Unpacking Translation Effects

Influences of Target Language Choice on Topic Modeling in Multilingual Environments

Nadezhda Ozornina / Mario Haim*

Machine translation is widely used in communication science to consolidate texts, not least for exploratory clustering approaches such as multilingual topic modeling. However, the impact of target language choice on topic modeling results remains unclear. This study examines these effects by (a) consolidating texts into one of the original document languages and (b) translating texts into an intermediary language not present in the dataset under study. To assess the effects, we use a corpus of parallel United Nations texts in Russian and German (N = 3,760). We compare the results of structural topic modeling after translating Russian texts into German, chosen as the original language, with consolidating the entire corpus into English as an intermediary language. The translation approaches are compared based on feature overlap, topical prevalence, and topical content. The findings show that intermediary-language translation yields a more symmetrical topic distribution and higher overlap in top words, but significantly reduces vocabulary size compared to consolidation into the original language. The results are replicated using a second bilingual journalistic corpus (N = 434) and validated across different numbers of topics. Finally, we discuss best practices for target language selection in multilingual topic modeling and situate them within the context of recent developments in computational communication science.

Key words: topic modeling, machine translation, multilingual text analysis, German, Russian, methods, computational communication science, computational social science

1. Introduction

With the increasing volume of journalistic and social media data, exploratory automated text analysis methods such as topic modeling, have gained prominence in communication science (Günther, 2022; van Atteveldt et al., 2022). Unsupervised approaches to clustering textual content have become especially valuable for studying multilingual contexts, as discussions on digital platforms increasingly take place in multiple languages simultaneously (Hase et al., 2021; Maier et al., 2022). In such cases, the ability to identify cross-case topics is crucial for gaining a comprehensive understanding of the issues under study and for facilitating quantitative comparison between cases (Lind et al., 2022). However, because topic modeling algorithms are typically designed for monolingual text collections, they often struggle to detect similar meanings across languages due to vocabulary misalignment (Chan et al., 2020). This poses a significant challenge for comparative analyses of large-scale textual data and cross-case comparability.

To enable the application of topic modeling in multilingual contexts, machine translation techniques combined with probabilistic topic modeling are commonly employed. Previous studies in communication science have demonstrated the ability of this approach to

* Nadezhda Ozornina, M.A., Ludwig-Maximilians-Universität München, Department of Media and Communication, Akademiestr. 7, 80799 München, Germany, Nadezhda.Ozornina@ifkw.lmu.de, https://orcid.org/0000-0002-8637-9394
Prof. Dr. Mario Haim, Ludwig-Maximilians-Universität München, Department of Media and Communication, Akademiestr. 7, 80799 München, Germany, haim@ifkw.lmu.de, https://orcid.org/0000-0002-0643-2299.

produce valid outputs for exploratory analyses when applied to texts in multiple languages (de Vries et al., 2018) and when using different translation tools (Reber, 2019; Licht et al., 2024). However, it remains unclear whether the choice of target language for machine translation impacts the outcomes of multilingual topic modeling and to what extent such effects are robust across different types of texts (Lucas et al., 2015; Maier et al., 2022). Understanding these potential effects is crucial for assessing the potential and limitations of machine translation for topic modeling and for developing best practices to ensure validity in multilingual exploratory text analysis.

This study investigates the impact of target language choice on topic modeling outcomes in multilingual text collections by comparing two strategies for target language selection. Specifically, it examines (a) consolidation into one of the original languages already present in the dataset and (b) translation into an intermediary language not included in the original text corpus. In doing so, our findings aim to advance methodological discussions on validation in automated text analysis (Baden et al., 2022) and to offer recommendations in the context of recent developments in machine translation and computational methods.

To provide a case study based on different language groups and media systems, the research employs Russian and German versions of official United Nations documents (N = 3,760). We apply structural topic modeling (STM, Roberts et al., 2019) after consolidating the texts into German, chosen as the original language, and compare the results with topic models based on translations into English as intermediary language. Drawing on previous evaluation strategies (de Vries et al., 2018; Reber, 2019), we assess the impact of translation on topic modeling results in three ways: feature overlap in document-feature matrices (DFMs), topical prevalence, and topical content. To ensure the broader applicability of our findings to communication science and to examine their robustness in more linguistically diverse settings, we replicate the analysis using a second, substantially different, bilingual journalistic corpus (N = 434). For both text types, the results are further validated across different numbers of topics.

2. Theoretical Background

This section provides background on multilingual topic modeling and the role of machine translation in enhancing its applicability across languages. Building on the discussion of how target language choice shapes topic modeling outcomes, we then develop the research questions.

2.1 Multilingual Topic Modeling

As an approach to unsupervised text analysis, topic modeling aims to uncover structures in large text collections by analyzing patterns of word co-occurrences (Blei, 2012; Jacobi et al., 2016). In standard probabilistic approaches, a generative model uses the words of a textual corpus to group documents into a predefined number of clusters while maximizing inter-cluster differentiation. These clusters have been shown to resemble latent topics consisting of words and documents. This probabilistic clustering is optimized using two matrices: one representing the distribution of topics across individual documents, indicating each topic's prevalence within a document, and the other representing the distribution of words across topics, indicating each word's prevalence within a topic (e.g., Haim, 2023; van Atteveldt et al., 2022). More recently, embedding-based approaches using models explicitly trained on multilingual corpora, such as BERTopic (Grootendorst, 2022), have also been employed to explore textual data by clustering feature-vector representations of multilingual content.

In communication science, topic modeling's ability to represent documents as mixtures of latent topics and to characterize these topics through words makes it a useful tool for exploring concepts such as frames, issues, and writing styles in large-scale textual data (Günther, 2022). Given the growing volume of multilingual content in digital environments, applying topic modeling algorithms to linguistically diverse contexts is becoming increasingly important. In such analyses, identifying cross-case topics provides a deeper understanding of the concepts under study and facilitates quantitative comparisons across different countries and research settings (Lucas et al., 2015; Hase et al., 2021).

Despite the growing relevance of case comparisons in multilingual text analysis (Lind et al., 2022), it remains unclear how current approaches to topic modeling can be validly applied to linguistically diverse settings. This challenge is particularly pronounced for probabilistic models, in which the identification of cross-case topics is hindered by the so-called "Babel problem" (Chan et al., 2020). This problem refers to the fact that similar concepts are represented by different words in different languages, which bag-of-words algorithms originally designed for monolingual data are unable to reconcile. Consequently, when probabilistic topic modeling is applied to multilingual data without prior consolidation, documents are likely to be grouped by language rather than by content, hindering the identification of topics across cases in multilingual text collections (for an example, see Lind et al., 2022). Applying monolingual embeddings to multilingual data poses related challenges, as vector representations trained on a single language cannot be reliably transferred to other linguistic contexts.

2.2 Mitigating Bias Through Machine Translation

One common approach to facilitate cross-case comparison for multilingual data is the use of machine translation algorithms (Lucas et al., 2015). In this process, the collected text corpus is first divided into subcorpora corresponding to the original languages, and each subcorpus is then consolidated into the chosen target language before further analysis. Machine translation is typically performed using tools such as DeepL or Google Translate (Reber, 2019), and the resulting translations are subsequently used as input for probabilistic topic modeling.

Nowadays, machine translation is a widely applied strategy that offers both advantages and disadvantages compared to other consolidation approaches, such as manual translation, multilingual dictionaries, or pre-training based on multilingual documents (see overviews in Lind et al., 2022 and Maier et al., 2022). In particular, it enables more comprehensive analyses than techniques based on multilingual dictionaries, which may be limited in capturing the full range of relevant meanings (Maier et al., 2022). Additionally, machine translation provides a faster and more cost-effective solution, requiring fewer resources than custom training of multilingual models based on parallel data (Lind et al., 2022). However, the reproducibility of machine translation is constrained by the "black-box" nature of existing algorithms, which are constantly updated (Chan et al., 2020), making researchers dependent on third-party tools and platforms. Moreover, translation costs can increase significantly with the volume of text, making it an expensive option for consolidating large multilingual corpora (Lucas et al., 2015; Reber, 2019). One recent approach to overcoming these limitations is the use of open-source translation tools such as OPUS-MT, which offer a more affordable and reproducible alternative to commercial providers (Licht et al., 2024).

With the advent of large language models, clustering based on multilingual embeddings has been explored as an alternative to combining machine translation with probabilistic topic modeling. Recent studies (e.g., Licht, 2023; Licht & Lind, 2023) demonstrate the capability of embedding models such as multilingual BERT to produce high-quality cross-

lingual outputs, making them an attractive strategy for aligning multilingual data without additional translation efforts. In this approach, the translation effort can be effectively frontloaded and integrated into large language models that are explicitly trained on multiple languages. However, unlike probabilistic topic modeling, multilingual embeddings may rely on partially transparent and comprehensible models, thus limiting researcher control, interpretability, and validation of produced outputs (Licht & Lind, 2023; Rinke et al., 2022).

In our study, we focus on the broader applicability of machine translation for probabilistic topic modeling. Here, key concerns relate to the quality of machine translation as input for automated text analysis. First, results may contain systematic errors in the translation of frequently occurring words, which can alter the representation of topics in the original data and assign the same concepts from different languages to separate topics (Lucas et al., 2015). Despite recent improvements in translation system quality (Chan et al., 2020), such errors still warrant consideration in automated analyses. Second, research in computational linguistics shows that, as languages vary in vocabulary richness and linguistic structure, machine translation may oversimplify the meaning of the original texts, leading to vocabulary loss, overgeneralization, and the omission of language-specific meanings and concepts across different language pairs (Kotait, 2024; Vanmassenhove et al., 2019). Together with systematic errors, these issues risk of the intended topical structure being "lost in translation" (de Vries et al., 2018), limiting the extent to which translated texts reflect the distribution and content of topics in the original corpus. Further investigation is needed to determine how prevalent these issues are when applying current translation systems for text consolidation.

2.3 Influences of Target Language Choice

Previous studies have examined the applicability of machine translation in various research settings, comparing texts written in different languages (de Vries et al., 2018), evaluating different translation tools (Reber, 2019; Licht et al., 2024), and assessing translation outcomes at the level of full texts versus DFMs (e.g., Lucas et al., 2015). Overall, findings suggest that machine translation provides valid inputs for multilingual topic modeling, with only minor translation errors and slight changes in topic prevalence and top words (for an overview, see Appendix A in supplementary materials at https://osf.io/qfch3). However, studies also indicate that translation quality may vary depending on the strategies employed and may perform differently across datasets and research settings (Maier et al., 2022; Reber, 2019), an aspect that has not been thoroughly explored.

Despite the general effectiveness of machine translation for topic modeling, there is a suggestion that translation quality issues may vary depending on the selected target language for text consolidation (Wang et al., 2022). Therefore, the choice of language into which texts are translated may influence the topic modeling outcomes (Lucas et al., 2015). However, the potential impact of target language choice has not yet been empirically evaluated, and English is commonly used for translation without consideration of possible effects on the meanings encoded in the original texts. Given English's status as a "lingua franca" in communication science (Baden et al., 2022; Lind et al., 2022), it is crucial to understand the limitations its usage may impose, particularly when consolidating texts from different languages. Moreover, the consequences of selecting alternative target languages for machine translation remain unclear in current research (Licht et al., 2024).

Building on previous discussions of target language choice (Koltsova & Pashakhin, 2020; Lucas et al., 2015), this study proposes two approaches to machine translation in multilingual text analysis, based on its relationship between the target language and the

languages present in the analyzed corpus. These strategies are illustrated using corpora consisting of texts in two languages.

The first strategy involves consolidating the texts into one of the two languages already present in the corpus, referred to as the *original language*. For example, if the dataset includes Russian and Ukrainian publications, researchers may select one of the corpus languages as the basis for text consolidation and translate the other language into it—for instance, translating the Ukrainian texts into Russian (Koltsova & Pashakhin, 2020). In this approach, translation is applied to only one part of the corpus, conserving translation resources and minimizing the loss of the original vocabulary. At the same time, this strategy introduces varying degrees of transformation across subcorpora, potentially compromising methodological equivalence between cases at the input stage of analysis (Baden et al., 2022; Licht & Lind, 2023). Because some portions of the corpus are translated while others remain unchanged, there is a risk of systematic bias favoring the untranslated texts in the resulting topic model (see Lucas et al., 2015). However, the effect of such "unequal" data consolidation on topic modeling outcomes has not yet been empirically evaluated.

An alternative strategy involves translating all documents into an *intermediary language* that is not present in the original dataset and serves as an external consolidation basis for multilingual text analysis. An example of this approach is provided by Lucas et al. (2015), who translated social media posts from Arabic and Chinese into English to facilitate a "symmetrical" transformation of both subcorpora. This consolidation strategy ensures that both sets of texts undergo an equal degree of machine translation, enabling cross-case equivalence at the input stage for topic modeling. While this approach mitigates the scenario in which one subcorpus remains untranslated, it requires more translation resources and results in a greater loss of the original text vocabulary. In addition, the outcomes of this strategy for different language pairs may systematically differ in quality due to varying degrees of linguistic similarity among the languages involved (Wang et al., 2022).

For well-resourced language pairs, such as English and German, machine translation quality may be higher than for less closely related languages, where original texts may undergo greater simplification and translations may contain more errors due to difficulties in conveying meaning across differences in morphology, scripts, and sentence structure. Recent experiments in computational linguistics illustrate this loss of linguistic richness, for example when translating Arabic texts into English (Kotait, 2024). Consequently, the need to force original texts "into the corset of English-like language structure" (Baden et al., 2022) highlights the limitations of using English as the target language in comparative settings and underscores the importance of a more detailed examination of the influence of target language choice on topic modeling.

Overall, the impact of target language choice on topic modeling results remains unclear. Although some considerations in communication science suggest that using an intermediary language may provide a more symmetrical transformation, whereas consolidating into one of the original languages may better preserve vocabulary, these assumptions have not been sufficiently examined. Therefore, we pose the following research question:

RQ1: How do topic modeling results of a multilingual corpus differ when (a) consolidating into one of the original languages versus (b) translating into an intermediary language?

2.4 Replicability Across Text Types

In addition to examining different translation strategies, there is a need to evaluate results across various settings, an aspect that has been rarely explored in previous studies on the

impact of machine translation on topic modeling. A particularly important consideration is the replicability of results across text types with varying linguistic variability, which can significantly influence translation quality and, consequently, topic modeling outcomes (Maier et al., 2022). Current methodological research on multilingual topic modeling is often conducted using official documents (e.g., de Vries et al., 2018), leaving it unclear whether these findings generalize to other text types commonly used in communication science. One such type is journalistic publications, which tend to be more diverse and metaphorical in language compared to official documents, making them more susceptible to translation errors (Maier et al., 2022). This leads to the second research question:

RQ2: How does machine translation impact topic modeling results when applied to official documents versus journalistic publications?

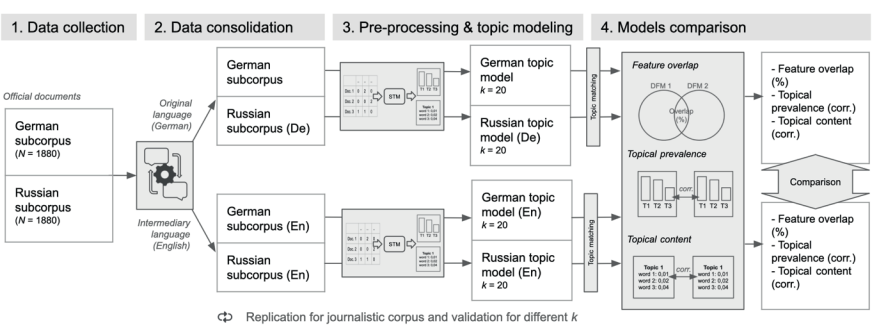
3. Methodology

In the following section, we describe the study design and methodological procedures applied. Reproducible scripts for each step are available in the supplementary materials of the publication.

3.1 Study Design and Case Selection

To investigate the impact of machine translation, we calculate and compare topic models after translating texts into different target languages across multiple text types (see Figure 1). For this study, we focus on official documents and journalistic publications written in Russian and German, allowing us to examine potential translation issues across different language branches (Fortson, 2011). This language selection also ensures sufficient data availability for both text types and serves as a prototypical case study for comparative research, representing differences in journalistic cultures and media systems (Hallin & Mancini, 2011; Hanitzsch et al., 2019). We illustrate consolidation into an original language by translating Russian texts into German and implement the intermediary language strategy by consolidating all texts into English.

Figure 1: Study Design



3.2 Data Collection

To assess the influence of target language choice on topic modeling results, we use a collection of openly available parallel texts from United Nations official documents (Eisele & Chen, 2010) in both Russian and German (N = 3,760), published as the part of the OPUS project (Tiedemann, 2012). Although no longer fully up to date, this corpus contains a sufficient number of UN resolutions in multiple languages and has proven its usefulness in methodological studies on multilingual text analysis (e.g., Windsor et al., 2019).

To enable comparison across text types, we replicate the UN corpus analysis using a corpus of journalistic publications (N = 434). These texts were obtained through web scraping of content on culture, lifestyle, and politics from various Russian media sources, supplemented with professional translations into German (for replication results, see Appendix C). Unlike UN publications, these journalistic texts are less standardized, incorporate more language-specific terms, and include much more recent data. For each text type, the corpus of parallel publications is divided into subcorpora written in Russian and German.

3.3 Machine Translation

For UN publications, text consolidation was performed using Google Translate, as it required fewer resources while providing quality comparable to other translation tools (Reber, 2019). The translation was carried out in May 2023 using the browser-based document translator. To implement consolidation into the original language, Russian UN documents were translated into German, while the German texts remained unchanged. For intermediary language strategy, both Russian and German texts were translated into English. The same procedure was applied to the second corpus of journalistic articles.

3.4 Pre-processing

As a result of applying both translation strategies to the UN publications, we obtained four subcorpora: original German, Russian translated into German, German translated into English, and Russian translated into English. Each subcorpus contained 1,880 documents and was preprocessed by removing URLs, numbers, punctuation marks, symbols, and language-specific stopwords with *spacyr* (Benoit & Matsuo, 2017). These steps are standard practice in topic modeling pre-processing and help to focus on meaningful features while removing noise for further analysis (Maier et al., 2018; van Atteveldt et al., 2022).

Next, we performed lemmatization using language-specific models chosen for their quality and computational efficiency (*en_core_web_sm* and *de_core_news_sm*). The outcomes of this step were evaluated using the type–token ratio (*TTR*) before and after lemmatization. *TTR* is a widely used linguistic metric for measuring lexical diversity, calculated as the ratio of unique tokens to the total number of tokens, with values closer to one indicating greater language variability (see Kettunen, 2014). We observed a similar reduction in *TTR* across the subcorpora, with German showing higher mean *TTR* values than English (after lemmatization: 0.549 vs. 0.444), consistent with findings from previous studies (Bentz & Kiela, 2014). These results indicate that the lemmatization process was appropriate for both languages, producing neither too many nor too few tokens relative to the original content.

Then, we applied vocabulary pruning of lemmatized data using common thresholds, removing terms that appeared in less than 0.5 % or more than 99 % of documents (Reber, 2019). At this stage, the four resulting subcorpora (original German, Russian translated into German, German translated into English, and Russian translated into English) were prepared for analysis and converted into separate DFMs containing between 4,373 and 8,397 features.

The same pre-processing steps were applied to the journalistic publications, resulting in four DFMs, each containing 217 documents (N features: 6,611–9,989). We also performed the same lemmatization checks using *TTR*, which produced results similar to those observed for the UN documents (mean *TTR* after lemmatization: 0.737 for German and 0.635 for English). These results further confirmed the higher linguistic variability of journalistic data compared to UN publications.

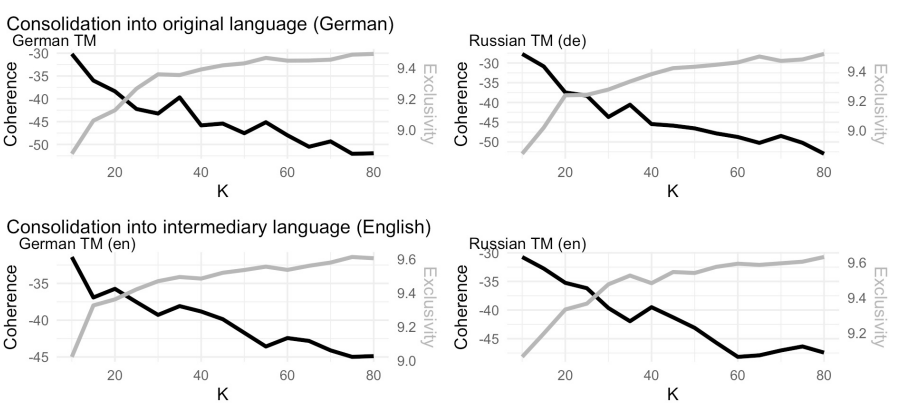
3.5 Topic Modeling

In the next step, we applied Structural Topic Modeling (STM; Roberts et al., 2019). STM has become a widely used technique for topic modeling in the social sciences, as it allows for the inclusion of covariates and is readily accessible through an *R* package. For each DFM derived from the UN corpus, we computed topic models with topic numbers ranging from 10 to 80 with incremens of 5. For each topic solution under each consolidation strategy, we automatically matched topics from the texts originally written in German and Russian using the method proposed by de Vries et al. (2018). In this process, each word was considered individually, and the topics from both models with the highest loading for that word were paired. Matching topics were then identified based on the frequency of such pairings.

To enable a more comprehensive analysis and labeling of topic models derived from the UN corpus, we focused on the model with $k = 20$ topics. This number was chosen based on statistical evaluations of coherence and exclusivity (Figure 2), which are widely applied in current studies (Bernhard-Harrer et al., 2025). Additionally, we manually examined and labeled the resulting topic matches based on their top words, which are provided in Appendix B. To ensure that the chosen topic solution was not unique, the model comparisons described in the following steps were conducted for all previously computed topic solutions, allowing the findings to be validated across different topic numbers.

For the journalistic corpus, we computed topic models with 5 to 40 topics to validate the results across different topic configurations. For detailed analysis, we focused on the solution with $k = 15$, reflecting the smaller size of the dataset and based on coherence and exclusivity metrics. The corresponding figure with statistical metrics and the main results for the journalistic publications is provided in Appendix C.

Figure 2: Coherence and Exclusivity for Different k (UN-Corpus)



3.6 Model Comparison

Following de Vries et al. (2018), differences between the translation strategies were evaluated using three metrics: feature overlap between DFMs, topical prevalence, and topical content. These metrics were calculated separately for each translation strategy, and the results were then compared to assess which target language produced more similar topic models.

Feature overlap assesses whether the translated texts share similar vocabulary by measuring the proportion of overlapping lemmas in the DFMs before the application of topic modeling. Higher overlap indicates greater similarity in the vocabularies of the consolidated texts. We also calculated the *TTR* to examine whether the observed differences reflect variations in vocabulary richness.

Topical prevalence measures the correlation of topic distributions across documents in the translated subcorpora originally written in German and Russian. This metric helps determine whether matched topics maintain similar proportions across the datasets. We additionally examined document-level topical prevalence, following previous studies (de Vries et al., 2018; Reber, 2019). These results were consistent with the corpus-level findings and are therefore not described in detail in the main text. While examining these outcomes (see Appendix D) could help identify documents driving translation misalignment, this aspect is beyond the scope of the current study, which focuses on general quantitative evaluation.

Finally, topical content quantifies the correlation of word distributions across topics from different subcorpora. This metric allows us to determine whether matched topics use the same words with similar frequency.

4. Results

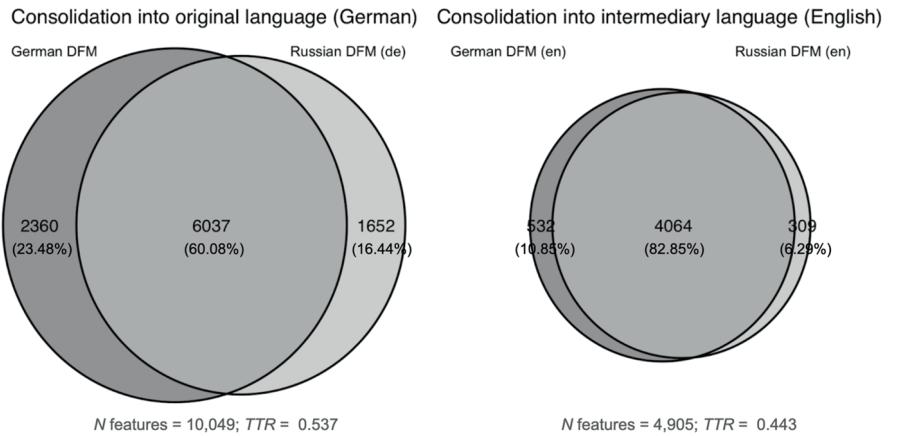
This section first presents the results for the corpus of official UN documents (k = 20). Subsequently, the results of the replication using journalistic publications are discussed in more detail (4.4).

4.1 Feature Overlap

Overall, consolidating UN publications into an intermediary language results in a higher degree of vocabulary overlap between the resulting subcorpora. When German and Russian texts are translated into English, more than 80 % of all DFM features are shared between the two translated versions (Figure 3, right-hand side). In contrast, when German texts remain unchanged and Russian texts are translated into German, just over half of the features (60.08 %) are shared between the resulting DFMs after pre-processing (Figure 3, left-hand side).

Another notable observation is that consolidating texts into the original language results in a considerably larger vocabulary in the data used for topic modeling. When texts are consolidated into German, the combined DFMs contain N = 10,049 unique lemmas—more than twice as the number found when both subcorpora are translated into English (N = 4,905), despite originating from the same collection of documents. The *TTR* values for models based on consolidation into German are also generally higher (0.54) compared to English translations (0.45). These results indicate that, in our case, translation into an intermediary language reduces vocabulary diversity more than the strategy of retaining part of the corpus in its original language for consolidation. Possible consequences and the generalizability of this issue are presented in the discussion section.

Figure 3: Unique DFM-Features and Their Overlap (UN-Corpus)



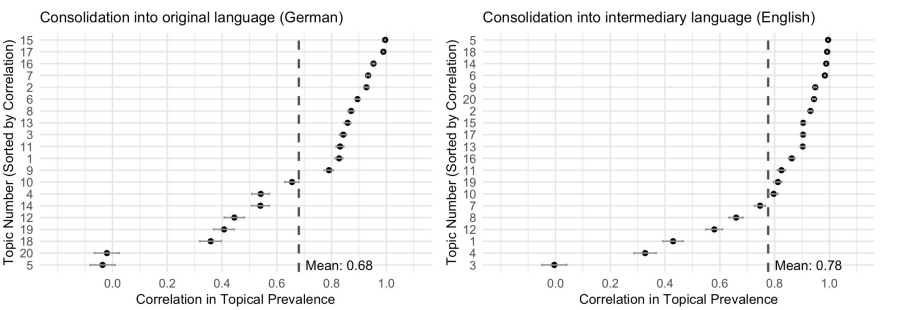
Note. The percentage is calculated based on the total number of features in the respective consolidation strategy.

4.2 Topical Prevalence

To further examine the outcomes at the level of topic models, we compared the translation strategies based on topical prevalence. For the UN documents, translating into an intermediary language produced a more similar distribution of topics across subcorpora than consolidation into an original language.

In particular, topics show a higher correlation in distribution across the translated Russian and German subcorpora when the texts are consolidated into English (M = 0.78, Figure 4, right-hand side) compared to consolidation into German (M = 0.68, Figure 4, left-hand side). Furthermore, in the English versions of the topic models, more than half of the matched topics have a correlation greater than 0.9, whereas in the German models, this is true for only five of twenty topic matches. Because we are working with parallel data, these results suggest that the intermediary language strategy is more likely to align similar topics in their distribution than consolidation into the original language.

Figure 4: Correlation in Topical Prevalence (UN-Corpus)

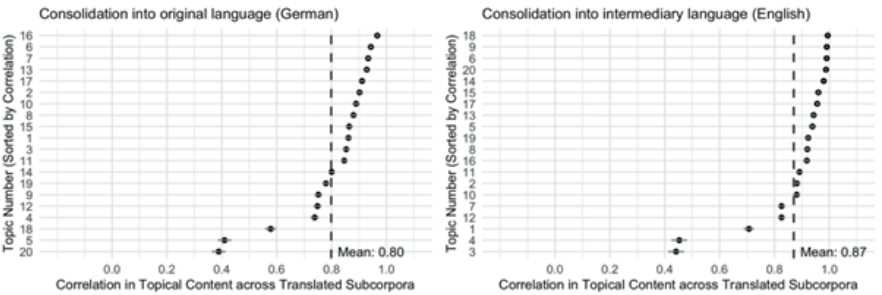


4.3 Topical Content

The results for topical content are consistent with the previous findings. We observe a higher level of similarity between the translated subcorpora in representing matched topics through words when consolidating the texts into the intermediary language.

For models based on translations into English (Figure 5, right-hand side), the correlations of word distributions for most matched topics range from 0.83 and 0.99, with a mean value of 0.87. For consolidation into German (Figure 5, left-hand side), the correlations are slightly lower (M = 0.80), though still indicating strong alignment. This suggests that translation into an intermediary language tends to produce topics using the same words with more similar frequencies compared to consolidation based on one of the original languages. Nevertheless, for both strategies, consolidation led to the generation of largely similar representations of parallel corpora in terms of topical content.

Figure 5: Correlation in Topical Content (UN-Corpus)



4.4 Replication Across Text Types

Comparing the results for UN publications with those for journalistic data, we find that the trends observed in official documents are also present in the journalistic corpus for k = 15 topics (see Table 1). First, for journalistic articles, intermediary-language translation results in greater DFM overlap (64.7 %) compared to consolidation into an original language (49.4 %), while reducing the number of features and *TTR* values used for topic modeling. Second, English-language translations exhibit a more symmetrical topic distribution (M = 0.42) and higher overlap in top words (M = 0.75), compared to consolidation into German (M = 0.25 for topical prevalence; M = 0.67 for topical content).

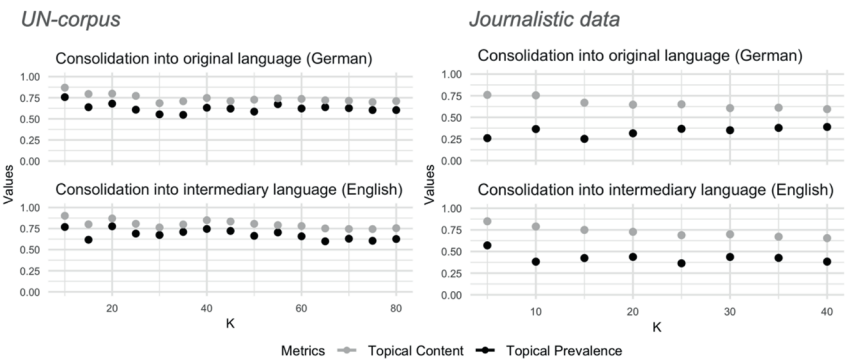
It is also evident that, across all three metrics, the similarities between models are lower for journalistic publications than for official documents. This is particularly noticeable in the mean correlation coefficients for topical prevalence, which do not exceed 0.5 for the journalistic corpus but range between 0.68 and 0.78 for UN publications. This suggests that, on average, matched topics are more similarly distributed in topic models of official documents, reflecting a more coherent vocabulary in the political discourse of UN resolutions. In contrast, the journalistic corpus contains a larger number of features and exhibits higher *TTR* values, indicating greater linguistic variability in journalistic articles despite the smaller corpus size.

Table 1: Results of Comparison Across Text Types

	Official documents *(k = 20)*		**Journalistic publications *(k = 15)***	
	Consolidation into original language (German)	Consolidation into intermediary language (English)	Consolidation into original language (German)	Consolidation into intermediary language (English)
Feature overlap	Overlap = 60.1% *N* Feat. = 10,049 *TTR* = 0.537	Overlap = 82.9% *N* Feat. = 4,905 *TTR* = 0.448	Overlap = 49.4% *N* Feat. = 12,967 *TTR* = 0.715	Overlap = 64.7% *N* Feat. = 8,211 *TTR* = 0.629
Topical Prevalence	Mean Corr. = 0.68	Mean Corr. = 0.78	Mean Corr. = 0.25	Mean Corr. = 0.42
Topical Content	Mean Corr. = 0.80	Mean Corr. = 0.87	Mean Corr. = 0.67	Mean Corr. = 0.75

These outcomes are consistent across different numbers of topics (*k*) for both text types, further highlighting the robustness of the findings (Figure 6). As shown, topical content and topical prevalence remain at similar levels within each text type, although the values are consistently lower for journalistic publications.

Figure 6: Mean Correlation Coefficients across Different Number of Topics



5. Discussion and Best Practices

For RQ1, we find that intermediary-language translation results in greater DFM feature overlap, a more symmetrical topic distribution, and higher overlap in top words compared to consolidation into an original language. This aligns with previous research (Lucas et al., 2015) and confirms that translating the entire corpus into an intermediary facilitates a more balanced representation of the analyzed data, as all subcorpora undergo the same level of content transformation.

However, translating texts into an intermediary language (English) results in a less diverse representation of features compared to direct consolidation into an original language (German). This reflects a noticeable vocabulary simplification in texts processed through

machine translation and supports previous findings (Kotait, 2024; Vanmassenhove et al., 2019), extending them to the new language pairs containing Russian, English, and German. The reason for this overgeneralization may be that, during the training of neural translation models such as Google Translate, frequent translations are reinforced while rare translations get suppressed, particularly when translating into a language with lower linguistic variability (Kettunen, 2014). Consequently, outputs from neural translation systems tend to be less diverse, potentially losing meaningful aspects and culture-specific items in subsequent analysis.

In the context of multilingual topic modeling, our statistical evaluation suggests that this simplification does not significantly impact model outcomes, likely because this method of automated text analysis is inherently focused on providing generalized representations of textual data. However, such simplification could be a critical concern for other types of automated analyses where individual word translations are important—for instance, in studies examining emotions across languages or categorizing text based on specific linguistic features and word components (see Windsor et al., 2019).

Regarding RQ2, we find that the impact of machine translation on journalistic publications follows patterns similar to those observed for official UN documents. For both text types, translation into an intermediary language leads to higher vocabulary overlap, greater similarity in topic prevalence, and more consistent topical content. These results demonstrate the applicability of our findings across different text types and provide additional replication for studies on multilingual analysis in communication science.

Nevertheless, it is important to note that similarities between the models from the Russian and German corpora are lower across all three metrics in the journalistic data. This may be due to both the smaller corpus size and the higher linguistic variability of journalistic texts (Maier et al., 2022). To investigate which features in the journalistic corpus might contribute to these discrepancies, we went through the top words for topics with prevalence and content correlations below the mean and examined their contextual usage.

The journalistic articles used in our study contain a broader range of lexical features than UN documents and include more context-specific, error-prone terms that can affect translation accuracy and lead to inconsistencies. For example, proper names and locations are often transcribed differently during corpus consolidation (e.g., "*Kirgistan*" in an original German text versus "*Kirgisistan*" in the translation from Russian into German), reducing vocabulary overlap between models. Vocabulary simplification is also evident, as context-specific, traditional, or slang terms may lack direct equivalents and are therefore replaced with simpler expressions. For instance, the Russian word *Khorovod*, referring to a traditional circle dance, is translated merely as "dance around" in English, losing its cultural and traditional significance. However, upon manual inspection of the top words, such issues were found to be relatively rare.

For future research, the findings of this study suggest that both target-language strategies can provide valid inputs for multilingual topic modeling. However, one strategy may be more or less suitable depending on the specific case, research objectives, and the languages present in the analyzed corpora.

Consolidation into the original language can be a useful strategy when studying cases in which the languages are closely related or belong to the same branch, such as Slavic or Nordic languages (Koltsova & Pashakhin, 2020; Licht et al., 2024). In such instances, translating into one of the languages already present in the corpus allows researchers to save translation resources while preserving the vocabulary richness of the data. This approach enables a more detailed examination and comprehensive representation of the analyzed

texts, including language-specific characteristics and original meanings that could be lost or blurred when translating into an intermediary language.

If the languages belong to different language groups, such as Chinese and Arabic (Lucas et al., 2015), consolidation into English as an intermediary language for automated analysis may be more appropriate. Here, English would serve as a common bridge between languages due to extensive training resources, making it a practical choice for intermediary translation and enabling a more symmetrical transformation of textual data. This strategy also allows researchers to achieve more aligned representations in terms of DFMs, topical prevalence, and topical content. However, it is important to keep in mind that translating into an intermediary language can reduce vocabulary richness and may omit certain details or nuances of meaning from the original data, which should be considered when interpreting the results of automated analysis.

Regardless of the strategy employed, it is important to ensure the quality of machine translation and to validate its outcomes. In research, this could be achieved through various approaches. For instance, one could manually examine translated documents and the top words in the topic models (see Lucas et al., 2015) to identify potential translation errors that could affect topic representations. Particular attention should be given to topics that appear prevalent only in one of the compared subcorpora. In such cases, it is essential to determine whether the topic is genuinely unique to that subcorpus or whether translation issues contributed to this outcome. Such input validation is especially important when analyzing journalistic publications and other text types relevant to communication science (e.g., social media posts; Maier et al., 2022), which often contain context-specific information and figurative language.

Beyond machine translation, the results of other text transformation steps applied prior to analysis should also be validated, as some words may be processed differently by standard pre-processing algorithms in different languages. For example, in our study, the name of the Russian social media platform *VKontakte* was lemmatized according to German verb rules, producing the non-existent word "*vkontaken*" and causing vocabulary misalignment in our topic models. However, we have not examined this issue systematically and relied only on the top words from topic models, so we can only suggest that such issues are rare. Nevertheless, we encourage researchers to consider the implications of target language choice and to address the potential limitations of machine translation in future studies.

6. Conclusion

Overall, the study demonstrates that both translation strategies examined provide a valid basis for multilingual topic modeling. Machine translation is therefore particularly valuable for analyzing large-scale data in the digital environments of communication science, enabling the identification of cross-case topics and supporting comparative analyses (Lind et al., 2022). Our comparison of strategies indicates that intermediary-language translation leads to greater overlap in features, topic distributions, and top words, but substantially reduces vocabulary diversity compared to consolidation into an original language. This pattern is also observed in journalistic publications, which exhibit higher linguistic diversity and underscore the need for careful validation of machine translations.

This study extends existing research on the impact of machine translation on topic modeling outcomes (de Vries et al., 2018; Reber, 2019; Maier et al., 2022) by examining the role of different translation strategies in automated text analysis. It also provides best practices for selecting an appropriate consolidation approach for large-scale text collections and validating the results of machine translation. More broadly, these findings highlight that thorough evaluation of analytical steps related to language alignment or cross-lingual con-

cept identification is crucial for establishing the validity of applied methods. For instance, this is particularly relevant for approaches based on outputs from generative models, which may overgeneralize interpretations of culture-specific phenomena across languages during training, forcing them into English-like patterns (Yoo et al., 2025).

The study has several limitations. First, only two languages (German and Russian) were analyzed, and in the original-language strategy, translation was conducted only into German. Expanding the study to include additional languages would help further assess the robustness of our findings. Second, the analysis is based on two text types derived from relatively small and partly outdated corpora. Future research could enhance generalizability by incorporating social media data and using larger, more diverse content. Finally, the study does not include detailed qualitative examination of resulting topics, which could provide deeper insights for future analyses and applies a limited, partly imprecise procedure of topic matching (de Vries et al., 2018). More comprehensive qualitative analysis and systematic examination of documents with low topic alignment could illuminate patterns in topic prevalence and clarify potential sources of translation-related issues.

Looking ahead, several aspects warrant further investigation. In particular, more modern approaches to automated analysis, such as word embeddings and transformer models (see Licht et al., 2024), should be explored in greater detail. These methods show promise due to their ability to incorporate contextual information from the textual data. However, they also present limitations, including the need for technical expertise, the potential for biases, higher computational demands, and challenges in interpretability due to the indirect relationship between input data and model outputs (Licht & Lind, 2023). Furthermore, while our study found that machine translation had minimal impact on topic modeling results, it would be valuable to examine its effects on other text analysis methods in greater depth (Windsor et al., 2019). Finally, as online environments become increasingly multimodal, future research could investigate text-image approaches to topic modeling (e.g. Torres, 2024) to assess their applicability in comparative settings.

References

Baden, C., Pipal, C., Schoonvelde, M., & Van Der Velden, M. A. C. G. (2022). Three Gaps in Computational Text Analysis Methods for Social Sciences: A Research Agenda. *Communication Methods and Measures*, *16*(1), 1–18. https://doi.org/10.1080/19312458.2021.2015574

Benoit, K., & Matsuo, A. (2017). *spacyr: Wrapper to the "spacy" "NLP" library* (1.2.1) [Software]. https://CRAN.R-project.org/package=spacyr [28.01.2026].

Bentz, C., & Kiela, D. (2014). *Measuring and Modelling Lexical Diversity across Languages* [Conference presentation]. 5th UK Cognitive Linguistics Conference, Lancaster, UK.

Bernhard-Harrer, J., Ashour, R., Eberl, J. M., Tolochko, P., & Boomgaarden, H. (2025). Beyond Standardization: A Comprehensive Review of Topic Modeling Validation Methods for Computational Social Science Research. *Political Science Research and Methods*, 1–19. https://doi.org/10.1017/psrm.2025.10008

Blei, D. M. (2012). Probabilistic Topic Models. *Communications of the ACM*, *55*(4), 77–84. https://doi.org/10.1145/2133806.2133826

Chan, C.-H., Zeng, J., Wessler, H., Jungblut, M., Welbers, K., Bajjalieh, J. W., van Atteveldt, W., & Althaus, S. L. (2020). Reproducible Extraction of Cross-lingual Topics (rectr). *Communication Methods and Measures*, *14*(4), 285–305. https://doi.org/10.1080/19312458.2020.1812555

de Vries, E., Schoonvelde, M., & Schumacher, G. (2018). No Longer Lost in Translation: Evidence that Google Translate Works for Comparative Bag-of-Words Text Applications. *Political Analysis*, *26*(4), 417–430. https://doi.org/10.1017/pan.2018.26

Eisele, A., & Chen, Y. (2010). MultiUN: A Multilingual Corpus from United Nation Documents. *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC)*, 2868–2872.

Fortson, B. W. IV. (2011). *Indo-European Language and Culture: An Introduction* (2nd ed.). John Wiley & Sons. https://www.wiley.com/en-us/Indo-European+Language+and+Culture%3A+An+Introduction%2C+2nd+Edition-p-9781405188968 [28.01.2026].

Grootendorst, M. (2022). BERTopic: Neural Topic Modeling with a Class-Based TF-IDF Procedure. arXiv. https://arxiv.org/pdf/2203.05794

Günther, E. (2022). *Topic Modeling: Algorithmische Themenkonzepte in Gegenstand und Methodik der Kommunikationswissenschaft.* [Topic Modeling: Algorithmic Topic Concepts in the Subject and Methodology of Communication Science.] Herbert von Halem Verlag. https://www.halem-verlag.de/produkt/topic-modeling/

Hallin, D. C., & Mancini, P. (Eds.). (2011). *Comparing Media Systems Beyond the Western World* (1st ed.). Cambridge University Press. https://doi.org/10.1017/CBO9781139005098

Haim, M. (2023). *Computational Communication Science: Eine Einführung.* [Computational Communication Science: Introduction.] Springer VS. https://link.springer.com/book/10.1007/978-3-658-40171-9

Hanitzsch, T., Hanusch, F., Ramaprasad, J., & De Beer, A. S. (Eds.). (2019). *Worlds of Journalism: Journalistic Cultures Around the Globe*. Columbia University Press. https://doi.org/10.7312/hani18642

Hase, V., Mahl, D., Schäfer, M. S., & Keller, T. R. (2021). Climate Change in News Media across the Globe: An Automated Analysis of Issue Attention and Themes in Climate Change Coverage in 10 Countries (2006–2018). *Global Environmental Change*, *70*, 102353. https://doi.org/10.1016/j.gloenvcha.2021.102353

Jacobi, C., van Atteveldt, W., & Welbers, K. (2016). Quantitative Analysis of Large Amounts of Journalistic Texts Using Topic Modelling. *Digital Journalism*, *4*(1), 89–106. https://doi.org/10.1080/21670811.2015.1093271

Kettunen, K. (2014). Can Type-Token Ratio be Used to Show Morphological Complexity of Languages? *Journal of Quantitative Linguistics*, *21*(3), 223–245. https://doi.org/10.1080/09296174.2014.911506

Koltsova, O., & Pashakhin, S. (2020). Agenda Divergence in a Developing conflict: Quantitative Evidence from Ukrainian and Russian TV Newsfeeds. *Media, War & Conflict*, *13*(3), 237–257. https://doi.org/10.1177/1750635219829876

Kotait, R. (2024). Richness Lost in Machine Translationese. *The Egyptian Journal of Language Engineering*, *11*(1), 66–85. https://doi.org/10.21608/EJLE.2024.267336.1064

Licht, H. (2023). Cross-Lingual Classification of Political Texts Using Multilingual Sentence Embeddings. *Political Analysis*, *31*(3), 366–379. https://doi.org/10.1017/pan.2022.29

Licht, H. & Lind, F. (2023). Going Cross-Lingual: A Guide to Multilingual Text Analysis. *Computational Communication Research*, *5*(2), 1–31. https://doi.org/10.5117/CCR2023.2.2.LICH

Licht, H., Sczepanski, R., Laurer, M., & Bekmuratovna, A. (2024). *No More Cost in Translation: Validating Open-Source Machine Translation for Quantitative Text Analysis.* OSF. https://doi.org/10.31219/osf.io/9trjs

Lind, F., Eberl, J., Eisele, O., Heidenreich, T., Galyga, S., & Boomgaarden, H. G. (2022). Building the Bridge: Topic Modeling for Comparative Research. *Communication Methods and Measures*, *16*(2), 96–114. https://doi.org/10.1080/19312458.2021.1965973

Lucas, C. G., Nielsen, R. A., Roberts, M. E., Stewart, B., Alex, S., & Tingley, D. (2015). Computer-Assisted Text Analysis for Comparative Politics. *Political Analysis*, *23*(2), 254–277. https://doi.org/10.1093/pan/mpu019

Maier, D., Baden, C., Stoltenberg, D., de Vries-Kedem, M., & Waldherr, A. (2022). Machine Translation vs. Multilingual Dictionaries Assessing Two Strategies for the Topic Modeling of Multilingual Text Collections. *Communication Methods and Measures*, *16*(1), 19–38. https://doi.org/10.1080/19312458.2021.1955845

Maier, D., Waldherr, A., Miltner, P., Wiedemann, G., Niekler, A., Keinert, A., Pfetsch, B., Heyer, G., Reber, U., Häussler, T., Schmid-Petri, H., & Adam, S. (2018). Applying LDA Topic Modeling in Communication Research: Toward a Valid and Reliable Methodology. *Communication Methods and Measures*, *12*(2–3), 93–118. https://doi.org/10.1080/19312458.2018.1430754

Reber, U. (2019). Overcoming Language Barriers: Assessing the Potential of Machine Translation and Topic Modeling for the Comparative Analysis of Multilingual Text Corpora. *Communication Methods and Measures*, *13*(2), 102–125. https://doi.org/10.1080/19312458.2018.1555798

Rinke, E. M., Dobbrick, T., Löb, C., Zirn, C., & Wessler, H. (2022). Expert-Informed Topic Models for Document Set Discovery. *Communication Methods and Measures*, *16*(1), 39–58. https://doi.org/10.1080/19312458.2021.1920008

Roberts, M. E., Stewart, B., & Tingley, D. (2019). stm: An R Package for Structural Topic Models. *Journal of Statistical Software*, *91*(2), 1–40. https://doi.org/10.18637/jss.v091.i02

Tiedemann, J. (2012). Parallel Data, Tools and Interfaces in OPUS. *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*, 2214–2218. http://www.lrec-conf.org/proceedings/lrec2012/pdf/463_Paper.pdf [28.01.2026].

Torres, M. (2024). A Framework for the Unsupervised and Semi-Supervised Analysis of Visual Frames. *Political Analysis*, *32*(2), 199–220. https://doi.org/10.1017/pan.2023.32

van Atteveldt, W., Trilling, D., & Arcila, C. (2022). *Computational Analysis of Communication*. Wiley Blackwell. https://cssbook.net/ [28.01.2026].

Vanmassenhove, E., Shterionov, D., & Way, A. (2019). Lost in Translation: Loss and Decay of Linguistic Richness in Machine Translation. *Proceedings of Machine Translation Summit XVII: Research Track*, 222–232. https://aclanthology.org/W19-6622/ [28.01.2026].

Wang, H., Wu, H., He, Z., Huang, L., & Church, K. W. (2022). Progress in Machine Translation. *Engineering*, *18*, 143–153. https://doi.org/10.1016/j.eng.2021.03.023

Windsor, L. C., Cupit, J. G., & Windsor, A. J. (2019). Automated Content Analysis across Six Languages. *PLOS One*, *14*(11), e0224425. https://doi.org/10.1371/journal.pone.0224425

Yoo, M. H., Kim, J., & Song, S. (2025). Multilingual Capabilities of GPT: A Study of Structural Ambiguity. *PloS one, 20*(7), e0326943. https://doi.org/10.1371/journal.pone.0326943