Machine-translated texts as an alternative to translated dictionaries for LIWC

Peter Boot
Huygens Institute for the History of the Netherlands

Author Note
Peter Boot 0000-0002-7399-3539
Correspondence concerning this article should be addressed to Peter Boot,
peter.boot@huygens.knaw.nl.

Machine-translated texts or translated dictionaries for LIWC

Abstract

Linguistic Inquiry and Word Count (LIWC) is a text analysis program developed by James Pennebaker and colleagues. At the basis of LIWC is a dictionary that assigns words to categories. This dictionary is specific to English. Researchers who want to use LIWC on non-English texts have typically relied on translations of the dictionary into the language of the texts. Dictionary translation, however, is a labour-intensive procedure. In this paper, we investigate an alternative approach: to use Machine Translation (MT) to translate the texts that must be analysed into English, and then use the English dictionary to analyse the texts. We test several LIWC versions, languages and MT engines, and consistently find the machine-translated text approach performs better than the translated-dictionary approach. We argue that for languages for which effective MT technology is available, there is no need to create new LIWC dictionary translations.

Introduction

Linguistic Inquiry and Word Count (LIWC) is a text analysis program developed by James Pennebaker and colleagues (Pennebaker, Boyd, Jordan, & Blackburn, 2015). LIWC counts the proportions of words in texts that belong to certain categories (grammatical, social, emotional, cognitive, biological and others). LIWC has been frequently used in psychological research, management studies, criminology, and also increasingly in a Digital Humanities context. Examples are the use of LIWC in an analysis of fiction (Piper, 2016) or of gender bias in newspapers (Wevers, 2019).

At the basis of LIWC is a dictionary that assigns words to categories. This dictionary is specific to English. Researchers that want to use LIWC on non-English texts have typically relied on translations of the dictionary into the language of the texts. Dictionary translation, however, is a labour-intensive procedure. In this paper, we look at an alternative approach: to use Machine Translation (MT) to translate the texts that must be analysed into English, and then use the English dictionary to analyse the texts. In a topic modelling context, MT has been shown to be an effective tool for transcending the language barrier (De Vries, Schoonvelde, & Schumacher, 2018). We compare a translated-dictionary and a translated-text approach and try to establish which one is more effective.

This article is structured as follows: we provide background for LIWC and LIWC dictionary translation, and for the use of MT in a text analysis context. In the method section, we discuss the research set-up, the choice of MT engines and the parallel corpora that we use. The results section provides results. In the analysis section we try to explain the differences between the translated-dictionary and translated-text approach. In a post-hoc analysis section we provide some results at the level of individual LIWC categories and we apply our results to the situation of a multilingual document collection. In the final section we draw some conclusions.

Background: LIWC and LIWC dictionaries

LIWC is a text analysis tool that has been developed in the context of a writing treatment for patients with traumatic experiences, and has been used for investigating cognitive style, personality and social integration (Pennebaker & Graybeal, 2001). As its creators have always admitted, LIWC is a simple tool, that does not take into account multiple meanings of words, context, sarcasm and many other features of text (Pennebaker, Mehl, & Niederhoffer, 2003). Research using the various LIWC categories, however, has proven to be very successful. The most recent overview is given by Tausczik and Pennebaker (2010).

Various versions of the LIWC programme and dictionary have been released (2001, 2007 and 2015). Both the categories in the dictionaries and the words that make up the categories have been subject to change. The latest version of the dictionary (Pennebaker et al., 2015) contains 76 categories and 6549 words (sometimes with a wildcard). A notable change in the latest version of the dictionary is the presence of a number of summary variables, such as analytical thinking and authenticity, derived from the other categories.

For the application of LIWC in languages other than English, the LIWC dictionary has been (fully or partially) translated into many languages, including Arabic (Hayeri, 2014), Chinese (Huang et al., 2012), French (Piolat, Booth, Chung, Davids, & Pennebaker, 2011), Italian (Alparone, Caso, Solano,

& Prezza, 2002), (Brazilian) Portuguese (the 2007 and 2015 dictionaries resp. Balage Filho, Pardo, & Aluísio, 2013; Carvalho et al., 2019), Russian (Kailer & Chung, 2011) and Japanese (Shibata, Wakamiya, Aramaki, & Kinoshita, 2016). In this article we use the Dutch 2007 and 2015 translations (Boot, Zijlstra, & Geenen, 2017; Van Wissen & Boot, 2017), the German 2001 and 2015 translations (Meier et al., 2019; Wolf et al., 2008) and the Spanish 2007 translation (Pennebaker Conglomerates, 2015).

Most of these translations were created by translating word by word the English LIWC dictionary. In some cases, the translators could base themselves on an earlier translation in their language. The Dutch 2007 dictionary for instance is based on the 2001 dictionary (Zijlstra, Van Meerveld, Van Middendorp, Pennebaker, & Geenen, 2004) Some translators started from an existing corpus of words in their languages, assigning these to LIWC categories (Andrei, 2014) or enriched the dictionary from an existing corpus (Gao, Hao, Li, Gao, & Zhu, 2013; Meier et al., 2019). Two translation teams opted for machine translation of the dictionary: a Catalan dictionary was created based on dictionaries in closely related languages (Massó, Lambert, Penagos, & Saurí, 2013); the Dutch 20015 translation was mostly based on a word by word translation of the English dictionary using Google translate (Van Wissen & Boot, 2017).

Some translators report no experimental data that validates their dictionary. Most translators, however, use a parallel corpus to validate their dictionary: they apply their own dictionary to the texts in their language and the English dictionary to the texts in English. Then they compute, for each of the LIWC categories, correlations and effect sizes between the two outputs. Some translators have also applied tests for internal consistency of the various categories (Meier et al., 2019). Other translation teams have applied their dictionary in actual research to establish its usefulness (e.g. Ramirez-Esparza, Pennebaker, García, & Suriá, 2007; Zijlstra, Van Middendorp, Van Meerveld, & Geenen, 2005), by checking consistency with e.g. other sentiment analysis measures (Balage Filho et al., 2013) or by checking the performance in text classification tasks (Carvalho et al., 2019).

Whatever the procedure used for creating a LIWC dictionary in a new language, it is a time-consuming effort. Even if translation of the words themselves is straightforward, there are many pitfalls. Languages use words differently. To give a few examples (Meier et al., 2019; Ramirez-Esparza et al., 2007; Van Wissen & Boot, 2017): (1) German capitalised *Sie* is a formal second-person pronoun, used in the singular and plural, while lower-case *sie* is equivalent to English *she* and *they*. As LIWC is not case-sensitive, the You, Shehe and They categories in German would become meaningless numbers if no precautions are taken. (2) Many Romance languages can omit the personal pronoun as the subject expressed by the verb's ending (Spanish *quiero*: 'I love'). LIWC results for personal pronouns in these languages will be much lower. The Spanish dictionary has responded by creating categories for verb + person + number, e.g. VerbI and VerbUs, and can therefore still give the number of, e.g. I-references. (3) Dutch has a group of highly frequent words called pronominal adverbs, that combine pronouns or adverbs with prepositions (compare English *wherein*). They sit uncomfortably between the LIWC categories Pronoun, Adverb and Preposition. (4) Many words are culture-specific or country-specific: the beverages in the LIWC Ingest category are American beverages, LSAT is an American test, etc. (5) most words have multiple meanings, and they may belong in different categories. Just one example: English *since* has a temporary meaning (*ever since*) and a causal one (*as*). These meanings have different equivalents in Dutch and should be assigned to different categories.

Machine-translated texts or translated dictionaries for LIWC

Given the efforts it takes to translate a dictionary, it is a natural question to ask whether the alternative procedure – translating the text and analysing it using the English-language dictionary – wouldn't be more efficient. We discuss this in the next section.

Background: Machine Translation and text analysis

That text analysis tools such as LIWC can be usefully applied to machine translated text is something that many researchers assume without comment. Agarwal, Xie, Vovsha, Rambow, and Passonneau (2011) for example do sentiment analysis on Twitter data translated through Google Translate without apparently asking whether the MT invalidates the procedure. There exists, however, a body of research that explicitly investigates this question. For topic modelling, De Vries et al. (2018) have shown that topic models derived from machine translated text are quite similar to those derived from manual translation. Reber (2019) showed that even MT of individual lemmas, rather than full text MT, can be an acceptable and cost-effective alternative. Lind, Eberl, Galyga, et al. (2019) mention cost as a reason to avoid MT and prefer polylingual topic modelling, where the bridge over the language gap is created by e.g. comparable documents in each language.

For sentiment analysis a comprehensive study is reported in (Araújo, Reis, Pereira, & Benevenuto, 2016). They machine translate datasets labelled for sentiment in nine different languages into English and apply twenty-one English sentiment detection tools. Many English-language tools perform quite well in predicting the labels. For those datasets where they also applied sentiment analysis tools for the original language, these did not perform better than the English language tools. This confirms earlier findings by Balahur and Turchi (2012), who showed that MT systems are mature enough to create sentiment analysis training data for other languages from English resources. In the field of subjectivity analysis Banea, Mihalcea, Wiebe, and Hassan (2008) already showed that MT was able to create training resources for other languages based on English ones.

In their introduction Araújo et al. (2016) write 'Most existing strategies in specific languages consist of adapting existing lexical resources, without presenting proper validations and basic baseline comparisons'. That is not entirely fair to most LIWC translators, who, as we saw above, certainly present validating data. But none of them have compared the effectiveness of translating the dictionary and translating the texts. The issue has been addressed for a number of other text analysis tools based on dictionaries (Lind, Eberl, Heidenreich, & Boomgaarden, 2019; Proksch, Lowe, Wäckerle, & Soroka, 2019). Both studies, however, assume machine translation of the dictionary, which in the LIWC context, as we saw, is the exception. Lind and colleagues, in the context of researching sentiment about migration issues, found that machine translated text works better and state: 'machine translation errors have subsequently larger consequences if researchers translate the sensitive instrument (…) rather than the corpus (…)'. Proksch and colleagues, on the contrary, investigating sentiment in legislative debate, prefer translating the dictionary: 'The shared nature of sentiment means both that we have much more confidence in translations of sentiment terms than we do in translations—particularly machine translations—of any substantive political topic'.

Specifically for LIWC – which measures much more than sentiment – the effectiveness of MT was researched by Windsor, Cupit, and Windsor (2019). They compare LIWC output of manually translated text and machine translated text for six very different languages, and found only relatively small differences. The background of their study is English-speaking politicologists wanting to use

foreign-language texts, and not that of a researcher who has texts in his or her own language that he or she needs to analyse. They ignore the possibility of using a translated LIWC dictionary.

That LIWC outputs should be more or less stable under human translation is something that has been assumed by the LIWC translators who validated their translation by comparing LIWC output on a corpus of translated texts. If, as Windsor and colleagues showed, machine translation is nearly as good as manual translation, it is at least plausible that the translated-text approach for LIWC should be effective.

Method

Research set-up

The most straightforward way of testing the suitability of a translated-text approach in applying LIWC would be to find a corpus in the working language, machine-translate it into English, apply LIWC to the original and translated corpus (using respectively the working language's LIWC dictionary and the English dictionary), and compare both outputs. If we assume the working language's translated LIWC dictionary is sufficiently validated, we could attribute differences between the LIWC outputs to limitations of the translated-text approach. However, LIWC dictionary translations are never perfect. As we noticed, they themselves are often validated by applying LIWC to a parallel corpus of working language texts manually translated from or into English and computing, for each of the LIWC categories, correlations and effect sizes between LIWC outputs. If we want to compare the suitability of the translated-dictionary and translated-text approaches we have to do a double comparison (see Figure 1). We start with a manually translated parallel corpus in English and the working language. We machine-translate the working language component of the parallel corpus into English (step 1). Then we apply appropriate LIWC dictionary to both components of the corpus and to the machine-translated files (step 2). We compare the outputs of both corpus components and the outputs of the MT English texts (step 3). Then we evaluate the results of the comparison (step 4). In the last step (5) we produce a report that can help us find possible explanations for the results.

We perform this procedure for multiple LIWC versions, multiple corpora, multiple languages, and multiple MT engines. In the following subsections we discuss the parallel corpora and the steps of our procedure. With respect to languages, we chose to use only languages that we read ourselves, in order to be able to inspect what may have gone wrong in the translation or in the application of LIWC. The languages that we test with are German, Spanish and Dutch. We were also limited to some extent by the availability of free and accessible MT software (see below).

We should note that, as mentioned in the background section, the latest version of LIWC includes a number of summary variables. For external dictionaries, however, the LIWC program does not compute these summary variables. This implies that in a translated dictionary approach the summary variables are typically unavailable, unless the dictionary translation is included in the LIWC program as an internal dictionary. So far, only the German LIWC 2015 translation has reached that status. Users of other translations have to follow the references to literature provided in the LIWC documentation (Pennebaker et al., 2015, p. 6), which are, however, not very conclusive. This is a limitation inherent in the translated dictionary approach.

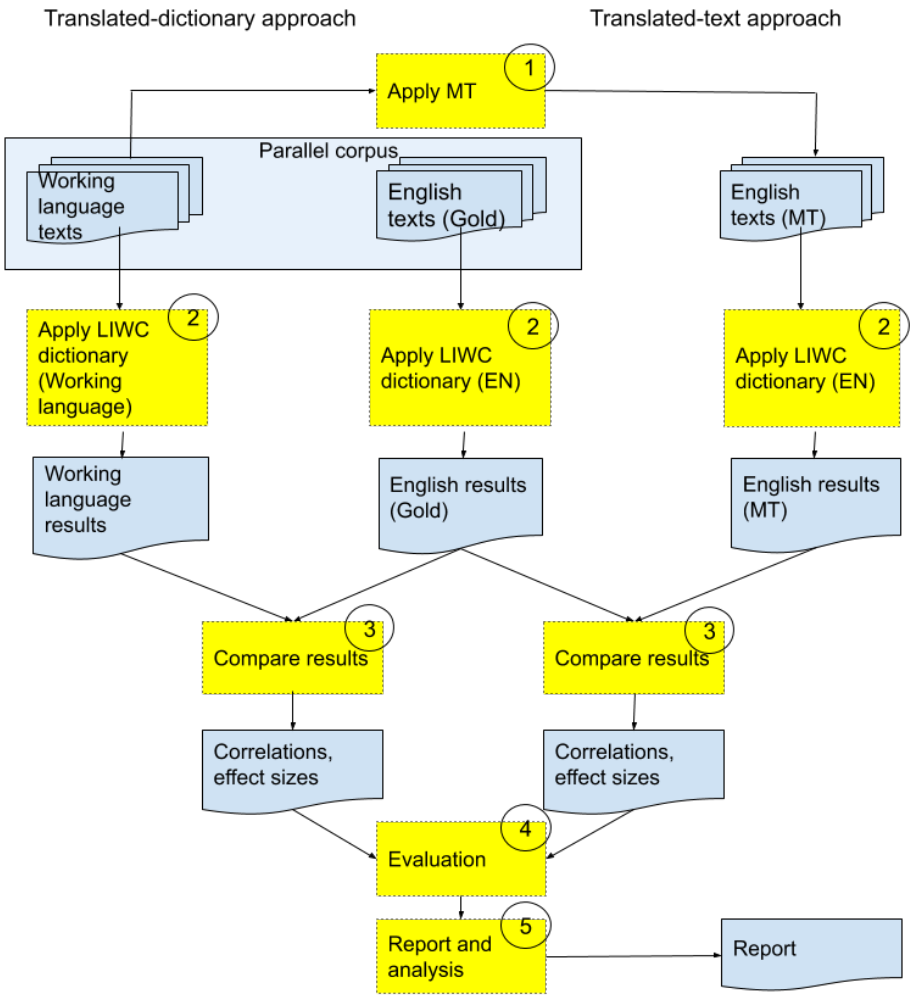

Figure 1. Comparing the translated-dictionary approach (left) and the MT translated-text approach (right).

Parallel corpora

The main parallel corpus that we used is the TED 2013 subtitles corpus (Tiedemann, 2012). We used a selection consisting of the first 300 (alphabetically) talks. In the various languages the exact number of talks that we use varies between 225 and 300, as not all talks are translated in all languages.

For Dutch, we also used the Dutch Parallel Corpus (DPC) (Macken, De Clercq, & Paulussen, 2011). DPC is a Dutch-English-French corpus (we used only the Dutch and English parts), balanced with respect to genre and translation direction. We use a subset of 720 texts from various fields.

Choice of MT engine (step 1)

We decided not to train a model ourselves but use only pre-trained models. We tried a number of MT systems but for some of these it proved hard get them to run or the output was disappointing. For comparability and replicability, we decided to use at least one open source MT system that would work for our three languages. For this, we selected Joshua (Post, Cao, & Kumar, 2015), a phrase-based statistical MT system that provides language packs for many language-pairs. For German, in addition we used Facebook's Fair system (Ng et al., 2019). Fair won the WMT19 competition for translation from German to English (Barrault et al., 2019). For Dutch, in addition to

Joshua we used Google's neural (NMT) and phrase-based machine translation systems (PBMT) (Wu et al., 2016).

Apply LIWC (step 2)

In order to automate our procedure as much as possible we did not use the LIWC programme itself for counting the words in the categories, as it cannot be scripted. Instead, we used the counting functionality of the LIWCtools scripts (Boot, 2016).

Comparison (step 3)

To evaluate the quality of LIWC results, we follow LIWC translators' best practices in computing correlations and effect sizes for each LIWC category. If the skewness of the score distribution is between −1.5 and 1.5, we compute the Pearson correlation, if otherwise we used Spearman's rank correlation coefficient. For the effect size, we take the difference of the means of the scores for the text collections that we compare and divide this by the square root of the mean of their variances. We compute the average correlations and effect sizes over LIWC categories.

Evaluation (step 4)

We consider average correlation (step 3) our main quality indicator, as it is sensitive to the properties of individual texts. The effect size, which measures base rate differences, is an additional indicator.

Report and analysis (step 5)

We produce a report that shows, for the LIWC categories with the largest differences between both procedures, a printout of a number of texts with high differences. We show the three text versions (the non-English text and the two English texts) and highlight the words in that LIWC category. Together, these reports help us understand whether the problem is in the dictionary, in the translation or in a mismatch between the languages and therefore in a sense unavoidable.

Results

We present the main results in Table 1. In all configurations the results are the same: in the translated text condition the correlations between the text pairs are higher and the effect sizes are lower. If equivalence to the validated English LIWC dictionary is the goal, the way to achieve that is to (machine-)translate the texts to English, independent of LIWC version and MT engine.

Table 1. Correlations and effect sizes for each configuration.

Run	Language	Corpus	MT engine	LIWC level	Translated dictionary		Translated text	
					Correlation (averages)	Effect size (averages)	Correlation (averages)	Effect size (averages)
1	Dutch	DPC	Joshua	2007	.78	.37	.88	.09

Machine-translated texts or translated dictionaries for LIWC

2	Dutch	DPC	Joshua	2015	.73	.58	.86	.09
3	Dutch	TED	Google-NMT	2007	.76	.44	.89	.19
4	Dutch	TED	Google-PBMT	2007	.76	.44	.89	.23
5	Dutch	TED	Joshua	2007	.76	.44	.87	.16
6	Dutch	TED	Joshua	2015	.74	.63	.86	.17
7	German	TED	Joshua	2001	.65	.94	.88	.16
8	German	TED	Joshua	2015	.71	1.02	.87	.20
9	German	TED	Fairseq	2001	.66	.94	.89	.16
10	German	TED	Fairseq	2015	.71	1.01	.90	.20
11	Spanish	TED	Joshua	2007	.67	.89	.93	.27

We should note some limitations in the figures in Table 1. For Spanish, the numbers do not take into account the categories for verb endings by person and number mentioned above. If we had added those to the corresponding pronominal categories, the translated dictionary approach would have scored a bit better. For German 2015, we did not use the optional script to disambiguate second and third person *Sie*. For Dutch, in the TED-talks corpus sometimes adjoining words have been merged.[1] But it is clear these limitations do not impact the overall patterns. For some results for specific categories, see below. For the full results, see the deposited results referred to in the Open Practices Statement below.

Discussion

When seeking to understand the apparent superiority of the translated-text approach over the translated-dictionary approach, we need to look into the factors that potentially explain a low correlation or high effect size between the measurements. For the translated-dictionary approach, the quality of the results is determined by two things: the quality of the manual translation of the texts and the quality of the LIWC dictionary translation. For the quality in the translated-text approach the relevant factors are the quality of the manual translation of the texts and the quality of the machine translation. A factor that might influence these qualities is the distance between languages. We briefly discuss each of these factors in turn, and then discuss a number of explanations for differences in the dictionary approach.

Factors determining quality

Manual translation: The quality of the manual translation influences the results of both approaches. In the translated-dictionary approach, because a difference between the texts in English and the working language means that even a perfect LIWC dictionary translation would give different results. In the translated text approach because any difference between the two measurements is caused by a machine translation following a manual translation or, alternatively, a manual translation and a machine translation of the same text. Because it influences the results of both approaches, the quality of the manual translation is not likely to be a relevant factor in explaining the differences between both approaches. We have no information about the relative quality of the TED translations

[1] For both Dutch and English, as a result of an error in our processing.

to Dutch, Germans and Spanish and the DPC translations. Anecdotally, the DPC translations are often quite free. On the other hand, the TED translations have been contributed by volunteers.

Machine translation: We chose to use Joshua as our baseline for the comparison. However, the linguistic quality of the Joshua translations is not very high. We give a single example: one fragment in Al Gore's talk 'New thinking on the climate crisis' reads 'I was reminded of that recently, by a woman who walked past the table I was sitting at, just staring at me as she walked past. She was in her 70s, looked like she had a kind face'. After being translated (into Dutch, German and Spanish respectively) and machine translated back into English again, the sentence reads 'I was recently recalled by a woman who walked along the table where i was, and to me as they walked past me by stared. she was more than 70, with a friendly face' (from Dutch), 'I was recently by a woman recalled the at the table, in which i was passed, and just as they passed. she was staring at around 70 and had a friendly face' (from German) and 'This reminded me recently a woman who spent by the bureau on which i was sitting, and i watched while passing. and something, and was about 70 a gentle expression' (from Spanish). We see that in coherence and syntactical correctness, the translations are not that good. However, much of the word meaning is retained, even though the texts have been translated twice. Apparently, this is enough for a word-based tool such as LIWC to be able to function. In addition it is also noteworthy that for German and Dutch, where we can compare Joshua's results with those of a state-of-the art neural MT system, the Joshua results are only slightly lower. Apparently, while the quality of the MT system has a measurable effect, its degree of perfection is not decisive for the adequacy of the LIWC results.

Dictionary translation: As we do not know the qualities of the texts' translations to the different languages, it is impossible to say something about the relative qualities of the LIWC dictionary translations in those languages. We can say, however, that in German the 2015 dictionary performs better (in the sense of agreeing with its English original) than the 2001 dictionary; in Dutch the 2007 dictionary does better than the (automatically translated) 2015 dictionary. Again, regardless of language, the results show that no dictionary translation can compete with the machine translation approach.

Linguistic distance: It seems plausible that a large distance between languages makes translation more difficult (Bernardo, 2010). The four languages that we study in this article are all Western, Indo-European languages. Nevertheless, Spanish, as a Romance language, is further from English than the Germanic languages Dutch and German, as appears from Table 2 (Ginsburgh, 2005). A priori we might therefore expect that the results for Spanish would be worse than for Dutch and German. For the translated text approach, however, the correlation for Spanish (run 11) is the highest that we have seen. We have an insufficient number of languages to say anything conclusive, but our results do not support the hypothesis that language distance should lead to lossy translation.

Table 2. Language distances based on Ginsburgh (2005)

	Dutch	German	Spanish
Distance from English	.39	.42	.76

Reasons for differences

In a further attempt to understand why the translated text-approach seems to do better, we looked at the reports produced in the last steps of our pipeline. For runs 1, 10 and 11 we manually analysed portions of the output of the reporting step, and looked at words where the dictionary approach led

to a different result between the working language and English. We identify four broad categories responsible for these differences: errors or unfortunate choices in the dictionary translation, errors or unfortunate choices in the English dictionary, differences between the languages and translators' choices.

Errors or unfortunate effects of choices in the translated dictionary: The majority of the differences results from unfortunate choices in translating the dictionary. The Spanish dictionary, for example has *padre* ('father') in the category religious, probably as a translation of the English 'priest'. However, 'padre' occurs much often in the family sense than in the religious sense. Similarly, in the Dutch 2007 dictionary, *alleen* is in the category sad, as a translation of 'alone'. However, 'alleen' is also an adverb, meaning 'just', 'simply'. In the German 2015 dictionary, *Gefühl* (primarily 'feeling', 'emotion') is included in the feel category, which is however meant for sensory expressions.

Errors or unfortunate effects of choices in the English dictionary: In the English 2015 dictionary, the word 'get' is included in the reward category. 'Get' however, can also mean 'become', which obviously ('he got bored') has nothing to do with rewards. In this case the translated dictionary approach will, correctly, not count the phrase in the reward category. In the relig category, English includes 'soul'. Use of the word in the musical sense is therefore also counted as religious. In financial texts, the word 'shares' counts as a positive emotion, just like 'securities', which also counts as a cognitive mechanism. In all these cases, the translated dictionary gets it right--however, in our procedure, the cases are counting in favour of the translated text approach, as the translation will often use the same words as the original English texts. This is an inevitable effect of assessing a tool's validity based on translations.

Differences between the languages: A major difference between Dutch and German on the one hand and English on the other is that Dutch and German have more compound words. English 'mass murder' translates to Dutch *massamoord*. 'Murder' and 'moord' are in the English and Dutch dictionaries, but including all compounds in the translated dictionary would be infeasible. The phrase therefore counts in English but not in Dutch. In Spanish, singular *su* and plural *sus* ('his', 'her' or 'their') is used depending on the number of the complement (*sus hijos*: 'his children' or 'their children'). There is no way the dictionary can assign *su* or *sus* to the shehe or they category without introducing errors. There are also many differences at the level of individual words and expressions.

Translator's choices: Translators have the freedom to avoid word-by-word translation if they believe the result is closer to the overall meaning of the whole. A few examples: 'this proposal is perfectly logical' became in Dutch 'valt er maar weinig in te brengen tegen het voorstel' (roughly: 'there is not much to be said against the proposal'), which includes prepositions, a social word, a quantifier, and more. In other cases a there is no exact equivalent and the translator's choice introduces a category, such as when English 'cronies' is translated by Dutch *kameraden*, which is included in the friends category. In some cases we note that translators tone down expressions that may be considered vulgar.

We did not attempt to quantify the number of occurrences for these categories. It would be very difficult, because in many cases multiple categories may be applicable. It is also hard for us to assess all of the reasons why the teams that developed LIWC or its translations choose to include or not a word in the dictionary. It would also not be very useful. What we can say is that errors or unfortunate choices in the translated dictionary and differences between the languages account for the lion's share of the differences in the translated dictionary approach. They usually do not cause differences in the translated text approach. Errors or unfortunate choices in the English dictionary occur less.

Logically, they can have no effect in the translated text approach. Translators' choices can cause differences in the translated dictionary approach, and in many cases the machine translation will introduce a similar wording in the translated text and thus create a difference in the translated txt approach as well. In other cases the machine translation will reverse the effect of the human translator.

The upshot is that limitations of the translated dictionaries and differences between the languages are responsible for most of the differences in the translated dictionary approach. In fact, most of the dictionary translations' limitations are caused by differences between the languages: it is the differences between the languages in meanings and usage patterns of words which make dictionary translation a task that inevitably introduces error.

Post-hoc analysis

Results for popular LIWC categories

The numbers we report in Table 1 are averages over all LIWC categories. Some categories traditionally score very low in this type of comparison, among others the informal categories (swear words, netspeak, fillers, etc.). To some extent at least this is a result of the fact that often used test corpora score very low on the informal language categories. It is interesting to look specifically at the results for some of the most frequently used LIWC categories.

As far as we know, the most recent overview of LIWC usage is (Tausczik & Pennebaker, 2010). The most frequently used categories reported there are WC (Word Count), I (first person singular pronouns), posemo, negemo, cogmech (cognitive mechanisms) and insight. We report the correlations and effect sizes for these categories for three runs in Table 3:

Table 3. Results for frequently used LIWC categories in a number of selected runs

Condition	Category	Translated dictionary		Translated text	
		Correlation	Effect size	Correlation	Effect size
5 (Dutch, TED, Joshua, 2007)	WC	.98	.18	.97	.11
	I	.99	.01	.99	.02
	posemo	.79	.22	.90	.17
	negemo	.82	.08	.91	.20
	cogmech	.79	.24	.88	.11
	insight	.81	.24	.92	.16
8 (German, TED, Joshua, 2015)	WC	.98	.27	.94	.35
	I	.99	.11	1.00	.06
	posemo	.79	.99	.94	.08
	negemo	.87	.23	.95	.11
	cogmech	.87	2.96	.87	.41
	insight	.84	.59	.92	.11
11 (Spanish, TED, Joshua, 2007)	WC	.98	.22	.99	.37

	I	.94	.95	.99	.07
	posemo	.80	.05	.96	.24
	negemo	.78	.05	.97	.19
	cogmech	.81	2.64	.94	.24
	insight	.78	1.64	.97	.32

While in the translated dictionary condition all correlations are satisfactory (above .78), in the translated text condition they are all above .90. For all categories, the reported correlations in the translated text condition are larger than those in the translated dictionary condition, except for the word counts in Dutch and German. We have no explanation for this exception.

A multilingual corpus

An extra complex situation is one where we have to analyse texts in multiple languages. This could arise for instance in an analysis of speeches in the European Parliament, where speakers can use their own languages, or an analysis of customer reviews for an international company. Nulty, Theocharis, Popa, Parnet, and Benoit (2016) for example analyse tweets from European politicians. They use what we have called a translated dictionary approach and apply to the tweets the LIWC dictionary of the appropriate language. So do Arroju, Hassan, and Farnadi (2015) in a personality prediction task. But if differences between languages make the translated text approach generally a better choice in a monolingual context, it seems likely that in a multilingual context the problems of the translated dictionary approach will be exacerbated, each dictionary embodying its own idiosyncratic choices in rendering English concepts in another language.

To quantify this effect, we use the TED corpus we used before. Now we compare first the application of language-specific LIWC dictionaries to the manually translated versions of TED talks, computing correlations and effect sizes. Then we compare the application of the English LIWC dictionaries to the machine translations into English. Finally we compute average correlations and effect size in both situations and compare these. We do this for two language pairs: Dutch and Spanish (LIWC 2007 dictionary) and Dutch and German (LIWC 2015). The results are given in Table 4.

Table 4. Cross-language effects

Language pair	LIWC version	Condition	Correlation	Effect size
Dutch - Spanish	2007	Translated text	.86	.12
		Translated dictionary	.61	.91
Dutch - German	2015	Translated text	.83	.07
		Translated dictionary	.65	1.01

To be clear: what we see in the translated dictionary approach is the result of the same texts (TED talks), being manually translated into two different languages and fed into the language-specific LIWC dictionary. The results (correlation .60 - .65 and effect size larger than .9) are unsatisfactory. The bad results must be due mostly to the application of the LIWC dictionaries, not to the quality of the textual translations. This is because these translations are also at the basis of the translated text approach: they are machine translated into English and then fed into the English LIWC dictionary. The results of this second approach are satisfactory (correlations .83 - .86 and effect size below .12). The conclusion is clear: when applying LIWC in a multilingual situation, translate the texts to English and do not apply the language-specific dictionaries.

Final considerations and conclusions

The results presented here seem very clear: at least for Western European languages a machine-translated text approach to LIWC is superior to a translated-dictionary approach. Given that the translation of the dictionary represents a significant effort, there seems to be no reason for continuing to use LIWC translations and especially for creating new LIWC dictionary translations. In this section we look at reasons why LIWC translations might still be useful.

Firstly, there may be language-pairs where machine translation has not yet reached a sufficient quality level. For languages with limited numbers of speakers, for which there is an insufficient amount of parallel text, machine translation to English may be unavailable. It is probably unlikely someone will make the effort of creating a LIWC dictionary for such a language, with one exception: historical language varieties. If a LIWC dictionary is available for a given language, it is relatively easy to extend the dictionary with historical word forms. Efforts have been reported for Dutch (Leemans, Van der Zwaan, Maks, Kuijpers, & Steenbergh, 2017). For German (in the context of a non-LIWC sentiment lexicon) Schmidt and Burghardt (2018) report that 'extension with historical linguistic variants consistently yields the strongest performance boost for all lexicons.' For researching historical language varieties, thus, the existence of a translated LIWC dictionary in the modern language version is clearly important.

Second, there may be a need for LIWC categories that capture word categories that English doesn't have. We mentioned above the pronominal adverbs prominent in Dutch and the distinction between informal and polite you in German. The Spanish dictionary has a category for subjunctive verb forms, which do not exist in English. Researchers may want to investigate the rates of occurrences of these language-specific phenomena and obviously a translated-text approach, which uses an English dictionary on an English text, will not reveal them. As an argument for using a translated LIWC dictionary, however, this is unconvincing. To determine these usage rates, it would be sufficient to create a specialized dictionary containing only these specific categories.

Third, the application of machine translation may not always be technically and financially feasible, depending on the skill profile of the researcher, the volume of text to be investigated, project funding and other factors. For a limited number of texts, 'manual' MT (feeding the texts into e.g. Google Translate's web interface) is practicable. For larger numbers of texts or longer texts, the researcher will need to do some simple programming to call e.g. the Google translate API. Depending on the project setup, IT staff may be available to help doing this. Use of the API is not free and as we saw, for some projects this is a reason to avoid MT (Lind, Eberl, Galyga, et al., 2019). Alternatively, the researcher or IT staff may download an open source MT toolkit such as Joshua. We found in this project that downloading and getting to run MT toolkits with precompiled language models often requires a non-trivial effort. In any case, when 'manual' MT becomes infeasible, using a translated text approach is more complicated than just running the LIWC program with the relevant dictionary.

Fourth, the argument we have been making in this article is based on the assumption that the validity of non-English LIWC results is to be determined by agreement with English LIWC results on the same texts. A case could be made that conformance to an English category system is not a suitable criterion to judge a research method for researching non-English texts. Different cultures employ different categories in thinking and talking about human behaviour and research practices should take that into account. This is no doubt true, but as an argument for translated LIWC

dictionaries it seems to shoot itself in the foot, as in that case it is unclear why one would translate the dictionary, with its embedded cultural assumptions, at all.

It does, however, point to a potential issue with the use of translated text for LIWC. Languages have different base rates for different word groups. For example, English uses less articles than German or Dutch, as shown by the LIWC base rates (Boot et al., 2017; Wolf et al., 2008). A good machine translation system takes these differences into account, as shown by the examples in Table 5. Applying a translated-text approach to a collection of Dutch texts will therefore result in generally lower scores on articles than the translated-dictionary approach. Probably, this does not pose a problem, as research is usually not interested in individual results but in differences between individuals.

Table 5. Median LIWC values for the *Articles* category, TED corpus.

Language	MT engine	LIWC version	Working language	English (original)	English (MT)
Dutch	Google NMT	2007	.091	.071	.080
German	FairSeq	2015	.108	.073	.078

In conclusion, we can say that for Western-European languages it is in general more effective to (machine-) translate non-English texts to English and apply an English LIWC dictionary than to apply a translated LIWC dictionary to the original texts. Even a dated MT system such as Joshua is good enough. The superiority of the translated-text approach is especially clear in situations where the non-English texts are in multiple languages. For the summary variables, the translated-text approach is really the only practicable. We have not tested with languages whose relationship to English is more remote, but given the word-based nature of LIWC scores, the progress in MT and the inevitability of compromise in translating the LIWC dictionary we would be surprised if for these languages the dictionary approach would fare better.

Acknowledgements

Preparatory work, the machine translations and some of the analysis were done by Yuying Ye and presented in (Ye & Boot, 2020).

Open Practices Statement

The LIWC dictionaries are available at http://ww.liwc.net. The parallel text corpora are available at http://opus.nlpl.eu/TED2013.php and https://taalmaterialen.ivdnt.org/download/tstc-dutch-parallel-corpus-niet-commercieel/. The dictionaries and corpora in the form that we used them are available from the author upon request. All computed and generated data, including the scripts for the analysis, are available in the DANS EASY repository at https://doi-org/10.17026/dans-xgp-nyp7. We do not report any preregistered experiments.

Competing interests

The author is one of the translators of the Dutch 2007 and 2015 LIWC dictionaries

References

Agarwal, A., Xie, B., Vovsha, I., Rambow, O., & Passonneau, R. J. (2011). *Sentiment analysis of twitter data*. Paper presented at the Proceedings of the Workshop on Language in Social Media (LSM 2011). https://www.aclweb.org/anthology/W11-0705.pdf

Alparone, F. R., Caso, S., Solano, L., & Prezza, M. (2002). Traduzione e adattamento al contesto linguistico italiano del 'Linguistic Inquiry and Word Count' (LIWC). *Emozioni: cultura, comunicazione, benessere*, 133-140.

Andrei, A. L. (2014). Development and Evaluation of Tagalog Linguistic Inquiry and Word Count (LIWC) Dictionaries for Negative and Positive Emotion. Retrieved from https://www.mitre.org/sites/default/files/publications/pr_14-3858-development-evaluation-of-tagalog-linguistic-inquiry.pdf

Araújo, M., Reis, J., Pereira, A., & Benevenuto, F. (2016). *An evaluation of machine translation for multilingual sentence-level sentiment analysis*. Paper presented at the Proceedings of the 31st Annual ACM Symposium on Applied Computing. https://dl.acm.org/doi/abs/10.1145/2851613.2851817

Arroju, M., Hassan, A., & Farnadi, G. (2015). *Age, gender and personality recognition using tweets in a multilingual setting*. Paper presented at the 6th Conference and Labs of the Evaluation Forum (CLEF 2015): Experimental IR meets multilinguality, multimodality, and interaction. http://ceur-ws.org/Vol-1391/57-CR.pdf

Balage Filho, P. P., Pardo, T. A. S., & Aluísio, S. M. (2013). *An evaluation of the brazilian portuguese liwc dictionary for sentiment analysis*. Paper presented at the In 9th Brazilian Symposium in Information and Human Language Technology, Fortaleza, Ceara. https://www.aclweb.org/anthology/W13-4829.pdf

Balahur, A., & Turchi, M. (2012). *Multilingual Sentiment Analysis using Machine Translation?* Paper presented at the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis, WASSA@ ACL 2012. https://www.aclweb.org/anthology/W12-3709.pdf

Banea, C., Mihalcea, R., Wiebe, J., & Hassan, S. (2008). *Multilingual subjectivity analysis using machine translation*. Paper presented at the Conference on Empirical Methods in Natural Language Processing. https://www.aclweb.org/anthology/D08-1014.pdf

Barrault, L., Bojar, O., Costa-jussà, M. R., Federmann, C., Fishel, M., Graham, Y., . . . Malmasi, S. (2019). *Findings of the 2019 Conference on Machine Translation (WMT19)*. Paper presented at the WMT 2019. https://www.aclweb.org/anthology/W19-5301.pdf

Bernardo, A. M. (2010). Translation as Text Transfer—Pragmatic Implications. *Linguistic Studies, 5*, 107-115.

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

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. doi:10.1075/dujal.6.1.04boo

Carvalho, F., Rodrigues, R. G., Santos, G., Cruz, P., Ferrari, L., & Guedes, G. P. (2019). Evaluating the Brazilian Portuguese version of the 2015 LIWC Lexicon with sentiment analysis in social networks. *Anais do VIII Brazilian Workshop on Social Network Analysis and Mining*. doi:0.5753/brasnam.2019.6545

De Vries, E., Schoonvelde, M., & Schumacher, G. (2018). No longer lost in translation: Evidence that Google Translate works for comparative bag-of-words text applications. *Political Analysis, 26*(4), 417-430. doi:10.1017/pan.2018.26

Gao, R., Hao, B., Li, H., Gao, Y., & Zhu, T. (2013). Developing simplified Chinese psychological linguistic analysis dictionary for microblog *Brain and Health Informatics* (pp. 359-368): Springer. doi:10.1007/978-3-319-02753-1_36

Ginsburgh, V. (2005). Languages, genes, and cultures. *Journal of Cultural Economics, 29*(1), 1-17. doi:10.1007/s10824-005-4074-7

Hayeri, N. (2014). *Does gender affect translation?: analysis of English talks translated to Arabic.* (Phd), Univ of Texas, Austin, TX. Retrieved from http://hdl.handle.net/2152/25082

Huang, C. L., Chung, C. K., Hui, N., Lin, Y. C., Seih, Y. T., Lam, B., & Pennebaker, J. W. (2012). The development of the chinese linguistic inquiry and word count dictionary. *Chinese Journal of Psychology, 54*(2), 185-201.

Kailer, A., & Chung, C. K. (2011). The Russian LIWC2007 dictionary. Austin, TX: Pennebaker Conglomerates. Retrieved from http://www.liwc.net

Leemans, I., Van der Zwaan, J. M., Maks, I., Kuijpers, E., & Steenbergh, K. (2017). Mining Embodied Emotions: A Comparative Analysis of Sentiment and Emotion in Dutch Texts, 1600-1800. *Digital Humanities Quarterly, 11*(4).

Lind, F., Eberl, J.-M., Galyga, S., Heidenreich, T., Boomgaarden, H. G., Jiménez, B. H., & Berganza, R. (2019). A Bridge Over the Language Gap: Topic Modelling for Text Analyses Across Languages for Country Comparative Research. Retrieved from https://www.reminder-project.eu/wp-content/uploads/2019/11/D8.7.pdf

Lind, F., Eberl, J.-M., Heidenreich, T., & Boomgaarden, H. G. (2019). When the Journey Is as Important as the Goal: A Roadmap to Multilingual Dictionary Construction. *International Journal of Communication, 13*, 21.

Macken, L., De Clercq, O., & Paulussen, H. (2011). Dutch parallel corpus: A balanced copyright-cleared parallel corpus. *Meta: Journal des traducteurs/Meta: Translators' Journal, 56*(2), 374-390. doi:10.7202/1006182ar

Massó, G., Lambert, P., Penagos, C. R., & Saurí, R. (2013). Generating New LIWC Dictionaries by Triangulation *Information Retrieval Technology* (pp. 263-271): Springer. doi:10.1007/978-3-642-45068-6_23

Meier, T., Boyd, R. L., Pennebaker, J. W., Mehl, M. R., Martin, M., Wolf, M., & Horn, A. B. (2019). "LIWC auf Deutsch": The Development, Psychometrics, and Introduction of DE-LIWC2015. *PsyArXiv*(a). doi:10.17605/OSF.IO/TFQZC

Ng, N., Yee, K., Baevski, A., Ott, M., Auli, M., & Edunov, S. (2019). *Facebook FAIR's WMT19 News Translation Task Submission*. Paper presented at the Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1). https://arxiv.org/abs/1907.06616

Nulty, P., Theocharis, Y., Popa, S. A., Parnet, O., & Benoit, K. (2016). Social media and political communication in the 2014 elections to the European Parliament. *Electoral Studies, 44*, 429-444. doi:10.1016/j.electstud.2016.04.014

Pennebaker Conglomerates. (2015). Spanish LIWC2007 Dictionary. Austin (Tx): Pennebaker Conglomerates. Retrieved from http://www.liwc.net

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

Pennebaker, J. W., & Graybeal, A. (2001). Patterns of natural language use: Disclosure, personality, and social integration. *Current Directions in Psychological Science, 10*(3), 90-93. doi:10.1111/1467-8721.00123

Pennebaker, J. W., Mehl, M. R., & Niederhoffer, K. G. (2003). Psychological aspects of natural language use: Our words, our selves. *Annual review of psychology, 54*(1), 547-577. doi:10.1146/annurev.psych.54.101601.145041

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. doi:10.1016/j.psfr.2011.07.002

Piper, A. (2016). Fictionality. *Journal of Cultural Analytics*. doi:10.22148/16.011

Post, M., Cao, Y., & Kumar, G. (2015). Joshua 6: A phrase-based and hierarchical statistical machine translation system. *The Prague Bulletin of Mathematical Linguistics, 104*(1), 5-16. doi:10.1515/pralin-2015-0009

Proksch, S. O., Lowe, W., Wäckerle, J., & Soroka, S. (2019). Multilingual sentiment analysis: A new approach to measuring conflict in legislative speeches. *Legislative Studies Quarterly, 44*(1), 97-131. doi:10.1111/lsq.12218

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

Reber, U. (2019). Overcoming language barriers: Assessing the potential of machine translation and topic modeling for the comparative analysis of multilingual text corpora. *Communication methods and measures, 13*(2), 102-125. doi:10.1080/19312458.2018.1555798

Schmidt, T., & Burghardt, M. (2018). *An Evaluation of Lxicon-Based Sentiment Analysis Techniques for the Plays of Gotthold Ephraim Lessing*. Paper presented at the Second Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature. https://www.aclweb.org/anthology/W18-4516.pdf

Shibata, D., Wakamiya, S., Aramaki, E., & Kinoshita, A. (2016). *Detecting Japanese patients with Alzheimer's disease based on word category frequencies*. Paper presented at the Clinical Natural Language Processing Workshop, Osaka, Japan. https://www.aclweb.org/anthology/W16-4211.pdf

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. doi:10.1177/0261927X09351676

Tiedemann, J. (2012). *Parallel Data, Tools and Interfaces in OPUS*. Paper presented at the Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'2012), Istanbul. https://www.aclweb.org/anthology/L12-1246/

Van Wissen, L., & Boot, P. (2017). *An Electronic Translation of the LIWC Dictionary into Dutch*. Paper presented at the eLex Conference 2017, Leiden. https://elex.link/elex2017/wp-content/uploads/2017/09/paper43.pdf

Wevers, M. (2019). *Using Word Embeddings to Examine Gender Bias in Dutch Newspapers, 1950-1990*. Paper presented at the 1st International Workshop on Computational Approaches to Historical Language Change, Florence. https://www.aclweb.org/anthology/W19-4712.pdf

Windsor, L. C., Cupit, J. G., & Windsor, A. J. (2019). Automated content analysis across six languages. *PloS one, 14*(11). doi:10.1371/journal.pone.0224425

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. doi:10.1026/0012-1924.54.2.85

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., . . . Macherey, K. (2016). *Google's neural machine translation system: Bridging the gap between human and machine translation*. Retrieved from https://arxiv.org/abs/1609.08144

Ye, Y., & Boot, P. (2020). *Machine Translation as an Alternative to Language-Specific Dictionaries for LIWC*. Paper presented at the DHBenelux 2020, Cloud. https://zenodo.org/record/3874671

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) [The Dutch version of 'Linguistic Inquiry and Word Count' (LIWC)]. *Gedrag & Gezondheid [Behaviour & Health], 32*, 271-281.

Zijlstra, H., Van Middendorp, H., Van Meerveld, T., & Geenen, R. (2005). Validiteit van de Nederlandse versie van de Linguistic Inquiry and Word Count (LIWC) [Validity of the Dutch version of Linguistic Inquiry and Word Count (LIWC)]. *Nederlands Tijdschrift voor Psychologie [Netherlands Journal of Psychology], 60*(3), 50-58. doi:10.1007/BF03062342