	55.64	55.40	-0.07	0.78	53.76	53.49	-0.09	0.75
Total pronouns	pronoun	11.10	10.06	-0.49	0.87	17.71	16.08	-0.56	0.89	14.41	13.07	-0.53	0.88
Personal pronouns	ppron	5.91	4.44	-0.91	0.92	9.97	8.77	-0.46	0.92	7.94	6.61	-0.63	0.92
1st pers singular	i	1.74	1.55	-0.22	0.95	2.99	2.82	-0.08	0.99	2.36	2.19	-0.11	0.97
1st pers plural	we	1.76	1.68	-0.10	0.94	2.32	2.19	-0.10	0.98	2.04	1.94	-0.10	0.96
2nd person (total)	you_total	0.43	0.46	0.05	0.89	1.18	1.84	0.61	0.81	0.80	1.15	0.41	0.85
2nd pers singular	you_sing	0.00	-	-	-	0.22	-	-	-	0.11	-	-	-
2nd pers plural	you_plur	0.05	-	-	-	0.16	-	-	-	0.10	-	-	-
2nd pers formal	you_formal	0.38	-	-	-	0.81	-	-	-	0.59	-	-	-
3rd person (other)	other	1.59	-	-	-	3.03	-	-	-	2.31	-	-	-
3rd pers singular	shehe	1.54	0.21	-2.72	0.51	2.93	0.76	-2.00	0.72	2.24	0.49	-2.23	0.62
3rd pers plural	they	0.88	0.54	-0.82	0.61	1.84	1.15	-0.82	0.70	1.36	0.84	-0.82	0.65
Impersonal pronouns	ipron	4.20	5.62	1.34	0.61	6.18	7.30	0.72	0.67	5.19	6.46	0.97	0.64
Articles	article	13.94	9.73	-2.82	0.63	10.83	7.35	-2.24	0.79	12.39	8.54	-2.52	0.71
Prepositions	prep	12.62	16.10	2.56	0.48	10.95	13.39	1.57	0.68	11.79	14.74	2.03	0.58
Auxiliary verbs	auxverb	7.19	7.69	0.37	0.70	8.49	8.97	0.32	0.65	7.84	8.33	0.34	0.67
Common Adverbs	adverb	3.71	3.51	-0.22	0.55	4.49	5.84	1.16	0.52	4.10	4.68	0.56	0.53
Conjunctions	conj	10.15	5.43	-4.08	0.58	12.46	7.30	-3.68	0.61	11.31	6.37	-3.87	0.59
Negations	negate	1.06	0.98	-0.17	0.91	1.34	1.29	-0.09	0.95	1.20	1.14	-0.12	0.93
Other Grammar													
Common verbs	verb	13.72	11.95	-0.92	0.71	16.18	16.41	0.10	0.74	14.95	14.18	-0.37	0.73
Common adjectives	adj	6.01	3.89	-2.05	0.53	7.18	4.21	-2.59	0.59	6.60	4.05	-2.33	0.56
Comparisons	compare	1.63	2.04	0.73	0.43	2.52	2.32	-0.27	0.66	2.07	2.18	0.17	0.54
Interrogatives	interrog	0.78	1.29	1.19	0.43	1.63	1.95	0.52	0.64	1.21	1.62	0.79	0.53
Numbers	number	1.76	1.75	-0.01	0.94	1.80	1.95	0.15	0.92	1.78	1.85	0.06	0.93
Quantifiers	quant	2.13	1.85	-0.46	0.43	3.10	2.38	-0.96	0.62	2.61	2.11	-0.73	0.52
Psychological Processes													
Affective processes	affect	5.00	4.62	-0.29	0.70	5.07	4.19	-0.60	0.76	5.03	4.41	-0.45	0.73
Positive emotion	posemo	3.33	3.27	-0.05	0.72	3.60	2.85	-0.67	0.70	3.46	3.06	-0.36	0.71
Negative emotion	negemo	1.64	1.29	-0.39	0.83	1.44	1.29	-0.17	0.84	1.54	1.29	-0.29	0.83
Anxiety	anx	0.23	0.25	0.07	0.82	0.23	0.24	0.04	0.78	0.23	0.24	0.06	0.80
Anger	anger	0.47	0.30	-0.38	0.66	0.36	0.33	-0.08	0.66	0.42	0.31	-0.24	0.66
Sadness	sad	0.54	0.22	-1.07	0.50	0.51	0.25	-0.81	0.65	0.53	0.23	-0.94	0.57
Social processes	social	12.65	8.11	-2.05	0.76	13.82	10.42	-1.15	0.85	13.24	9.27	-1.53	0.80
Family	family	0.38	0.06	-0.79	0.53	0.54	0.29	-0.42	0.81	0.46	0.18	-0.57	0.67
Friends	friend	0.39	0.19	-0.60	0.52	0.26	0.17	-0.40	0.58	0.33	0.18	-0.52	0.55
Female references	female	0.46	0.38	-0.18	0.87	0.23	0.51	0.38	0.77	0.35	0.44	0.16	0.82
Male references	male	1.06	0.79	-0.41	0.84	0.69	0.74	0.07	0.86	0.88	0.77	-0.15	0.85
Cognitive processes	cogproc	16.61	10.43	-2.85	0.74	17.38	11.66	-2.29	0.82	17.00	11.04	-2.55	0.78
Insight	insight	2.73	2.15	-0.78	0.45	2.92	2.52	-0.43	0.75	2.83	2.33	-0.59	0.60
Causation	cause	2.33	2.04	-0.42	0.58	2.87	2.02	-1.11	0.45	2.60	2.03	-0.78	0.52
Discrepancy	discrep	2.03	1.88	-0.23	0.52	2.53	1.49	-1.50	0.59	2.28	1.69	-0.89	0.56
Tentative	tentat	2.59	1.53	-1.52	0.69	3.72	2.51	-1.26	0.76	3.15	2.02	-1.37	0.72
Certainty	certain	3.64	1.73	-2.48	0.56	3.21	1.41	-2.52	0.53	3.43	1.57	-2.50	0.55
Differentiation	differ	3.49	2.34	-1.25	0.80	4.34	3.11	-1.31	0.78	3.91	2.73	-1.28	0.79
Perceptual processes	percept	1.71	1.07	-1.09	0.20	2.49	2.77	0.22	0.73	2.10	1.92	-0.20	0.46
See	see	0.49	0.44	-0.18	0.37	1.22	1.25	0.03	0.86	0.85	0.84	-0.02	0.62
Hear	hear	0.90	0.40	-1.05	0.21	0.70	0.93	0.30	0.69	0.80	0.67	-0.22	0.45
Feel	feel	0.18	0.16	-0.13	0.05	0.35	0.41	0.16	0.46	0.27	0.29	0.07	0.26
Biological processes	bio	2.15	0.65	-1.99	0.59	3.81	1.84	-1.32	0.78	2.98	1.25	-1.55	0.68
Body	body	0.24	0.14	-0.51	0.28	0.78	0.60	-0.25	0.82	0.51	0.37	-0.31	0.55
Health	health	0.50	0.32	-0.36	0.75	0.90	0.78	-0.13	0.91	0.70	0.55	-0.21	0.83
Sexual	sexual	0.16	0.04	-0.59	0.68	0.27	0.10	-0.50	0.79	0.22	0.07	-0.54	0.74
Ingestion	ingest	0.16	0.18	0.05	0.51	0.32	0.35	0.05	0.87	0.24	0.26	0.05	0.69
Drives	drives	14.63	10.82	-1.99	0.63	9.99	8.09	-0.87	0.81	12.31	9.46	-1.40	0.72
Affiliation	affiliation	3.96	3.39	-0.47	0.84	3.29	3.18	-0.08	0.93	3.62	3.28	-0.26	0.89
Achievement	achiev	4.29	1.79	-2.69	0.52	4.20	1.50	-3.07	0.61	4.24	1.64	-2.87	0.57
Power	power	5.65	4.65	-0.68	0.61	2.03	2.37	0.36	0.67	3.84	3.51	-0.27	0.64
Reward	reward	0.82	1.01	0.43	0.48	0.66	1.21	1.12	0.49	0.74	1.11	0.79	0.48
Risk	risk	0.96	0.81	-0.31	0.63	0.63	0.49	-0.36	0.72	0.79	0.65	-0.33	0.67
Time orientations													
Past focus	focuspast	2.40	2.07	-0.39	0.71	3.72	3.89	0.09	0.94	3.06	2.98	-0.06	0.82
Present focus	focuspresent	4.47	8.16	2.91	0.53	5.60	11.16	2.54	0.73	5.04	9.66	2.68	0.63
Future focus	focusfuture	1.62	1.21	-0.74	0.45	1.00	1.11	0.21	0.60	1.31	1.16	-0.28	0.53
Relativity	relativ	16.77	12.31	-2.41	0.54	17.22	13.67	-1.49	0.77	17.00	12.99	-1.89	0.65
Motion	motion	1.81	1.39	-0.71	0.20	2.09	2.09	0.01	0.65	1.95	1.74	-0.31	0.43
Space	space	10.81	7.35	-2.47	0.49	10.29	7.32	-1.71	0.73	10.55	7.34	-2.05	0.61
Time	time	5.11	3.59	-1.28	0.77	5.74	4.44	-1.00	0.80	5.42	4.02	-1.13	0.78
Personal concerns													
Work	work	7.88	5.00	-1.66	0.66	3.71	2.55	-0.73	0.85	5.80	3.77	-1.21	0.75
Leisure	leisure	0.47	0.31	-0.41	0.44	1.36	0.90	-0.49	0.82	0.91	0.61	-0.46	0.63
Home	home	0.24	0.19	-0.17	0.60	0.25	0.30	0.16	0.80	0.24	0.24	0.01	0.70
Money	money	1.24	1.20	-0.04	0.73	0.86	0.67	-0.24	0.87	1.05	0.93	-0.13	0.80
Religion	relig	0.81	0.13	-1.53	0.34	0.35	0.19	-0.35	0.89	0.58	0.16	-0.94	0.61
Death	death	0.16	0.11	-0.17	0.74	0.21	0.18	-0.08	0.87	0.19	0.15	-0.12	0.81
Informal language	informal	0.59	0.20	-1.34	0.17	1.54	0.47	-1.52	0.54	1.06	0.34	-1.47	0.36
Swear words	swear	0.00	0.00	-0.04	0.40	0.03	0.03	-0.01	0.43	0.02	0.02	-0.01	0.41
Netspeak	netspeak	0.04	0.01	-0.27	0.21	0.20	0.08	-0.33	0.36	0.12	0.05	-0.33	0.29
Assent	assent	0.11	0.10	-0.09	0.30	0.26	0.16	-0.36	0.73	0.18	0.13	-0.27	0.51
Nonfluencies	nonflu	0.00	0.09	1.04	0.03	0.04	0.19	0.91	0.56	0.02	0.14	0.97	0.29
Fillers	filler	0.00	0.00	-0.06	-	0.03	0.01	-0.29	0.24	0.02	0.01	-0.20	0.24

Note. Calculations are based on the two DE-ENG parallel corpora Europarl and TED Talks (see section above for corpora description). Grand means are the unweighted means of relative word count (% of total words) for these two corpora. Effect sizes *g* were computed as Hedges' g for the mean base rate differences in English vs. German texts. Correlations are Pearson correlations between the DE- and ENG-LIWC2015 dictionaries.

[10] Please note that, for clarity of presentation, the Table is not in DINA4 format. When printing the pdf version page, scaling options can be used to adjust the oversized pages to paper size.

A Note About Spoken and Written German Text

LIWC may be used for the analysis of both written text and transcribed, spoken text. In both cases, the accuracy of the LIWC output depends upon the quality of text samples that you provide. A good general overview on how to properly prepare your text files for LIWC analysis, including organizing, cleaning and naming the files, is provided in the Operator's Manual of LIWC2015 (Pennebaker, Booth, Boyd & Francis, 2015). What follows here is a guide on how to deal with typing conventions and oral transcriptions specific to German text samples and crucial steps that need to be taken prior to analysis, such as choosing the right encoding of the files.

Analyzing German Text Samples

1. Preprocessing and Things to Keep in Mind

Whenever you correct or clean your text files prior to analysis with LIWC, it is helpful to keep in mind your goals in analyzing the data. LIWC can only count words that are in its dictionaries.

Please be also aware that misspellings, colloquialisms, abbreviations and foreign words are usually also not going to be captured by the internal German dictionaries of LIWC. An exception includes common abbreviations (evtl, etc) and common foreign words (Anglicisms such as crazy, chatten, Follower) that are also frequently used in certain German language-contexts. With the introduction of the netspeak category, a few abbreviations and English expressions commonly used in the social media context have found their way into the DE-LIWC2015 dictionary. As an example, many shorthand interpersonal communication markers (e.g. lol, 4ever, asap) are part of the DE-LIWC2015's "netspeak" and "informal" dictionaries and do not need to be altered.

We here provide an overview of items that you should be aware of, some of which are specific to the analysis of texts in the German language. For some of these considerations, it makes sense to plan for these aspects prior to your analysis, as some of them may require preprocessing or particular methods taken during transcription.

Encoding of text files

It is **extremely** important to ensure that your input texts have the same encoding as the DE-LIWC2015 dictionary: **utf-8**.

Prior to starting any analyses of your text files, a crucial but often-forgotten step is to check the encoding of your text files. Particularly when working with non-English datasets, the encoding of these text files can be very much spread out and ranging from us-ascii, windows-1252, iso-8859-1, and utf-16. If you were to run all of these files through something like LIWC2015 using the German dictionary, which is encoded in utf-8, you would run into problems. Most importantly, you would not receive an error message about the encoding of your text files, rather the encoding mismatch might just go unnoticed and bias your results. LIWC would miss words that it should be detecting, and other issues would arise due to the encoding mismatch (a problem that can be very difficult to spot with the naked eye, let alone diagnose).

We here recommend two programs that can help you dealing with different encodings of text samples in a two-step approach:

ExamineTXT (Boyd, 2018a). In a first step, this program will show you the encoding of each input file (can also be used on other text-based files, like CSV or json) so that you can plan accordingly. This program is helpful if you are unsure of the encoding of your texts prior to analysis. As an important note, the detection of your files' encoding is based on *heuristics* and may not always be 100% accurate. For example, if you have a folder that contains 100 text files, this software might identify 90 of them as having the utf-8 encoding, and 10 of them as having the us-ascii encoding. If you know that your texts all originate from the same source, it would be safe to assume that they all possess the same encoding (e.g., utf-8). This brings us to the next program:

TranscodeTXT (Boyd, 2018c). This program will allow you to convert the encoding of text-based files to the encoding of your choice and, like ExamineTXT, works with other text-based files, such as CSV. Before running your analysis with the DE-LIWC2015, you will want to change all of your input texts' encoding into utf-8 if they are not already using this encoding.

Both of these helper programs are available for free from the following website:

> https://toolbox.ryanb.cc

Preprocessing of formal language
In one of the sections above ("DE-LIWC2015: Special Adaptations to German Automated Word Count Analysis"), we introduced an automatic preprocessing solution for formal second person pronouns. This feature is optional and may be of interest to you, should your samples contain formal conversations (e.g. therapy transcripts) and should you be interested in pronoun use.

Spellcheck
Remember that DE-LIWC2015 is designed primarily to capture common German words. Proper names, scientific or technical terms, or rarely-used words are not parts of the internal dictionary. This means, that for these words, it does not matter so much whether they are spelled correctly in your text samples. However, you may want to correct spelling errors on common words to get more precise LIWC-results.

> ***Helpful tip:*** Do not worry too much about spelling. It has previously been demonstrated that LIWC is robust to natural frequencies of spelling errors in German texts (Wolf et al., 2008).
> The general rule of thumb is that the more files and words you have, the less it matters. If your study contains 100 people who wrote about 1,000 words each, even a misspelling rate of 1% in your files will probably not matter much. Just make sure that systematic misspellings are corrected.

Regional spelling variants
For the German LWIC2015, different regional spelling conventions are included in the dictionary. For example, all words in the dictionary containing a "ß" are also included

in their alternative "ss"-spelling, which conforms to the spelling conventions in Switzerland and Liechtenstein.

"Umlaute"
The "Umlaute" are treated in a similar way: all dictionary words containing "ä", "ö" or "ü" are additionally included in their alternative "ae", "oe" and "ue"-spelling. Therefore, if you have participants in your sample who wrote texts using a foreign keyboard, there is no need to adapt these "Umlaute", since the DE-LIWC2015 will recognize both versions.

Abbrevations
Meaningful abbreviations should be spelled out. Note that some of the abbrevations commonly used in German are included in the dictionary and are not needed to be spelled out (*bzw*, *evtl*). However, beware that the "Words per sentence" (WPS) category is based on the number of times that end-of-sentence markers are detected, which included periods (.), question marks (?) and exclamation points (!). One potential problem with abbreviations in texts (e.g. Dr., u.A., i.O.) is that they may lead to the count of multiple sentences unless the periods are removed.

Compound nouns
One additional challenge that the German language poses to computerized word count programs is its frequent use of compound words that consist of several nouns attached at each other, such as "Unabhängigkeitserklärung", "Liebeskummer" or "Haustürschlüssel". While the main goal of LIWC is to count frequently-used words in psychologically meaningful categories, an emphasis on more specific word stems (*) has been put in the development of the DE-LIWC2015 dictionary, in order to not track words that may ultimately be unrelated to the word stem.

While some of the most common compound words are included in the DE-LIWC2015 dictionary, less common compound words may not be counted. This means that words such as "Lebensgefährte" (partner) are part of the LIWC dictionary categories, while other rarely used and more extreme forms of compound words won't be recognized by LIWC.

2. Transcribing Oral Exchanges: Special Problems

To meet needs that are specific to oral communication, the DE-LIWC2015 has adopted certain conventions specific to spoken language. Whenever transcribing oral exchanges, there are several things to be aware of and these conventions may make your life easier for subsequent LIWC analyses. Table 9 provides an overview of the DE-LIWC2015 "filler" and "nonfluency" categories, which are more common in spoken language.

1. Nonfluencies.
The DE-LIWC2015 dictionary contains several markers of nonfluency, which usually refer to utterances that are not per se meaningful (i.a. "*äh", "hach"*). If you are interested in nonfluency markers, you will want to use consistent spelling of these utterances in your transcripts and keep them close to the words that will be caught in this category (Table 9).

Some very specific forms may not be caught, but may be changed into forms that are part of the nonfluency-dictionary (e.g., *aaaaaah* should be adapted to "*äh"* in the transcripts if used as a nonfluency). In a similar manner, stuttering may be captured by replacing the stuttering part of a phrase to a nonfluency marker, e.g. "i-ich bin" might be changed into "*äh*, ich bin".

Table 9. Overview of "Nonfluency" and "Filler" Words in the DE-LIWC2015

Nonfluency		Filler
ach	seufzen	blabla*
aeh	seufzer	blah
aehm*	seufzt seufzte	dings*
ah	seufzten	gell
äh	seufztest	glaub
ähm*	tss*	halt
ehm		idk
hä		keineahnung
hach		na
hae		naja
hm		ne
hmm*		odersoaehnlich
huh		odersoähnlich
mh		quasi
mhh*		sozusagen
mm		tja
mmm*		wa
och		wasweißich
oehm		wasweissich
oh		weißnich*
ohh*		weißtdu
öhm		weissnich*
puh		weisstdu
seufze		wieauchimmer

2. Assents.

One other important thing we would like to point out is the word "*mhm*"*, which is part of the "assent" category. On the other hand, similar forms such as "*mm", "mhh*"* are nonfluency markers. Should any of these categories be of interest to your analyses, please note the following: When transcribing any of these utterances, please make sure to transcribe them in a way they will be counted in the right category: The spelling should be "*mhm"* when referring to agreement or affirmation ("assent"), whereas "*mm"* or "*mhhh"* would be the right choice when reflecting thinking behavior ("nonfluency").

3. Fillers.

Everyday speech is typically fraught with rather meaningless filler words. For automatized text analysis, the tricky part about this is that these filler words are often composed of important words in the other LIWC categories. For the German language in particular, fillers often represent small phrases rather than single words. An overview of the DE-LIWC2015 "filler" category is provided in Table 9. Please be aware of the following:

Keine Ahnung. As in, “Damals waren wir… keine Ahnung…20 Jahre alt…"

If you want it to be counted as a filler word, change it to one word in your samples -> *keineahnung*. “…*keineahnung* 20 Jahre alt...”

Oder so ähnlich. As in, " Vielleicht als wir zusammen zur Schule gingen, als wir noch klein waren, oder so ähnlich? "

Change to one word: *Odersoähnlich*.

4. Dialect.
The DE-LIWC2015 has been designed to effectively track words in Standard German. Therefore, its dictionary entries correspond to Standard German orthography, and, with some very few exceptions, do not contain dialect words. Exceptions that are covered by the DE-LIWC2015 include colloquial expressions that are frequently used in online social media (e.g. *“ne”, “nem”, “nen”, “ner”, “aufm”, “wasn”*).

However, spoken language often involves dialect words. For the transcription of oral exchanges, we recommend adopting Standard German orthography rules wherever possible for the sake of LIWC analyses. This is especially advised for pronouns: Expressions such as “*isch”, “icke”* or *“ig”* should be transcribed as *“ich”*.

5. Transcribers’ comments.
In general, LIWC2015 is designed for expressions in the form of natural language. However, transcripts often include other information such as remarks, tags or annotations from transcribers, e.g. [nervöses Lachen], [zittrige Stimme], [Flüstern]. For the sake of LIWC analyses, the general recommendation is to remove these.

In cases where the transcriber was unable to understand a certain word or passage, it is recommended to just leave the unheard part out instead of including written comments [e.g. "can't understand" or "?"] (c.f. Pennebaker, Booth, et al., 2015).

Other LIWC Dictionary Translations

Apart from German, the LIWC dictionaries have been translated into several other languages, including the following:

Spanish (Ramírez-Esparza, Pennebaker, García, & Suriá Martínez, 2007)

French (Piolat, Booth, Chung, Davids, & Pennebaker, 2011)

Italian (Agosti & Rellini, 2007)

Dutch (Boot et al., 2017; Van Wissen & Boot, 2017)

Serbian (Bjekić, Lazarević, Živanović, & Knežević, 2014)

Russian (Kailer & Chung, 2011)

Malay (Ahmad, Lutfi, Kushan, Khairuddin, Zolkeplay, Rahmat, & Mishan, 2017)

Chinese (e.g. Huang et al., 2012)

To our knowledge, most of these translations have relied on the LIWC2001 or 2007 version so far, with the exception of the Dutch and Chinese dictionaries for which both a LIWC2007 and LIWC2015 version exist.

References

Ahmad, Z., Lutfi, S. L., Kushan, A. L., Khairuddin, M. H., Zolkeplay, A. F., Rahmat, M. H., & Mishan, M. T. (2017). Construction of the Malay language psychometric properties using LIWC from facebook statuses. *Advanced Science Letters*, *23*(8), 7911–7914. https://doi.org/10.1166/asl.2017.9607

Agosti, A., & Rellini, A. (2007). *The Italian LIWC Dictionary*. Technical report, LIWC. net, Austin, TX.

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

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

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

Boyd, R. L. (2017a). MEH: Meaning Extraction Helper (Version 1.4.20) [Software]. Available from https://meh.ryanb.cc

Boyd, R. L. (2017b). Psychological text analysis in the digital humanities. In S. Hai-Jew (Ed.), *Data analytics in the digital humanities* (pp. 161–189). New York: Springer International Publishing.

Boyd, R. L. (2018a). ExamineTXT (Version 1.10) [Software]. Available from https://toolbox.ryanb.cc

Boyd, R. L. (2018b). TextEmend (Version 1.01) [Software]. Available from https://toolbox.ryanb.cc

Boyd, R. L. (2018c). TranscodeTXT (Version 1.01) [Software]. Available from https://toolbox.ryanb.cc

Boyd, R. L., & Pennebaker, J. W. (2015). A way with words: Using language for psychological science in the modern era. In C. Dimofte, C. Haugtvedt, & R. Yalch (Eds.), *Consumer psychology in a social media world* (pp. 222–236). New York: Routledge Publishers. https://doi.org/10.4324/9781315714790

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

Brown, W. (1910). Some experimental results in the correlation of mental abilities[1]. *British Journal of Psychology, 1904-1920*, *3*(3), 296–322. https://doi.org/10.1111/j.2044-8295.1910.tb00207.x

Carey, A. L., Brucks, M. S., Küfner, A. C. P., Holtzman, N. S., Deters, F. G., Back, M. D., … Mehl, M. R. (2015). Narcissism and the use of personal pronouns revisited. *Journal of Personality and Social Psychology*, *109*(3), e1–e15. https://doi.org/10.1037/pspp0000029

Caton, S., Hall, M., & Weinhardt, C. (2015). How do politicians use Facebook? An applied Social Observatory. *Big Data & Society*, *2*(2), 2053951715612822. https://doi.org/10.1177/2053951715612822

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

Cohen, J. (1988), *Statistical Power Analysis for the Behavioral Sciences, 2nd Edition.* Hillsdale, N.J.: Lawrence Erlbaum. https://doi.org/10.1016/C2013-0-10517-X

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

Durlak, J. A. (2009). How to select, calculate, and interpret effect sizes. *Journal of pediatric psychology*, *34*(9), 917-928. https://doi.org/10.1093/jpepsy/jsp004

Friederich, H.-C., Brockmeyer, T., Wild, B., Resmark, G., de Zwaan, M., Dinkel, A., … Herzog, W. (2017). Emotional Expression Predicts Treatment Outcome in Focal Psychodynamic and Cognitive Behavioural Therapy for Anorexia Nervosa: Findings from the ANTOP Study. *Psychotherapy and Psychosomatics*, *86*(2), 108–110. https://doi.org/10.1159/000453582

Horn, A. B., Holzgang, S., & Rosenberger, V. (2017). Couples coping with the transition to retirement: Interpersonal emotion regulation and wellbeing in daily life. *The European Health Psychologist,* Vol. 19 (Supp.). Retrieved July 24, 2018, from http://ehps.net/ehp/index.php/contents/article/view/2532

Horn, A. B., Kneisler, L., Schuster, H., & Traue, H. C. (2010). Subjective illness representations in depressive disorder: A longitudinal study. [Subjektive Krankheitskonzepte bei depressiven Stoerungen. Laengsschnittstudie einer rehabilitativen Massnahme.]. *Zeitschrift für Gesundheitspsychologie, 18*(1), 40-51. https://doi.org/10.1026/0943-8149/a000006

Horn, A. B., & Maercker, A. (2016). Intra- and interpersonal emotion regulation and adjustment symptoms in couples: The role of co-brooding and co-reappraisal. *BMC Psychol, 4*(1), 51. https://doi.org/10.1186/s40359-016-0159-7

Horn, A. B., & Mehl, M. R. (2004). Expressives Schreiben als Copingtechnik: ein Überblick über den Stand der Forschung. *Verhaltenstherapie*, *14*(4), 274–283. https://doi.org/10.5167/uzh-100060

Horn, A. B., Pössel, P., & Hautzinger, M. (2011). Promoting adaptive emotion regulation and coping in adolescence: a school-based programme. *Journal of Health Psychology, 16*(2), 258-273. https://doi.org/10.1177/1359105310372814

Huang, C.-L., Chung, C. K., Hui, N., Lin, Y.-C., Seih, Y.-T., Lam, B. C., … Pennebaker, J. W. (2012). The development of the Chinese linguistic inquiry and word count dictionary. *Chinese Journal of Psychology*.

Jakubíček, M., Kilgarriff, A., Kovář, V., Rychlý, P., & Suchomel, V. (2013, July). The TenTen corpus family. *In 7th International Corpus Linguistics Conference CL* (pp. 125-127).

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*(2), 125–143. https://dx.doi.org/10.1177/0261927X13502654

Kailer, A., & Chung, C. K. (2011). The Russian LIWC2007 Dictionary. *Austin, TX: LIWC. Net*.

Karan, A., Wright, R. C., & Robbins, M. L. (2017). Everyday emotion word and personal pronoun use reflects dyadic adjustment among couples coping with breast cancer. *Personal Relationships*, *24*(1), 36–48. https://doi.org/10.1111/pere.12165

Karins, K., MacIntyre, R., Brandmair, M., Lauscher, S., & McLemore, C. (1997). CALLHOME German Transcripts. *LDC97T15*.

Koehn, P. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. *Conference Proceedings: The Tenth Machine Translation Summit* (pp. 79--86), Phuket, Thailand: AAMT.

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. *Frontiers in Psychology*, *4*. https://doi.org/10.3389/fpsyg.2013.00863

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*(2), 211. https://dx.doi.org/10.1037/0033-295X.104.2.211

langdetect. (n.d.). Retrieved May 9, 2018, from https://pypi.org/project/langdetect/

Lenhard, W. & Lenhard, A. (2016). *Berechnung von Effektstärken.* Retrieved July 18, 2018, from https://www.psychometrica.de/effektstaerke.html. Dettelbach: Psychometrica. https://doi.org/10.13140/rg.2.1.3478.4245

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

Mehl, M. R., Robbins, M. L., & Holleran, S. E. (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*(2), 30–50. https://dx.doi.org/10.2458/v3i2.16477

Meier, T., Boyd, R.L., Milek, A., Pennebaker, J.W., Mehl, M.R., Martin, M., Wolf, M., & Horn, A.B. (In prep). Lost in Translation? How Psychological Signatures are Morphed by Gender during Translation. Retrieved from https://osf.io/jvp6r/ [pre-registration of data analysis]

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

Neysari, M., Bodenmann, G., Mehl, M. R., Bernecker, K., Nussbeck, F. W., Backes, S., … Horn, A. B. (2016). Monitoring pronouns in conflicts temporal dynamics of verbal communication in couples across the lifespan. *GeroPsych: The Journal of Gerontopsychology and Geriatric Psychiatry*, *29*(4), 201–213. https://doi.org/10.1024/1662-9647/a000158

Nowson, S., & Gill, A. J. (2014). Look! who's talking?: Projection of extraversion across different social contexts. In *Proceedings of the 2014 ACM Multi Media on Workshop on Computational Personality Recognition* (pp. 23–26). ACM. https://doi.org/10.1145/2659522.2659530

Pennebaker, J. W. (2011). The Secret Life of Pronouns: What Our Words Say about Us. NY: Bloomsbury Press.

Pennebaker, J. W., & Beall, S. K. (1986). Confronting a traumatic event: toward an understanding of inhibition and disease. *Journal of Abnormal Psychology*, *95*(3), 274. https://dx.doi.org/10.1037/0021-843X.95.3.274

Pennebaker, J. W., Booth, R. J., Boyd, R. L., & Francis, M. E. (2015). *LIWC 2015 Operator's Manual*. Austin, TX: Pennebaker Conglomerates Inc.

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.

Pennebaker, J. W., Chung, C. K., Frazee, J., Lavergne, G. M., & Beaver, D. I. (2015). When Small Words Foretell Academic Success: The Case of College Admissions Essays. *PLOS ONE*, *9*(12), e115844. https://doi.org/10.1371/journal.pone.0115844

Pennebaker, J. W., & Francis, M. E. (1999). *Linguistic Inquiry and Word Count (LIWC): A computer-based text analysis program*. Mahwah, NJ: Erlbaum.

Pennebaker, J. W., Francis, M. E., & Booth, R. J. (2001). *Linguistic inquiry and word count (LIWC): LIWC2001*. Mahway: Lawrence Erlbaum ASsociates.

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

Ramirez-Esparza, N., Chung, C. K., Kacewicz, E., & Pennebaker, J. W. (2008). The psychology of word use in depression forums in English and in Spanish: Testing two text analytic approaches. In E. Adar, M. Hurst, T. Finin, N. Glance, N. Nicolov, & B. Tseng (Eds.), *Proceedings of the 2008 International Conference on Weblogs and Social Media* (pp. 102–108). Menlo Park, CA: AAAI Press.

Ramírez-Esparza, N., Pennebaker, J. W., García, F. A., & Suriá Martínez, R. (2007). La psicología del uso de las palabras: Un programa de computadora que analiza textos en español.

Ritter, R. S., Preston, J. L., & Hernandez, I. (2014). Happy tweets: Christians are happier, more socially connected, and less analytical than atheists on Twitter. *Social Psychological and Personality Science*, *5*(2), 243–249. https://doi.org/10.1177/1948550613492345

Robbins, M. L., Karan, A., Lopez, A. M., & Weihs, K. L. (2018). Naturalistically observing non-cancer conversations among couples coping with breast cancer. *Psycho-Oncology.* Advance online publication. https://doi.org/10.1002/pon.4797

Schiller, A., Teufel, S., Stöckert, C., & Thielen, C. (1999). *Guidelines für das Tagging deutscher Textcorpora mit STTS [Guidelines for tagging German corpora of written language with STTS].* (Technical Report). Stuttgart, Germany: Institut für maschinelle Sprachverarbeitung [Institute for Machine Language Processing].

Schoch-Ruppen, J., Ehlert, U., Uggowitzer, F., Weymerskirch, N., & La Marca-Ghaemmaghami, P. (2018). Women's Word Use in Pregnancy: Associations With Maternal Characteristics, Prenatal Stress, and Neonatal Birth Outcome. *Frontiers in Psychology*, *9*, 1234. https://doi.org/10.3389/fpsyg.2018.01234

Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., … Ungar, L. H. (2013). Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach. *PLoS ONE*, *8*(9), e73791. https://doi.org/doi:10.1371/journal.pone.0073791

Sonnenschein, A. R., Hofmann, S. G., Ziegelmayer, T., & Lutz, W. (2018). Linguistic analysis of patients with mood and anxiety disorders during cognitive behavioral therapy. *Cognitive Behaviour Therapy*, *47*(4), 315–327. https://doi.org/10.1080/16506073.2017.1419505

Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology, 1904-1920*, *3*(3), 271–295. https://doi.org/10.1111/j.2044-8295.1910.tb00206.x

Stanton, A. M., Meston, C. M., & Boyd, R. L. (2017). Sexual Self-Schemas in the Real World: Investigating the Ecological Validity of Language-Based Markers of Childhood Sexual Abuse. *Cyberpsychology, Behavior, and Social Networking*, *20*(6). https://doi.org/10.1089/cyber.2016.0657

Tackman, A. M., Sbarra, D. A., Carey, A. L., Donnellan, M. B., Horn, A. B., Holtzman, N. S., . . . Mehl, M. R. (2018). Depression, negative emotionality, and self-referential language: A multi-lab, multi-measure, and multi-language-task research synthesis. *Journal of Personality and Social Psychology.* Advance online publication. https://dx.doi.org/10.1037/pspp0000187

Tausczik, Y. R., & Pennebaker, J. W. (2010). The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *Journal of Language and Social Psychology*, *29*(1), 24–54. https://doi.org/10.1177/0261927X09351676

Van Wissen, L., & Boot, P. (2017). An Electronic Translation of the LIWC Dictionary into Dutch. In *Electronic lexicography in the 21st century: Proceedings of eLex 2017 conference* (pp. 703–715). Lexical Computing.

Wolf, M., Chung, C. K., & Kordy, H. (2010). Inpatient treatment to online aftercare: E-mailing themes as a function of therapeutic outcomes. *Psychotherapy Research*, *20*(1), 71–85. https://doi.org/10.1080/10503300903179799

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*(2), 85–98. https://dx.doi.org/10.1026/0012-1924.54.2.85

Wolf, M., Sedway, J., Bulik, C. M., & Kordy, H. (2007). Linguistic analyses of natural written language: Unobtrusive assessment of cognitive style in eating disorders. *International Journal of Eating Disorders*, *40*(8), 711–717. https://doi.org/10.1002/eat.20445

Wolf, M., Theis, F., & Kordy, H. (2013). Language use in eating disorder blogs: Psychological implications of social online activity. *Journal of Language and Social Psychology*, *32*(2), 212–226. https://doi.org/10.1177/0261927X12474278

Other Selected Articles, based on 2001 Version of DE-LIWC

Back, M. D., Schmukle, S. C., & Egloff, B. (2009). Predicting Actual Behavior From the Explicit and Implicit Self-Concept of Personality. *Journal of Personality and Social Psychology*, *97*(3), 533–548. https://doi.org/10.1037/a0016229

Bentley, R. A., Acerbi, A., Ormerod, P., & Lampos, V. (2014). Books average previous decade of economic misery. *PLoS ONE*, *9*(1). https://doi.org/10.1371/journal.pone.0083147

Brockmeyer, T., Kulessa, D., Hautzinger, M., Bents, H., & Backenstrass, M. (2015). Mood-incongruent processing during the recall of a sad life event predicts the course and severity of depression. *Journal of Affective Disorders*, *187*, 91–96. https://doi.org/10.1016/j.jad.2015.08.010

Dang-Xuan, L., & Stieglitz, S. (2012). Impact and diffusion of sentiment in political communication - An empirical analysis of political weblogs (pp. 427–430). Presented at the ICWSM 2012 - Proceedings of the 6th International AAAI Conference on Weblogs and Social Media.

Eisenwort, B., Till, B., Hinterbuchinger, B., & Niederkrotenthaler, T. (2014). Sociable, Mentally Disturbed Women and Angry, Rejected Men: Cultural Scripts for the Suicidal Behavior of Women and Men in the Austrian Print Media. *Sex Roles*, *71*(5–8), 246–260. https://doi.org/10.1007/s11199-014-0395-3

Gemmiti, M., Hamed, S., Wildhaber, J., Pharisa, C., & Klumb, P. L. (2017). Pediatric Consultations: Negative-Word Use and Parent Satisfaction. *Journal of Pediatric Psychology*, *42*(10), 1165–1174. https://doi.org/10.1093/jpepsy/jsx061

Gieselmann, A., & Pietrowsky, R. (2016). Treating procrastination chat-based versus face-to-face: An RCT evaluating the role of self-disclosure and perceived counselor's characteristics. *Computers in Human Behavior*, *54*, 444–452. https://doi.org/10.1016/j.chb.2015.08.027

Greving, H., & Sassenberg, K. (2015). Counter-regulation online: Threat biases retrieval of information during Internet search. *Computers in Human Behavior*, *50*, 291–298. https://doi.org/10.1016/j.chb.2015.03.077

Haefeli, J., Lischer, S., & Haeusler, J. (2015). Communications-based early detection of gambling-related problems in online gambling. *International Gambling Studies*, *15*(1), 23–38. https://doi.org/10.1080/14459795.2014.980297

Haug, S., Gabriel, C., Flückiger, C., & Kordy Dr., H. (2010). Resource activation with psychotherapy patients. Effectiveness of a minimal intervention in Internet chat groups. *Psychotherapeut*, *55*(2), 128–135. https://doi.org/10.1007/s00278-009-0657-7

Hellmann, J. H., Adelt, M. H., & Jucks, R. (2016). No Space for Others? On the Increase of Students' Self-Focus When Prodded to Think About Many Others. *Journal of Language and Social Psychology*, *35*(6), 698–707. https://doi.org/10.1177/0261927X16629521

Jacobi, C., Kleinen-von Königslöw, K., & Ruigrok, N. (2016). Political News in Online and Print Newspapers: Are online editions better by electoral democratic standards? *Digital Journalism*, *4*(6), 723–742. https://doi.org/10.1080/21670811.2015.1087810

Küfner, A. C. P., Back, M. D., Nestler, S., & Egloff, B. (2010). Tell me a story and I will tell you who you are! Lens model analyses of personality and creative writing. *Journal of Research in Personality*, *44*(4), 427–435. https://doi.org/10.1016/j.jrp.2010.05.003

Lindner, A., Hall, M., Niemeyer, C., & Caton, S. (2015). BeWell: A sentiment aggregator for proactive community management (Vol. 18, pp. 1055–1060). Presented at the Conference on Human Factors in Computing Systems - Proceedings. https://doi.org/10.1145/2702613.2732787

Lumma, A.-L., Böckler, A., Vrticka, P., & Singer, T. (2017). Who am I? Differential effects of three contemplative mental trainings on emotional word use in self-descriptions. *Self and Identity*, *16*(5), 607–628. https://doi.org/10.1080/15298868.2017.1294107

Lumma, A.-L., Valk, S. L., Böckler, A., Vrtička, P., & Singer, T. (2018). Change in emotional self-concept following socio-cognitive training relates to structural plasticity of the prefrontal cortex. *Brain and Behavior*, *8*(4). https://doi.org/10.1002/brb3.940

März, A., Schubach, S., & Schumann, J. H. (2017). "Why Would I Read a Mobile Review?" Device Compatibility Perceptions and Effects on Perceived Helpfulness. *Psychology and Marketing*, *34*(2), 119–137. https://doi.org/10.1002/mar.20979

Morales, M. R., & Levitan, R. (2017). Speech vs. text: A comparative analysis of features for depression detection systems (pp. 136–143). Presented at the 2016 IEEE Workshop on Spoken Language Technology, SLT 2016 - Proceedings. https://doi.org/10.1109/SLT.2016.7846256

Peters, J., Wiehler, A., & Bromberg, U. (2017). Quantitative text feature analysis of autobiographical interview data: Prediction of episodic details, semantic details and temporal discounting /631/378/2649 /631/477/2811 article. *Scientific Reports*, *7*(1). https://doi.org/10.1038/s41598-017-14433-6

Proyer, R. T., & Brauer, K. (2018). Exploring adult Playfulness: Examining the accuracy of personality judgments at zero-acquaintance and an LIWC analysis of textual information. *Journal of Research in Personality*, *73*, 12–20. https://doi.org/10.1016/j.jrp.2017.10.002

Przyrembel, M., & Singer, T. (2018). Experiencing meditation – Evidence for differential effects of three contemplative mental practices in micro-phenomenological interviews. *Consciousness and Cognition*, *62*, 82–101. https://doi.org/10.1016/j.concog.2018.04.004

Rauthmann, J. F., & Sherman, R. A. (2016). Measuring the situational eight diamonds characteristics of situations; an optimization of the RSQ-8 to the S8∗. *European Journal of Psychological Assessment*, *32*(2), 155–164. https://doi.org/10.1027/1015-5759/a000246

Rohr, M. K., Wieck, C., & Kunzmann, U. (2017). Age differences in Positive feelings and their expression. *Psychology and Aging*, *32*(7), 608–620. https://doi.org/10.1037/pag0000200

Rosenbach, C., & Renneberg, B. (2015). Remembering rejection: Specificity and linguistic styles of autobiographical memories in borderline personality disorder and depression. *Journal of Behavior Therapy and Experimental Psychiatry*, *46*, 85–92. https://doi.org/10.1016/j.jbtep.2014.09.002

Rudolph, A., Schröder-Abé, M., Riketta, M., & Schütz, A. (2010). Easier when done than said! Implicit self-esteem predicts observed or spontaneous behavior, but not self-reported or controlled behavior. *Journal of Psychology*, *218*(1), 12–19. https://doi.org/10.1027/0044-3409/a000003

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

Sousa, T., Flekova, L., Mieskes, M., & Gurevych, I. (2015). Constructive feedback, thinking process and cooperation: Assessing the quality of classroom interaction (Vol. 2015-January, pp. 2739–2743). Presented at the Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH.

Stieglitz, S., & Dang-Xuan, L. (2011). The role of sentiment in information propagation on Twitter - An empirical analysis of affective dimensions in political tweets. Presented at the ACIS 2011 Proceedings - 22nd Australasian Conference on Information Systems.

Stieglitz, S., & Dang-Xuan, L. (2012a). Impact and diffusion of sentiment in public communication on Facebook. Presented at the ECIS 2012 - Proceedings of the 20th European Conference on Information Systems.

Stieglitz, S., & Dang-Xuan, L. (2012b). Political communication and influence through microblogging - An empirical analysis of sentiment in Twitter messages and retweet behavior (pp. 3500–3509). Presented at the Proceedings of the Annual Hawaii International Conference on System Sciences. https://doi.org/10.1109/HICSS.2012.476

Tondorf, T., Kaufmann, L.-K., Degel, A., Locher, C., Birkhäuer, J., Gerger, H., … Gaab, J. (2017). Employing open/hidden administration in psychotherapy research: A randomized-controlled trial of expressive writing. *PLoS ONE*, *12*(11). https://doi.org/10.1371/journal.pone.0187400

Zimmer, B., Moessner, M., & Kordy, H. (2010). The communication of chronically III patients in an internet chat for aftercare of inpatient psychosomatic treatment. *Rehabilitation*, *49*(5), 301–307. https://doi.org/10.1055/s-0030-1262848

Zimmermann, J., Brockmeyer, T., Hunn, M., Schauenburg, H., & Wolf, M. (2017). First-person Pronoun Use in Spoken Language as a Predictor of Future Depressive Symptoms: Preliminary Evidence from a Clinical Sample of Depressed Patients. *Clinical Psychology and Psychotherapy*, *24*(2), 384–391. https://doi.org/10.1002/cpp.2006

Zimmermann, J., Wolf, M., Bock, A., Peham, D., & Benecke, C. (2013). The way we refer to ourselves reflects how we relate to others: Associations between first-person pronoun use and interpersonal problems. *Journal of Research in Personality*, *47*(3), 218–225. https://doi.org/10.1016/j.jrp.2013.01.008

Acknowledgements

During her work on this project, Tabea Meier was a pre-doctoral fellow of LIFE (International Max Planck Research School on the Life Course; participating institutions: MPI for Human Development, Humboldt-Universität zu Berlin, Freie Universität Berlin, University of Michigan, University of Virginia, University of Zurich).

Special thanks go to Vanessa Infanger, who significantly contributed to the DE-LIWC2015 dictionary development.

We are also indebted to Jara Francalancia and Moira Rudolphi, who supported us in early phases of the dictionary development.

Funding

Tabea Meier and Andrea B. Horn are affiliated with, and Mike Martin is the director of the University Research Priority Program "Dynamics of Healthy Aging" at the University of Zurich, Switzerland.

Financial support by the Jacobs Foundation, the Swiss National Science Foundation (SNF PMPDP1_164470) and the Faculty of Arts and Social Sciences, University of Zurich helped to conduct research reported in this manual.

Preparation of this manuscript was also aided by grants from the National Institute of Health 5R01GM112697-02), John Templeton Foundation (#48503 and #61156) and the National Science Foundation (IIS-1344257). The views, opinions, and findings contained in this manual are those of the authors and should not be construed as position, policy, or decision of the aforementioned agencies, unless so designated by other documents.