An Electronic Translation of the LIWC Dictionary into Dutch

Leon van Wissen[1], **Peter Boot**[2]

[1]Vrije Universiteit Amsterdam, The Netherlands
[2]Huygens ING, Amsterdam, The Netherlands
E-mail: l.van.wissen@vu.nl, peter.boot@huygens.knaw.nl

Abstract

LIWC (Linguistic Inquiry and Word Count) is a text analysis tool developed by social psychologists but now widely used outside of psychology. The tool counts words in certain categories, as defined in an accompanying (English-language) dictionary. The most recent version of the dictionary was published in 2015. We present a pipeline for the automatic translation of LIWC dictionaries into Dutch. We first make an automated translation of the LIWC 2007 version and compare it to the manually translated version of this dictionary. Then we use the pipeline to translate the LIWC 2015 dictionary. We also present the provisional Dutch LIWC 2015 dictionary that results from the pipeline. Although a number of categories require further work, the dictionary should be usable for most research purposes.

Keywords: Machine translation; Linguistic Inquiry and Word Count (LIWC); Google Translate

1. Introduction

LIWC (Linguistic Inquiry and Word Count, often pronounced 'Luke') is a lexical resource developed by social psychologist James Pennebaker and his team at the University of Texas (Pennebaker et al., 2001). Its lexical information is stored in a dictionary that groups English words into categories with psychological significance, such as emotions, cognitive processes, life concerns, social words and several categories of function words. This dictionary can be used in an application that processes a collection of texts and outputs the relative frequencies of words belonging to the categories in each of the texts. The distribution of those categories in the text can give insight into the psychological state of its author or can reflect an author's personal condition. The LIWC dictionary has been published in multiple versions (notably Pennebaker et al., 2001, 2007, 2015b) and the dictionary has been translated into many languages, mostly using the 2001 version as reference.

The 2015 version of LIWC introduces several new categories and sizable amounts of new words into existing categories, improving and fine graining the program's results. To use the capabilities of the LIWC 2015 program for Dutch text analysis, a Dutch version of the dictionary with the same structure and categories needs to be available. In this paper we therefore present an automated translation of the 2015 version of the LIWC dictionary into Dutch. The 2001 and 2007 versions were both manually translated into Dutch (Zijlstra et al., 2004; Boot et al., 2017). Since the process of manual translation is very labour-intensive, the experiment of trying an automated process is an obvious one. Our provisional translation is, as far as we know, the first LIWC translation based on the 2015 dictionary.

We show a method to automatically translate an English LIWC dictionary into Dutch, by using a pipeline of machine translation and combining part-of-speech tagging with different dictionary expansions through lexica. We first make an automated translation of the LIWC 2007 version and compare it to the manually translated version of this dictionary. The result of the procedure can then be used to evaluate the translation process and to translate the LIWC 2015 dictionary. We developed the pipeline by testing

it on the same corpus that was used in the evaluation of the manually translated version (Boot et al., 2017). Finally, we evaluate the method on the Dutch and English portion of the Dutch Parallel Corpus (Paulussen et al., 2013). For this, we use and extend the evaluation scripts and the Python LIWCtools script (Boot, 2016) that assisted the manual translation. The pipeline, as well as the lexical resources we use, are (in so far as the license allows for it) available in our GitHub repository.[1]

2. Background

2.1 LIWC dictionary

The LIWC program has been designed to work with multiple dictionaries, allowing users to input their own research- or language-specific data files. The program counts the occurrences of words in texts, based on the words contained in its dictionary. It does not take into account the words' context, nor does it do word sense disambiguation.[2] Usually, words will only be included in the dictionary under the category that is relevant for their most frequently used word sense. By the standards of computational linguistics, the program is very simple indeed. Still, it is a widely used research tool (see Tausczik & Pennebaker, 2010, for examples), also widely used outside its original field of psychology.

The LIWC dictionary (Pennebaker et al., 2015a) consists of a number of categories (identified by a number and label) and a number of words or terms, assigned to one or more of these categories. Terms are words or strings ending in the '*' wildcard. As the dictionary contains the term *administrat**, the LIWC program will count *administrator* and *administrative* in categories assigned to *administrat**. In Figure 1 an example of the dictionary layout is shown.

In the 2015 dictionary, there is a possibility to take into account multi-word expressions, though it is used only a few times. The LIWC categories are organised into partial hierarchies. The function word category contains the category of pronouns, which contains the category of personal pronouns, which contains the category of personal pronouns for the first person singular. There are also hierarchies for, among others, social words, for emotions, cognitions, biology, and, new in 2015, drives (a.o. achievement, risk, power).

The content and number of categories in the LIWC dictionaries has increased over the years. While the 2001 dictionary contained 2,319 words, the 2007 version contained 4,487 words and the 2015 version 6,549. The number of categories has been more or less stable (68 categories in 2001, 64 categories in 2007, 76 in 2015). However, both in 2007 and in 2015, a number of categories have disappeared and a number of new ones were created. New words have been added to existing categories, but words have also been removed from categories.

2.2 LIWC translation

The English LIWC dictionary has been translated into many languages, among others German (Wolf et al., 2008), French (Piolat et al., 2011), Spanish (Ramírez-Esparza et al.,

[1] https://github.com/LvanWissen/liwc-translation

[2] From the content of the 2015 English dictionary, it appears there might be a way of taking into account previous words' content or category. If this works, it would be an undocumented feature, and apparently only used to distinguish the various uses of (American) English *like*.

```
80          drives (Drives)
81          affiliation (Affiliation)
82          achieve (Achievement)
83          power (Power)
%
additional  21
address     112
adds        25  80 84  91
adequa*     80  82
adjust*     50  56
administr*  80  83 110
```

Figure 1: Example layout of a LIWC dictionary taken from the 2015 internal dictionary. The upper part of the excerpt shows categories (by number) and their definition. The lower part lists words and terms that are each assigned to one or multiple categories. The term *adequa** as well as all the words from a text starting with this string are for example assigned the 'drives (Drives)' and the 'achieve (Achievement)' categories.

2007) and Chinese (Gao et al., 2013). Translating a LIWC dictionary is not as straightforward as finding one or multiple equivalents for the English words. We mention three general complications. (i) Because words are assigned to multiple categories, the translator will have to check which equivalents fit into which categories. This led the creators of the Dutch 2007 translation to translate a word multiple times, for each of the categories in which it appeared. (ii) Another complication is presented by the wildcards: before an entry such as *manag** is translated, it has to be expanded into *manager*, *management*, *manageable*, *manage*, etc. (iii) Finally, in some cases, translating the dictionary requires finding corresponding words in a different culture. The Dutch 2007 translation for example includes names of Dutch labour unions in the category 'work', and Dutch beverages in the category 'leisure'.

Other problems are related to specific ways in which languages differ from English. In Romance languages, verbs are conjugated into many different forms. Do all of these forms have to be included in the dictionary? Because the subject of the sentence can often be deduced from the verb form, these languages use less personal pronouns than English does. To what extent does the translation need to take that into account? For Dutch a significant difference from English is its use of composite words: the English dictionary contains the entries *drug* and *addict**, but the Dutch equivalent of *drug addict* is a composite word *drugsverslaafde*, which would not necessarily appear in the dictionary when translating individual words.

Because of this, the translation of an LIWC represents a significant amount of work. The Dutch upgrade of the 2001 translation to 2007 took eight years. Yet, all translations known to us were compiled manually, except the translation into Catalan (Massó et al., 2013). Masso and his colleagues created a Catalan LIWC dictionary by automatically translating LIWC dictionaries from other Romance languages into Catalan. The main focus of their efforts is in assigning the words in the translation to the correct categories. They do not report an evaluation of their dictionary on a (parallel) corpus.

3. Translation procedure

We have developed a translation pipeline to translate an English LIWC dictionary into Dutch, which consists of the following steps:

3.1 Initalisation

The LIWC internal English dictionary is read and stored into a data structure that is listing words and their respective categories in a machine readable form. The categories from the source term are copied as is, with the exception of the function word categories (see below).

3.2 Wildcard expansion

Terms ending in an asterisk (*), which represent every word form in a text that starts with the preceding string, are resolved by looking for matching words in the Google n-gram corpus (Brants & Franz, 2006).[3] We use the frequency list of the unigram model. In order to remove noise, we only extract words that have a minimal frequency of 750,000 (which scales the corpus down to 46,717 tokens).

3.3 Translation

All words are sent for translation to the Google Translate interface[4] for a word to word translation. Since the online translations are bound to change due to improvements in the algorithm or user contributions and corrections, we store the translations to replicate and backtrack the procedure, if necessary.

3.4 Filtering

To prevent non-existing (malformed or not translated) Dutch words from entering the dictionary, words that are returned from the translation query are removed if they do not occur as token entry in the e-Lex corpus (NTU (Nederlandse Taalunie) [Dutch Language Union], 2006). We also discard any multiword expressions returned by the online translation.

3.5 Tagging

All translations are in this step tagged with part of speech information by TreeTagger (Schmid, 1994). The POS tags are converted to LIWC (function word) categories which are then added to the word's category information. We implement a conversion from POS tags to LIWC categories by using rules of the type shown in Figure 2.

3.6 Adding lemmas

In the same call, TreeTagger returns a lemmatised form of a word, which we recursively also tag, convert to LIWC functional categories using the same table and add to the dictionary as a separate entry.

[3] This corpus dates from 2006 and contains approximately 1 trillion words from the web from mostly English web pages. It is available online through the Linguistic Data Consortium (LDC).

[4] https://translate.google.com/. Although translating to Dutch was already possible for a long time, Google recently updated the system to include Dutch in its new Neural Machine Translation (Wu et al., 2016).

POS	description	LIWC-category	
adj	adjective	21	adjective
adv	adverb	13	adverb
conjcoord	coord. conjunction	14	conj
det_art	article	10	article
det_indef	indefinite pronoun	2,9	pron,ipron
det_poss	possessive pronoun	2	pron
int	interjection	125	filler

Figure 2: Example from a set of POS tags and their corresponding LIWC function word categories. We apply this mapping after tagging the words.

3.7 Adding other word forms

As a final step, we further extend the dictionary with word forms from a lemma list (NTU (Nederlandse Taalunie) [Dutch Language Union], 2015), which we again tag and add to the dictionary with both functional and content categories. If the word already exists, the category information is merged so that there exists only one entry in the resulting dictionary.

3.8 Handling function words

Since translating pronouns by a (statistical) machine translation system is known to be harder than translating content words due to differences in the way a language deals with pronouns (Guillou et al., 2016), we have chosen to exclude most function words from the translation process described above. We fill these categories based on the POS-tagging in the e-Lex lexicon[5] (NTU (Nederlandse Taalunie) [Dutch Language Union], 2006). We query the lexicon and ask it to return a list of all words meeting specified POS and category criteria. We retrieve for instance all first person singular pronouns by asking for all words that have POS equal to 'VNW' (voornaamwoord [=pronoun]) with categories '1' (first person) and 'ev' (enkelvoud [=singular]). The output is given in Figure 3. We add all those words to the dictionary, in the 'I'-category.[6]

mijzelf, m'n, mezelve, ik, mij, ikzelf, mijne me, eigen, mijn, waterdragen, 'k, mijns

Figure 3: List of first person singular pronouns from e-Lex for the 'I' category in LIWC.

3.9 Remove function words from content categories

We use similar lookups for words that we only allow in a certain category. Translation artifacts, faulty translations or inconsistencies in the lexicon can for example put a determiner inside one of the content categories, and its high frequency would have a large effect on the category scores. We specify for example that all determiners from e-Lex may only occur in the 'det' category of the LIWC dictionary.

[5] Formerly the TST-lexicon. The e-Lex lexicon is a Dutch lexicon (we use the one-word version) that contains over 600,000 word forms in ca. 200,000 entries with POS and category information (e.g. gender and number).

[6] The word 'waterdragen' (i.e. 'carry water', 'domestic service') is obviously an error. e-Lex is constructed from several other corpora that have been annotated semi-automatically and as such can contain errors. However, the problems that we found are minor.

3.10 Extending hierarchy

The LIWC dictionary has a hierarchical structure. As a final step in the translation pipeline we extend the scope of terms by also adding the parent category to its categories. This means that we also add a word that is part of the 'health' category (category id 72) to the parent 'bio' (category id 70) category. We use the completion function of LIWCtools (Boot, 2016) for this step, which takes the existing English dictionary as a model and projects its structure onto the newly translated Dutch one.

3.11 Wrap-up

When the translation is complete, the dictionary is stored in a format that can be used in the official LIWC program.

3.12 Manual correction

Although the dictionary that is created in the automatic procedure performs acceptably (see the sections below), errors are inevitable. The more frequent words among the errors have a measurable effect on the outcome. We decided to add a manual correction step to remove those from the dictionary. What we did was to compute, for each LIWC category and for both the Dutch and English dictionary, a list of the words that accounted for more than 1.5% of the hits in that category. For most categories, this produces a list of ca. 10 to 15 words. For the English words, we manually checked whether their main translation(s) occurred in the generated dictionary. If not, we added them. For the Dutch words, we checked whether these words belonged in the category. If not, we removed them. We also did a superficial inspection of the translated dictionary and corrected some of the more obvious errors.

4. Evaluation procedure

4.1 Corpus

The translation pipeline was designed, developed and tested on the same set of parallel Dutch and English texts that was used by Boot et al. (2017). The test corpus includes letters of Vincent Van Gogh, documents from the European parliament, TED-talk subtitles and Bible books. This corpus is also used to test the efficacy of the manual corrections to the dictionary.

In order to avoid the risk of overfitting to this development corpus, we use a separate corpus for the final evaluation of the dictionary. Here we use the Dutch Parallel Corpus (DPC, Paulussen et al., 2013).[7] From the test and evaluation corpora, we remove files with a low word count (<1,000) to prevent small files from influencing the results.

4.2 Calculations

We use the count functionality of LIWCtools to replicate the textual analysis function of the official LIWC software. Each Dutch text from the DPC is processed using the

[7] A corpus built from Dutch and English texts coming from a broad range of fields such as finance, science, culture and communication.

translated dictionary. Its English equivalent is processed by the English dictionary. The result is a table containing the coverage (expressed in relative frequency) per dictionary category (columns) for each individual processed file (rows). A sample is shown in Figure 4 below.

Filename	**function**	**pronoun**	**ppron**	**i**
education/dpc-vla-001191.txt	0.479	0.079	0.041	0.000
education/dpc-vla-001172.txt	0.482	0.05	0.029	0.000
education/dpc-mis-001909.txt	0.488	0.069	0.046	0.001
institutions/dpc-bal-001241.txt	0.54	0.142	0.088	0.011
institutions/dpc-gim-002525.txt	0.424	0.076	0.051	0.005

Figure 4: Example of the output that is created after processing text files from the parallel corpus. Shown is an excerpt of the data that shows five processed files (rows) and the share of several categories (columns) of the total amount of words of the text file. The format of the file is very close to the output of the official LIWC program.

We then calculate a correlation score and effect size (Cohen, 1992) for the corresponding columns (e.g. the function words in the Dutch texts with the function words in the English texts). Based on whether the data are normally distributed, either a Pearson or a Spearman correlation measure is used. For both English and Dutch we also compute the median, minimum and maximum frequencies.

The target values for our automatic translation are those of the Dutch manual (gold) translation of LIWC 2007. This translation achieved an average correlation of 0.77 with the English dictionary (effect size 0.39) on the DPC.[8]

5. Evaluation for the 2007 LIWC dictionary

We evaluate our automatic approach by comparing the correlation coefficient and effect sizes between the English 2007 dictionary and the manual translation with those for the English dictionary and the automatic translation.

As mentioned above, evaluating the manually translated 2007 dictionary on the DPC corpus results in an average correlation score of 0.77 (effect 0.39). Our automatically translated 2007 dictionary, without a manual correction step, scores a bit less with an average correlation coefficient of 0.72 (effect 0.72). Our translation does especially well for the function word categories with most correlations above 0.80. Only the impersonal pronouns category ('ipron') scores much lower compared to the manual translation. This is probably due to the word *niet* [=not] being included in the translation, which accounts for ca. 40% of the 'ipron' category. The adverb category is problematic too, as it has an effect size of 6.21. This is because a number of prepositions ended up in this category.

For the content word categories, some actually do better than the manual translation, e.g. 'home'. Given the large numbers of words in these categories, it is hard to say what is the cause of this improvement. The categories 'inclusive', 'body', 'ingest', 'time' and 'leisure' score lower on correlation. For the 'body' category, this is probably largely due

[8] This comparison and performance test was already done when the Dutch 2007 dictionary was presented (Boot et al., 2017). The translators then achieved a correlation of 0.80 (effect size: 0.35) on their test set. We did this comparison again on our own evaluation corpus.

to the ambiguous words *haar* [=hair, her] and *enkel* [=ankle, solely]. In other cases it is impossible to point to a few words to explain an unsatisfactory result. Some other categories do not score that well in the manual translation either (e.g. 'feel' and 'motion'). For the 'swear' category, this might be due to a lack of testing material in the corpus.

From preliminary testing, we know that a manual correction step can improve the result of the automatic 2007 translation with ca. 0.04 (correlation) and -.20 (effect size). That would bring us quite close to the results of the manual translation.

6. Evaluation for the 2015 LIWC dictionary

6.1 Procedure

For the automatic translation of the 2015 dictionary, we do not have the manual translation to compare the results. What we do have is the possibility to compare the results with that of the English dictionary on our test corpus. We first do an automatic translation and test the result against the test corpus, then add a manual correction and test again against the test corpus. Finally, we evaluate the end result against the evaluation corpus.

6.2 Results

Table 1 shows the average correlations and effect sizes for the different conditions. The initial automatic 2015 translation scores somewhat lower than the automatic 2007 translation. While most categories perform somewhere between acceptably and very well, the informal word categories perform very bad. The correction step does have a measurable effect, an effect that is largely retained when testing against the evaluation corpus.

Dictionary	**Corpus**	**Correlation d**	**Effect size r**
Automated 2015 translation	Test corpus	0.69	0.88
Automated 2015 translation with correction	Test corpus	0.73	0.52
Automated 2015 translation with correction	Evaluation corpus	0.73	0.59

r: correlation, d: effect size (Cohen's d).

Table 1: Average correlation coefficients and effect sizes for the Dutch LIWC 2015 dictionary.

6.3 Results by category

The numbers shown in Table 2 below give the results by category of the corrected dictionary on the evaluation corpus. The table should provide researchers with the information necessary to decide which LIWC 2015 categories should work the same in a Dutch-language context as in an English context.

By and large, the function word categories perform very well. Exceptions are the new categories 'adjectives', 'comparatives' ('greater', 'greatest', etc.) and 'interrogatives' ('where', 'how', etc.). For the adjectives, the explanation may be that the translation contains many more adjectives than the original; for the interrogatives, the explanation may be that in both languages these words can also occur as adverbs or pronouns. These categories clearly need more work, as does the category of quantitative words, which scores inexplicably low.

Some of the function word categories profited significantly from the manual correction, such as 'shehe' where we removed the male possessive pronoun *zijn*, as it is more frequently used as a verb (*to be*). For other categories we added words missing in the translation, such as the demonstrative pronouns that should have been in the impersonal pronouns category.

The psychological categories of emotion, social words and cognitive words again perform rather well. From the 'insight' category, maybe we should have removed *worden* [=become], which is translated correctly, but also serves as a passive auxiliary verb in Dutch. From 'friends', maybe we should have removed the word *kennis* which in Dutch is *acquaintance* as well as *knowledge*. The biological categories are less satisfactory, without clear culprits. In contrast, the new categories under 'drives' ('affiliation', 'power', 'reward', 'risk') perform generally well.

From the 'time orientation' group, 'focusfuture' could perform better. We might try to remove the verb *gaan* [=to go] which is often but certainly not always used to express a focus on the future. The categories from the 'personal concerns' group do generally well. But as noted, the informal categories perform very poorly. This was also true, though not quite to this extent, in the manual LIWC 2007 translation. The results are probably to some extent due to the test and evaluation corpora, that are heavily oriented to written language, and certainly do not contain terms from the netspeak category (a category where Dutch borrowed lots of terms from English). Another issue is probably that the translation engine will have been trained on written language. There are also some problems with the English categories: the 'nonfluencies' category for instance contains the word *well*, which is responsible for 85% of the category count, but of course has many other uses besides its use as a nonfluency. And, finally, in these categories cultural differences may play an important role. For example, Dutch often uses names of illnesses as swear words (Fletcher, 1996).

	Word counts						**Equivalence statistics**	
Category	**English**			**Dutch**			***r***	***d***
	Median	***Min***	***Max***	***Median***	***Min***	***Max***		
Word count	2,179	1,003	122,206	2,169	999	128,338	0.99*	0.00
Linguistic dimensions								
function words	46.60	27.96	60.67	51.33	32.37	62.62	0.94	0.94
pronoun	7.55	1.34	22.05	9.24	1.86	23.25	0.97	0.37
ppron	3.30	0.05	16.48	3.72	0.17	15.90	0.98	0.11
I	0.23	0.00	7.67	0.19	0.00	7.73	0.95*	0.03
we	0.56	0.00	3.99	0.52	0.00	4.24	0.97	0.07
you	0.18	0.00	2.40	0.26	0.00	2.33	0.91*	0.07
shehe	0.20	0.00	7.09	0.82	0.00	8.89	0.80*	0.30
they	0.51	0.00	3.31	0.75	0.00	5.85	0.79*	0.50
ipron	3.85	0.86	7.58	4.20	1.10	7.69	0.86	0.18
article	9.25	5.01	17.06	12.12	6.79	17.96	0.81	1.48
prep	15.29	9.83	20.39	16.48	12.49	21.51	0.74	0.78
auxverb	6.22	2.04	10.12	5.77	1.72	9.09	0.77	0.32
adverb	3.06	0.65	6.92	6.07	1.60	11.54	0.83	1.90

	Word counts						Equivalence statistics	
Category	**English**			**Dutch**			***r***	***d***
	Median	***Min***	***Max***	***Median***	***Min***	***Max***		
conj	5.24	2.18	8.22	6.03	2.87	9.46	0.85	0.74
negate	0.70	0.00	2.42	0.86	0.00	3.10	0.85	0.37
Other grammar								
verb	10.48	3.67	20.20	11.54	4.43	20.00	0.90	0.47
adj	4.43	1.76	7.83	6.28	3.19	10.63	0.52	1.84
compare	2.41	0.93	5.74	3.35	1.99	7.05	0.50	1.54
interrog	1.06	0.12	2.59	0.84	0.06	3.00	0.45	0.33
number	0.95	0.13	6.90	1.12	0.00	5.53	0.79*	0.16
quant	1.87	0.58	3.45	1.76	0.45	5.30	0.31	0.10
Psychological processes								
affect	4.12	0.97	9.50	2.62	0.79	7.09	0.81	1.13
posemo	2.80	0.44	7.35	1.70	0.47	4.94	0.77	1.07
negemo	1.03	0.00	7.02	0.83	0.00	5.30	0.85*	0.42
anx	0.21	0.00	2.59	0.18	0.00	1.27	0.71*	0.23
anger	0.19	0.00	3.81	0.15	0.00	2.30	0.82*	0.30
sad	0.19	0.00	1.16	0.19	0.00	0.84	0.57*	0.17
social	6.64	0.22	18.00	6.55	1.28	16.85	0.95	0.06
family	0.05	0.00	3.89	0.07	0.00	4.04	0.80*	0.16
friend	0.15	0.00	1.22	0.12	0.00	1.21	0.59*	0.20
female	0.07	0.00	7.21	0.48	0.00	7.92	0.67*	0.32
male	0.31	0.00	7.03	1.15	0.19	7.36	0.87*	0.55
cogproc	8.58	2.51	16.69	10.07	5.15	16.29	0.84	0.72
insight	1.78	0.39	3.98	2.41	0.94	4.99	0.56	1.04
cause	1.77	0.49	4.30	1.29	0.34	4.03	0.74	0.82
discrep	0.96	0.07	3.35	1.97	0.34	5.46	0.77	1.42
tentat	1.57	0.15	6.49	1.80	0.45	4.55	0.78*	0.30
certain	1.15	0.19	2.88	1.23	0.19	3.14	0.73	0.18
differ	2.08	0.09	5.76	2.44	0.52	5.45	0.88	0.33
percept	1.29	0.07	7.25	0.99	0.04	4.71	0.87*	0.39
see	0.54	0.00	5.88	0.43	0.00	3.39	0.74*	0.36
hear	0.30	0.00	3.45	0.23	0.00	2.77	0.89*	0.20
feel	0.24	0.00	2.05	0.22	0.00	2.48	0.61*	0.21
bio	0.82	0.00	7.06	0.51	0.05	4.71	0.75*	0.48
body	0.17	0.00	3.29	0.16	0.00	2.49	0.69	0.16
health	0.41	0.00	5.09	0.23	0.00	3.11	0.62*	0.48
sexual	0.00	0.00	2.63	0.00	0.00	1.60	0.60*	0.16
ingest	0.15	0.00	2.94	0.10	0.00	1.43	0.64*	0.32
drives	7.75	2.62	16.28	5.49	1.94	12.50	0.88	0.89
affiliation	1.86	0.00	7.28	1.47	0.08	6.11	0.90	0.26
achieve	1.76	0.15	4.75	1.40	0.19	3.49	0.80	0.61
power	3.12	1.20	9.75	2.09	0.61	7.50	0.76	0.86
reward	1.02	0.07	2.63	0.78	0.00	2.67	0.65	0.61

	Word counts						Equivalence statistics	
Category	**English**			**Dutch**			***r***	***d***
	Median	***Min***	***Max***	***Median***	***Min***	***Max***		
risk	0.58	0.00	4.31	0.44	0.00	2.60	0.71*	0.40
Time orientation								
focuspast	2.22	0.49	10.76	3.38	1.38	10.93	0.87*	0.57
focus present	6.52	1.96	11.71	9.39	3.37	15.64	0.66	1.30
focusfuture	0.97	0.15	4.04	1.85	0.52	4.24	0.63*	1.29
relativ	13.87	7.74	19.28	13.97	9.56	19.26	0.75	0.06
motion	1.62	0.32	4.60	1.42	0.29	2.97	0.55	0.53
space	7.89	3.43	13.71	7.70	4.45	12.18	0.76	0.15
time	4.24	1.63	7.75	5.12	2.50	7.44	0.71	0.82
Personal concerns								
work	4.82	0.53	14.60	2.77	0.46	9.22	0.85	0.95
leisure	0.43	0.00	4.48	0.26	0.00	3.36	0.80*	0.45
home	0.19	0.00	2.02	0.11	0.00	2.18	0.64*	0.37
money	1.14	0.00	9.22	0.71	0.00	5.95	0.92*	0.55
relig	0.09	0.00	6.02	0.03	0.00	3.54	0.68*	0.24
death	0.06	0.00	2.23	0.03	0.00	1.80	0.81*	0.20
informal	0.17	0.00	2.99	1.23	0.07	3.86	0.31*	2.22
swear	0.00	0.00	0.40	0.00	0.00	0.26	0.42*	0.22
netspeak	0.00	0.00	2.99	0.21	0.00	3.38	0.35*	0.80
assent	0.03	0.00	0.60	0.00	0.00	0.40	0.50*	0.43
nonflu	0.07	0.00	0.43	0.00	0.00	0.09	0.17*	1.51
filler	0.00	0.00	0.13	0.95	0.07	3.31	0.20*	2.71

r: correlation, d: effect size (Cohen's d).

*: Correlations with * were computed using Spearman's rank correlation coefficient.

Table 2: Results of equivalence test on translated Dutch and English dictionary.

7. Conclusion

We presented a pipeline for automatic translation of the LIWC dictionary from English into Dutch. The result of a comparison between an automatic translation of the 2007 and the manually translated version shows that the automatic translation is nearly as good as the manual one when looking at the correlation coefficients. When repeating this translation procedure for the new 2015 dictionary, we are able to produce a dictionary with an average correlation coefficient of 0.69 (effect 0.88) to the English dictionary. Manual correcting boosts these numbers to 0.73 (effect 0.59), a score that is again very close to the one reached by the manual (2007) translation.

We should note that the correlations for the informal word categories (netspeak, swear words, etc.) are considerably less satisfactory. There are a number of underperforming categories as well among the psychological processes and function words. Still, the au-

tomatic translation as a whole performs well. This is all the more remarkable as our automatic translation does not take into account some of the aspects of translation that we discussed in the Background section 2.2 and that a human translator will care about, such as the assignment of translated words to the fitting LIWC category or the use of composite words in Dutch.

Given the fact that a manual translation of an LIWC dictionary is a very time-consuming task, the automatic translation should therefore be considered a serious alternative, at least for those languages for which a sufficient number of linguistic resources is available. Further improvement (manual or automatic) is always possible.

As is unavoidable in any automatic treatment of language, the translated dictionary does contain errors. However, given the fact that the categories contain many words, most only responsible for a tiny fraction of the total of words in its category, errors are not necessarily problematic. It is also in the nature of a tool such as LIWC, that does not do word sense disambiguation, that words are occasionally misclassified. In spite of the errors, the resulting (provisional) dictionary should be usable for most research purposes. We invite researchers in psychology, digital humanities and other fields to validate its usability in the context of practical research.

8. Acknowledgements

We thank Isa Maks for the suggestion that led to this paper and we thank Sem Zweekhorst for his contributions to the manual corrections.

9. References

Boot, P. (2016). LIWCtools. Tools for working with LIWC dictionaries (Version 0.0.1). URL https://github.com/pboot/LIWCtools.

Boot, P., Zijlstra, H. & Geenen, R. (2017). The Dutch translation of the Linguistic Inguiry and Word Count (LIWC) 2007 dictionary. *Dutch Journal of Applied Linguistics*, 6(1).

Brants, T. & Franz, A. (2006). Web 1T 5-gram corpus version 1.1. *Google Inc.*

Cohen, J. (1992). A power primer. *Psychological bulletin*, 112(1), p. 155.

Fletcher, W.H. (1996). Come down with cholera: Disease names in Dutch strong language. *Canadian Journal of Netherlandic Studies*, 17, pp. 231–239.

Gao, R., Hao, B., Li, H., Gao, Y. & Zhu, T. (2013). Developing simplified Chinese psychological linguistic analysis dictionary for microblog. In *International Conference on Brain and Health Informatics*. Springer, pp. 359–368.

Guillou, L., Hardmeier, C., Nakov, P., Stymne, S., Tiedemann, J., Versley, Y., Cettolo, M., Webber, B. & Popescu-Belis, A. (2016). Findings of the 2016 WMT shared task on cross-lingual pronoun prediction. In *Proceedings of the First Conference on Machine Translation (WMT16), Berlin, Germany. Association for Computational Linguistics*. pp. 525–542.

Massó, G., Lambert, P., Penagos, C.R. & Saurí, R. (2013). Generating New LIWC Dictionaries by Triangulation. In *Asia Information Retrieval Symposium*. Springer, pp. 263–271.

NTU (Nederlandse Taalunie) [Dutch Language Union] (2006). e-Lex Version 1.1.1.

NTU (Nederlandse Taalunie) [Dutch Language Union] (2015). URL http://taalunieversum.org/sites/tuv/files/downloads/dutch_lemmas.txt.

Paulussen, H., Macken, L., Vandeweghe, W. & Desmet, P. (2013). Dutch parallel corpus: A balanced parallel corpus for Dutch-English and Dutch-French. In *Essential Speech and language technology for Dutch*. Springer, pp. 185–199.

Pennebaker, J.W., Booth, R.J., Boyd, R. & Francis, M.E. (2015a). Linguistic Inquiry and Word Count: LIWC2015 Operator's Manual. URL https://s3-us-west-2.amazonaws.com/downloads.liwc.net/LIWC2015_OperatorManual.pdf.

Pennebaker, J.W., Boyd, R.L., Jordan, K. & Blackburn, K. (2015b). The development and psychometric properties of LIWC2015. URL http://www.liwc.net.

Pennebaker, J.W., Chung, C.K., Irel, M., Gonzales, A., Booth, R.J. & Framework, T.L. (2007). The Development and Psychometric Properties of LIWC2007. URL http://www.liwc.net/LIWC2007LanguageManual.pdf.

Pennebaker, J.W., Francis, M.E. & Booth, R.J. (2001). Linguistic inquiry and word count: LIWC 2001. *Mahway: Lawrence Erlbaum Associates*, 71.

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), pp. 145–159.

Ramírez-Esparza, N., Pennebaker, J.W., García, F.A., Suriá Martínez, R. et al. (2007). La psicología del uso de las palabras: Un programa de computadora que analiza textos en español. *Revista Mexicana de Psicología*, 24(1), pp. 85–99.

Schmid, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of International Conference on New Methods in Language Processing*. p. 154.

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), pp. 24–54.

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), pp. 85–98.

Wu, Y., Schuster, M., Chen, Z., Le, Q.V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K. et al. (2016). Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. *arXiv preprint arXiv:1609.08144*.

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