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

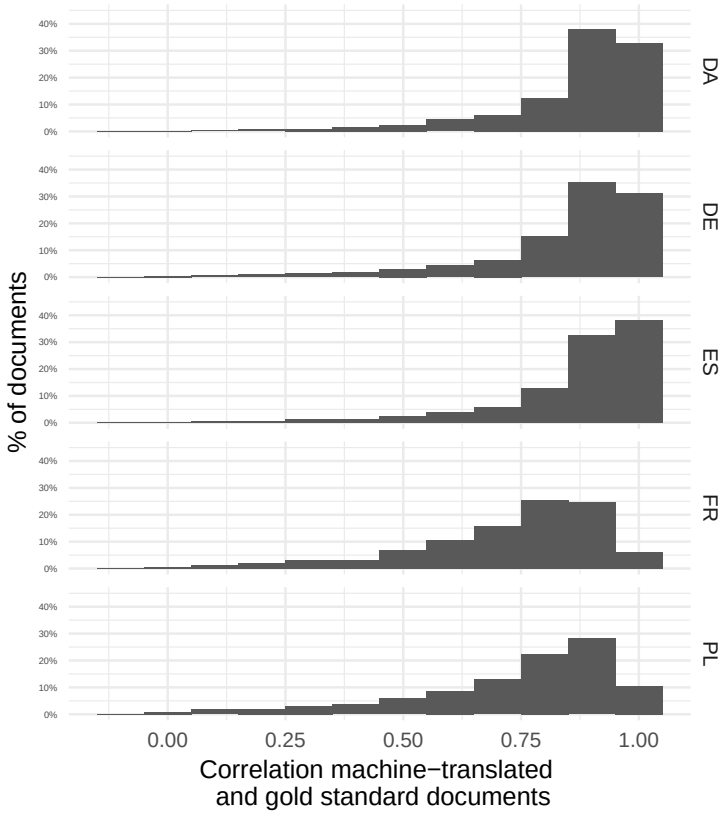
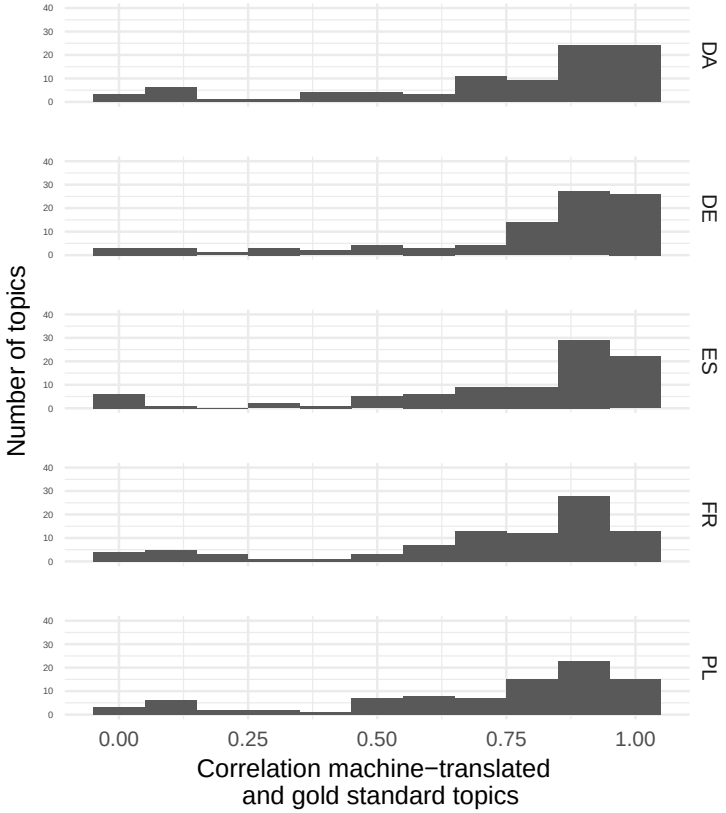
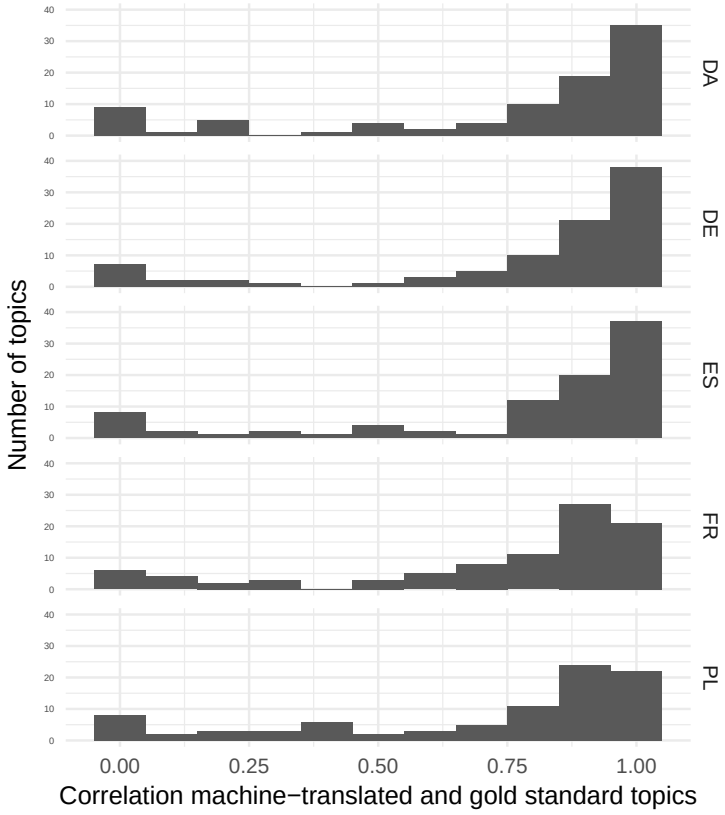


Figure B.2: Similarity of corpus-level topical prevalence with unequal number of topics



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)

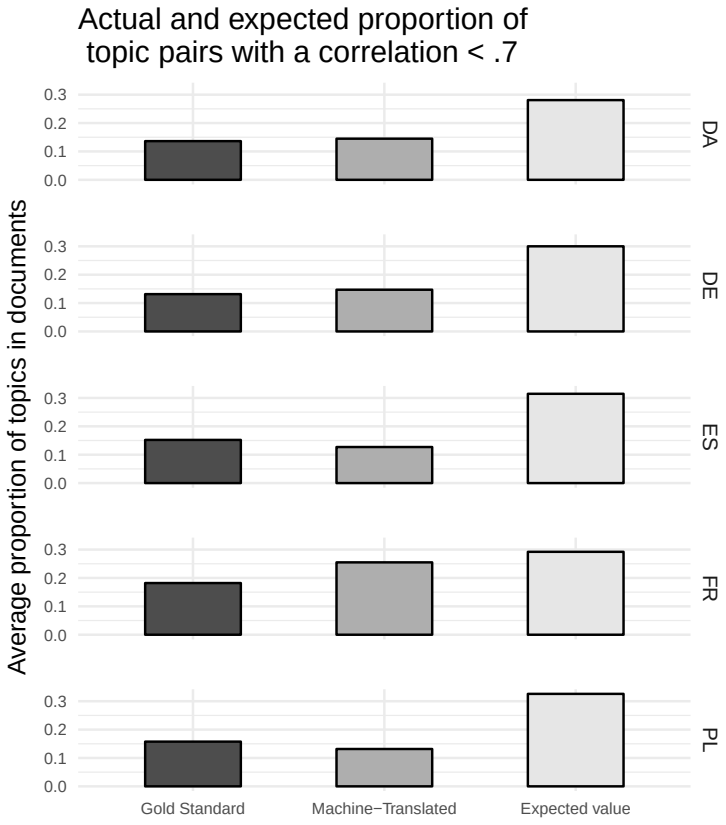


Figure D.2: Average proportion of topics with correlation <.7 in documents (unequal number of topics)