No Longer Lost in Translation: Evidence that Google Translate Works for Comparative Bag-of-Words Text Applications[1]

Erik de Vries†, Martijn Schoonvelde‡, Gijs Schumacher§

† Department of Media and Social Sciences, University of Stavanger
email: erik.devries@uis.no
‡ Department of Political Science and Public Administration, Vrije Universiteit
email: h.j.m.schoonvelde@vu.nl
§ Department of Political Science, University of Amsterdam
email: g.schumacher@uva.nl

Automated text analysis allows researchers to analyze large quantities of text. Yet, comparative researchers are presented with a big challenge: across countries people speak different languages. To address this issue, some analysts have suggested using Google Translate to convert all texts into English before starting the analysis (Lucas et al., 2015). But in doing so, do we get lost in translation? This paper evaluates the usefulness of machine translation for bag-of-words models – such as topic models. We use the *europarl* dataset and compare term-document matrices as well as topic model results from gold standard translated text and machine-translated text. We evaluate results at both the document and the corpus level. We first find term-document matrices for both text corpora to be highly similar, with minor differences across languages. What is more, we find considerable overlap in the set of features generated from human-translated and machine-translated texts. With regards to LDA topic models, we find topical prevalence and topical content to be highly similar with again only small differences across languages. We conclude that Google Translate is a useful tool for comparative researchers when using bag-of-words text models.

[1] *Authors' note:* Replication code and data are available at the *Political Analysis* Dataverse (De Vries, Schoonvelde and Schumacher, 2018) while the Supplementary materials for this article are available on the *Political Analysis* web site. The authors would like to thank James Cross, Aki Matsuo, Christian Rauh, Damian Trilling, Mariken van der Velden and Barbara Vis for helpful comments and suggestions. For conducting this research, Schoonvelde & Schumacher received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 649281, EUENGAGE

Introduction

Automated text analysis is like a gold rush. Many researchers have noticed its potential and are now using methods such as topic modeling, scaling and sentiment analysis to analyze political texts (for an overview see Grimmer and Stewart, 2013). But researchers interested in cross-country comparisons face a problem: people speak different languages. In order to make comparisons across countries, researchers first need to translate texts from several languages into one. On the plus side, nowadays this can be automated by using machine translation, such as, for example, Google Translate. But does the meaning of these texts get lost in Google translation? That is, do we lose (too much) information if we Google Translate texts before we analyze them? Or does doing so leave us like the poor souls who journeyed west for gold but were left with nothing?

This paper evaluates the usefulness of machine translation for automated bag-of-words models.[2] We identify and evaluate four reasons why the meaning of a text may get lost in translation. First, a general problem occurs when words or stems in machine-translated documents are translated differently than those in gold standard documents, leading to different term-document matrices (TDMs).[3] We evaluate this issue by comparing the overlap between gold standard and machine-translated TDMs. Other translation problems relate more specifically to LDA topic modeling, a popular bag-of-words model that identifies the topics in a corpus, and assigns documents and words to these topics. In this case, translation issues may arise because (1) topics in the machine-translated corpus may be assigned to different documents than in the gold standard corpus, (2) machine-translated documents are assigned to different topics than gold standard documents and (3) a topic in the machine-translated corpus consists of different words than the same topic in the gold standard corpus. We evaluate each issue by systematically comparing topic models estimated using machine-translated documents with those estimated using human-translated (gold standard) documents.

[2]The goal in this paper differs from much work in computational linguistics or Natural Language Processing (NLP), as that type of research is mostly concerned with syntax, readability and the correct use of grammar in translations (e.g., Scarton and Specia (2014), Kaljahi and Samad (2015), Aharoni (2015)). In contrast, this paper compares bag-of-words vectors and topic models that are based on them. Both are used regularly in applications of automated text analysis in the social sciences.

[3]Throughout the paper, we use the terms *bag-of-words vectors* and *term-document matrices* (TDMs) interchangeably.

To set up our comparisons, we use the *europarl* dataset (Koehn, 2005), which contains the official transcriptions of debates in the European Parliament both in English and in most other official languages of the EU. From this dataset we take debate transcriptions in Danish, German, Spanish, French and Polish for the period of January 2007 to November 2011. Delivered by professional translators, these official transcriptions serve as our gold standard.[4] We first compare the bag-of-words vectors of each document in the machine translation and the gold standard translation. We then compare the output of the LDA topic models in three ways: topical prevalence at the document-level, topical prevalence at the corpus level and topical content at the corpus level.[5]

We find that TDMs for both sets of data are highly similar, with significant but minor differences across languages. What is more, we find considerable overlap in the set of features (stems) appearing in human- and machine-translated texts. With regards to LDA topic models, at both the document and the corpus levels we find topical prevalence to be similar with only small differences across languages, and we find topical content to strongly overlap as well. These findings suggest that Google Translate does in fact generate useful TDMs, and, what is more, it deals successfully with the above-mentioned risks of machine translation when estimating topic models. We conclude that Google Translate is a useful tool for researchers who use or want to use bag-of-words text models for comparative questions.

Background

Numerous bag-of-words based studies have analyzed machine-translated texts, yet little is known about the quality of machine translations and its impact on subsequent analyses. Generally, authors either assume machine-translated text to be suitable for their purposes or they do not pay attention to the issue at all. For example, Agarwal et al. (2011) use Twitter data which was machine-translated by an unidentified commercial source, but they do not address the possibility that

[4] Partly because of thorough quality requirements, the costs of hiring professional translators in the European Union are high, by some estimates as much as €2 per EU inhabitant per year (see `http://ec.europa.eu/dgs/translation/faq/index_en.htm`). A gold standard indeed.

[5] Topical prevalence refers to which topics appear in a document or in the corpus as a whole (i.e., topic distributions), whereas topical content refers to what words constitute a topic (i.e., word distributions). (Lucas et al., 2015).

machine-translation may have influenced their results. Schwarz, Traber and Benoit (2017) use Google Translate in the multilingual Swiss context. While these authors describe the machine-translation process in more detail, they do not discuss comparisons between different machine-translation strategies, or the quality of their translations.

To be clear, we do not imply that machine-translation is not useful for analyzing texts in multiple languages. As Lotz and Van Rensburg (2014) show, developments in machine-translation systems are moving fast and their quality is clearly increasing with time. Balahur and Turchi (2014) give a comprehensive account of using machine-translated text for automated analyses in the context of sentiment analysis, and Courtney et al. (2017) find that machine-translated newspaper articles can be reliably categorized by human coders. But while these contributions are highly relevant, they do not evaluate the implications of machine-translation for bag-of-words methods more generally. The same is true for Lucas et al. (2015), who write extensively about the possible pitfalls of analyzing machine-translated text but do not evaluate its quality empirically. Adding to this line of research, this paper systematically evaluates both the bag-of-words approach in general and topic modeling in particular.

Another issue relevant to this study concerns the impact of specific languages and language groups on machine translation quality. For example, machine-translated texts may be of better quality when translated from French to English than when translated from Polish to English. There are two reasons for this. First, some language pairs are simply more easily translated than others (Koehn and Monz, 2006). Furthermore, larger parallel corpora are available to train machine-translation models for some language pairs than for others (e.g., there is more parallel data available for French and English than there is for Polish and English). To examine this possibility we include in our analysis languages from different language groups: French and Spanish (belonging to the Italic language group), German and Danish (belonging to the Germanic language group), and Polish (belonging to the Balto-Slavic language group).[6]

[6]While there are different topologies of language groups in the field of linguistics, this paper uses the topology described in Gray and Atkinson (2003).

Data & Measurement

To evaluate the quality of machine-translation, we need to compare its output to gold standard translations of identical documents.[7] The *europarl* dataset (Koehn, 2005) contains parallel corpora which allow us to set up such comparisons. The dataset consists of official European Parliament debate transcriptions in most of the official EU languages.[8] The *europarl* dataset ranges from April 1996 until November 2011, but some Slavic language translations were included only since January 2007. Because of that, we focus our analysis on translations between 2007 and 2011.

The primary purpose of the *europarl* dataset is to train, test and improve machine translation algorithms (e.g. Koehn, 2005; Popescu-Belis et al., 2012; Loaiciga, Meyer and Popescu-Belis, 2014). The data is available in both the raw form and as text files with sentence-aligned language pairs. We use the raw data, because the sentence-aligned text files do not distinguish between different dates and debate chapters. The raw data files are organized per session (typically one day) and chapter. Each chapter is a different item on the agenda (e.g a debate, questioning of EU official or vote) of a session.[9] When estimating topic models, we consider each chapter to be a single document, because each chapter in a session concerns a specific agenda item. Each agenda item may in turn consist of multiple topics.

Methods

Figure 1 shows the steps we take to compare machine-translated and gold standard documents. In both cases we start with identical non-English texts, which have been translated into English, either

[7]Replication code and data are available at the *Political Analysis* Dataverse (De Vries, Schoonvelde and Schumacher, 2018) while the Supplementary materials for this article are available on the *Political Analysis* web site.

[8]Contributions to debates in the European Parliament can be either in English or in one of the other official EU member state languages. Contributions in those languages are then translated—by official translators—in *all* other recognized EU languages either directly or indirectly through English. What we consider our gold standard data of the debates is the English corpus which consists of a) English contributions, 2) contributions in one of the EU languages translated into English by official translators. What we consider our machine-translated data consists of Google translations into English of these same contributions in 1) one of our 5 languages, and 2) in other EU languages that have been translated into these 5 languages either directly or indirectly by official translators.

[9]Because the provided data is not exactly the same for all languages (e.g. chapter 5 in the session of 04-01-2007 might be present in the English but not in the German data, while the German data does contain other chapters from that same session), we had to match all language pairs (EN-DA, EN-DE, EN-ES, EN-FR, EN-PL) by checking for the presence of each chapter in each session for both languages. This results in between 2148 (DE) and 2347 (FR) chapters per language pair.

through Google Translate or through EU-employed expert translators (*Step 1*). These translations are preprocessed and turned into TDMs (*Step 2*) on which we then estimate a topic model (*Step 3*). We then compare the similarities of the TDMs, the topical prevalence at the level of individual documents and the corpus at large, and the topical content (*Step 4*). In what follows we discuss each step in more detail.

Step 1. Machine-translation & Google Translate

We use Google Translate as the specific machine-translation service to evaluate the performance of machine-translated texts in bag-of-words analyses. We chose Google Translate because of its translation quality, which is top-tier when compared to other online machine translating services (Hampshire and Salvia, 2010). We translated the texts using the Google Website Translator plugin which can translate web pages. To able to use this plugin we converted the raw text data to bare html web pages. The translation process took place in August and September 2016.[10] We have translated the texts into English, because machine translation algorithms are expected to perform best when translating to and from English.[11]

Step 2. Preprocessing & generating TDMs

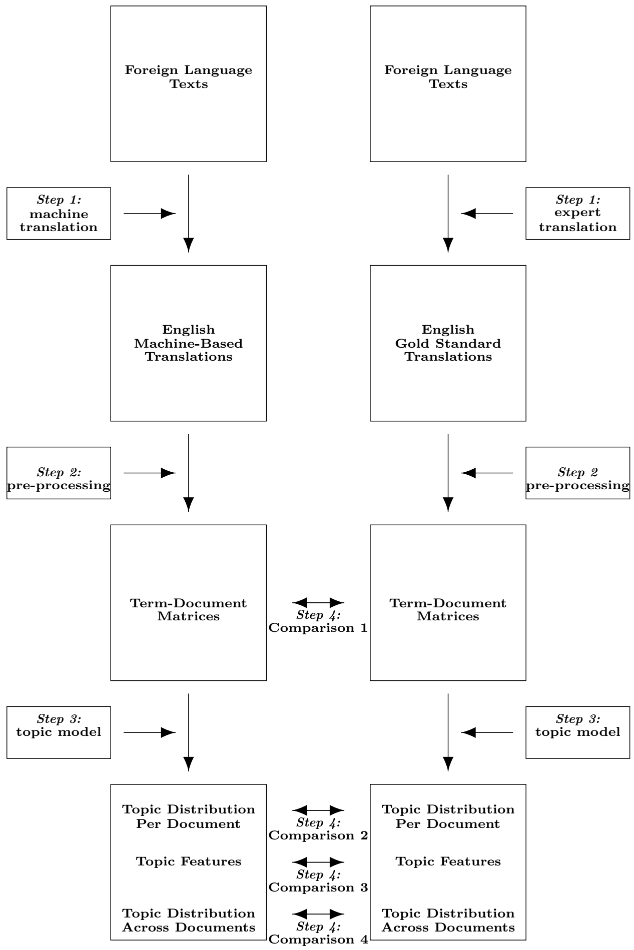
When using bag-of-words models, it is common to preprocess the data in order to remove noise. In our case we have removed punctuation, numbers and general stopwords, and all remaining words have been lowercased and stemmed. The preprocessing steps on both the gold standard and machine-translated texts are identical, and were applied to the translated texts.[12] To perform these preprocessing steps, we used both Python and R libraries. For stemming, stopword removal,

[10] We conducted the translations before Google rolled out their deep learning/neural network-based translation algorithms. Because this may have improved the quality of Google Translate, our results are likely to be conservative.

[11] The reason for this is that English is the *lingua franca* of the internet and, by consequence, most translations are from or to English. This produces large parallel corpora between English and other languages. Machine translation algorithms are trained on these models, and typically the more data the better the performance of the algorithm.

[12] Recent research shows that seemingly innocuous preprocessing steps might impact the outcome of (unsupervised) automated text analyses (Denny and Spirling, 2016; Greene et al., 2016). Our comparison, however, is between gold standard and machine-translated texts on which we applied identical preprocessing steps. Although we cannot we certain, we do not expect these preprocessing steps to have had a systematically different impact on both corpora. We should note, however, that, in general, the removal of stopwords will influence word and topic distributions within a topic model, and this will also apply to the model results we present here. But since stopwords contain–by definition–no topical content we expect their removal to have had minimal substantive implications.

Figure 1: Research design



Note: This figure shows the different steps of our research design. In both cases we start with non-English texts, which have been translated into English, either through Google Scholar or through EU-employed expert translators (*Step 1*). The English translations are then preprocessed and turned into TDMs (*Step 2*), on which we then estimate a topic model (*Step 3*). We then compare our four different outcome variables (*Step 4*). The comparisons are the following:

Comparison 1: document-to-document comparison TDM similarity;
Comparison 2: document-to-document comparison of topic distributions (topical prevalence);
Comparison 3: topic-to-topic comparison of stem weights (topical content);
Comparison 4: topic-to-topic comparison of topic distribution (topical prevalence).

number removal, lowercasing, and punctuation removal, we used regular expressions in Python and the *NLTK* package (Bird, Klein and Loper, 2009). To create the TDMs we switched to R and the *quanteda* package (Benoit and Nulty, 2013).[13] We will compare the TDMs of the machine-translated and gold standard documents and we also use them as input for the topics models described below. Readers primarily interested in our analysis of the TDMs may decide skipping the next section, which is contains more technical details regarding the specification of our topic models.

Step 3. Fitting topic models

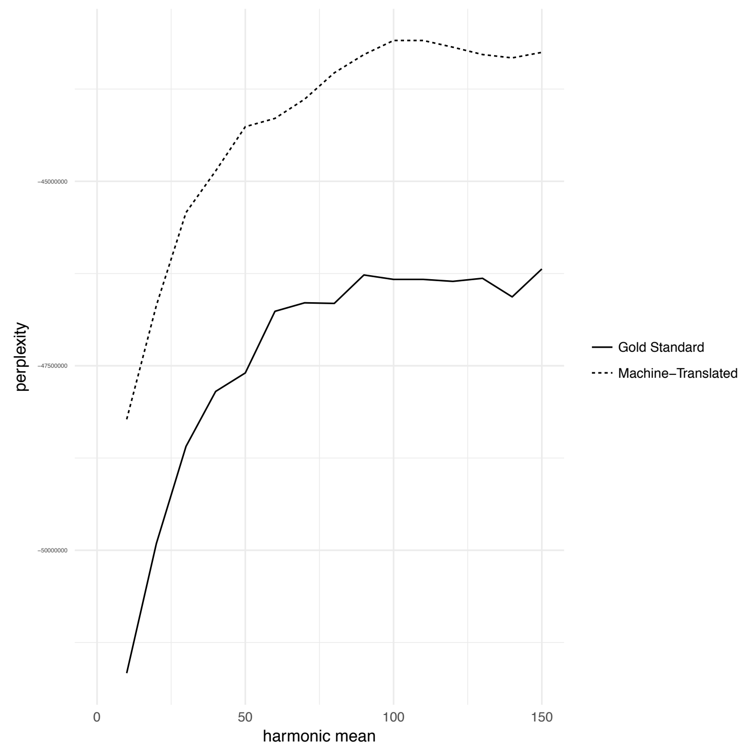
To assess the quality of machine-translated texts, we estimated topic models on the gold standard and machine-translated texts separately using the LDA algorithm (Blei, Ng and Jordan, 2003) and Gibbs sampling. For this we used the LDA function in the R *topicmodels* package (Hornik and Grün, 2011). LDA is a generative model. It takes the words in each text as input and then estimates the topical prevalence and topical content in the corpus. To run the model researchers need to set a few parameters: the number of topics in the corpus, the model seed, burn-in time, the number of iterations and which and how many iterations to use for the final model. To ensure that differences between a model based on the gold standard corpus and a model based on the machine-translated corpus are solely the result of language differences between these corpora, the parameters for the topic models based on gold standard translations and machine-translations were kept identical. This means that the number of topics was kept constant, and a fixed seed was used – based on the sys.time variable – as suggested by Hornik and Grün (2011). This seed (1473943969) has been used for all models described below. Furthermore, the burnin (1000) and number of iterations (300) were also kept constant. The algorithm keeps every 100th model and returns the model with the highest posterior likelihood (the best-fitting model). Consequently, all variation between the models – when the model parameters are kept the same – results from differences caused by the translation process.

The most important parameter to set is the number of topics in the topic model. This is crucial

[13] We switched from Python to R for practical reasons only: we started this project using Python because of the large amount of scraping and cleaning required. We then switched to R because we considered it more suitable for the modeling and plotting tasks at hand.

because the number of topics affects the distribution of words over topics (topical content) and the distribution of topics over documents (topical prevalence). When the number of topics changes so do these distributions. It was practically infeasible to run and optimize the number of topics for each language pair. Also all language pairs are based on roughly similar data from the same time period. Therefore the optimum number of topics for all models was determined based on the French dataset. This is the largest gold standard and machine-translated dataset. We estimated the best-fitting number of topics by evaluating the model harmonic mean of models that contain between 10 and 150 topics, in increments of 10. The model harmonic mean indicates the extent to which word counts in the documents used to construct the model match the word distributions in the model itself. Put differently, it indicates the extent to which the model accurately describes the distribution of words in the documents. In this case, a larger harmonic mean indicates that the model fits the data better. The results of the optimization runs are displayed in figure 2. The gold standard model has an optimum of 90 topics. After 90 topics adding more topics does not improve model fit. The machine-translated model peaks at 100 topics. To isolate the effect of language differences between gold standard and machine-translated texts it is important to choose the same number of topics for both models. Therefore we settled for 90 topics. That said, we also evaluated comparisons of models with 90 topics for the gold standard models and 100 topics for machine-translated models. This produced results almost identical to the topic model comparisons with 90 topics. These results are available in Appendix.

Figure 2: Model harmonic mean



Our next challenge is to match the topics generated by the gold standard and machine-translated models. This is because the topic order in both models may differ (i.e. topic 1 in the machine-translated model may match best with, for example, topic 2 in the gold standard model). Our matching procedure is as follows: for each stem we find the highest loading in the machine-translated topic model and the gold standard topic model. For example, take the stem "agricultur". This stem loads highest on (is most important in) topic 12 of the machine-translated model, and topic 45 in the gold standard model. This results in a 12-45 topic pairing for that specific stem. We subsequently count the topic pairings of all shared stems. We match topics based on the highest count of topic pairings. For example, we pair topic 12 of the machine-translated model with topic 45 in the gold standard model because they have the highest number of important, shared stems like the stem "agricultur" (See the Appendix for a numerical example of our topic matching procedure).[14] Using this procedure we matched 90 topics for the German corpus and 89 topics for

[14]It is of course possible that a topic in the machine-translated model is matched to several different topics in the

all other languages.[15,16]

Step 4. Comparing term-document matrices & topic models

We will make four different comparisons, which vary on two dimensions: stems versus topics and documents versus corpora (see Table 1). The comparison of TDMs takes place at the level of stems and documents (*Comparison 1* in Figure 1). Furthermore, we report three comparisons based on our topic models, all of which give us evidence how much the matched topics in the machine-translated and the gold standard topics overlap in content and prevalence. We evaluate topical content by means of stem loadings per topic pair (*Comparison 3* in Figure 1). We evaluate topical prevalence by means of topic distributions over document-pairs (*Comparison 2* in Figure 1), and topic distributions across the corpus at large (*Comparison 4* in Figure 1).

Table 1: Comparisons between gold standard and machine-translated data

	Stems	*Topics*
Document level	stem counts per document pair	topic distribution per document pair
Corpus level	stem loadings per topic pair	topic distribution per topic pair

It is important to evaluate results at both the document and the corpus level because the former only speak to how similar individual documents are being characterized by the topic model (i.e., the extent to which topical prevalence for gold standard and machine-translated documents is

gold standard model. For example, while “agricultur” is matched 12-45, it could be that the stem ”farmer” loads highest on topics 12 and 33, resulting in two different topic pairings for topic 12 in the machine-translated model (namely 12-45 and 12-33). In those cases, we use the topic combination with the highest number of topic pairings, while ignoring the other. This results in topic pairs that always consist of the two topics that share their highest loading words with each other.

[15] The reason that not all topics can be matched for all languages is because when every shared stem loads higher on another topic in the same model, there are simply no stems to base a match on. We can again take “agricultur” as an example. This stem is the most important (highest loading) in both topic 12 (word loading: 0.12) and 19 (word loading: 0.09). Yet our procedure only registers on which topic “agricultur” loads highest (which in this case is topic 12). So topic 19 will not be matched to another topic based on this stem alone. If not a single stem loads highest on a topic, then that topic cannot be matched and we discard it. In practice this means that while the unmatched topics might have some substantive importance, all their stems are – by design – more important to other topics.

[16] Even though our matching procedure worked well, we should note there are other possible ways to match topics, using for example the Hungarian algorithm (see e.g., Chuang et al., 2013; Roberts, Stewart and Airoldi, 2016).

similar). However, such a comparison does not tell us how similar the fitted topics themselves are. For example, both the gold standard and machine-translated document might have a high topic loading on topic 1, making them highly similar on the document level, but if topic 1 is about cars in the gold standard topic model and about trees in the machine-translated model, then document-level similarity does not tell us much. While the chances of this happening are slim, structural and consistent translation errors by Google Translate might cause such differences. As a consequence, the level of topical similarity does say something about the quality of the translation. We thus need comparisons on both the document and corpus level.

Our outcome measure for the TDM comparisons is different from that of the topic model comparisons. For the TDM comparisons, we use cosine similarity because – in contrast to correlation – it takes into account the absolute differences in values. This is relevant for comparing TDMs because of our goal of knowing how similar the counts of all TDM features per document pair are to each other. Cosine similarity varies between 0 and 1, with the latter indicating a perfect match (i.e., two identical vectors). For the topic model comparisons, correlations are a more suitable similarity measure because they detect trends rather than absolute values. This is important because we will make comparisons between different models.[17] Correlations vary between -1 and 1, with the latter indicating a perfect linear positive relationship, and the former indicating a perfect linear negative relationship.

Results

This section contains the results of our four comparisons, starting with the TDM analysis, and continuing with the topic model analyses.

Comparing TDMs

We first compare – at the document-level – machine-translated and gold standard bags-of-words to each other, using the built-in similarity function in the *quanteda* R package (Benoit and Nulty,

[17] As discussed before, changing the number of topics influences both the document-level and corpus-level topic distributions as well as stem distributions per topic, and because of that absolute values are no longer meaningful.

2013). Figure 3 displays the distribution of the cosine similarity scores for each language. Most notably, the average similarity between the gold standard documents and their machine-translated counterparts is very high (M=0.92, SD=0.07). Furthermore, more than 92% of all document pairs achieve a cosine similarity score of 0.80 or higher. These results show that the TDMs of machine-translated and gold standard documents are very similar. Very often the stems in the machine-translated and gold standard documents occur with (approximately) the same frequency.

Figure 3: Distribution of cosine similarity per language pair

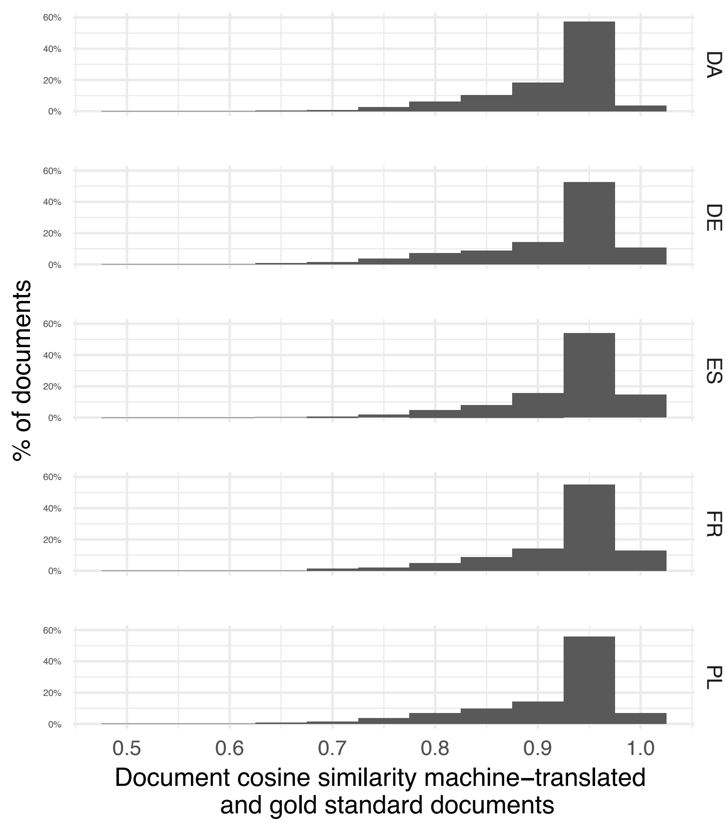


Table 2 shows the means and standard deviations for document cosine similarity scores per language. The differences between languages are tiny: the lowest mean cosine similarity (Polish = 0.913) is only 0.016 smaller than the highest mean cosine similarity (Spanish = 0.929). The French and Spanish documents have significantly higher average cosine similarities than the overall mean (French: t=7.07, p<0.001; Spanish: t=5.11, p<0.001), but the size of these differences is, again, very small (French: 0.005 and Spanish: 0.009). The Danish, Polish and German cosine similarities between documents pairs are not significantly different from the overall mean.

Table 2: Cosine similarity distribution per language

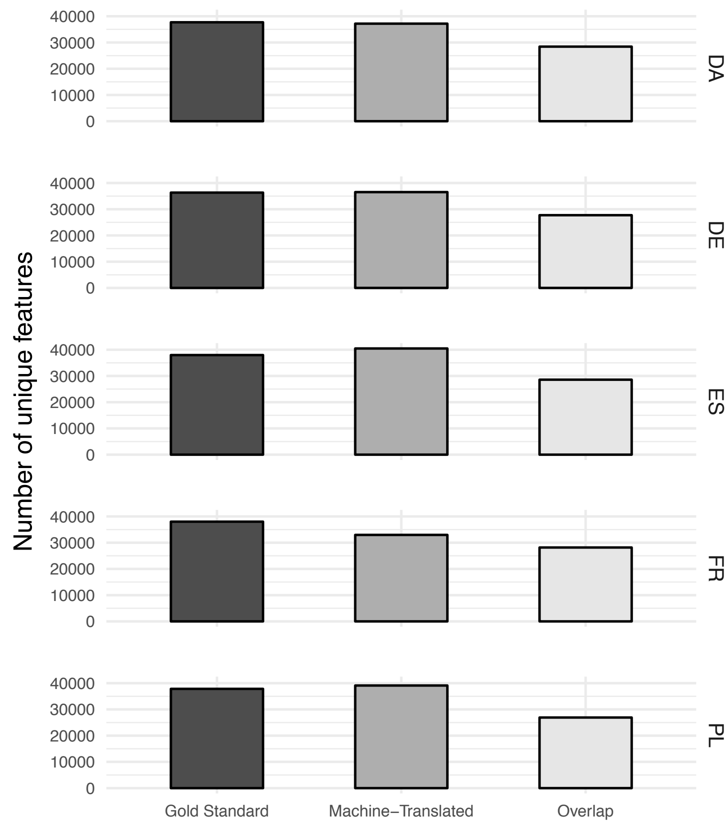
Language	N	Mean	St. Dev.	Min	Max
Danish	2,301	0.915	0.063	0.549	0.992
German	2,148	0.915	0.074	0.488	0.991
Spanish	2,335	0.929	0.059	0.483	0.991
French	2,347	0.925	0.064	0.564	0.989
Polish	2,338	0.913	0.073	0.475	0.989
Total:	11,469	0.919	0.066	0.475	0.992

Note: Statistically significant but substantively small difference between languages (ANOVA results: F(4, 11464) = 27.855, $\rho < 0.001$, $\eta^2 = 0.010$)

We also consider the total number of unique stems (features), as well as the number of shared stems between the gold standard and machine-translated TDMs. The higher the number of shared stems, the more overlap there is. Figure 4 shows that the shared features of the TDMs of the gold standard and machine-translated documents overlap to a large degree (about 75% or higher). The number of shared features is also quite similar for each language (DA, 28431; DE, 27732; ES, 28578; FR, 28162; PL, 26916). The same goes for the features that are unique to either the gold standard or machine-translated TDMs.

The exception is French, and to a lesser extent Spanish. In the Spanish case, more unique features are present in the machine-translated than in the gold standard texts, which indicates that Google Translate adds new features to the texts (by using different English translations for the same Spanish word). Similarly, French translations are simplified (different French words are translated

Figure 4: Unique TDM features for gold standard and machine-translated corpora



Reading example: For the French language, the amount of overlapping features is around 28,000, while the total number of features is around 33,000 for the machine-translated documents and around 38,000 for the gold standard documents.

as the same English word).[18] However, regardless of these differences, both the substantial overlap among features and the high cosine similarity scores for both Spanish and French show that their machine-translated and gold standard TDMs are highly similar.

Comparing topic models

Each document in our corpus is about one or more topics. Do the topic models with the machine-translated text as input assign the same topics to a document as the topic models with the gold standard translated texts? Figure 5 displays for each language how similar topical prevalence is for each pair of gold standard and machine-translated documents (based on an equal number of topics; for the comparison between unequal number of topics, see the Appendix). These correlations denote the extent to which topical prevalence in individual gold standard and machine-translated documents overlaps. The higher the correlation the more overlap.[19] It shows that document-level topical prevalence is similar for gold standard and machine-translated corpora, with on average – across all languages – 65% of document pairs having a topic distribution correlation of 0.8 or higher. Put differently, a particular document is likely to be assigned to identical topics regardless of whether it was machine-translated or gold standard translated.

That said, there are statistically significant differences between languages (see Table 3).[20] Tables 3 (equal number of topics) break down mean topical prevalence for each language, as well as their standard deviations. The highest mean topic distribution per document pair is obtained for Spanish (0.83), and the lowest for French (0.75). Again, the absolute differences are small, and across languages it appears that topical prevalence at the level of individual documents is similar.

Each topic in our data is discussed in several documents. Are these the same documents in the topic models of the machine-translated text and the gold standard translations? To evaluate

[18] It would be very interesting to see if these unique features are actually caused by inaccurate – but in meaning similar – translations. However, due to the automated nature of all the analyses conducted here, this is not within the scope of the current paper.

[19] Aside from a very small number of documents with negative correlation in case of an unequal number of topics, the overall distributions of correlations in the comparison of unequal models and equal models are similar. This indicates that changing the number of topics towards the optimum for both datasets does not affect document topic distribution scores much.

[20] Note that t-values for all languages are significant regardless of comparing equal or unequal numbers of topics, so the statistical significance of differences cannot be attributed to a specific language.

Figure 5: Similarity of document-level topical prevalence with equal number of topics

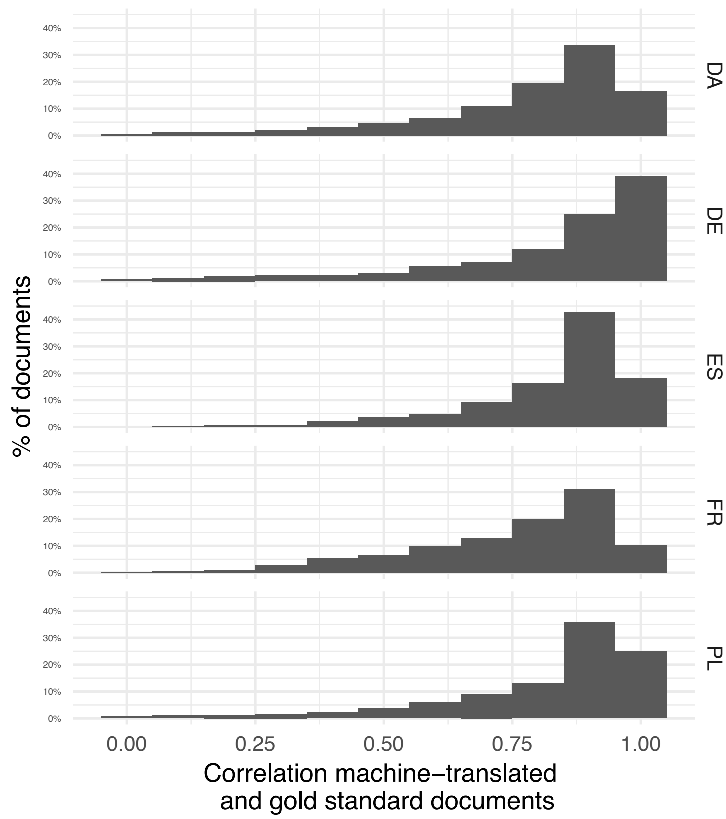


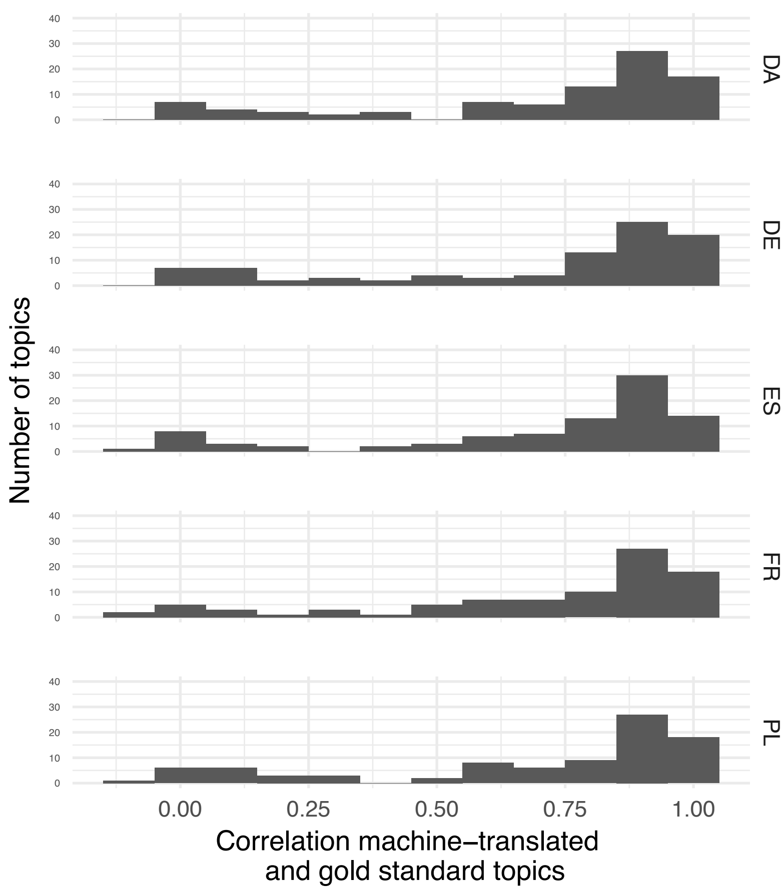
Table 3: Similarity of document-level topical prevalence with equal number of topics

Statistic	N	Mean	St. Dev.	Min	Max
Danish	2,301	0.783	0.202	−0.031	0.998
German	2,148	0.824	0.216	−0.031	0.999
Spanish	2,335	0.826	0.165	0.028	0.997
French	2,347	0.753	0.194	−0.043	0.996
Polish	2,338	0.809	0.206	−0.031	0.998
Total	11469	0.799	0.197	−0.043	0.999

Note: ANOVA results: $F(4, 11464) = 56.414$, $\rho < 0.001$, $\eta^2 = 0.019$

this we calculate the correlations between the topical prevalence of each topic in the gold standard and the machine-translated documents (Figure 6 show the results of 446 topic distribution comparisons.[21] As in the case of document-level topic distributions, these corpus-level correlations are generally quite similar, having a mean of 0.69. This indicates that on average topics are similarly distributed across all documents in the corpus. This indicates that a topic is likely to be distributed similarly across documents, regardless of whether these documents where machine-translations or gold standard translations of the same source.

Figure 6: Similarity of corpus-level topical prevalence with equal number of topics

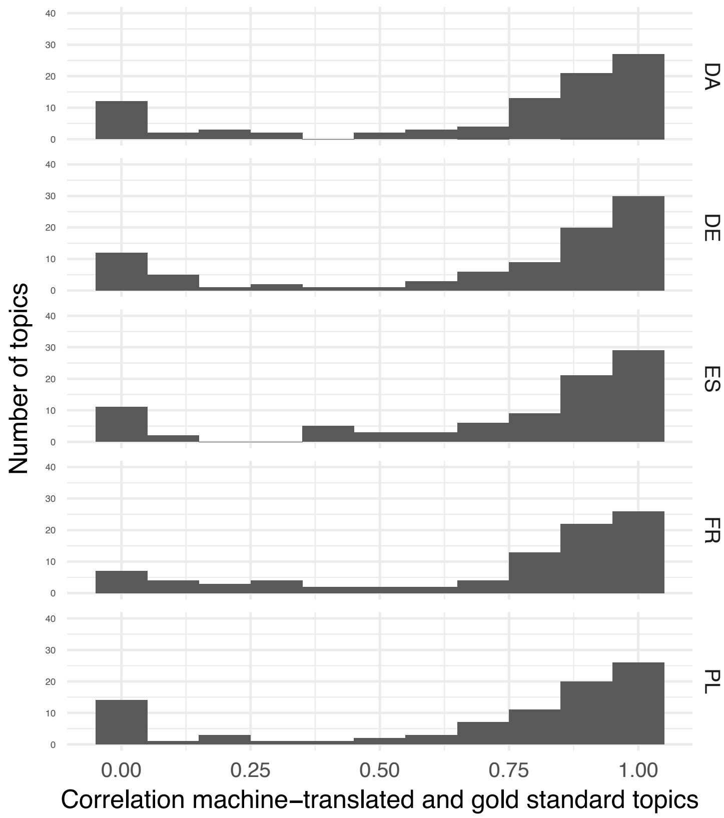


Overall descriptives: N=446, M=0.699, SD=0.321.

Finally, we also compare the similarity in the content of paired topics. To do so, we analyze for

[21] 90 comparisons for German, and 89 comparisons for Danish, Spanish, French and Polish, summing to 446.

Figure 7: Similarity of topical content with equal number of topics



Overall descriptives: N=446, M=0.708, SD=0.345

each topic pair the stem loadings of all shared features in the gold standard and machine-translated TDMs. The results are presented in Figure 7. Again, the average correlation is about 0.70 across languages indicating that topical content, as measured by the distribution of stem loadings, is similar for both the machine-translated and the gold standard corpora.[22] That implies that topics are discussed using the same terms in both the machine translation and gold standard translation documents.

[22]A question that remains is the spike in topic correlations on the low end of figures 6 and 7. The reason is that these are topic pairs that contain very few documents. As such these differences are unlikely to affect the topic model output of theoretical interest. Results discussed in the Appendix present evidence for this explanation.

Conclusion

The results in this paper support the claim that Google Translate is a useful tool for researchers using bag-of-words text models for comparative questions. We first found TDMs for machine translations and gold standard translations to be highly similar, with substantively small differences across languages. What is more, we found considerable overlap in the set of features (stems) generated from both corpora. With regards to LDA topic models, at both the document and the corpus levels we found topical prevalence to be generally similar with only small differences across languages. Furthermore, we found topical content to be highly similar.

Do our findings extend to other bag-of-words approaches such as position scaling or sentiment analysis? If a topic model with 90 models using machine-translated documents is highly similar to the topic model with the gold standard documents, we believe it to be very likely that a 2-dimensional or 3-dimensional scaling model can be similarly reproduced. In addition, for sentiment analysis translation is already used. Sentiment dictionaries are sometimes translated from English to other languages without validation. This is problematic since the specific meaning of words is more relevant. Some words may be translated in such a way that they lack emotional content, while other words may gain emotional content in translation. As long as these translation issues are random, the problem of the identification of false positives or false negatives is reduced when sentiment scores are aggregated over entire documents. Then again, we do not quite know whether these translation issues are random or not. We leave these issues for future work.

References

Agarwal, Apoorv, Boyi Xie, Ilia Vovsha, Owen Rambow and Rebecca Passonneau. 2011. Sentiment analysis of twitter data. In *Proceedings of the workshop on languages in social media*. Association for Computational Linguistics pp. 30–38.

Aharoni, Roee. 2015. "Automatic detection of machine translated text and translation quality estimation." PhD thesis, Department of Computer Science, Bar-Ilan University Ramat Gan, Israel.

Balahur, Alexandra and Marco Turchi. 2014. "Comparative experiments using supervised learning and machine translation for multilingual sentiment analysis." *Computer Speech & Language* 28(1):56–75.

Benoit, Kenneth and Paul Nulty. 2013. "Quanteda: quantitative analysis of textual data." `http://quanteda.io`. An R library for Managing and Analyzing Text.

Bird, S., E. Klein and E. Loper. 2009. *Natural language processing with Python*. O'Reilly Media, Inc.

Blei, David M, Andrew Y Ng and Michael I Jordan. 2003. "Latent dirichlet allocation." *Journal of Machine Learning Research* 3(1):993–1022.

Chuang, Jason, Sonal Gupta, Christopher Manning and Jeffrey Heer. 2013. Topic model diagnostics: assessing domain relevance via topical alignment. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*. pp. 612–620.

Courtney, Michael, Michael Breen, Iain McMenamin and Gemma McNulty. 2017. "Deductive validation of machine translation for text analysis in comparative politics." Working Paper, Dublin City University.

De Vries, Erik, Martijn Schoonvelde and Gijs Schumacher. 2018. "Replication Data for: No Longer Lost in Translation: Evidence that Google Translate Works for Comparative Bag-of-Words Text Applications." `http://dx.doi.org/10.7910/DVN/VKMY6N`.

Denny, Matthew James and Arthur Spirling. 2016. "Assessing the consequences of text preprocessing decisions." *Available at SSRN 2849145* .

Gray, Russell D and Quentin D Atkinson. 2003. "Language-tree divergence times support the Anatolian theory of Indo-European origin." *Nature* 426:435–439.

Greene, Zac, Andrea Ceron, Gijs Schumacher and Zoltán Fazekas. 2016. "The nuts and bolts of automated text analysis. Comparing different document pre-processing techniques in four countries." Open Science Framework.

Grimmer, Justin and Brandon M Stewart. 2013. "Text as data: the promise and pitfalls of automatic content analysis methods for political texts." *Political Analysis* 21(3):267–297.

Hampshire, Stephen and Carmen Porta Salvia. 2010. "Translation and the internet: evaluating the quality of free online machine translators." *Quaderns: revista de traducció* 17:197–209.

Hornik, Kurt and Bettina Grün. 2011. "Topicmodels: An R package for fitting topic models." *Journal of Statistical Software* 40(13):1–30.

Kaljahi, Zadeh and Rasoul Samad. 2015. "The role of syntax and semantics in machine translation and quality estimation of machine-translated user-generated content." PhD thesis, Dublin City University.

Koehn, Philipp. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT summit*. Vol. 5 pp. 79–86.

Koehn, Philipp and Christof Monz. 2006. Manual and automatic evaluation of machine translation between european languages. In *Proceedings of the Workshop on Statistical Machine Translation*. Association for Computational Linguistics pp. 102–121.

Loaiciga, Sharid, Thomas Meyer and Andrei Popescu-Belis. 2014. English-French verb phrase alignment in Europarl for tense translation modeling. In *LREC*. pp. 674–681.

Lotz, Susan and Alta Van Rensburg. 2014. "Translation technology explored: Has a three-year maturation period done Google Translate any good?" *Stellenbosch Papers in Linguistics Plus* 43:235–259.

Lucas, Christopher, Richard A Nielsen, Margaret E Roberts, Brandon M Stewart, Alex Storer and Dustin Tingley. 2015. "Computer-assisted text analysis for comparative politics." *Political Analysis* 23(2):254–277.

Popescu-Belis, Andrei, Thomas Meyer, Jeevanthi Liyanapathirana, Bruno Cartoni and Sandrine Zufferey. 2012. Discourse-level annotation over europarl for machine translation: connectives and pronouns. In *Proceedings of the eighth international conference on Language Resources and Evaluation (LREC)*. Number EPFL-CONF-192582.

Roberts, Margaret E, Brandon M Stewart and Edoardo M Airoldi. 2016. "A model of text for experimentation in the social sciences." *Journal of the American Statistical Association* 111(515):988–1003.

Scarton, Carolina and Lucia Specia. 2014. Document-level translation quality estimation: exploring discourse and pseudo-references. In *The 17th Annual Conference of the European Association for Machine Translation*. pp. 101–108.

Schwarz, Daniel, Denise Traber and Kenneth Benoit. 2017. "Estimating Intra-Party Preferences: Comparing Speeches to Votes." *Political Science Research and Methods* 5(2):379–396.

Online Appendix for: No Longer Lost in Translation: Evidence that Google Translate Works for Comparative Bag-of-Words Text Applications

Erik de Vries [*] Martijn Schoonvelde[†] Gijs Schumacher [‡]

A An Example of the Topic Matching Procedure

As an example of our topic matching procedure, consider Table A.1. It shows the 10 highest loading words for 5 matching topics in the French gold standard and machine-translated models. The correlation score reported at the bottom indicates to what extent the word stem loadings match between the matched gold standard and machine-translated topics. As one can see, most of the topic pairs are highly similar, and can be interpreted as being similar. For example topic pair 3-25 can be interpreted as concerning the possible admission of Turkey to the EU, and enlargement of the EU in general. Similarly, topic pair 4-70 can be interpreted as a topic about procedure in the European Parliament, but not about any societal topic. In contrast, topic pair 5-23 are an obvious mismatch, with only the stems "totalitarian" and "crime" linking them (summarized by the low correlation of stem loadings).

In addition, Table A.2 shows excerpts from two documents in both the gold standard and machine-translated French dataset. These excerpts show for topic pairs 2-58 and 3-25 the extensive similarity between the gold standard and machine-translated documents. The bold text indicates the most important words for that specific topic, and coincides with the contents of Table A.1. Document similarity shows the cosine similarity scores of the gold standard and machine-translated TDMs for these specific documents. It indicates to what extent the documents consist of the same word stems.

[*]PhD student, Department of Media and Social Sciences, University of Stavanger
[†]Postdoctoral Researcher, Department of Political Science and Public Administration, Vrije Universiteit
[‡]Assistant Professor, Department of Political Science, University of Amsterdam

Table A.1: Topic matching example from the French dataset

Topic no. in gold standard model	1	2	3	4	5
Most important (highest loading) words per topic	polit	health	turkey	mr	european
	elect	diseas	access	vote	today
	democrat	patient	countri	presid	histori
	democraci	healthcar	croatia	amend	europ
	countri	care	negoti	group	year
	govern	treatment	progress	rule	totalitarian
	parti	prevent	reform	would	parliament
	support	cancer	turkish	resolut	crime
	presid	servic	enlarg	ask	peopl
	peopl	medic	process	procedur	symbol
Topic no. in machine-translated model	64	58	25	70	23
Most important (highest loading) words per topic	countri	health	turkey	vote	cuba
	presid	diseas	access	mr	crime
	govern	prevent	croatia	presid	victim
	polit	cancer	negoti	amend	totalitarian
	peopl	vaccin	countri	group	regim
	right	care	progress	would	communist
	situat	peopl	reform	parliament	cuban
	elect	treatment	turkish	ask	histori
	author	research	enlarg	paragraph	memori
	human	fight	process	propos	communism
Correlation of stem loadings within the topic pair	0.75	0.88	0.97	0.95	0.50

Table A.2: Comparison between gold standard and machine-translated texts

	Gold standard excerpt	Machine-translated excerpt	Document similarity
Topic 2-58	But even though screening is important, I think that Community action against **cancer** must cover a wider range of topics. For example: **health** information and data on the **cancer** burden that will highlight inequalities and best practices across Europe; **prevent**ative measures and **health** promotion on topics such as tobacco, nutrition and alcohol; best practices on **treatment** and integrated **cancer care**, such as palliative **care**; bringing together expertise through European reference networks; providing investment in infrastructure through the Structural Funds; and support for **cancer** research at Community level.	As important as screening, I think that Community action against **cancer** must cover a wider area. For example: **health** information and data on the **cancer** burden that will highlight inequalities and best practices across Europe; **prevent**ative measures and **health** promotion on topics such as smoking, diet and alcohol; best practices on **treatment** and integrated **care**, such as palliative **care**; the gathering of knowledge and skills on the European reference networks; infrastructure investments through the Structural Funds; and support research against **cancer** at the community level.	0.974
Topic 3-25	In the meantime, I would like to briefly mention a few points in this phase of **Turkey**'s **access**ion **negoti**ations. We are of the opinion that the recent elections in **Turkey** demonstrated the wish of the **Turkish** people for democracy, stability - both political and economic - and **progress**. We also welcome how the elections were conducted, the high voter turnout and the improved representativeness of the new **Turkish** Parliament. The Presidency shares the views and concerns of this House regarding **Turkey**'s **reform process**. We believe that the new Government enjoys increased legitimacy and a clear mandate that should enable decisive steps to be made in advancing and broadening the **reform process** in **Turkey**.	Meanwhile, let me briefly address a few points at this stage of the **access**ion **negoti**ations with **Turkey**. The recent elections in **Turkey**, we believe, demonstrated the desire for democracy, stability - both political and economic - and **progress** of the **Turkish** population. We also welcome the way in which these elections were held, the high rate of participation and better representation of the new **Turkish** Parliament. The Presidency shares the opinion and concerns of this House regarding **Turkey**'s **reform process**. We believe that the new Government enjoys increased legitimacy and a clear mandate, which should achieve breakthroughs in terms of **progress**ion and expansion of the **reform process** in **Turkey**.	0.986

Note: The topic numbers represent topics in the gold standard French dataset. Words printed in bold are of high importance to the topic (see table A.1)

B Figures of topic model output with unequal number of topics

Figure B.1: Similarity of document-level topical prevalence with unequal number of topics

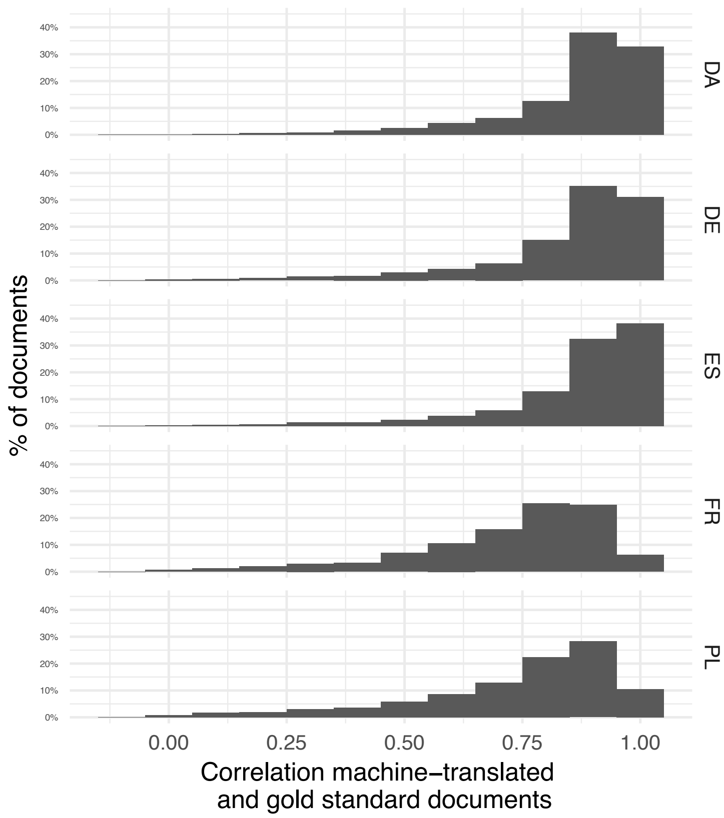
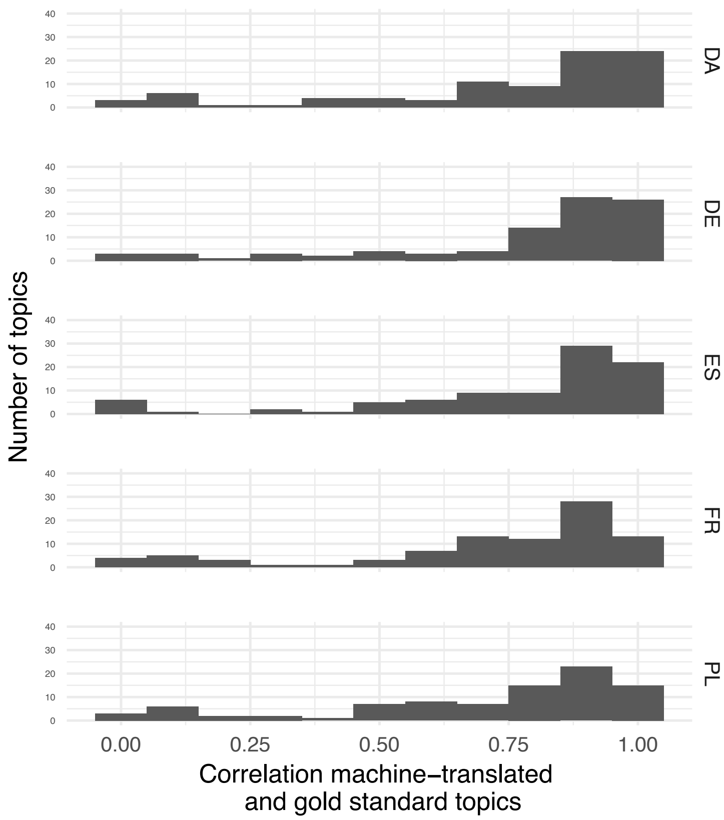
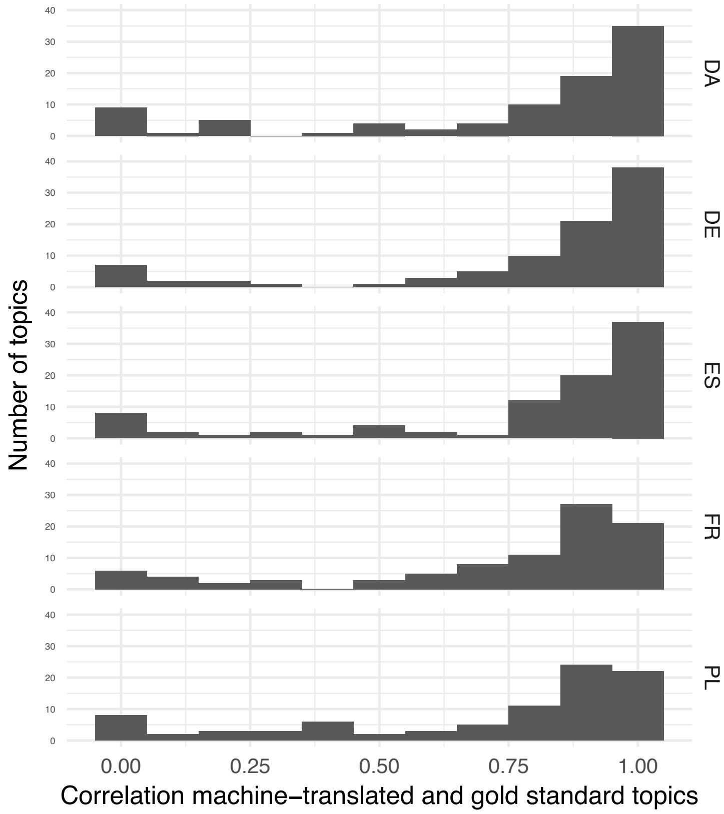


Figure B.2: Similarity of corpus-level topical prevalence with unequal number of topics



Overall descriptives: N=449, M=0.740, SD=0.280

Figure B.3: Similarity of topical content with unequal number of topics



Overall descriptives: N=449, M=0.747, SD=0.315

C Tables of topic model output with unequal number of topics

Table C.3: Similarity of corpus-level topical prevalence with unequal number of topics

Statistic	N	Mean	St. Dev.	Min	Max
DA	2,301	0.859	0.161	−0.039	0.998
DE	2,148	0.842	0.181	−0.051	0.998
ES	2,335	0.860	0.168	−0.030	0.998
FR	2,347	0.727	0.201	−0.047	0.998
PL	2,338	0.740	0.216	−0.035	0.994
Total	11469	0.806	0.185	−0.051	0.998

Note: ANOVA results: $F(4, 11464) = 294$, $\rho < 0.001$, $\eta^2 = 0.093$

D Analysis of poorly matching topic pairs

Why is there a spike in topic correlations on the low end of figures 6 and 7? And why does this spike appear in the topic-level topic comparisons but not so much in the document-level topic comparisons? One explanation can be found in figures D.1 and D.2, which show for topic pairs with a correlation of less than 0.70 how much these topics are on average present in documents (range 0-1) for both gold standard and machine-translated models. In addition, the expected proportion of topic pairs with a correlation below 0.70 is also plotted, assuming that all topics have on average an equal share in documents. The most notable difference between the plots for models with an equal and unequal number of topics is that the average expected proportion of these topics in documents is lower with an unequal number of topics. This is explained by the fact that with different numbers of topics, matches between topics are made more easily, as at least 10 of the topics from the machine-translated model are dropped by design. Furthermore, it shows that in general the proportion of topic pairs with a correlation below 0.70 decreases.

This figure show that, generally, there is a large difference between the observed and expected proportion of these topics in documents, implying that topic pairs with relatively low correlation are not commonly present in documents, and as such not so much relevant for estimating the topic models. One result that deviates from this interpretation concerns the relatively small difference between the observed and expected topic proportions for French machine-translated texts in the comparison of models with an equal number of topics. However, this difference becomes larger, and more in line with the observations for other languages, when looking at the comparison of models with an unequal number of topics. This is also evidence that supports the assumption that when using machine-translated text in topic models, choosing the optimum number of topics based on the actual data is the way to go.

Figure D.1: Average proportion of topics with correlation < 0.70 in documents (equal number of topics)

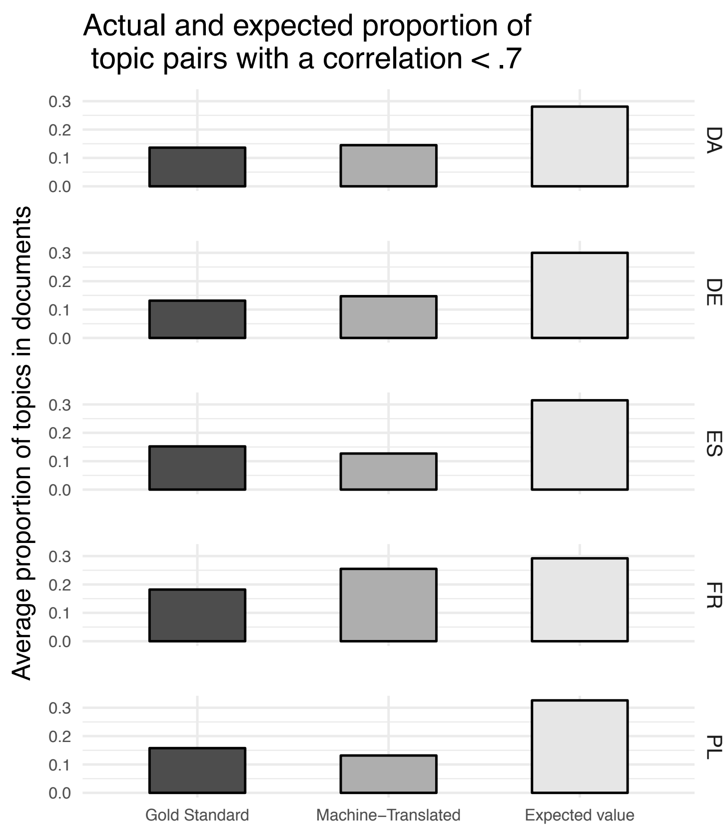


Figure D.2: Average proportion of topics with correlation <.7 in documents (unequal number of topics)