Amsterdam University Press

COMPUTATIONAL COMMUNICATION RESEARCH 5.2 (2023) 1–31
HTTPS://DOI.ORG/10.5117/CCR2023.2.2.LICH

Going cross-lingual: A guide to multilingual text analysis

Hauke Licht
University of Cologne, Cologne Center for Comparative Politics

Fabienne Lind
University of Vienna, Department of Communication

Abstract
Text-as-data methods have revolutionized the study of political behavior and communication, and the increasing availability of multilingual text collections promises exciting new applications of these methods in comparative research. To encourage researchers to seize these opportunities, we provide a guide to multilingual quantitative text analysis. Responding to the unique challenges research faces in multilingual analysis, we provide a systematic overview of multilingual text analysis methods developed for political and communication science research. To structure this overview, we distinguish between separate analysis, input alignment, and anchoring approaches to cross-lingual text analysis. We then compare these approaches' resource intensiveness and discuss the strategies they offer for approaching measurement equivalence. We argue that to ensure valid measurement across languages and contexts, researchers should reflect on these aspects when choosing between approaches. We conclude with an outlook on future directions for method development and potential fields of applications. Overall, our contribution helps political and communication scientists to navigate the field of multilingual text analysis and gives impulses for their wider adoption and further development.

Keywords: multilingual text analysis, text-as-data, computational text analysis, comparative methods

Introduction

Many communication and political scientists turn to text-as-data methods to study political behavior and communication (cf. Grimmer & Stewart, 2013; Schoonvelde et al., 2019; van Atteveldt & Peng, 2018) and the increasing availability of multilingual text collections promises exciting new applications of these methods in comparative research (Lucas et al., 2015). Cross-lingual text analysis is, for example, indispensable when researchers

want to compare political behavior and communication across multiple countries (Baden & Tenenboim-Weinblatt, 2016; Barberá et al., 2022; Gattermann, 2018; Strömbäck et al., 2021), in the international news media (Baden & Tenenboim-Weinblatt, 2016; Baum & Zhukov, 2019; Gattermann, 2018), or in multilingual contexts (Ruedin, 2013; Zelenkauskaite & Balduccini, 2017).

This paper provides a guide to cross-lingual text analysis methods for computational political and communication scientists. The ability to analyze texts across languages with automated methods bears the potential to move quantitative comparative research in these fields forward. However, many scholars who set out to seize these opportunities face multilingualism as a challenge in their research (Baden, Pipal, et al., 2022; Dolinsky et al., 2022). Accordingly, multilingual quantitative text analysis methods have become a very active field of research in both disciplines (e.g., Chan et al., 2020; Courtney et al., 2020; de Vries et al., 2018; Glavaš et al., 2017a, 2017b; Licht, 2023; Lind et al., 2022; Lind et al., 2019; Lind et al., 2021; Lucas et al., 2015; Maier et al., 2021; Proksch et al., 2019; Reber, 2019; Windsor et al., 2019). While overall a positive development, this growth has also resulted in an increasingly complex and interdisciplinary landscape. This makes it difficult for newcomers to orient themselves in the literature. And even for experienced researchers, the speed of methodological developments makes it hard to stay up to date with the latest advancements. This paper responds to the resulting need for orientation, overview, and guidance.

We begin by discussing the methodological implications of applying text-as-data methods in cross-lingual research. We argue that to enable valid inferences from cross-lingual comparisons researchers should strive to obtain similar measurements for documents whose content is similar at a conceptual level – even if these documents are written in different languages.

We then turn to an overview of the existing options researchers have to analyze multilingual corpora. We present a systematization that distinguishes between three main approaches: separate analysis, input alignment, and anchoring. The separate analysis approach means to split a multilingual corpus into its language-specific subcorpora and analyze each subcorpus separately in its original language. The input alignment approach is to represent the documents that go into the text analysis with a common denominator. This can be done by machine-translating them into a single language or by representing them in a multilingual embedding space. The anchoring approach constrains a text model to produce identical measurements for bridging observations (e.g., topically comparable documents or multilingual lexica) when fitting it to multilingual data. Our systematization

groups existing contributions according to how they approach cross-lingual measurement and thus organizes methods for different text analysis tasks, such as document classification or text scaling, within a common framework.

Finally, we discuss the considerations researchers should factor into their decision when choosing an approach for their application. We first compare the three approaches with regard to the resource investments their implementation demands in applied research. We then complement this comparison by discussing to what extent and how each of the three approaches enables researchers to facilitate and assess measurement validity across languages, paying particular attention to challenges arising in the measurement of context-dependent concepts.

In sum, we make three contributions to the literature on multilingual text analysis and quantitative comparative research. First, our guide aids researchers in navigating a very active field of research. Existing text-as-data overview articles devote little to no attention to the challenges arising in cross-lingual applications (Boumans & Trilling, 2016; Gentzkow et al., 2019; Grimmer et al., 2022; Grimmer & Stewart, 2013; Schoonvelde et al., 2019; van Atteveldt & Peng, 2018). And other contributions limit their discussion to alternatives for specific analytical tasks, such as topic modeling (Lind et al., 2021; Reber, 2019), dictionary analysis (Lind et al., 2019; Proksch et al., 2019; Ruedin, 2013), or supervised text classification (Courtney et al., 2020; Glavaš et al., 2017a; Licht, 2023; Lind et al., 2022). Our contribution fills this gap by providing an overview of the various options scholars have to leverage multilingual corpora to address their research questions.

Second, our guide facilitates the application of multilingual text analysis methods by discussing central considerations scholars should factor into their decision between approaches. Since the seminal contribution of Lucas et al. (2015), the existing literature has mainly revolved around resource efficiency considerations. We add cross-lingual measurement equivalence and context-sensitive measurement as two central methodological challenges that researchers face when applying text-as-data methods to multilingual corpora.

Third, our article highlights key methodological questions researchers should engage with when conducting multilingual quantitative text analysis. By doing so, we seek to encourage a critical engagement with the underlying assumptions and limitations of existing methods and to point out areas for further methodological development.

Why cross-lingual text analysis is challenging

The common goal of analysts who apply text-as-data methods in their research is to answer substantively interesting questions about the political and social world (Gentzkow et al., 2019; Grimmer & Stewart, 2013; Schoonvelde et al., 2019; van Atteveldt & Peng, 2018). One of the biggest promises of *multilingual* quantitative text analysis is to enable researchers to leverage text materials written in different languages in the comparative analysis of political behavior and communication (Lucas et al., 2015).

Some of the challenges that researchers confront when applying text-as-data methods in cross-lingual analyses are similar to those they face in monolingual analysis. Text-as-data analyses generally require researchers to operationalize their theoretical concepts of interest with manifest numerical data.[1] Researchers thus need to extract textual features from the documents in their corpus that are indicative of the concept in the focus of their study while, at the same time, reducing the complexity of human language (Grimmer & Stewart, 2013). For example, when applying count-based methods like dictionaries, the *Wordfish* text scaling algorithm (Slapin & Proksch, 2008), or topic models (Blei et al., 2003; Quinn et al., 2010), researchers commonly represent documents as so-called "bag of words" (i.e., in a document-term matrix, cf. Grimmer & Stewart, 2013, p. 273). This approach represents documents without paying attention to word's context and implies that researchers need to choose how to select and simplify the words and phrases used across documents. The overall purpose of these choices is to facilitate automated methods' ability to detect the conceptually relevant signals in the data. However, as these choices can be consequential for measurement validity (cf. Denny & Spirling, 2018), achieving this goal in monolingual analysis commonly requires domain expertise and knowledge about word choice and general language use in the corpus under study.

In applications of text-as-data methods to *multilingual* corpora, researchers are further confronted with the fact that documents are written in different languages. This creates two additional challenges. First, documents' bag-of-words text representations are not directly comparable in

[1] In practice, this involves obtaining numeric representations of the documents in a corpus under study and to define, infer, or "learn" a mapping between these inputs and the conceptual outcome space. Depending on researchers prior knowledge of the possible outcome space (cf. Grimmer & Stewart, 2013; Quinn et al., 2010), researchers can either apply unsupervised methods to infer this mapping from unlabeled data (e.g., topic modeling or text scaling), supervised methods to learn it from (human-)annotated data (e.g., supervised text classification), or define it according to their domain expertise and case knowledge (e.g., dictionary analysis).

	Sentence text	Subtopic
Doc_1	Asylum seekers do not burden the community and the social system	welfare
Doc_2	Asylsuchende belasten das Gemeinwesen nicht [Asylum seekers do not burden the community]	welfare
Doc_3	Das System der Einschüchterung führt zu mehr Gewalt [The system of intimidation leads to more violence]	security

Note: Text in brackets shows English translations of sentences in German.

Table 1: Three example sentences illustrating the content of a multilingual corpus.

their original languages because languages' vocabularies mismatch. Second and related, languages' structural differences warrant different preprocessing choices to facilitate cross-lingually comparable measurement.

Before we further explain these challenges, we introduce a running example designed to make our discussion more accessible. The application in our running example is one that has been widely studied by communication and political scientists: quantifying how migration is discussed by political parties and the news media (e.g., Dancygier & Margalit, 2020; Helbling et al., 2010; Strömbäck et al., 2021). For the sake of simplicity, our running example assumes that we are only dealing with one English and two German sentences (Table 1) that are part of a larger corpus of news articles. A typical quantitative text analysis task would be to categorize the topics of these sentences and thus to categorize the subtopics of the migration discourse. For example, documents 1 and 2 could be labeled as "welfare-related" and document 3 as "security-related" migration discourse.

The first methodological challenge that arises in multilingual quantitative text analysis is that texts written in different languages are represented with different vocabularies. Accordingly, many of the words that exist in one language do not exist in the other language(s).[2] And conversely, words that exist with the same spelling in several languages typically have very different meanings.[3] Chan et al. (2020, p. 285) refer to this as the "Tower of Babel" problem (cf. Maier et al., 2021).

Table 2 illustrates the Babel problem for the four example sentences in

[2]In fact, semantically similar words and phrases typically do not co-occur with each other in different languages (Zhang et al., 2010).

[3]Examples of so-called "false friends" are the English word "gift" (which exists in other languages but means, for example, *poison* in German and *married* in Norwegian) and "home" (means *mold* in Finnish and *man* in Catalan; source).

	Tokens																					
	asylum	seekers	do	not	burden	the	community	and	social	system	asylsuchende	belasten	das	gemeinwesen	nicht	der	einschüchterug	führt	zu	mehr	gewalt
Doc_1	1	1	1	1	1	2	1	1	1	1											
Doc_2											1	1	1	1	1						
Doc_3										1			1			1	1	1	1	1	1

Table 2: Bag-of-words representations of our three example sentences.

our running example. Say we want to prepare them for automated classification and represent them with bag-of-words count vectors. The terms that make up documents 1 and 2 are sorted into different columns. Thus, although these documents are *conceptually* very similar, they have not a single term in common. Document 3, on the other hand, is conceptually relatively different from the other three documents, but it shares terms with documents 1 and 2.

A second and related challenge arising in multilingual quantitative text analysis is that there are systematic differences in the composition of words and phrases between languages (cf. Shababo & Baden, 2023).[4] To illustrate this point based on our running example, consider the term "asylum seeker."[5] One would generally expect that the occurrence of this term provides an important signal when wanting to automatically identify migration-related topics in a text corpus. However, while "asylum seeker" is represented with an (open-compound) noun phrase in English, in German and other Germanic languages it is a compound word ("Asylsuchender").[6] Such structural differences between languages have practical implications for the researchers in our running example. Depending on the language, terms indicative of the concept a researcher wants to measures might be referred to with words that span one, two, or more[7] tokens. This implies that if

[4]This holds even for the rules that govern whether or not, and if so, how, social-scientifically relevant information like gender, time orientation, plurality, actor-patient relations, etc. are encoded.

[5]We thank one of the anonymous reviewers for suggesting this example.

[6]This holds, too, for other Germanic languages like Danish ("Asylansøger"), Dutch ("Asielzoeker"), Norwegian Bokmål ("Asylsøker"), or Swedish ("Asylsökande").

[7]In (West) Slavic languages, for example, the entity "asylum seekers" is referred to with terms that span multiple words ("Žadatel o azyl" in Czech, "Osoba ubiegająca się o azyl" in

the researchers in our running example adopt bag-of-words methods, they need to adapt their preprocessing choices to each language in the corpus to ensure that their quantitative text model or instrument is able to tap into the signal provided by the occurrence of the term "asylum seekers." Our example thus shows that multilingualism requires researchers to pay great attention to preprocessing choices when attempting to measure the same concept in different languages.

Overall, our running example illustrates that multilingualism makes it much more difficult than in monolingual applications to define or "learn" how documents' text representations relate to one's quantity of interest at a conceptual level. Hence, if "language barriers" such as those illustrated in Table 2 are not overcome in some way, these text representations are of little use to, for example, learn the relationship between term frequencies and sentences' subtopics.

In cross-lingual applications, text analysts who want to obtain indicators of political and communicative behavior thus need to attempt that the outputs of their measurement procedure are "aligned" at a conceptual level across languages. Only then is a scaling or classification in one language comparable with that in another language (Lucas et al., 2015, p. 262). The strategies we discuss next are all about overcoming this problem.

Existing approaches to analyzing multilingual corpora

Navigating existing options to analyze multilingual corpora with text-as-data methods is not easy. The pioneering article by Lucas et al. (2015) is a good starting point. Lucas et al. (2015) present a machine translation-based approach to cross-lingual topic modeling. One of their significant contributions in relation to our article is that they identify three "common approaches" to analyzing multilingual corpora (p. 261f.). However, they focus on implementing only one of these approaches and thus offer little additional discussion to clarify the practical and methodological implications of the other two approaches. What is more, the methodological literature on the quantitative analysis of multilingual political text corpora has grown considerably since 2015.

Hence, we build on Lucas et al. (2015) and offer a systematization that accommodates a large variety of text analysis tasks and contributions. Our systematization emphasizes the commonalities between methods and an-

Polish, and "Žiadateľ o azyl" in Slovak).

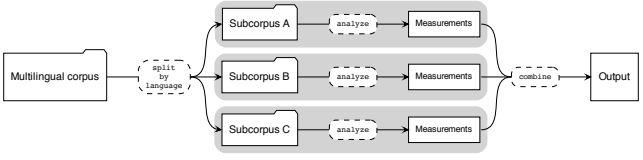


Figure 1: Illustration of the separate analysis strategy to multilingual text analysis.

alytical strategies in how they approach measurement from multilingual corpora. This abstraction allows us to organize methods developed for *different text analysis tasks*, such as topic classification or text scaling, within *a common framework* and to discuss the practical and methodological consequences of their implementation.

Our systematization distinguishes between three approaches for dealing with multilingual corpora in comparative text-as-data applications: the *separate analysis* of language-specific subcorpora; *input alignment* through machine translation or multilingual text embedding; and *anchoring*, which relies on external information such as bilingual lexica, parallel texts, or topically comparable documents as "bridging observations." Below, we discuss and illustrate each approach in greater detail.

Separate analysis

The first approach is to split a multilingual corpus into several, language-specific subcorpora and to analyze each subcorpus separately (see Figure 1). The resulting measurements can then be combined across languages in one dataset for downstream analyses. Accordingly, applying the separate analysis approach requires adapting and implementing the measurement procedure for each language in a corpus.[8] The idea that motivates the separate analysis approach is that to obtain measurements that are aligned at a conceptual level across languages, researchers should work with the original text materials to, for example, avoid introducing bias or getting semantics lost in translation. Analyzing texts separately in their original languages presents a way to realize this ambition.

In the case of our running example, a team of researchers that wants to

[8]In cross-country comparative research, the target corpus is also often analyzed country by country, even if the documents from some countries are written in the same language (Lehmann & Zobel, 2018).

measure how prevalent migration-related subtopics (e.g., welfare) are in their multilingual corpus could perform manual coding. This would mean that they need to collect annotations ("codings") for documents in their original languages. When they cannot recruit coders fluent in multiple languages they would need to split the target corpus into its language-specific subsets and distribute the documents in each to coders that are proficient in the given language. They could then combine the resulting annotations and directly estimate their quantities of interests, such as how prevalent the welfare subtopic is in the migration-related documents published by different news outlets (Schmidtke, 2018; A. R. Schuck et al., 2014). This strategy has been adopted by influential large-scale human coding projects like the *Comparative Manifestos Project* (Lehmann et al., 2022), or the *European Election Media Study* (A. Schuck et al., 2010) as well as in numerous independent studies (cf. Klüver & Bäck, 2019; Lehmann & Zobel, 2018).

Alternatively, the team of researchers could task native language speakers with annotating only subsets of each language-specific subcorpus. They could then use these annotations to train language-specific supervised text classifiers and apply the resulting classifiers to label the remaining, unannotated documents in their original languages (cf. Theocharis et al., 2016).

Another strategy that might involve fewer upfront resource investments would be to develop one dictionary per language and search for relevant keywords in the English- and German-language subcorpora separately (Lind et al., 2019; Proksch et al., 2019). This strategy, too, has been adopted by applied researchers to study various phenomena (e.g., Rooduijn et al., 2014).

And if the researchers do not or cannot define the migration-related subtopics (e.g., welfare) before seeing the data, they could fit one topic model to the German-language and another one to the English-language portion of their corpus (e.g., Ceron et al., 2020). However, this would require that they qualitatively align the extracted topics across languages *post hoc* (cf. Chan et al., 2020; Maier et al., 2021). Importantly, whatever strategy the team of researchers adopts, the numeric representations of the German- and English-language documents in their corpus (e.g., document-term matrices) would need to be analyzed separately.

Input alignment

The second approach involves finding a "common denominator" that enables the joint quantitative analysis of documents across languages (cf. Lind et al., 2021). To enable such cross-lingual analysis, researchers need to convert their multilingual textual *inputs* into numeric representations that are

directly comparable across languages. There are currently two ways to achieve such "alignment" of texts' input representations: machine translation and multilingual embedding.

Machine translation

Machine translation means transferring the documents in a multilingual corpus into a single target language by using a (neural) translation model such as those provided by *Google Translate* or *DeepL*. As noted by Lucas et al. (2015, p. 268), translation creates overlap in the vocabulary used to represent the documents in a multilingual corpus and thus allows aligning their text representations. Hence, researchers often rely on machine translation to enable the joint analysis of documents in their multilingual corpus. For example, Barberá et al. (2022) translate the tweets of world leaders to allow their manual coding and supervised classification in English.

Coming back to our running example, Table 3 shows that if researchers would machine-translate the text of the German documents in their corpus to English, semantically equivalent terms of documents 1 and 2 (e.g., "community" and "gemeinwesen") would be represented with the same token (i.e., "community"). Consequently, applying an English dictionary or training a supervised classifier on documents' English versions would allow recognizing these documents' conceptual similarities.

There are *two alternative techniques* to align the quantitative representations of documents in a multilingual corpus through machine translation: full-text and token translation (cf. Lucas et al., 2015; Reber, 2019). The first technique, depicted in Figure 2a, is to translate documents' *full texts* into the target language. Full-text translating documents before preprocessing, researchers can represent the documents in their originally multilingual corpus with a set of tokens from only one language, which enables them to apply standard (monolingual) bag-of-words text analysis methods (cf. Lucas et al., 2015; Reber, 2019). Moreover, the evidence presented by Courtney et al. (2020) suggests that full-text translation enables reliable content analysis of multilingual corpora when coders speak only the target language.

The alternative is to translate only the text features obtained by tokenizing documents in their original languages.[9] Figure 2b shows that this "token translation" approach involves translating only the set of unique words and phrases (or "tokens") obtained by preprocessing documents in their original

[9]We use the term "tokenization" here to refer to the common practice of splitting documents into their constituent words and phrases (n-gram tokens).

	Sentence text	English translation	Subtopic
Doc_1	Asylum seekers do not burden the community and the social system	Asylum seekers do not burden the community and the social system	welfare
Doc_2	Asylsuchende belasten das Gemeinwesen nicht	Asylum seekers do not burden the community	welfare
Doc_3	Das System der Einschüchterung führt zu mehr Gewalt	The system of intimidation leads to more violence	security

(a) example sentences and English translations

	Tokens															
	asylum	seekers	do	not	burden	the	community	and	social	system	of	intimidation	leads	to	more	violence
Doc_1	1	1	1	1	1	2	1	1	1	1						
Doc_2	1	1	1	1	1	1	1									
Doc_3						1				1	1	1	1	1	1	1

(b) bag-of-words representations of sentence texts' English translations

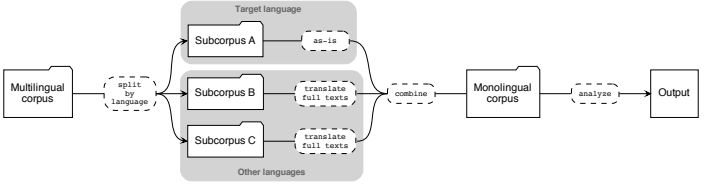
Table 3: English translations of German sentences along with originally English sentences (top) and bag-of-words representations of the English versions of all example sentences (bottom).

languages.[10] Combining documents' translated bag-of-words representations across language-specific subcorpora, in turn, results in a monolingual document-term matrix that researchers can analyze with count-based methods. The token translation approach has been shown to be reliable for topic modeling (cf. de Vries et al., 2018; Reber, 2019) and document similarity analysis (Düpont & Rachuj, 2022).

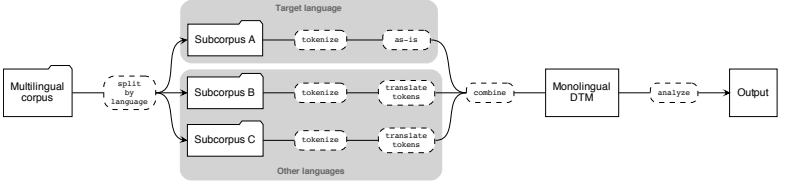
Multilingual embedding

An alternative strategy to transfer the documents in a multilingual corpus to a "common denominator" is multilingual embedding. Simply put, in natural language processing (NLP), embedding means to represent a collection of

[10] Others refer to this approach as "word-by-word" or "term-by-term" translation (cf. de Vries et al., 2018; Lucas et al., 2015; Reber, 2019; van der Veen, 2022).



(a) full-text translation



(b) token translation

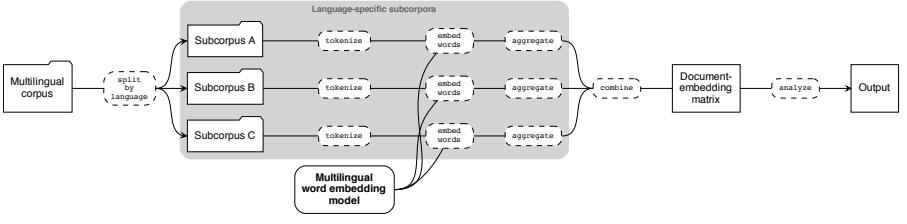
Figure 2: Illustration of translation approaches to input alignment.

words, sentences, or documents in a dense, real-valued vector space (cf. Rodriguez & Spirling, 2022).[11] The goal of text embedding methods is to "discover" this space. Specifically, these methods learn to represent text items (e.g., words, sentences, or documents) with vectors whose location in the embedding space reflects their linguistic similarities. Consequently, similar text items are placed close in the embedding space, and dissimilar text items are placed further apart. Multilingual text embedding models enable this similarity-based text representation across languages (cf. Conneau et al., 2020; Ruder et al., 2019) and thus provide for a translation-free approach to input alignment (cf. Chan et al., 2020; Glavaš et al., 2017a; Licht, 2023).

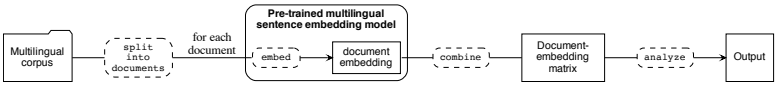
As shown in Figures 4 and 5, implementing the multilingual embedding approach to input alignment involves processing the documents in a multilingual corpus through a (pre-trained) embedding model. For example, researchers can rely on multilingual *word* embeddings to align textual inputs across languages (e.g., Chan et al., 2020; Glavaš et al., 2017a).[12] And if the

[11] In contrast to typically sparse word count-based representations such as document-term matrices.

[12] However, researchers need to take some additional steps to obtain document representations in this case. The first need to tokenize their documents in their original languages. Then look up the embedding for each token. And finally, aggregate tokens' embeddings at



(a) multilingual word embedding



(b) multilingual sentence embedding

Figure 3: Illustration of multilingual embedding approaches to input alignment.

documents in the multilingual corpus are sentence-like, evidence presented by Licht (2023) suggests that pre-trained multilingual *sentence* embedding models allow high-quality cross-lingual alignment. Researchers can then use documents' representations in the multilingual embedding space as inputs to quantitative text analysis instead of their (translated) bag-of-words count vectors. In line with this approach, researchers also increasingly rely on multilingual language models like mBERT (e.g., (Greene & Sylvester, 2022; Laurer, 2023) that implicitly align text inputs across languages.

Several contributions adopt this approach to enable cross-lingual quantitative text analysis. Chan et al. (2020) present a cross-lingual topic modeling method that takes multilingual word embeddings as primary inputs. Further, Glavaš et al. (2017b) present a method to scale documents' multilingual embedding representations that obtains quantities analogous to those of the widely-applied *Wordfish* and *Wordscores* models.[13] Glavaš et al. (2017a) and Licht (2023) evaluate multilingual word and sentence embedding approaches for supervised text classification, respectively.

Accordingly, the team of researchers in our running example could adopt

the document-level into a single representation, for example, by computing the element-wise (weighted) mean across embedding dimensions.

[13](Goist, 2020, Chapter 2) proposes an alternative embedding-based multilingual scaling approach that shares some interesting ideas with the topic modeling method proposed by Chan et al. (2020).

the approach proposed by Chan et al. (2020) to discover migration-related topics in the target corpus that can be compared across languages. Alternatively, if they were collecting topic annotations for the migration-related sentences in their corpus, they could apply the multilingual embedding approaches discussed by Licht (2023) or Glavaš et al. (2017a). In all these cases, the documents in their corpus would first be transferred into the multilingual embedding space, and documents' embeddings would then be used as inputs for unsupervised topic modeling or supervised classification.

It is worth noting, however, that researchers cannot apply well-established count-based quantitative text analysis models when using embeddings as input representations. Existing topic modeling and text scaling methods, for example, infer their respective quantities of interest by modeling differences in the number of times words and phrases occur across documents. Since embeddings are real-valued vectors, researchers cannot directly apply these methods to documents' embeddings representations. Yet, as discussed above, research by Chan et al. (2020), Glavaš et al. (2017a, 2017b), and Licht (2023) show that by now there exist such alternatives for the most common quantitative text analysis tasks.

Anchoring

There is a third approach we would like to discuss: anchoring. Simply put, anchoring is performed by using "bridging observations". Bridging observations are, for example, parallel documents that exist or have been translated into several languages (e.g., Koehn, 2005; Tiedemann, 2012) or multilingual lexica that map words in different languages with similar meaning. Anchoring constrains the measurements that a text model produces for bridging observations' to be similar when fitting the model. This explicitly incentivizes cross-lingual comparability of the measurements obtained for documents in the multilingual corpus under study.

Before discussing this approach in detail, we need to add a note of caution. To the best of our knowledge, the anchoring approach has thus far only been implemented for LDA-like multilingual topic modeling (cf. Lind et al., 2021). The argument can be made that both dictionary translation[14] and the manual analysis of multilingual content by coders fluent in multiple languages[15] leverage anchoring strategies. But we are not aware of

[14] See, for example, Maier et al. (2021), who present dictionary translation as an anchoring strategy.

[15] Proksch et al. (2019) discuss how coders fluent in multiple languages can function as "bridges" since they implicitly transfer the measurement instrument to different languages.

research in communication or political science that implements the anchoring approach for supervised text classification. We include anchoring in our discussion nevertheless to increase the visibility of this idea in the literature on multilingual text analysis.

The insight that motivates the idea of anchoring is that cross-lingual quantitative text analysis presents an *identification problem* (Proksch et al., 2019).[16] Typically, every document in the corpus from which a researcher wants to generate measurements is recorded in only one language. This makes it difficult to place the documents in a single shared, cross-lingual measurement scale or space. Anchoring the measurements obtained for bridging observations – that is, constraining them to be identical – is intended to mitigate this problem. Specifically, bridging observations are selected to provide information about documents' *conceptual* similarity across languages. This information, in turn, can be leveraged in the measurement model or procedure to facilitate cross-lingual measurement equivalence – that conceptually similar documents receive similar measurements.

For example, the bridging observations used in the polylingual topic model (PLTM) proposed by Mimno et al. (2009) to align topics across languages are documents that are assumed to have comparable distributions over latent topics (i.e., that are "topically comparable"). In political and communication science, such documents could, for example, be press releases, debate transcripts, or news articles that are published in different languages (cf. Lind et al., 2021).[17] The PLMT estimates first one comparable word distribution per topic and language for the bridging observations and then a topic distribution for the documents in the multilingual corpus. Consequently, each estimated topic can be described with a distribution over language-specific vocabularies, and any document written in one of the languages included when fitting the model can be represented in the cross-lingually aligned topic space discovered by the PLMT. For example, Lind et al. (2021) implement this model to identify the most prevalent topics for a corpus of English, German, and Spanish migration-related documents.

[16] Proksch et al. (2019) draw parallels to the identification problem confronted in scaling the positions of legislators in separate chambers of a legislature.

[17] The two types of bridging observations contrasted by Lind et al. (2021) are the use of comparable documents (*EuroNews* articles published in English, German, and Spanish) and of synthetic parallel data (a subset of the corpus was translated so that each document is available in all languages).

How to choose between approaches

The above discussion documents a wealth of methods, techniques, and strategies for cross-lingual quantitative text analysis. The assessment that there 'exists little or no support for cross-lingual comparison' Lucas et al. (2015, p. . 259) arrived at about seven years ago thus seems no longer accurate.

Yet, the fact that scholars today have several options to quantitatively analyze multilingual corpora in their research also raises an important practical question: How to choose between approaches in a given research application? We believe that there is no one-size-fits-all solution. None of the three approaches is generally better suited for cross-lingual analyses. Instead, we argue that the three approaches presented above have specific advantages and limitations that make them more or less suited for different researcher applications.

To facilitate the employment of the three approaches discussed in Section 3, we compare them in view of considerations that should factor into researchers' decision-making. First, we compare the three approaches with regard to the resource input their implementation requires. Second, we compare the extent to which the three approaches enable researchers to facilitate and validate measurement equivalence.

Resource considerations

Quantitative text analysis methods vary greatly in the amount of time, financial resources, and manual labor researchers need to invest in implementing them (cf. Barberá et al., 2022; Grimmer & Stewart, 2013; Quinn et al., 2010). Such resource considerations have been a major driver of methods development in the quantitative text analysis literature, and Dolinsky et al. (2022) show that they are an important determinant of applied researchers' choices between methods. To help researchers choose between the three approaches in their specific applications, below we discuss the different resource requirements their implementation creates.

Separate analysis

Separate analysis requires researchers to adapt, execute, and validate the measurement procedure for each language present in their corpus. In practice, this implies a duplication of researchers' efforts for every additional language they included in their analyses. Importantly, this added effort arises

beyond the need to adapt preprocessing procedures to different languages we have discussed in Section 2. Manual content analysis is the extreme case in this regard because researchers need to adapt and validate the coding instrument (Lucas et al., 2015, p. . 261) and recruit linguistically qualified human coders for all languages present in the target corpus Krippendorff (2004, pp. 127-9). Likewise, supervised text classification presupposes access to a pool of human coders fluent in multiple languages or another multilingual instrument that allows labeling documents in their original languages (Lind et al., 2022). A similar added cost arises in multilingual dictionary analysis since researchers depend on linguistically qualified domain experts or machine translation to adapt and validate their keywords list to new languages (Lind et al., 2019; Proksch et al., 2019) if they do not speak all the languages present in their target corpus. In contrast, adopting the separate analysis approach when applying unsupervised methods, such as topic modeling or text scaling, requires qualitative *post hoc* alignment of the measurements obtained from multilingual corpora (e.g., Ceron et al., 2020). In sum, separate analysis typically requires the highest additional resource investments per language compared to the other two approaches. Accordingly, separate analysis is generally considered 'complex, labor-intensive, and costly' (Reber, 2019, p. 102) and likely discourages multilingual analysis altogether (Baden, Dolinsky, et al., 2022).

Input alignment

In contrast, the resources required to implement the input alignment approach relate mainly to the necessary skill set and expenses for alignment and validation. However, these requirements differ between the full-text translation, token translation, and multilingual embedding variants. These differences are most nuanced during preprocessing. The full-text machine translation approach requires relatively little linguistic knowledge because researchers can tokenize all documents in a language they are familiar with and for which reliable preprocessing tools exist (cf. Baden, Dolinsky, et al., 2022). In contrast, as noted in Section 3.2, implementing the token translation approach requires that researchers preprocess all documents in their original languages. A similar difference exists between the two multilingual embedding variants. Pre-trained sentence embedding and multilingual Transformer models rely on built-in tokenizers, and researchers thus can (and need to) "outsource" key preprocessing decisions. In contrast, multilingual *word* embedding requires preprocessing and tokenizing documents in

their source languages.[18] Thus, the token translation and multilingual word embedding approaches require comparatively more attention to languages' structural differences during preprocessing than their full-text translation and sentence embedding counterparts. However, we emphasize that certain knowledge of all languages in an analysis should always be ensured, even when using the input alignment approach, in order to be able to monitor individual processing steps and to critically review and interpret results.

Beyond these differences in the linguistic knowledge and extra attention required during preprocessing, a general practical advantage of the machine translation approach is that implementing it is relatively easy compared to multilingual embedding. When relying on a commercial service like *Google Translate* or *DeepL*, researchers can rely on web applications or existing software packages.[19] Alternatively, they can rely on a pre-trained open-source machine translation model like M2M (Fan et al., 2021) or OPUS-MT (Tiedemann & Thottingal, 2020). However, the latter option presupposes moderate Python programming skills (cf. Licht, 2023; van der Veen, 2022). In contrast, applying the multilingual embedding approach further demands some understanding of deep learning and natural language processing methods.

However, full-text machine translation can be very expensive when translating large multilingual corpora with a commercial machine translation service. It is thus not surprising that some research has focused on evaluating the more cost-efficient token translation technique (cf. Düpont & Rachuj, 2022; Reber, 2019). However, this technique risks translation errors. A simple plug-and-play alternative to commercial translation services is to rely on publicly available open-source models like M2M or OPUS-MT (cf. Licht, 2023). Yet, implementing this option is not entirely cost-free either since it requires access to GPU processing resources.[20] Moreover, open-source machine translation models have not yet been evaluated to a similar extent as their commercial counterparts.[21] The multilingual embedding alternative compares favorably in this regard, as there are several open-source models available for public use.[22] And since embedding models compute much

[18]The reason is that pre-trained static word embeddings (e.g., fasttext) have a fixed vocabulary, and researchers need to process their document accordingly to enable a look-up of words' embeddings in the pre-trained embedding matrix (cf. Chan et al., 2020, and footnote 14).

[19]e.g., https://cran.r-project.org/web/packages/googleLanguageR

[20]GPUs (Graphics Processing Units) are specialized hardware components designed for parallel processing tasks. In deep learning, GPUs significantly accelerate the training process by processing vast amounts of data through neural networks' numerous weight matrices while performing multiple mathematical operations simultaneously.

[21]The results presented by Licht (2023) suggest that open-source machine translation models are indeed a viable alternative to their commercial counterparts.

[22]for pre-trained sentence embedding models see https://www.sbert.net/docs/pretrained_

faster than machine translation models, the availability of GPU processing resources is typically not too constraining in practice.

Anchoring

Compared to the input alignment approach, the anchoring approach has the advantage that documents enter the measurement procedure in their original languages. Hence, if external "bridging observations" are available (cf. Lind et al., 2021), researchers do not need to invest resources into (machine) translation. However, this advantage comes at the cost that researchers require greater linguistic knowledge of the languages they analyze because – as discussed in Section 2 – they might need to adapt their preprocessing choices. In these regards, the anchoring approach is on par with the separate analysis, token translation, and multilingual word embedding approaches.

Moreover, there are currently two major practical constraints that tend to hinder the adoption of the anchoring approach in applied research. First, as mentioned in Section 3.3, to the best of our knowledge, anchoring has not been implemented for supervised text classification(see the papers cited in Lucas et al. 2015, p. 261f.; and in Lind et al. 2021). Second, there exist only a few parallel political text corpora, most notably EuroParl (Koehn, 2005) and the multi-UN corpus (Tiedemann, 2012). While these corpora have already powered innovative methods research in the field of multilingual quantitative text analysis (e.g., de Vries et al., 2018; Windsor et al., 2019), they do not fit the needs of applied researchers that study other domains like social media or the news media.

Facilitating and validating equivalence and context sensitivity

As discussed in Section 2, a central goal when applying text-as-data methods to multilingual corpora is to obtain measurements that are aligned at a conceptual level across languages (Lucas et al., 2015). We relate this notion to the idea of cross-lingual *measurement equivalence* (cf. van Deth, 1999) and have explained that it implies that researchers should strive to obtain similar measurements for documents whose content is similar at a conceptually level even if these documents are written in different languages.

Below, we will discuss how the three approaches outlined in Section 3 enable researchers to pursue this goal and assess whether they have attained

models.html#multi-lingual-models; for multilingual word embeddings see, e.g., https://fasttext.cc/docs/en/aligned-vectors.html and Ruder et al. (2019)

it. However, we first introduce an important conceptual distinction that will add some important nuance to our discussion: the distinction between the semantic and pragmatic dimensions of language.

The *semantic* perspective focuses on the literal meaning of a text. In this regard, measurement equivalence aims to identify texts with similar meanings across languages. So far, this perspective has been the main focus of our discussion. The *pragmatic* perspective, in turn, puts the social meaning of language in the spotlight. It emphasizes the importance of context in the production and comprehension of texts (e.g., Kahditani, 2022).

We argue that it is often not only the semantics but also the pragmatic dimension of language use that matters for cross-lingual measurement equivalence in quantitative text analysis. For instance, phenomena like hate speech or populist rhetoric may be expressed differently in different countries (cf. Esser & Pfetsch, 2020). Similarly, in the case of our running example, differences in countries' political systems, migration histories, and media systems can result in differences in the words, phrases, and frames political actors or the media use to express ideas related to migration (Eberl et al., 2018). Therefore, to compare texts conceptually across these countries, researchers must select context-specific vocabulary that indicates their target concept in each country.

Accordingly, researchers should take into account the socio-political contexts of text when comparing them (Lind et al., 2022). In the case of multilingual text analysis, this means that failing to accommodate contexts' specifics in addition to languages' differences can threaten measurement validity and ultimately result in wrong empirical conclusions (Adcock & Collier, 2001).

This raises the question of how the different approaches we have summarized in Section 3 allow us to facilitate and validate measurement equivalence across languages *and* contexts. Specifically, which procedures can researchers adopt in the measurement process to facilitate equivalence, and which techniques can be applied to assess the success of these efforts *ex post*?

Unfortunately, to the best of our knowledge, there do not yet exist well-established strategies for facilitating and validating measurement equivalence in multilingual text analysis in a way that is sensitive to both language and context differences (for a proposed framework see Baden, Dolinsky, et al., 2022). This current lack of clear guidelines limits our ability to directly compare the three approaches in terms of their "validatability." We thus instead focus on synthesizing and adding to an ongoing debate, hop-

ing that this will aid researchers in choosing the approach that fits best the needs of their practical application as much as it motivates scholars to further research this topic. Specifically, we focus here on whether the measurements obtained across languages and contexts adequately capture the target concept of a given study.[23]

Separate analysis

Facilitating semantic and pragmatic equivalence is difficult with the separate analysis approach. The measurement instruments or text models a researcher applies to their language- and context-specific subcorpora have no information about the measurements obtained for inputs in other languages and contexts. Accordingly, potential information about which documents contain conceptually comparable content is *not* explicitly shared across languages and contexts during the measurement process.

In the worst case, the quantities obtained from the different subcorpora can thus reflect different concepts that are hardly comparable. One way to end up in this worst-case scenario is to fail to account for languages' structural (linguistic) differences during preprocessing.[24] Similarly, separately analyzing a multilingual corpus that records documents from different contexts with an unsupervised method (e.g., one topic model per subcorpus) means that there is no guarantee that separately fitted models extract the same latent dimension or topics from language-specific and context-specific subcorpora (cf. Chan et al., 2020; Lind et al., 2021; Maier et al., 2021). Ensuring that outputs align across languages and contexts is thus difficult.

Accordingly, researchers can only intervene *before* the analysis (when creating their measurement instrument) and *after* it (when interpreting the results). As discussed in Section 2, during preprocessing, this means adopting procedures that accommodate languages' structural differences. In the instrument design step of the research process, this means that researchers select dictionary keywords or labeled training data that are representative of the languages and contexts their data covers. For the interpretation step, this means that researchers, for example, align the topics learned by separate topic models in accordance with their language and context knowledge. As these examples show, all ways to facilitate equivalence across languages and

[23]The procedures for assessing convergent and discriminant validity commonly operate at the level of (aggregate) measurement outputs (cf. Adcock & Collier, 2001) and thus do not depend on the approach taken to overcome language barriers.

[24]In the case of our running example, a researcher who simply tokenizes documents at white spaces would segment noun phrases like "asylum seeker" into two words although the term maps conceptually to the German compound words "Asylsuchende."

contexts with the separate analysis approach are indirect. Hence, a common recommendation is to implement the measurement procedure as consistently across languages and contexts as possible (cf. Esser & Vliegenthart, 2017; Krippendorff, 2004).

With regard to the validation step of the measurement process, adopting a separate analysis approach requires researchers to validate measurements for each language and context.[25] One technique to do this is to validate measurements against human-annotated materials, an external indicator, and/or a benchmark that are representative of all languages and contexts covered by the target corpus. If no external indicators or benchmarks exist, researchers should instruct and train their coders to label evaluation materials in a language-sensitive and context-sensitive way. Generally, researchers can move back and forth between concept definition, measurement instrument design, measurement, and validation to improve the language- and context-sensitivity of their measurement instrument (Lind et al., 2019).

Input alignment

The idea motivating the input alignment approach is that translation or multilingual embedding makes documents' text representations comparable across languages. If semantic equivalence is indicated by whether two documents with similar content receive similar measurements, translating or embedding them goes part of the way to ensure this property because it "removes" the cross-lingual dimension from the comparison.

However, input alignment does *not* guarantee output alignment. Just because documents are translated into the same language or embedded in a joint vector space does not automatically imply that measurements are comparable between languages on a semantic level (cf. Lucas et al., 2015; Maier et al., 2021). As discussed in Section 2 and reemphasized in Section 4.1, one important caveat of the token translation and multilingual word embedding variants is that preprocessing choices can influence how well documents' vocabularies and text representations are aligned across languages. Input alignment should thus not be understood as a silver bullet but as a way to facilitate semantic measurement equivalence.

That said, a number of studies indeed suggest that this facilitating character can be maintained for text types and analytical tasks typically studied in

[25] If languages are not synonymous with cases, the corpus split can be done in a comparative design based on cases and based on languages. For example, a comparison of Canada and Switzerland may mean that five measurement instruments are developed and validated (Canadian context: French and English, Swiss context: French, German, Italian).

political and communication science. For example, de Vries et al. (2018) show that the document-term matrices obtained from machine-translated documents are very similar to those obtained by tokenizing human-translated full texts of the same documents and that the topic models fitted to these matrices yield very similar document-topic and topic-word estimates (see also Courtney et al., 2020; Licht, 2023; Reber, 2019; Windsor et al., 2019).

Similarly, several studies evaluating the multilingual embedding strategy present evidence supporting the assumption that existing embedding models yield well-aligned document representations. Chan et al. (2020), for example, show that dimensionality reduction techniques applied to word embedding-based, multilingual document representations recover meaningful and analytically useful document clusters. Licht (2023), in turn, shows that multilingual sentence embedding enables reliable supervised classification compared to the more common approach of training bag-of-words classifiers using full text-translated texts. This finding suggests that the alignment quality of the pre-trained sentence embedding models he evaluates is at least not worse than that achieved by full-text machine translation.

Moving to pragmatic equivalence, when using machine translation or multilingual embedding, to the best of our knowledge there exist no well-established methods to validate measurements' sensitivity to documents' case contexts. However, there are ways to facilitate context sensitivity. To begin with, researchers should inform themselves about the corpora used to pre-train available machine translation, embedding models, or Transformer models. Ideally, they should select a model that was pre-trained on texts whose language use resembles those the research wants to analyze. We would expect that this helps to minimize the measurement error introduced by the fact that the translation or embedding model being used is not adapted well enough to the language use in the corpus under study. Further, in the case of multilingual word embeddings models, researchers can train their own context-specific models from scratch or continue to train an existing embedding model on their corpus (cf. Rodriguez & Spirling, 2022).[26]

With regard to validation, researchers that have evaluated the input alignment approach have concentrated on a comparison of model outputs to a benchmark, like measurements obtain from human-translated texts (de Vries et al., 2018) or labels obtained by annotating documents in their source languages (cf. Licht, 2023). We second this approach and encour-

[26]Such "domain adaptation" may allow to align the embedding space more strongly with the pragmatic language use in the corpus under study. In contrast, relying on commercial machine translation services prevents researchers from adopting the underlying model.

age researchers to compare the outputs of their models and measurement instruments with suitable benchmarks that reflect experts' language and context knowledge – even if they train or fit their models on monolingual text representations or multilingual embeddings. To validate the outputs of a topic modeling fitted on machine-translated texts, for example, researchers can read the original version of translated documents that were rated as being highly representative for specific topics and compare the original with the translated document (e.g., Lucas et al., 2015).

With this recommendation in mind, it is important to stress that implementing the multilingual embedding approach to input alignment can complicate validation because using documents' embeddings as inputs hinders interpretability. When using documents' representations in an embedding space as inputs to a quantitative text analysis, making an assessment of the relation between textual inputs and model output is difficult because the relation between textual inputs and their embeddings is non-linear. We thus believe that to facilitate the wider adoption of the multilingual embedding approach to input alignment, more research should focus on developing and evaluating approaches to render embeddings-based text analysis methods interpretable.

Anchoring

Turning to the anchoring approach, we find that the idea of using bridging observations to incentivize output alignment is a very promising approach to facilitating measurement equivalence. The polylingual topic model evaluated by Lind et al. (2021), for example, explicitly incentivizes cross-lingual alignment of measurements by constraining the topic distributions of parallel or topically comparable documents to be similar. This contrasts, for example, with the input alignment approach that hinges on the assumption that input alignment through multilingual embedding or translation results in output alignment.

Further, anchoring-based methods that maintain a multilingual representation of model outputs allow researchers to directly assess how their model maps measurements across language barriers. The polylingual topic model, for example, estimates topic representations in each of the input languages. This provides researchers with a direct way to validate whether the topic-word distributions in different languages map equivalent semantic entities.

However, it is important to stress that any anchoring-based method's built-in capability of inducing output alignment does *not* guarantee equiva-

lence. For one, this is because researchers cannot avoid identifying conceptually sensitive preprocessing procedures if the languages in their corpus differ structurally (see Section 2). For another, anchoring does not guarantee equivalence because the alignment quality depends on the "quality" of the bridging observations being used. Unfortunately, we are not aware of research that would allow us to be more specific about what makes "good" bridging observations. Thus, more research is needed to better understand what types and amounts of bridging observations enable anchoring strategies to facilitate cross-lingual measurement equivalence. These questions could be examined based on existing parallel political text corpora but should be expanded to domains where such resources are lacking.

Further, it is currently difficult to assess how such anchoring-based methods perform in measuring context-dependent concepts. In theory, the fact that the polylingual topic model learns language-specific word-topic distributions should give it some leeway to discover language-specific patterns in the data. However, more research is needed to understand whether this feature of anchoring-based topic models, in fact, facilitates context-sensitive measurement. In addition, researchers should try to apply the anchoring idea to other text analysis tasks like scaling or supervised classification.

Conclusion

While communication and political scientists increasingly turn to text-as-data methods to study political behavior and communication, employing these methods to analyze multilingual text collections presents them with challenges they would not face in monolingual analyses (Baden, Dolinsky, et al., 2022; Lucas et al., 2015). We have sought to enable researchers to overcome these hurdles by providing a guide to multilingual text analysis methods. The main goals of our contributions are to aid researchers in navigating, applying, and further developing methods for multilingual quantitative text analysis.

We have emphasized that the main challenges in multilingual text analysis are to bridge language barriers and to facilitate cross-lingual measurement equivalence. We have then discussed the three options researchers currently have to address these challenges: the *separate analysis* of language-specific subcorpora; *input alignment* through machine translation or multilingual text embedding; and *anchoring* through external "bridging observations" such as bilingual lexica, parallel texts, or topically comparable documents. By synthesizing existing methodological research within a unified framework, our discussion provides newcomers and established researchers

who seek to navigate this rapidly evolving and dynamic field with orientation and guidance.

To encourage wider adoption of existing methods in applied research, we have discussed the considerations researchers should factor into their decision when choosing an approach for their application. In particular, we have emphasized that the three approaches differ in the resources their implementation requires and in how they facilitate and enable to assess the validity of cross-lingual measurements.

We conclude by highlighting some limitations of our article that point out our omissions but also directions for future methods research and development. While our focus has been on the use of multilingual text analysis methods in comparative research, these methods can also be applied for other purposes. For example, researchers can use English measurement instruments (e.g., a domain-specific dictionary) to analyze documents written in a so-called "low resource" language that is underrepresented in applied text-as-data research (cf. Lauscher et al., 2020; Wu & Dredze, 2020). As a point in case, Windsor et al. (2019) show that dictionaries developed for the English language can be applied to obtain valid measurements of machine-translated documents originally written in various non-Germanic languages. Similarly, researchers can apply multilingual text analysis techniques to analyze documents that contain text in multiple languages (so-called "code-switching") which is, for example, typical for citizen-produced texts such as online blogs, comments, and social media posts (e.g., Zelenkauskaite & Balduccini, 2017).

Second, beyond our discussion of the work by Glavaš et al. (2017a), Chan et al. (2020), and Licht (2023) on the multilingual embedding approach to input alignment, we have not discussed how recent innovations in deep learning and natural language processing are about to impact the political and communication sciences. We emphasize that an increasing number of studies will leverage large pre-trained multilingual language models like mBERT (Devlin et al., 2019) or XLM-R (Conneau et al., 2020) in their research (Greene & Sylvester, 2022; Laurer, 2023).

Third, multilingual text analysis presents many more challenges than those we have focused on in our discussion. One important challenge is that involving third-party tools in multilingual text analysis can cause "hidden costs." Relying on commercial services for machine translation, for example, imperils reproducibility (cf. Chan et al., 2020). Using pre-trained multilingual language models or the sentence embedding models evaluated by Licht (2023), in turn, comes with the costs of a large carbon footprint (cf. Bender

et al., 2021), potential bias (cf. Bender et al., 2021), and their unreliability in low-resource languages (Wu & Dredze, 2020). Researchers need to keep these costs in mind and should consider how they impact the results of their analyses.

References

Adcock, R., & Collier, D. (2001). Measurement validity: A shared standard for qualitative and quantitative research. *American Political Science Review*, *95*(3), 529–546.

Baden, C., Dolinsky, A., Lind, F., Pipal, C., Schoonvelde, M., Shababo, G., & Velden, M. A. v. d. (2022). *Integrated standards and context-sensitive recommendations for the validation of multilingual computational text analysis* (OPTED Deliverable No. D6.2). Retrieved 2023, from https://opted.eu/fileadmin/user_upload/k_opted/OPTED_Deliverable_D6.2.pdf

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.

Baden, C., & Tenenboim-Weinblatt, K. (2016). Convergent news? a longitudinal study of similarity and dissimilarity in the domestic and global coverage of the israeli-palestinian conflict. *Journal of Communication*, *67*(1), 1–25.

Barberá, P., Gohdes, A. R., Iakhnis, E., & Zeitzoff, T. (2022). Distract and divert: How world leaders use social media during contentious politics. *The International Journal of Press/Politics*, *0*, 19401612221102030.

Baum, M. A., & Zhukov, Y. M. (2019). Media ownership and news coverage of international conflict. *Political Communication*, *36*(1), 36–63.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Boumans, J. W., & Trilling, D. (2016). Taking stock of the toolkit: An overview of relevant automated content analysis approaches and techniques for digital journalism scholars. *Digital Journalism*, *4*(1), 8–23.

Ceron, A., Gandini, A., & Lodetti, P. (2020). Still 'fire in the (full) belly'? anti-establishment rhetoric before and after government participation. *Information, Communication & Society*, *24*(10), 1460–1476.

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.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, É., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2020). Unsupervised cross-lingual

representation learning at scale. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 8440–8451.

Courtney, M., Breen, M., McMenamin, I., & McNulty, G. (2020). Automatic translation, context, and supervised learning in comparative politics. *Journal of Information Technology & Politics*, *17*(3), 208–217.

Dancygier, R., & Margalit, Y. (2020). The evolution of the immigration debate: Evidence from a new dataset of party positions over the last half-century. *Comparative Political Studies*, *53*(5), 734–774.

Denny, M. J., & Spirling, A. (2018). Text preprocessing for unsupervised learning: Why it matters, when it misleads, and what to do about it. *Political Analysis*, *26*(2), 168–189.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv:1810.04805 [cs]*.

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.

Dolinsky, A., Schoonvelde, M., & van der Velden, M. A. C. G. (2022). Multilingualism in computational text analysis methods: Evidence from a pre-registered survey experiment. *2022 Annual Meeting of the European Political Science Association*.

Düpont, N., & Rachuj, M. (2022). The ties that bind: Text similarities and conditional diffusion among parties. *British Journal of Political Science*, *52*, 1–18.

Eberl, J.-M., Meltzer, C. E., Heidenreich, T., Herrero, B., Theorin, N., Lind, F., Berganza, R., Boomgaarden, H. G., Schemer, C., & Strömbäck, J. (2018). The european media discourse on immigration and its effects: A literature review. *Annals of the International Communication Association*, *42*(3), 207–223.

Esser, F., & Pfetsch, B. (2020). Comparing political communication: A 2020 update. In D. Caramani (Ed.), *Comparative politics* (pp. 336–358). Oxford University Press.

Esser, F., & Vliegenthart, R. (2017). Comparative research methods. In *The international encyclopedia of communication research methods* (pp. 1–22). John Wiley & Sons, Ltd.

Fan, A., Bhosale, S., Schwenk, H., Ma, Z., El-Kishky, A., Goyal, S., Baines, M., Celebi, O., Wenzek, G., & Chaudhary, V. (2021). Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*, *22*(107), 1–48.

Gattermann, K. (2018). Mediated personalization of executive european union politics: Examining patterns in the broadsheet coverage of the european commission, 1992–2016. *The International Journal of Press/Politics*, *23*(3), 345–366.

Gentzkow, M., Kelly, B., & Taddy, M. (2019). Text as data. *Journal of Economic Literature*, *57*(3), 535–574.

Glavaš, G., Nanni, F., & Ponzetto, S. P. (2017a). Cross-lingual classification of topics in political texts. *Proceedings of the Second Workshop on NLP and Computational Social Science*, 42–46.

Glavaš, G., Nanni, F., & Ponzetto, S. P. (2017b). Unsupervised cross-lingual scaling of political texts. *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, 688–693.

Goist, M. (2020). *The radical right in parliament: A new method and application for studying political text in multiple languages* (Dissertation). The Pennsylvania State University.

Greene, Z., & Sylvester, C. (2022). Mr BERT goes to parliament: A supervised approach to classifying parliamentary speech in europe. *2022 Annual Meeting of the European Political Science Association*.

Grimmer, J., Roberts, M. E., & Stewart, B. M. (2022). *Text as data: A new framework for machine learning and the social sciences*. Princeton University Press.

Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis, 21*(3), 267–297.

Helbling, M., Hoeglinger, D., & Wüest, B. (2010). How political parties frame european integration. *European Journal of Political Research, 49*(4), 495–521.

Kahditani, A. J. (2022). The function of pragmatics in translation and the pragmatic challenges translators face. *Journal of Language and Linguistics in Society, 2*(5), 48–56.

Klüver, H., & Bäck, H. (2019). Coalition agreements, issue attention, and cabinet governance. *Comparative Political Studies, 52*(13), 1995–2031.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. *Proceedings of Machine Translation Summit X: Papers*, 79–86.

Krippendorff, K. (2004). *Content analysis: An introduction to its methodology* (Second edition). Sage.

Laurer, M. (2023). *Lowering the language knowledge barrier - investigating deep transfer learning and machine translation for multilingual analyses of political texts* (Working paper). https://osf.io/c2fym/

Lauscher, A., Ravishankar, V., Vulić, I., & Glavaš, G. (2020). From zero to hero: On the limitations of zero-shot language transfer with multilingual transformers. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 4483–4499.

Lehmann, P., Burst, T., Matthieß, T., Regel, S., Volkens, A., Weßels, B., Zehnter, L., & Sozialforschung, W. B. F. (2022). Manifesto project dataset. Retrieved 2022, from https://manifesto-project.wzb.eu/doi/manifesto.mpds.2022a

Lehmann, P., & Zobel, M. (2018). Positions and saliency of immigration in party manifestos: A novel dataset using crowd coding. *European Journal of Political Research, 57*(4), 1056–1083.

Licht, H. (2023). Cross-lingual classification of political texts using multilingual sentence embeddings. *Political Analysis, 31*(3), 366–379.

Lind, F., Eberl, J.-M., 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.

Lind, F., Eberl, J.-M., Heidenreich, T., & Boomgaarden, H. G. (2019). When the journey is as important as the goal: A roadmap to multilingual dictionary construction. *International Journal of Communication*, *13*(0), 21.

Lind, F., Heidenreich, T., Kralj, C., & Boomgaarden, H. G. (2021). Greasing the wheels for comparative communication research: Supervised text classification for multilingual corpora. *Computational Communication Research*, *3*(3).

Lucas, C., Nielsen, R. A., Roberts, M. E., Stewart, B. M., Storer, A., & Tingley, D. (2015). Computer-assisted text analysis for comparative politics. *Political Analysis*, *23*(2), 254–277.

Maier, D., Baden, C., Stoltenberg, D., De Vries-Kedem, M., & Waldherr, A. (2021). Machine translation vs. multilingual dictionaries assessing two strategies for the topic modeling of multilingual text collections. *Communication Methods and Measures*, *16*, 19–38.

Mimno, D., Wallach, H. M., Naradowsky, J., Smith, D. A., & McCallum, A. (2009). Polylingual topic models. *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, 880–889.

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

Quinn, K. M., Monroe, B. L., Colaresi, M., Crespin, M. H., & Radev, D. R. (2010). How to analyze political attention with minimal assumptions and costs. *American Journal of Political Science*, *54*(1), 209–228.

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.

Rodriguez, P. L., & Spirling, A. (2022). Word embeddings: What works, what doesn't, and how to tell the difference for applied research. *The Journal of Politics*, *84*(1), 101–115.

Rooduijn, M., de Lange, S. L., & van der Brug, W. (2014). A populist zeitgeist? programmatic contagion by populist parties in western europe. *Party Politics*, *20*(4), 563–575.

Ruder, S., Vulić, I., & Søgaard, A. (2019). A survey of cross-lingual word embedding models. *Journal of Artificial Intelligence Research*, *65*, 569–631.

Ruedin, D. (2013). The role of language in the automatic coding of political texts. *Swiss Political Science Review*, *19*(4), 539–545.

Schmidtke, H. (2018). Elite legitimation and delegitimation of international organizations in the media: Patterns and explanations. *The Review of International Organizations*, *14*(4), 633–659.

Schoonvelde, M., Schumacher, G., & Bakker, B. N. (2019). Friends with text as data benefits: Assessing and extending the use of automated text analysis in political science and political psychology. *Journal of Social and Political Psychology*, *7*(1), 124–143.

Schuck, A., Xezonakis, G., Banducci, S., & De Vreese, C. H. (2010). *European Parliament Election Study 2009, Media StudyEuropean Parliament Election Study 2009, Media Study* (Version 1.0.0). GESIS Data Archive. https://search.gesis.org/research_data/ZA5056?doi=10.4232/1.10203

Schuck, A. R., Vliegenthart, R., & De Vreese, C. H. (2014). Who's afraid of conflict? the mobilizing effect of conflict framing in campaign news. *British Journal of Political Science*, *46*(1), 177–194.

Shababo, G., & Baden, C. (2023). Language matters: How linguistic differences impact computational text analysis methods. *73rd ICA Annual Conference*.

Slapin, J. B., & Proksch, S.-O. (2008). A scaling model for estimating time-series party positions from texts. *American Journal of Political Science, 52*(3), 705–722.

Strömbäck, J., Meltzer, C. E., Eberl, J.-M., Schemer, C., & Boomgaarden, H. G. (2021). *Media and public attitudes toward migration in europe: A comparative approach*. Routledge.

Theocharis, Y., Barberá, P., Fazekas, Z., Popa, S. A., & Parnet, O. (2016). A bad workman blames his tweets: The consequences of citizens' uncivil twitter use when interacting with party candidates. *Journal of Communication, 66*(6), 1007–1031.

Tiedemann, J. (2012). Parallel data, tools and interfaces in OPUS. *Proceedings of the Eighth International Conference on Language Resources and Evaluation*, 2214–2218.

Tiedemann, J., & Thottingal, S. (2020). OPUS-MT – building open translation services for the world. *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, 479–480.

van der Veen, M. (2022). Translation for the rest of us: The value of word-level machine translation. *COMPTEXT 2022*.

van Atteveldt, W., & Peng, T.-Q. (2018). When communication meets computation: Opportunities, challenges, and pitfalls in computational communication science. *Communication Methods and Measures*, *12*(2), 81–92.

van Deth, J. (1999). Equivalence in comparative political research. In J. van Deth (Ed.), *Equivalence in comparative politics* (pp. 1–19). Routledge.

Windsor, L. C., Cupit, J. G., & Windsor, A. J. (2019). Automated content analysis across six languages. *PloS One*, *14*(11), e0224425.

Wu, S., & Dredze, M. (2020). Are all languages created equal in multilingual BERT? *Proceedings of the 5th Workshop on Representation Learning for NLP*, 120–130.

Zelenkauskaite, A., & Balduccini, M. (2017). "information warfare" and online news commenting: Analyzing forces of social influence through location-based commenting user typology. *Social Media + Society, 3*(3), 205630511771846.

Zhang, D., Mei, Q., & Zhai, C. (2010). Cross-lingual latent topic extraction. *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, 1128–1137.