TRANSLATION FOR THE REST OF US

WORD-LEVEL TRANSLATION FOR AUTOMATED TEXT ANALYSIS

A. Maurits van der Veen[*]
Department of Government
William & Mary
Williamsburg, VA 23185
maurits@wm.edu

June 14, 2022

ABSTRACT

Automated machine translation has been improving in quality rapidly for some years, and now approaches that of professional human translation. However, for large volume translations, the best methods remain costly, in terms of money, computational resources, or time. In contrast, word-level translation is both free and fast, simply mapping each word in a source language deterministically to a target language. While the resulting translations are ungainly, this matters less for automated text analysis than one might suspect. In this paper, I discuss how to create high-quality word-level translation dictionaries, and apply some of the most common text analysis methods — sentiment analysis, dictionary-based content analysis, and topic modeling — to translations generated using those dictionaries, comparing them to translations by professional human translators and state-of-the-art neural machine translation. I show that while machine translation comes closer to human translation, the word-level translation results are comparatively quite close. The translation dictionaries as well as the code used to generate them are available on Github.

Keywords machine translation · word embeddings · text-as-data · computational social science

1 Introduction

Social science scholarship often relies on texts. Speeches, debates, media coverage, memoirs, tweets, and many other texts are key to our understanding of the world and the people around us. Frequently, the sheer volume of relevant texts makes it infeasible for individual scholars to read everything of interest. Accordingly, automated text analysis has rapidly become a popular approach (Grimmer and Stewart, 2013). For comparative analyses across linguistic borders, translation becomes essential. Machine translation offers a way to do this automatically, and recent advancements in neural machine translation produce translations of impressive quality, approaching those produced by professional human translators (Graham et al., 2019).

Google Translate is perhaps the best known automated translation engine available to the general public. Previous research has shown that applying automated text analysis techniques to translations produced by Google Translate produces results that are comparable to those from performing the same techniques on human translations of the same texts (de Vries et al., 2018). DeepL is another, more recent machine translation service (`www.deepl.com`). While such services are invaluable for collections of a few hundred or several thousand texts, their costs rise rapidly for applications featuring tens of thousands of texts or more, as is increasingly common for text analysis studies. Open-source alternatives exist (Tiedemann and Thottingal, 2020), producing comparable-quality translations with no direct cost, although the resources required to run such alternatives may not come cheaply.

Fortunately, many of the most widely used text analysis methods do not require smoothly flowing, grammatically correct translations. In fact, many techniques take a “bag of words” approach, which ignores grammar and word order

[*]Thanks to the wonderful undergraduate research assistants at STAIR lab (stair.wm.edu) who contributed to the original development of this project.

altogether. For such applications, the crucial consideration is whether key terms are translated correctly. In this paper, I show that word-level translation can produce translations whose suitability for common text analysis techniques approaches that of high quality neural machine translation or even human translation, across a range of languages. Moreover, it produces such translations both rapidly and at essentially no cost, making it feasible to compare translations into different target languages.

The great advantage of word-level translation is that any word only needs to be translated once, before being added to the translation dictionary. The paper also discusses how to produce good translations for individual words. This can be done in two ways: by translating words one at a time using a commercial service such as Google Translate or Deepl, or by taking advantage of aligned word embeddings, which approximate multilingual, high-dimensional semantic spaces. In such aligned spaces, translating a word simply means identifying the target language word whose location in the shared space is closest to that of the source language word. In this paper, I adopt a combination of the two approaches.

The paper begins with a brief overview of competing approaches to machine translation. Next, I outline how tor produce high-quality, word-level translation dictionaries and provide summary information about the translation dictionaries presented here. The third section assesses the quality of machine translation, both neural and word-level, using a European Parliament speech corpus with speeches translated by professional human translators into all official languages of the European Union. I assess translation quality by investigating how closely the results on three standard text analysis tasks — sentiment analysis, measuring individual emotions in text, and topic modeling — compare among the professional translation into English, state-of-the-art neural machine translation, and word-level translations produced. The translation dictionaries, along with replication code, are available on Github.[2]

2 The appeal of word-level machine translation

Recent years have brought rapid advancements in machine translation, both at the word and document level. For document-level translation, neural machine translation produces the best results. Already in 2018, analyses showed that texts translated using Google Translate produced term-document matrices that were comparable to those produced from texts translated by professional human translators, and that applying a technique such as topic modeling to the automatically translated texts generated similar results (de Vries et al., 2018). The quality of Google Translate's output has only increased since then, and other commercial services have appeared offering similarly high-quality translation, most notably DeepL (`www.deepl.com`). However, commercial translation services come at a cost. Even though both Google Translate and DeepL are very reasonably priced, the cost of translating the volumes of texts now regularly used in automated text analysis studies can quickly mount to thousands or even tens of thousands of dollars.

Open-source technologies for neural machine translation (Junczys-Dowmunt et al., 2018; Klein et al., 2018, 2020) eliminate these costs, and with sufficient training can produce translations that approach the quality of Google Translate or DeepL. Moreover, pre-trained translation engines for a growing number of language pairs can be found online, obviating the need for users to collect their own training data for and engage in the time-consuming training of the translation engines. In this paper, I use pre-trained models from OPUS-MT, which are available for many language pairs (Tiedemann and Thottingal, 2020). Even the computational resources such technologies require do not need to be acquired directly, as Google's Colab makes it possible to run them online. However, the time demands of producing translations remain considerable. Moreover, while such approaches have become increasingly accessible even to those with comparatively little coding experience, they do require basic coding skills and a familiarity with different programming environments.

The corpora used in the validation tests in this paper total about 700,000 short parliamentary speeches, adding up to 165 million words. This makes it a medium-size dataset for automated text analysis work. Translating them using a commercial neural machine translation engine such as DeepL would cost around $25,000 total. In contrast, using pretrained models from OPUS-MT, a comparatively fast, pre-trained open machine translation model (Tiedemann and Thottingal, 2020), is free. However, performing these translations took about 240 hours on Google Colab, and required frequent user intervention. In contrast, translating the same corpora using dictionaries costs nothing and takes a matter of minutes. Moreover, as this paper shows, the resulting translations are of sufficient quality for many common applications.

3 Method

All leading automated machine translation approaches build on the insight that one can "know a word by the company it keeps" (Firth, 1957, p. 11). Examining large quantities of texts makes it possible to place words (and phrases) in

[2] `https://github.com/amaurits/translation4tru/`

a multidimensional space which encapsulates meaning, but also part-of-speech, word form, etc. (Gärdenfors, 2004; Mikolov et al., 2013a). Moreover, it is possible to map these spaces to one another across languages. Such mappings encode high-quality information about the most similar words in different languages (Chen and Cardie, 2018; Conneau et al., 2017; Joulin et al., 2018; Mikolov et al., 2013b).

Just as words can be placed in a semantic space, so can longer pieces of text, including entire sentences or even paragraphs. This is the basic intuition of how neural machine translation works: a source language input sentence is transformed into an internal representation, which in turn is transformed back into the closest representation in the target language. OPUS-MT is built on the Marian-NMT toolkit (Junczys-Dowmunt et al., 2018), which provides transformer-based neural machine translation. Pre-trained models are available for a wide range of language pairs, including additional ones produced as part of the Tatoeba translations challenge (Tiedemann, 2020) (`https://github.com/Helsinki-NLP/Opus-MT`). These can generally be run fairly straightforwardly using either `easyNMT` (`https://github.com/UKPLab/EasyNMT` or `CTranslate2` (`https://github.com/OpenNMT/CTranslate2`). In this paper, I use the the standard Opus-MT models, as accessed through `easyNMT`, except for Portuguese and Greek, for which there there is no pre-trained model accessible through `easyNMT` for direct translation to English. For those two languages I use the Tatoeba models, as accessed through `CTranslate2`. [3]

To generate word-level translation dictionaries, it is possible simply to feed such open-source translation models individual words. However, the processing time for individual words is comparable to that for complete texts of hundreds of words, which makes it impractically slow to use these models for such a purpose. Moreover, since the models are trained on texts, not individual words, feeding them single words produces poor results. Arbitrarily connecting words together into "sentences" produces nonsense and thus even worse translations (even when the words are set apart by quotes, commas, or periods). Fortunately, this is where the need to translate a word just once makes the commercial services a viable alternative.

To demonstrate the feasibility of developing good word-level translation dictionaries for additional languages cheaply, I adopted an upper cost limit of \$25, on average, per language translated (\$250 for translating the 10 European languages used here into a single target language, English). This makes it possible to translate roughly the 100,000 most common words in each languages using Deepl. To identify the most common words in each language, I rely on the python module `wordfreq`, by Luminoso (Speer et al., 2018). This provides frequency data for the top 300,000 words (or more) for all of the languages used in this paper except for Danish and Greek. For Danish, I generate a frequency dictionary based on more than 1 million Danish newspaper articles, from the newspaper *Politiken*, 1997-2018. For Greek, I derive a frequency dictionary from the complete text of Greek Wikipedia, as available through the Internet Archive.[4]

Translating the top 100,000 words is not necessarily sufficient to reliably produce good word-level translations. While 100,000 words sounds like a lot, this excludes many words that are comparatively common, because every distinct word form counts: differences in word endings (plurals, verb conjugations, etc.) or capitalization all constitute different 'words'. I adopt two techniques to cover words beyond the most common 100,000. First, I leverage those 100,000 translations to extend the dictionary directly, by looking through the remainder of the common word lists for words whose translation can be inferred from the first 100,000: those with the same spelling but different capitalization, those that differ only in word ending (such as endings denoting masculine/feminine or singular/plural), and those that are compound words constructed from two or more separate words for which we have a translation already.[5]

Some language-specific additional heuristics further expand the dictionary. For instance, in Germanic languages, compound words are often formed by inserting the letter 's' between two words. Thus, if a new word is composed of two words already in the dictionary connected by an 's', we can translate its individual pieces and simply ignore the 's'. In addition, in German, words that should have an umlaut sometimes replace that umlaut by an 'e' following the vowel in question: Oesterreich for Österreich, for example.

Second, I expand the dictionaries by adding new translations derived from aligned word embedding spaces, as described at the top of this section. I use the mapping process proposed by Joulin et al. to map individual language spaces to one another 2018.[6] While this process is slower and lower-quality than using DeepL,[7] it is cost-free. Moreover, it is possible

[3] It is possible to go from Portuguese or Greek to English via an intermediate language in easyNMT, but the resulting translations are of poorer quality than direct word-level translation.

[4] `https://archive.org/details/elwiki-20220420`

[5] To be sure, some compound words mean something different than each of the component words do separately, but this is comparatively unusual.

[6] I use the pre-mapped Wikipedia-based word embeddings made available by Facebook at `https://fasttext.cc/docs/en/aligned-vectors.html` (accessed 2022-04-30), but applying the mapping to other languages is straightforward and gives comparable results.

[7] The original version of the dictionaries was produced using the aligned embedding process only. While this produced good word-level translations, the dictionaries created using DeepL are unambiguously better, across the board.

to improve translation range and quality further by using some of the same heuristics described above: generalizing across word endings, decomposing compound words, etc.

Mapping words across aligned embeddings does introduce a few additional wrinkles, driven by a some common features of embeddings: word locations are not evenly distributed across the embedding vector space; embeddings almost always include words that are not actually part of a language (misspellings, foreign words, etc.); and the correct translation of a word in the source embedding may be missing from the target embedding. The appendix outlines how I tackle these issues.

3.1 Supplementing or substituting translations

It is trivial to supplement any word-level translation dictionary with preferred translations for particular terms. An important disadvantage of even the very best word-level translation for text analysis purposes is that for each source language word we need to pre-select a single target language word without knowing anything about the context in which the source language word occurs. This is particularly problematic for words with multiple meanings (polysemy). For example, for ‘sink’, do we choose a translation for one of its verb meanings (‘descend’, or ‘go down below a surface’), or for its main noun meaning (a ‘fixed basin with a water supply and a drain’)? Whichever we choose, it is certain to be hilariously wrong at least part of the time.

However, this problem is easily addressed by substituting in specific translations. Thus, to translate a plumber’s manual using our general purpose dictionary, we could simply supply our own domain-specific list of plumbing-related translations, in which ‘sink’ is always translated to the target language version of the ‘fixed basin’ meaning. Similarly, to translate texts related to the European Union, as I do below, translation quality could be increased further by substituting specific translations of EU-related terms, such as those available in the European Union terminology dataset (https://iate.europa.eu/home). I do not do so here, since not all domains have similar multilingual glossaries available to swap in. However, as a general rule it is advisable to verify the correct translation of key domain-specific terms in the dictionary and to override the general purpose translations if necessary.

3.2 Translation dictionary contents

I generate translation dictionaries from 10 different European languages to English. The languages were selected based on several criteria. First, the languages of the five largest EU member states are included: French, German, Italian, Polish, and Spanish. Dutch is added as the sixth most widespread national language after these five, being spoken both in the Netherlands and parts of Belgium. Next, I add Danish as a representative of the Nordic (Northern Germanic) languages (performance of word-level Danish-English translation will be a good proxy for comparable Swedish-English and Norwegian-English translation). I also add Finnish, a language with almost no shared linguistic history with English, to pose a particularly difficult translation challenge, and Greek, a language which uses a different script, representing a different kind of challenge because every letter needs to be translated, even in words with shared roots. Finally, I add Portuguese as an instance of a language that has much in common with the three other Romance languages already included, but for which fewer important translation resources are available. In particular, there is no direct bilingual translation model available for Portuguese-English in `easyNMT`. In addition, the available word frequency data for Portuguese is more limited.

The words in the `wordfreq` module provide the initial list of words to include in the translation dictionary. The 100,000 most common of these are translated directly using DeepL, at an average cost of less than $25 per language, on average. Other words from the wordfreq list (or the comparable newspaper- or Wikipedia-based lists for Danish and Greek) are translated using the heuristics outlined above. Remaining untranslated words that exceed a basic occurrence threshold are then translated using aligned word embeddings. As an implicit test of how much value the latter translation method adds, no aligned embedding translations are included for Greek. The results suggest that this comes at little cost — instead, the primary value of the aligned embedding alternative derives from the possibility it offers of bypassing commercial translation services altogether, and thus having no up-front financial costs associated with a translation project.

The initial dictionaries were supplemented by a small number of translations for words common in European Parliament debates but not in general frequency lists. Table 1 shows the resulting size of the translation dictionaries into English for the 10 languages included here. The table is organized by language group, with the Romance and Germanic languages shown first. Within those language groups, the order is alphabetic. In addition to the total size, the table shows the percentage of words in the dictionary that were translated using DeepL, either directly or using one or more heuristics, along with comparable percentages for the aligned embedding method. These four categories account for over 90% of each language’s dictionary. "Words" copied directly from source to target comprise the remainder; these include punctuation, numbers, proper names, as well as words that are common in the target language and hence are assumed to be words borrowed from that language.

Table 1: Translation dictionary size & composition

language	Dictionary size	% DeepL direct	% Deepl heuristic	% aligned direct	% aligned heuristic
French	825,722	24.2	18.6	48.8	1.8
Italian	734,805	27.7	17.8	45.4	1.6
Portuguese	295,480	39.6	31.4	27.2	0.5
Spanish	754,088	26.0	20.4	40.9	2.8
Danish	277,322	50.3	45.0	4.5	0.0
Dutch	1,231,127	17.0	34.1	39.3	2.3
German	1,580,763	13.7	22.1	56.9	2.2
Finnish	684,016	16.9	64.1	6.5	4.2
Greek	226,085	51.8	48.2	0.0	0.0
Polish	395,154	30.8	49.7	14.8	1.8

4 Validation

In order to validate translation quality, it is essential to have a corpus of parallel texts in the source and target languages. One widely used such corpus for European languages is Europarl, which provides parallel text in 21 official languages of the European Union, drawn from the legislative record of the European Parliament (Koehn, 2005). Legislative debates in the European Parliament (EP) are, by law, translated into the primary official national language of each member state by professional translators. As such, the translation is of a very high quality. I use release v7 of the corpus, which includes speeches through November 2011. More specifically, I use EuroparlExtract, a cleaned and parsed version of the corpus which makes it easier to identify the original language of any speech (Ustaszewski, 2019; Graën et al., 2014).

The EuroparlExtract corpus stores texts by source & target language, with one speech per file. Each line in the file, generally corresponding to a sentence, has source text and target text, separated by a tab. Due to the corpus creation process for EuroparlExtract, the source text for a given speech can appear with small variations depending on the target language, either at the sentence or the speech level. To enable direct comparisons across languages, therefore, I use only speeches or sentences whose source language contents are the same across all the target languages.

I compile two different corpora. The first, *Sourcespeeches*, includes only speeches whose original language is the source language, along with their professional translation into English. This corpus eliminates possible biases introduced by prior translation from another language, but it is (obviously) not parallel across the ten languages. The second corpus, *Allspeeches*, includes all complete speeches for which the source language appears identically across the target languages in the dataset. This produces parallel corpora across all ten languages to be translated into English, along with a 'gold standard' corpus of speeches that were either originally given in English or else translated into that language by professional translators.

Table 2 shows, for each language, the *Sourcespeeches* corpus sizes in speeches, total words, and distinct words. For the *Allspeeches* corpus, the number of speeches is constant across all languages, so the table only shows total and distinct words.[8] After the target language, English, the remaining languages are shown in the same order as in Table 1.

To offer a first comparison of the quality of different translation approaches, the inset below shows the same original Danish sentence, as translated by the European Union's professional human translators, by Google Translate, by the Opus-MT model used here, and using word level translation dictionaries. The final entry shows the word level translation into English, but starting from the Italian translation produced by the European Parliament's professional translators (rather than the Danish original).

[8] Words with different capitalization are considered distinct. Many distinct words are proper names, so the count of distinct words is not a measure of the vocabulary size of 'actual' (i.e. not proper name) words in a corpus or language.

Table 2: Word count info for *Sourcespeeches* and *Allspeeches* corpora

	Source speeches			All speeches (n=35,783)	
Language	# speeches	# words	# distinct	# words	# distinct
English				6,810,150	(50,069)
French	19,586	6,245,683	(65,977)	7,193,857	(60,110)
Italian	11,419	2,857,803	(55,805)	6,599,909	(75,649)
Portuguese	11,926	2,634,029	(50,375)	6,927,596	(77,526)
Spanish	9,171	3,092,418	(52,382)	7,274,272	(70,953)
Danish	2,414	631,365	(32,702)	6,257,976	(112,384)
Dutch	9,886	2,965,638	(70,382)	7,014,985	(95,463)
German	21,374	5,796,590	(136,055)	6,424,088	(129,658)
Finnish	3,089	578,767	(80,845)	4,624,400	(237,745)
Greek	3,833	863,135	(48,736)	6,876,634	(99,797)
Polish	4,933	736,423	(57,929)	5,695,828	(137,839)

Corpus speech and word counts, with number of distinct words in parentheses.

Danish original
Kernen i de problemer, vi beskæftiger os med her, er den demografiske ubalance i verden.

EU human translation
At the core of the problems that we are concerned with here is the demographic imbalance in the world.

Google Translate
At the heart of the issues we are dealing with here is the demographic imbalance in the world.

OPUS-MT
The core of the problems we are dealing with here is the demographic imbalance in the world.

Word-level translation, from Danish original
Core in ones problems we employs ourselves with where is the demographical imbalance in world

Word-level translation, from professional Italian translation
The center of problems of including there whaaaa occupying in this headquarter c is it imbalance demographic in world

While both of the word-level translations are clunky and difficult to grasp at first glance, it is important to note that they do include the three substantive key words in the sentence: “demographic”, “imbalance”, and “world”. Interestingly, too, the word-level translations retain the word “problems”, as does the EU’s official translation, whereas Google substitutes “issues”. The direct word-level translation also retains the word “core” (as do the EU’s official translation and the translation using OPUS-MT), while Google substitutes “heart” and the word-level translation from Italian replaces it by “center”. In short, in terms of correctly translating key words that might be of relevance in automated text analysis, the word level translation of this example performs quite well.

To test the overall translation quality, I perform three separate automated text analyses on each of the three corpora outlined above. In each case, I pool the human-translated and machine-translated texts into a single corpus, run the analysis, and assess the similarity of the results at the level of individual texts, by original language. Here, I present the results for the *Sourcespeeches* corpus, which includes only speeches in their original language. These represent a more realistic proxy for likely applications of the translation dictionaries. On the other hand, the results for different languages are not fully comparable since the source texts differ. Results for the *Allspeeches* corpus, which do permit direct comparison and are, reassuringly, quite similar, appear in the appendix.

4.1 Sentiment analysis

I use the sentiment analysis method `MultiLexScaled`, a lexicon-based approach (van der Veen and Bleich, 2021), which builds on eight widely-used sentiment analysis lexica, and calibrates them against a representative corpus of representative newspaper texts from the United Kingdom. Figure 1 presents two different measures of the comparison between professional human translation and machine translation. The first group of bars shows the percentage of texts, by language, for which the sentiment polarity (positive or negative) matches. Since we are also often concerned with relative gradations in sentiment (more or less positive; more or less negative), the second group shows the correlation between the sentiment scores for human and word-level translation, respectively. In each pair of bars, the left, darker bar shows the performance of word-level translation; the right, lighter bar shows the results of translating using a trained Opus-NMT model.

The languages are clustered by family, in the same order as in the tables above: four Romance languages, three Germanic languages, followed by Finnish, Greek, and Polish on their own. The horizontal dotted gray line in the chart at 0.75 represents a level of coding agreement that is often used as a threshold in both automated and manual coding. The word-level translation comfortably exceeds this threshold for each language as regards sentiment polarity, with the worst-performing language, Dutch, at 0.81 for coding agreement. As one would expect, neural machine translation performs better still, with two important exceptions: Portuguese and Greek. These are the two languages for which `easyNMT` does not provide a pre-trained Opus-NMT model. Instead, I use a trained Opus-NMT model available through `cTranslate2`.[9]

The striking drop-off in performance for these two languages underscores the importance of using the best available pre-trained neural machine translation model, but also of the risk that such a model may not be readily available. While the open-source NMT set-up produces polarity and correlation similarity scores that are a few percentage points higher on average, that gain might be outweighed, for some applications, by the risk of inadvertently using a poorly trained NMT model.

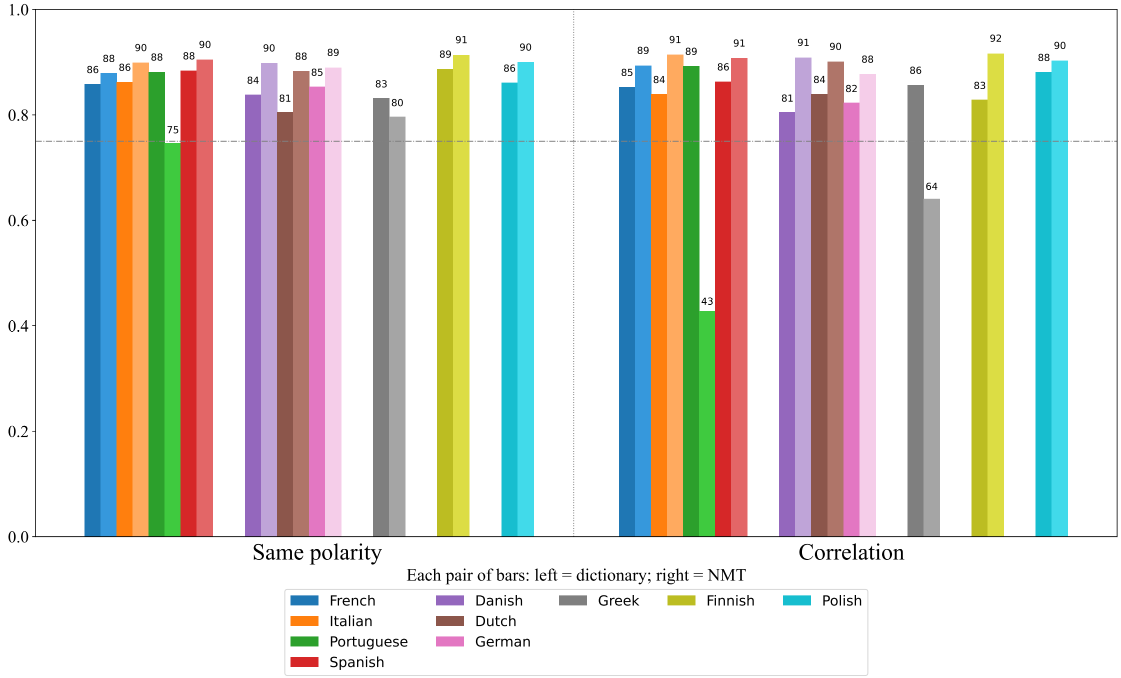

Figure 1: Sentiment analysis performance on original source translation.

As a practical illustration of the value of word-level translation, I translate a small corpus of 284 Dutch newspaper headlines about economics into English, using both Google Translate and the dictionary introduced here, and then apply the same sentiment analysis method to both sets of translated results. This corpus was used by van Atteveldt et al. 2021 to evaluate different sentiment analysis methods. The Google-translated corpus produces headlines whose sentiment polarity (negative, neutral, and positive) is classified correctly for 53.5% of the headlines; with the dictionary-translated corpus, one additional headline is translated correctly, for a score of 53.9%. Needless to say, this difference is not statistically significant, and 54% is not a particularly strong performance (though headlines, with their shorthand style, are notoriously difficult to classify automatically (Marcoci, 2014)). However, this performance does exceed that of the best-performing Dutch-language dictionary-based sentiment analysis method identified in (van Atteveldt et al., 2021), and it offers proof that word-level translation need not incur any performance loss compared to state-of-the-art neural machine translation.

[9] Two-step translation, through an intermediate language, is possible through easyNMT, but produces even worse results.

4.2 Emotions in text

To measure emotions in text, I use the `WM_emotion` lexicon (van der Veen, 2022). The lexicon contains word lists for each of the 8 emotions in Plutchik's well-known wheel of emotions: anger, anticipation, disgust, fear, joy, sadness, surprise, and trust 2001. I use the lexicon to count the number of words in each emotion category in each speech, both in its human-translated form and in the machine-translated version. This test is more difficult than the sentiment analysis test, since the individual emotion dictionaries range in size from roughly 200 words (for fear) to roughly 500 words (for trust), compared to thousands of words in the sentiment analysis lexica.

To compare human translation against machine translation, I compare the eight emotion counts produced for each version of a text in two ways. First, I consider the set of eight values as a single vector, and measure the cosine similarity between the two vectors. In addition, I compare the emotions individually, calculating the correlation between the human translation counts and the machine translation counts for each, and then averaging those eight correlation values to get an overall measure. The results are shown in Figure 2.

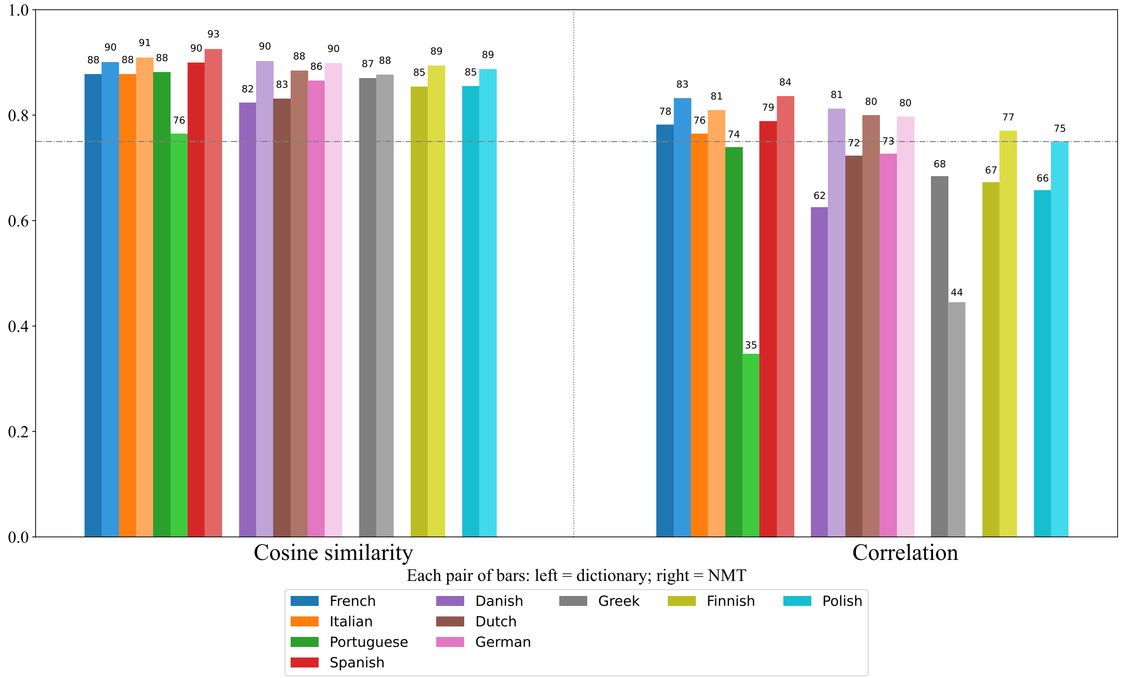

Figure 2: Measuring 8 emotions in text: cosine similarity & correlations of emotion tallies.

Again, the overall similarity between emotion measurements on the human translations and those on the word-level translations is quite high, as shown in the left-hand set of bars. For word-level translation, cosine similarity scores range from 0.82 (Danish) to 0.90 (Spanish); for open-source machine translation, ignoring the outlier that is Portuguese, figures are higher still, ranging from 0.88 (Dutch) to 0.93 (Spanish again). For individual emotions, correlations are lower. As already noted, this is likely due in part to the comparatively small dictionary sizes for these emotions. More generally, it is simply very difficult to identify specific emotions in texts. Moreover, while neural machine translation does come closer to human translation, the performance improvement is comparatively minor, except in the case of Danish. (And, as before, the neural machine translations produced for Portuguese and Greek do notably worse).

As for specific emotions, the greatest differences between the human and machine translations (i.e. the lowest correlations) occur for anger, disgust, and surprise, all of which produce average correlations (across the 10 languages) below 0.7. In contrast, the other five emotions all feature correlations meeting or exceeding the 0.75 level, with the average for fear an impressive 0.87. This variation nicely illustrates the importance of testing any dictionary-based measurement technique on a sample set of texts. In the present case, relying on word-level translation and a dictionary of fear words would produce good results, but for anger, the results would be less reliable.

4.3 Topic modeling

I run a topic model on a pooled corpus containing, for each text or speech, the EU's official translation into English, along with the automated translations (word-level as well as NMT) of the source language versions. The resulting corpora contain either 2 (*Sourcespeeches*) or 11 (*Allspeeches*) versions of each text. For each corpus, I run a topic modeling analysis, using Non-Negative Matrix Factorization (NMF) (Lee and Seung, 1999), selecting the optimal number of topics by measuring the average topic coherence across all topics in the model using corpus-specific embeddings (Greene and Cross, 2017).

For the *Sourcespeeches* corpus, the optimal number of topics is 26. The top 6 words for each topic are shown in Table 3. NMF topic models almost always contain a few topics that aggregate common non-substantive features of texts, such as standard text headers, stopwords, etc. In the case of legislative debates, various formalities and procedural terms also get included here. Of the 26 topics, 7 fall into this category. In Table 3, these non-substantive topics, as judged from their top words, are greyed out. They are not included in the similarity calculations reported below.

Table 3: Topics in the *Sourcespeeches* corpus

Topic	Top 6 words
	think, believe, madam, want, point, ladies
International	international, peace, united, people, military, government
EU institutions	council, parliamentx, commission, treaty, decision, presidency
	weather, man, president, woman, colleagues, got
	president, summer, advice, today, politics, versus
EU legislation	directive, legal, protection, proposal, rules, information
Voting	vote, debate, place, statements, written, 12
Budget	budget, financial, funds, eur, budgetary, fund
Agriculture	agricultural, food, agriculture, production, farmers, products
	om, left, al, europa, president, board
Member states	member, states, eu, state, national, union
Voting (2)	voted, favor, report, resolution, vote, voting
Energy	energy, nuclear, renewable, climate, emissions, gas
Socio-economic	social, employment, workers, labor, people, economic
	gives, ue, written, fur, board, 00
	know, east, plus, president, board, everybody
EP procedures	item, committee, behalf, report, statement, affairs
Development	development, policy, regional, regions, cohesion, important
Rights	rights, human, fundamental, respect, freedom, charter
EU legislation - amendments	amendment, oral, vote, paragraph, amendments, group
Turkey	turkey, turkish, accession, cyprus, negotiations, membership
Women and gender	women, men, equality, violence, gender, children
International relations	countries, trade, agreement, eu, agreements, world
Transportation	transport, air, road, rail, safety, traffic
Internal EU relations	thank, commissioner, excellent, work, rapporteur, colleagues
Economics and finance	economic, euro, monetary, growth, stability, crisis

As with the emotion analysis, I assess the similarity between the human and machine translations in two ways: across all topics, and by topic (averaged across the various topics to produce a single number). Figure 3 shows the results. On average, cosine similarity across the 20 topics, all ten languages score 0.85 or higher. Correlations at the individual topic level are only marginally lower, with Danish lowest at 0.81. Once again, the open-source neural machine translation shows even closer similarity to the human translation, with all scores over 0.92, with the exception, again of the two outliers Portuguese and Greek.

4.4 Comparisons across identical speeches

It is possible that some topics are particularly associated with specific languages. For example, the Polish fishing sector is rather smaller than is the case for some of the other countries included here. To verify that the results shown in Figure 3 are not driven by language-specific topics, I conduct the same analysis on the pooled *Fullspeeches* corpus. For this corpus, the optimal number of topics was 20, of which 7 are substantive, as shown in Table 4. Figure 4 shows the associated performance levels, juxtaposed against those for the original source language corpus.

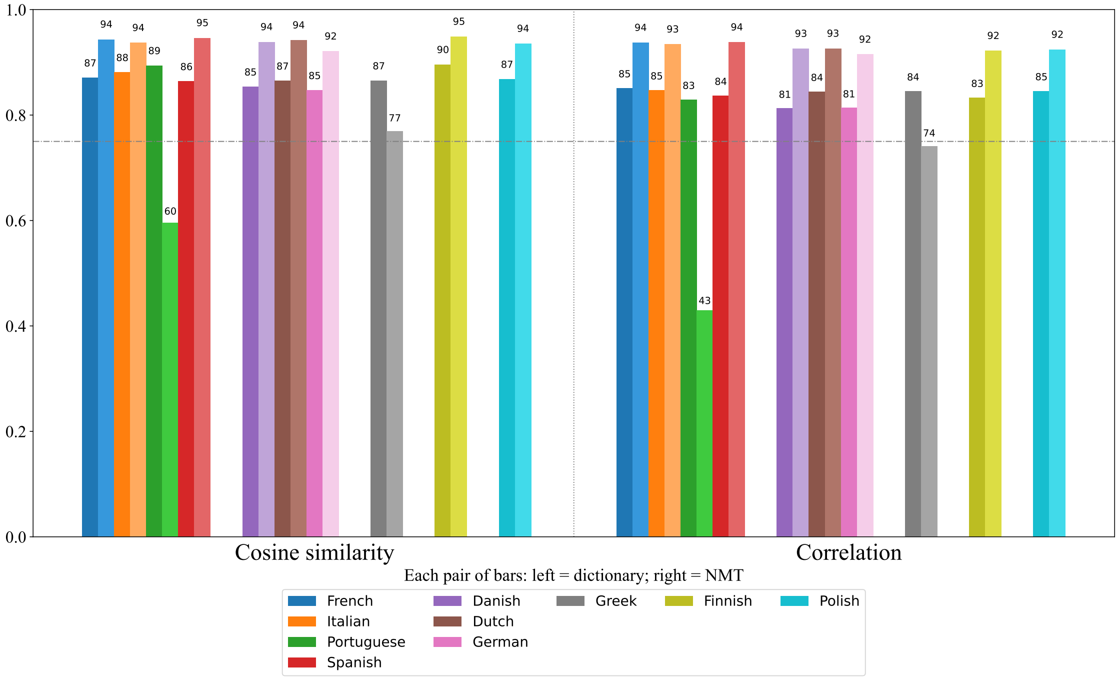

Figure 3: Topic modeling of original source speeches: cosine similarity & correlations of topic weights

Table 4: Topics in the *Allspeeches* corpus

Topic	Top 6 words
	uh, sir, association, got, committee, wanted
	able, chairman, sig, eu, ind, welcome
Development	development, social, strategy, economic, education, policy
Member states	states, members, member, union, state, national
Rights	rights, human, democracy, freedom, country, democratic
	om, left, committee, al, regarding, europa
	venerable, spokesperson, europe, member, desire, mixed
	gives, ue, fur, board, turn, 00
	know, east, plus, everybody, ue, embargo
Women and gender	women, men, equality, children, violence, life
Internal market	market, consumers, protection, directive, products, companies
	weather, need, woman, able, got, president
Voting	vote, report, voted, amendment, written, parliamentx
Internal EU relations	president, question, thank, today, parliamentx, commissioner
Energy	energy, nuclear, climate, efficiency, emissions, renewable
EP legislative activities	concerning, committee, purpose, activities, mp, reports
International relations	agreement, trade, countries, agreements, international, negotiations
Agriculture	food, agricultural, agriculture, farmers, production, products
EP procedures	order, day, commission, statement, bears, statements
Economics and finance	financial, crisis, budget, economic, fund, euro

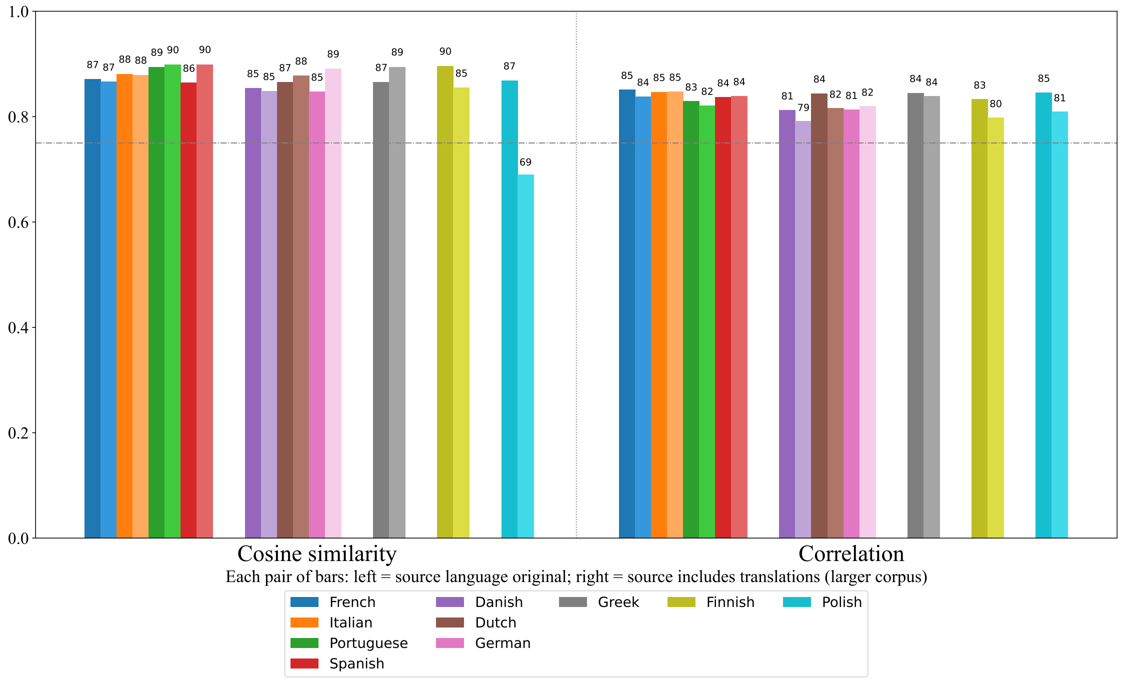

Figure 4: Topic modeling of original source & all speeches: cosine similarity & correlations of topic weights

The results show that, if anything, texts that have already been translated by human translators into another European language are easier to translate automatically than texts in the original source language. Only Polish forms a surprising exception, but even then only in terms of overall cosine similarity across topics; for individual topic correlations, the results are in line with those for the other languages.

5 Conclusion

When resources and time permit, there is no doubt that neural machine translation is to be preferred over word-level translation. However, resource and time often do not permit, and in those cases, the availability of high-quality word-level translation dictionaries can be invaluable. In addition, scholars may wish to run preliminary analyses on multilingual corpora before deciding whether to invest the necessary time or money. Others still may have laboriously developed a particular set of coding categories (for instance in the form of dictionaries) which have been validated in English, and may want to compare the scores for English-language texts on those categories to those of texts originally in other languages.

The present paper makes two important contributions on this front. First it outlines an affordable method to automatically generate high-quality word-level translation dictionaries from individual word translation from a commercial service, combined with relying on aligned word embeddings. The dictionaries used in this paper are available online, and are regularly updated and expanded. Second, the paper performs a number of comparisons between professional human translation, on the one hand, and word-level translation using the dictionaries introduced here as well as open-source neural machine translation using pre-trained models, on the other hand. These demonstrate that the latter both produce comparable results on standard text analysis tasks compared to human translation and, more importantly, that the performance loss from relying on word-level rather than neural machine translation is mostly minor. The comparisons presented here should assuage fears that the ungainliness of word-level translations implies they must also produce poorer or incompatible results when used as inputs for automated text analyses.

Automated translation steadily continues to improve in quality, but the highest quality translations will continue to require money, resources, or time for the foreseeable future. Moreover, for many of the analyses social scientists might want to conduct, it is not obvious that incremental improvements in translation quality will make much of a difference for the eventual findings. Accordingly, there will continue to be a need for high-quality word-level translation. Hopefully the dictionaries and method introduced here will help address that need.

References

Chen, X. and Cardie, C. (2018). Unsupervised Multilingual Word Embeddings. arXiv: 1808.08933.

Conneau, A., Lample, G., Ranzato, M., Denoyer, L., and Jégou, H. (2017). Word Translation Without Parallel Data. arXiv: 1710.04087.

de Vries, E., Schoonvelde, M., and 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.

Firth, J. (1957). A synopsis of linguistic theory 1930-1955. In *Studies in linguistic analysis*, pages 1–32. Philological Society, Oxford, UK.

Graham, Y., Haddow, B., and Koehn, P. (2019). Translationese in Machine Translation Evaluation. arXiv: 1906.09833.

Graën, J., Batinic, D., and Volk, M. (2014). Cleaning the Europarl Corpus for Linguistic Applications. In *Konvens 2014, Hildesheim, 8 October 2014 - 10 October 2014.*, Hildesheim. Stiftung Universität Hildesheim.

Greene, D. and Cross, J. P. (2017). Exploring the Political Agenda of the European Parliament Using a Dynamic Topic Modeling Approach. *Political Analysis*, 25(1):77–94.

Grimmer, J. and Stewart, B. M. (2013). Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political Analysis*, 21(3):267–297.

Gärdenfors, P. (2004). *Conceptual Spaces: The Geometry of Thought*. A Bradford Book, Cambridge, Mass., revised edition edition.

Joulin, A., Bojanowski, P., Mikolov, T., Jegou, H., and Grave, E. (2018). Loss in Translation: Learning Bilingual Word Mapping with a Retrieval Criterion. arXiv: 1804.07745.

Junczys-Dowmunt, M., Grundkiewicz, R., Dwojak, T., Hoang, H., Heafield, K., Neckermann, T., Seide, F., Germann, U., Aji, A. F., Bogoychev, N., Martins, A. F. T., and Birch, A. (2018). Marian: Fast Neural Machine Translation in C++. arXiv: 1804.00344.

Klein, G., Hernandez, F., Nguyen, V., and Senellart, J. (2020). The OpenNMT Neural Machine Translation Toolkit: 2020 Edition. In *Proceedings of the 14th Conference of the Association for Machine Translation in the Americas*, volume 1, pages 102–109.

Klein, G., Kim, Y., Deng, Y., Nguyen, V., Senellart, J., and Rush, A. M. (2018). OpenNMT: Neural Machine Translation Toolkit. arXiv: 1805.11462.

Koehn, P. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. In *Proceedings of Machine Translation Summit X: Papers*, volume 5, pages 79–86, Phuket, Thailand.

Lee, D. D. and Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791.

Marcoci, S. (2014). Some Typical Linguistic Features of English Newspaper Headlines. *Linguistic & Philosophical Investigations*, 13:708–714. Publisher: Addleton Academic Publishers.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient Estimation of Word Representations in Vector Space. arXiv: 1301.3781.

Mikolov, T., Le, Q. V., and Sutskever, I. (2013b). Exploiting Similarities among Languages for Machine Translation.

Plutchik, R. (2001). The Nature of Emotions: Human emotions have deep evolutionary roots, a fact that may explain their complexity and provide tools for clinical practice. *American Scientist*, 89(4):344–350.

Radovanovic, M., Nanopoulos, A., and Ivanovic, M. (2010). Hubs in Space: Popular Nearest Neighbors in High-Dimensional Data. *Jooournal of Machine Learning Research*, 11:2487–2531.

Schnabel, T., Labutov, I., Mimno, D., and Joachims, T. (2015). Evaluation methods for unsupervised word embeddings. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 298–307, Lisbon, Portugal. Association for Computational Linguistics.

Speer, R., Chin, J., Lin, A., Jewett, S., and Nathan, L. (2018). Wordfreq v2.2 (by Luminoso Insight).

Tiedemann, J. (2020). The Tatoeba Translation Challenge – Realistic Data Sets for Low Resource and Multilingual MT. pages 1174–1182.

Tiedemann, J. and Thottingal, S. (2020). OPUS-MT – Building open translation services for the World. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 479–480, Lisbon, Portugal.

Ustaszewski, M. (2019). Optimising the Europarl corpus for translation studies with the EuroparlExtract toolkit. *Perspectives*, 27(1):107–123.

van Atteveldt, W., van der Velden, M. A. C. G., and Boukes, M. (2021). The Validity of Sentiment Analysis:Comparing Manual Annotation, Crowd-Coding, Dictionary Approaches, and Machine Learning Algorithms. *Communication Methods and Measures*, pages 1–20.

van der Veen, A. M. (2022). Measuring discrete emotions in text: A word-mapping approach.

van der Veen, A. M. and Bleich, E. (2021). Automated sentiment analysis for the social sciences: A domain-independent, lexicon-based approach.

6 Appendix

6.1 Using aligned mappings to generate translation dictionaries

In principle, translation is a simple matter of finding a word's location in the source language space and identifying the closest word to that location in the target language space. The standard metric for finding the closest word is cosine distance (equal to the vector dot product for vectors normalized to length 1). However, this metric fails to take into account that word locations are not evenly distributed across the vector space. As Radovanovic et al. have shown, high dimensionality inevitably produces 'hubs', which are nearest neighbours of many words 2010. More significantly, the distribution of words across vector space varies by language: a hub in one language will not necessarily be a hub in another. As it turns out, words tend to be closer to words in another language that have a similar frequency, but not necessarily a similar meaning. Moreover, the problem is worse for more frequent words, which are often among the ones for which translation errors will have the largest impact. Schnabel et al. show that this pattern is driven by the algorithms that generate the word embeddings, rather than by the "intrinsic properties of natural language" 2015, p. 10.

Fortunately, this also means that it is a problem for which we can make adjustments. I use the improved distance metric proposed in 2017 by Conneau et al., which adjusts for the degree of 'hubness' of various candidate neighbours 2017, p. 4. This cross-domain similarity local scaling (CSLS) adds two adjustments to the standard cosine metric: one for the average distance to the nearest target-language neighbours of the source-language word ($r_T(Wx_s)$), and one for the average distance to the nearest source-language neighbours of the target-language translation candidate ($r_S(y_t)$), where the Wx_s indicates that the source embedding for x has been mapped to the same space as the target embedding. The CSLS metric is:

$$CSLS(Wx_s, y_t) = 2cos(Wx_s, y_t) - r_T(Wx_s) - r_S(y_t) \tag{1}$$

I follow Conneau et al. in using 10 as the number of nearest neighbours to consider 2017, p. 4. The impact of using the CSLS distance metric can be dramatic: words that are not even among the 25 closest neighbours using 'standard' cosine distance regularly become the closest neighbour under CSLS. More importantly, the top candidates using CSLS are virtually always better translations for the source word than are the top candidates using cosine distance.

In addition, the same techniques used to expand the DeepL translations apply here. In particular, for words that are not in the embedding, I check for variations in capitalization, accents (diacritical marks), and spelling, including word endings (for example, adjectives in Italian can generally end in four different vowels, without a change in meaning). I also apply the same segmentation techniques. If none of these options produce a translation, the source word more often than not is a proper name, for which a straight copy is the appropriate 'translation.' Finally, candidate translations that contain characters that do not form part of the target language are simply ignored.[10]

For each word translated, I keep track of the way it was translated. This makes it straightforward to update or change just a particular part of the translation dictionary containing translations generated in a specific way. For instance, to replace embedding-derived translations by additional DeepL translations, we can simply select all words with the relevant translation codes. A full list of different translation codes appears in Table 5.

6.2 Comparing results for the *Sourcespeeches* and *Allspeeches* corpora

Figure 5 and Figure 6 display, for the sentiment analysis and emotion measurement tasks, respectively, the results for the *Sourcespeeches* and *Allspeeches* corpora, juxtaposed. In each pair of bars, the left, darker bar shows the score for the former, while the right, lighter bar shows the results for the latter. As is clear from the figures, the performance is nearly identical for the two. This establishes that differences across languages in the paper were not driven by the fact that scores were derived from distinct, non-overlapping corpora. Here, the right-hand bar of each pair is directly comparable to each of the other right-hand bars.

[10] For example, the closest neighbour to the Dutch word 'consequent' in the English model is the Japanese kanji for 'mansion', rather than the actual English translation, 'consistent'.

Table 5: Translation codes

Code	Explanation
-1	Unknown
0	Trusted (use to directly supply domain-specific, gold-standard translations)
30	Deepl direct
31	Deepl lower-cased
32	Deepl umlaut replaced by <vowel>e
33	Deepl hyphenated segments
34	Deepl hyphenated segments, lower-cased
44	Deepl word starts with lead hyphen (ignore hyphen)
45	Deepl word ends with trailing hyphen (ignore hyphen)
35	Deepl capitalized segments
36	Deepl capitalized segments, lower-cased
37	Deepl segments combinatorially identified
38	Deepl segments joined by connecting letter (ignore letter)
39	Deepl segments, with final segment stemmed
40	Deepl segments, with final segment stemmed & lower-cased
41	Deepl plural constructed with 's (Dutch only)
42	Deepl definite article suffix (Danish)
43	Deepl ending modified (singular/plural, masculine/feminine, etc.)
1	Aligned direct
2	Aligned lower-cased
7	Aligned accent-stripped
8	Aligned accent-stripped, lower-cased
9	Aligned spell-corrected
10	Aligned spell-corrected, lower-cased
19	Aligned ending modified
20	Aligned ending modified, lower-cased
11	Aligned hyphenated segments
12	Aligned hyphenated segments, lower-cased
13	Aligned capitalized segments
14	Aligned capitalized segments, lower-cased
15	Aligned segments combinatorially identified
16	Aligned segments accent-stripped
21	Aligned segments with final segment stemmed
17	Aligned segments stemmed
3	Copied over - punctuation
4	Copied over - contains digits
5	Copied over - common word in target language
6	Copied over - common word in target language, lower-cased
18	Copied over - no translation identified & no feasible heuristics

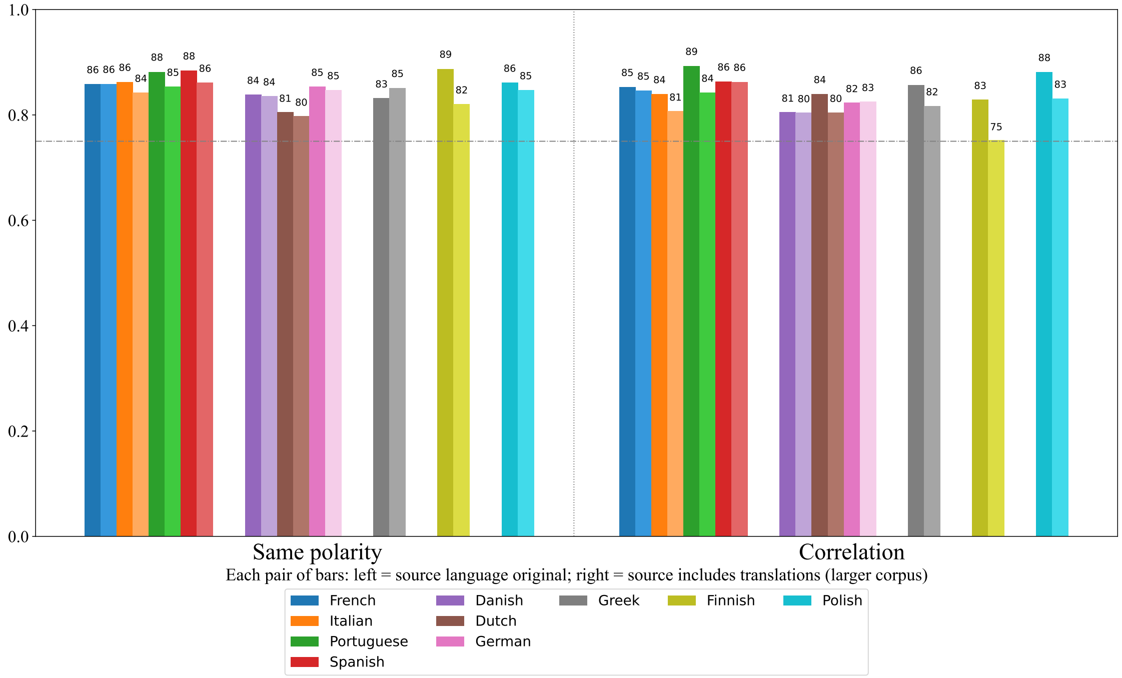

Figure 5: Topic modeling of original source & all speeches: cosine similarity & correlations of topic weights

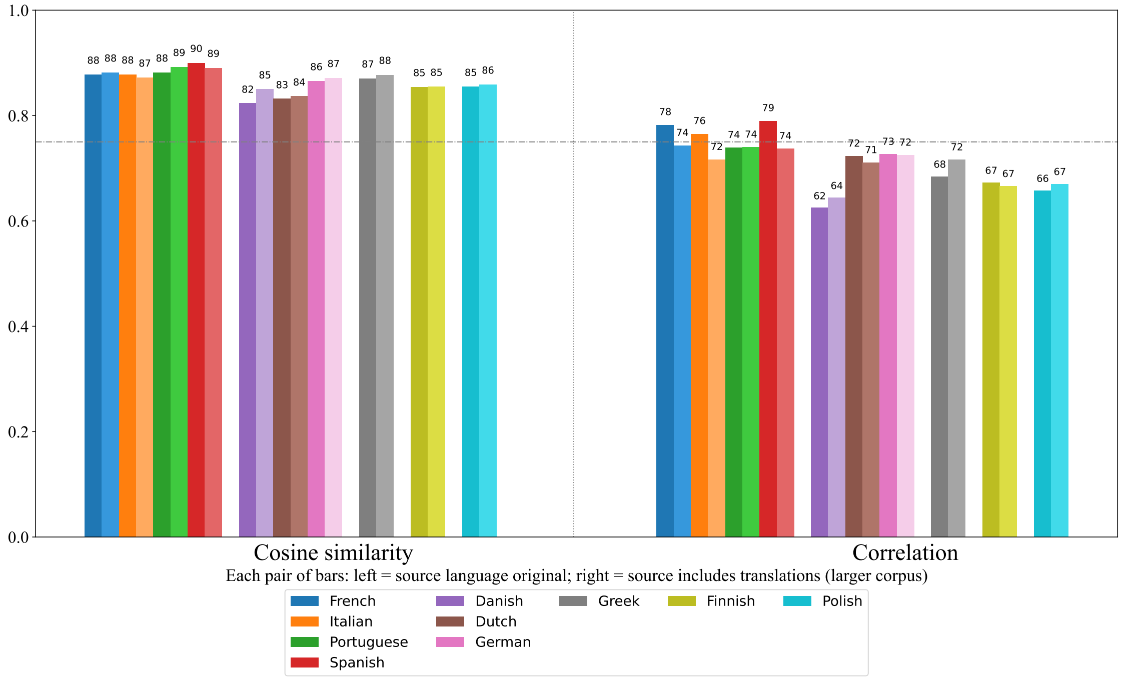

Figure 6: Topic modeling of original source & all speeches: cosine similarity & correlations of topic weights