www.reminder-project.eu

REMINDER

ROLE OF EUROPEAN MOBILITY AND ITS IMPACTS
IN NARRATIVES, DEBATES AND EU REFORMS

A Bridge Over the Language Gap: Topic Modelling for Text Analyses Across Languages for Country Comparative Research

WORKING PAPER

Authors: Fabienne Lind
Jakob-Moritz Eberl
Sebastian Galyga
Tobias Heidenreich
Hajo G. Boomgaarden
Beatriz Herrero Jiménez
Rosa Berganza

Published: November 2019

universität wien

Universidad Rey Juan Carlos

REMINDER

ROLE OF EUROPEAN MOBILITY AND ITS IMPACTS
IN NARRATIVES, DEBATES AND EU REFORMS

A Bridge Over the Language Gap: Topic Modelling for Text Analyses Across Languages for Country Comparative Research

Authors: Fabienne Lind[1], Jakob-Moritz Eberl[1], Sebastian Galyga[1], Tobias Heidenreich[1], Hajo G. Boomgaarden[1], Beatriz Herrero Jiménez[2], Rosa Berganza[2]

[1] University of Vienna
[2] Universidad Rey Juan Carlos

Submitted: August 2019

Paper prepared as part of the REMINDER project

www.reminder-project.eu

Correspondence address:

Fabienne Lind, University of Vienna, Department of Communication
Rathausstraße 19, 1010 Vienna, Austria, Mail: fabienne.lind@univie.ac.at

This project has received funding from the European Union's Horizon 2020
research and innovation programme under grant agreement No 727072

An Introduction

The European Union (EU) is a prime example for a multilingual democratic international organisation with equally multilingual political and media environments. The EU was founded on the principle of multilingualism, and the number of official languages grew with the inclusion of new member states. Nowadays, the EU has 24 official languages as well as over sixty indigenous regional and/or minority languages. Among other objectives, the EU aims to protect member states' rich linguistic diversity, and to promote language learning in Europe so as to facilitate free movement (of students/workers/other people) across member states. Furthermore, the EU's policy is to seek to communicate with citizens in their own native languages, and accordingly, a vast number of texts produced by the EU and its member states every day come in multiple languages.[1]

However, the sheer number of texts and their sometimes multilingual character make the understanding of such texts and their reception in different member state contexts a task that is far beyond what human readers are capable of processing. This is where computer-assisted text analysis methods may prove fruitful. These are commonly divided into dictionary methods, semi-automated (or supervised) and automated (or unsupervised) approaches. We will here focus on the latter, as such a methodology does not necessarily require a basic understanding of all languages in a given corpus.

Topic modelling, as a bottom-up text mining approach, has become more and more popular in the social sciences, as it facilitates the discovery of themes in large quantities of textual data with comparably little effort. Yet, so far, automated methods of content analysis (such as topic modelling), are usually applied to text documents in just one language – mostly English (Boumans & Trilling 2016; Pang & Lee 2008). This usage does neither correspond with the digital availability of texts in different languages, nor with the plethora of

[1] See https://europa.eu/european-union/about-eu/eu-languages_en

substantive comparative research interests. Moreover, it favours large languages and countries with more developed research infrastructures, corroborating already existing divides between, for example, Western and Eastern European countries (e.g., Directorate-General for Research and Innovation 2014: 11; also Eberl et al. 2018).

Recent topic modelling contributions in the social sciences have dealt with the challenge of multilingual data (i.e., the analysis of a corpus that includes multiple languages) by (1) relying on expensive machine translation routines, consolidating the data based on a target language and thus analysing data in just one language rather than many (e.g. Lucas et al. 2015), or (2) running separate topic models for each language, instead of just one, with all the subsequent methodological difficulties and limitations in comparing the topic model results (e.g. Heidenreich, Lind, Eberl, & Boomgaarden, forthcoming).

Outside of the social sciences, researchers have found ways to bridge the language gap without the use of expensive translation and without having to process the different languages separately (Vulić, De Smet, Tang, & Moens 2015). Mimno, Wallach, Naradowsky, Smith, and McCallum (2009), for example, proposed the "Polylingual Topic Model" that makes use of document connections such as online links between documents, which are not directly translated but cover the same topics (i.e. Wikipedia articles on the same topic in different languages).

Against this background, this paper presents a comprehensive overview of different methodological strategies to conduct computer-assisted text analysis across languages. In contrast to previous work (see Lind et al. forthcoming),[2] where we concentrated on dictionary methods (a top-down approach), here we shift the focus to topic modelling and discuss its usability for comparative research. The paper proceeds as follows: (1) We will give a general overview of the intricacies of computer-assisted text analysis of multilingual

[2] See also Deliverable 8.2.

data and (2) present the basic topic modelling approach. We then turn to a more elaborate introduction of topic models for multilingual text corpora, where we present (3) more basic approaches, and (4) more advanced topic modelling approaches. We end with a discussion of the potential and limitations inherent to topic modelling strategies applied to multilingual texts.

For this paper, we use data generated by Work Package 8 of the REMINDER project – whenever suitable – for hands-on examples.

Computer-Assisted Text Analysis across Languages

Questions about differences and similarities between different countries/regions/systems are at the centre of comparative research. When such comparisons are made on the basis of text documents, the methodical handling of language becomes the focus. This is especially true when relevant textual data is written in more than one language. To get to the heart of the problem: the different original languages of the texts prevent simple, immediate comparison and the direct application of common analytical methods. These problems are non-trivial. Formulating the challenge more clearly: to achieve meaningful results with a comparative text analytical study, different languages with their diverse intricacies and complexities need to be accommodated first.

When relying on dictionary methods or supervised methods these accommodation efforts focus mostly on the required pre-defined input. The input, keywords (i.e. for dictionary methods) or annotated material (i.e. for supervised methods) must be created for each language individually, considering country-/region-/language-specific particularities.

Alternatively, it is possible to rely on (machine) translation to consolidate the language of the input or the language of the textual material.[3]

Both methods share a deductive or top-down character and are useful for studying already defined categories in languages that one has a basic understanding of (Boumans and Trilling 2016: 14/15). This means that researchers have to have a good knowledge of their case of interest beforehand. However, such knowledge is not always available, and concepts may have to be adapted several times during the process. To avoid this problem, other approaches to computer-assisted text analysis – such as topic modelling – follow a bottom-up approach instead. Here, accommodating the diverse intricacies and complexities of different languages takes place using other strategies. In order to understand them, we will now first give a short introduction to regular topic modelling, i.e. a methodology originally designed for modelling monolingual text.

Classic Topic Modelling: An Overview

Topic modelling is an unsupervised machine-learning approach and as such an inductive (i.e., bottom-up) way of automatically discovering coherent themes in a text corpus (e.g., Boumans & Trilling 2016; Jacobi, van Atteveldt, & Welbers 2016). These themes or “topics” are clusters of words that are more likely to appear together in documents within a corpus. Theoretically, “topics” are hard to define (Maier et al. 2018: 95). As Günther and Domahidi (2017: 3057) point out, topic modelling is not an automated measure of topic categories as traditionally understood in a manual content analysis (e.g., different policy issues). In its original sense, the term “topic” stems from psycholinguistic approaches to how texts are comprehended and describes “what is being talked/written about” (Günther and Domahidi 2017: 3, citing Brown and Yule 1983: 73). Lacking a more substantive theoretical definition,

[3] For a comprehensive discussion of these methodological strategies see for example Proksch and colleagues (2019) and Lind and colleagues (forthcoming).

then, the “topic” retrieved remains primarily an empirical assessment (Jacobi et al. 2016: 91), which has to be connected to theoretical frameworks such as framing theory or political/societal “issues” (Maier et al. 2017, 2018). In short, a topic model helps the researcher to inductively “capture the salient themes that run through the collection” (Blei, Lawrence, & Dunson 2010: 55) of documents.

In the social sciences, topic modelling has primarily been used to identify issues or frames discussed in text corpora, as well as to measure their salience and track their development over time. Among many different applications, it has been used to analyse news media articles (Jacobi et al. 2016; Krestel & Mehta 2010), social media content (Paul & Dredze 2014; Nguyen & Shirai 2015; Pennacchiotti & Gurumurthy 2011), parliamentary press releases (Greene and Cross 2017; Grimmer 2010; Nguyen & Shirai 2015), or scholarly publications (e.g., Günther and Domahidi 2017; Mann, Mimno, & McCallum 2006). Investigating migration in particular, studies have analysed the framing of refugees in European news media (Heidenreich et al. forthcoming), the substantive topics of comments below YouTube Videos about the European refugee crisis (Lee & Nerghes 2018), or discursive strategies in immigration-related discussions in a Swedish Facebook group (Merrill & Åkerlund 2018).

When referring to a topic model without further specification, one probably uses one of the so-called “basic” models. There is the Probabilistic Latent Semantic Analysis (PLSA) model (Hofmann 1999), which we will briefly refer to at a later point, and the widely used Latent Dirichlet Allocation (LDA) model (Blei, Ng, & Jordan 2003). Without going into too much detail – an accessible explanation of the assumptions and specifications of LDA can be found in Blei (2012) – topic models like LDA are algorithms designed to extract the “topics” that pervade a text collection. A topic in LDA is empirically defined as being a distribution over a fixed vocabulary (Blei 2012: 78). Furthermore, LDA assumes that all the documents in a text corpus share the same set of topics. The documents are however distinct from each other, as each document exhibits the topics in different proportions. LDA takes a Bayesian approach. The task of the LDA algorithm is to infer the hidden topic structure from the documents, which is the statistical task of computing the conditional distribution of the hidden variables given the documents of a corpus.

In general, topic modelling has proven highly fruitful for further development by researchers. It has led to a wealth of different topic models allowing the accommodation of diverse research interests. As such, it has been further developed to allow the incorporation of covariates on the document level (e.g. the Structural Topic Model [STM], Roberts, Brandon, & Tingley 2014). Of particular note for our interests here are current research efforts that seek to enhance the applicability of LDA for text analysis across languages so as to make it fit for answering comparative research questions.

That is to say, "vanilla"[4] LDA is of limited use for a cross-country research design, where the goal often is to understand if and how frequent the same or similar topics are discussed in different countries. To show how futile topic modelling of multilingual data can be without further model specification, we present a brief analysis and two illustrations below.

For demonstration purposes, we apply vanilla LDA topic modelling on multilingual text data. We take a multilingual corpus of Spanish, English, German, Swedish, Polish, Hungarian, and Romanian news articles about migration (*n* = 44,328), define – for example – seven (*K* = 7) as the number of topics that we believe to be present in the corpus, and, perform LDA modelling[5]. The results are displayed in Figure 1 below.

[4] Classic LDA (Blei et al. 2003).

[5] Details on the text documents, pre-processing steps and model specifications: the corpus consists of Spanish, English, German, Swedish, Polish, Hungarian, and Romanian news articles about migration published between January 1, 2014 and December 31, 2017. The corpus was gathered with validated search strings, duplicated articles were removed, a stratified sample was drawn (for details see: Eberl et al., 2019). The cleaned multilingual corpus contains Spanish (*n* = 3,204), English (*n* = 8,627), German (*n* = 10,386), Polish (*n* = 3,512), Swedish (*n* = 4,826), Hungarian 9,575(*n* = 9,575), and Romanian (*n* = 3,198) articles. Pre-processing steps: Lemmatization and part of speech tagging using R package udpipe (Wijffels 2018), we selected only nouns in

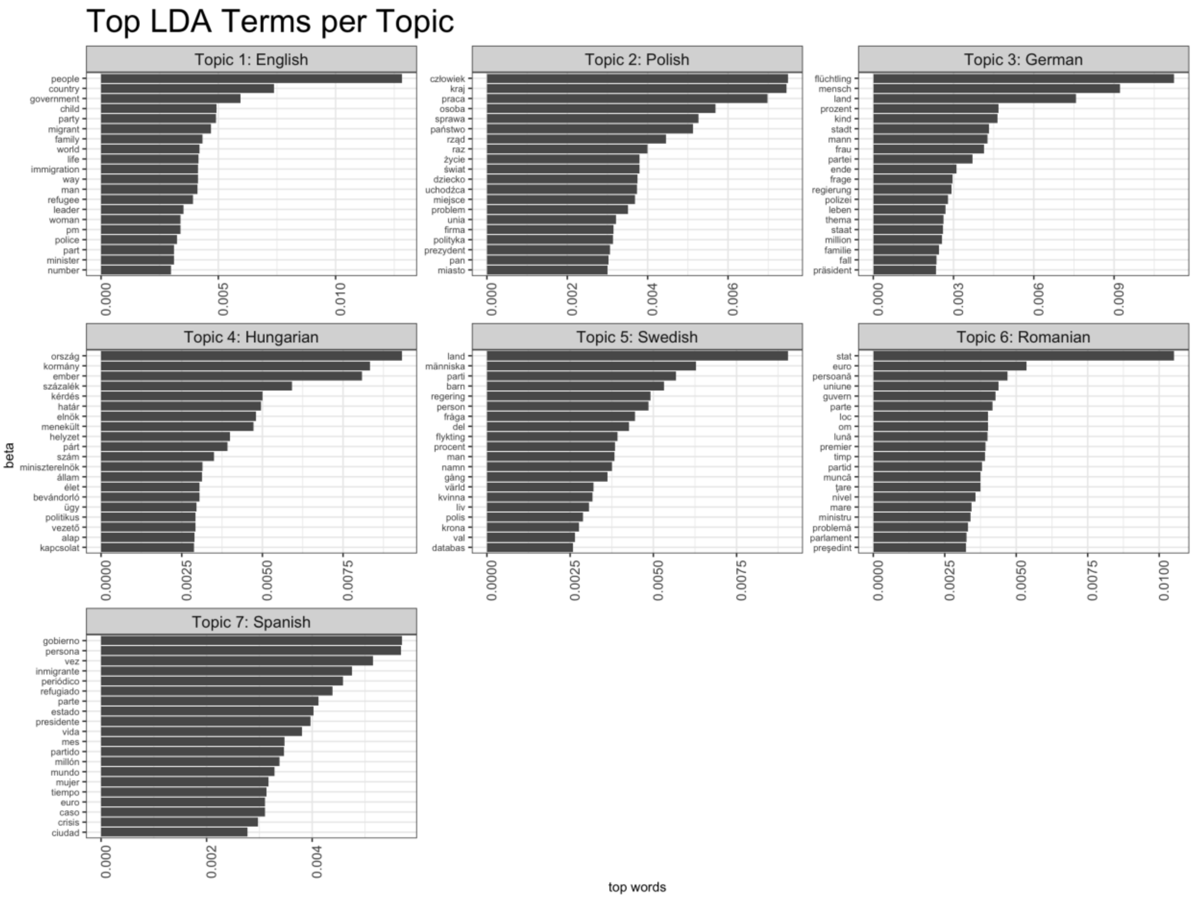


Figure 1. Top 20 terms per topic using a basic LDA model on a multilingual corpus.

Figure 1 mainly shows how the top terms of each topic contain terms in one language at a time. For example, English words are clustered in Topic 1 and Polish words in Topic 2, etc. The topics do therefore not represent substantive themes throughout the topics, but the

their lemmatized form (for a similar approach see: Jacobi et al. 2015), tokens were lowercased, remaining punctuation, urls, and numbers were removed; relative pruning (see Maier et al. 2018): features that appeared in more than 99% or less than 5% of all documents were removed. LDA model specifications: *K* = 7 topics; Gibbs sampling method, 1000 iterations (see Mayer et al. 2018).

different languages, in which the texts within the corpus are written. We thus named the topics according to the respective language they capture.

The R Package LDAvis (Sievert & Kenneth 2014) allows for another visualization of the same topic model (Figure 2). Interestingly, in this case topics (i.e., languages) with overlap in vocabulary (Romance languages: Spanish, Romanian), (Germanic languages: Swedish, German, English) are put closer together. Conversely, it makes sense that the topic for “Polish”- the only Slavic language - is displayed further away. Knowing that Hungarian is a Uralic language with little connection to the Indo-European languages (here, Romance, Germanic and Slavic languages) explains why the Hungarian topic is most distant from the other topics.

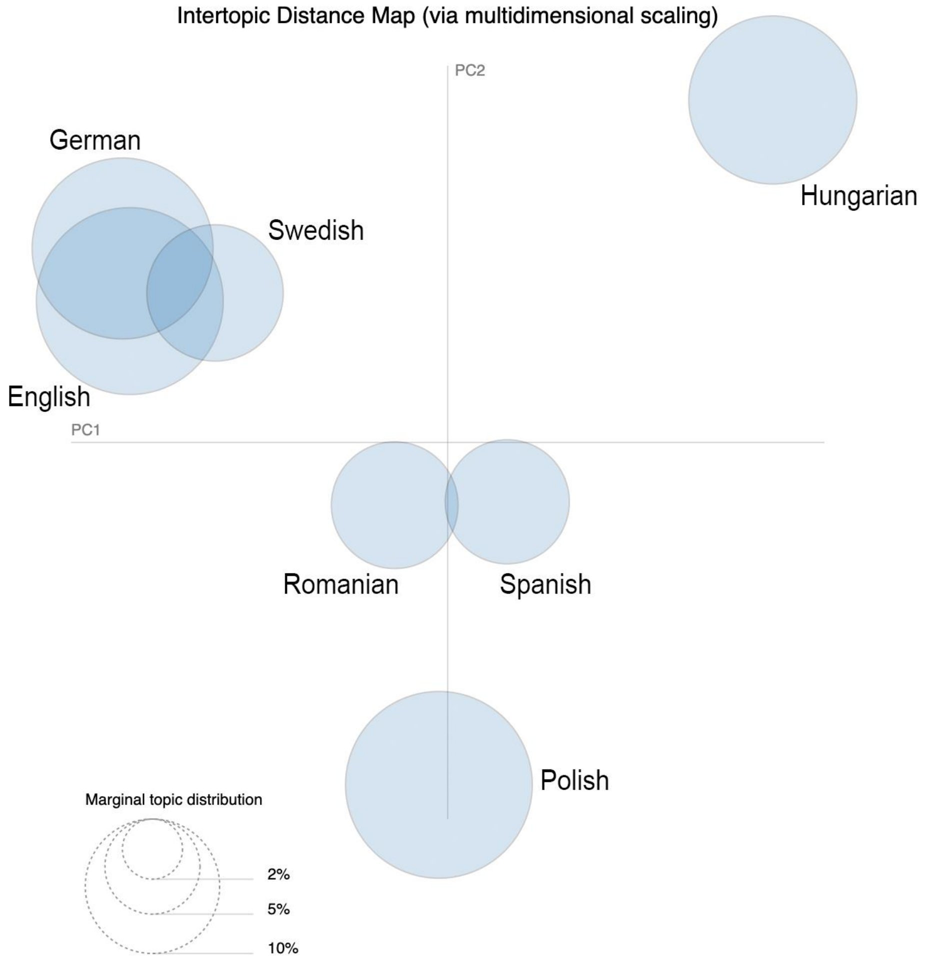


Figure 2. Intertopic distance map created with the R package LDAvis. The visualization implements a complex distance measure called Jensen-Shannon divergence on the two main principal components that arise from the list of words a topic contains. This figure allows for two main interpretations: First, it shows how dominant each topic is in the corpus of data. Second, it shows the relation of the topics regarding their similarity. Each circle in the plot represents an individual topic with the size of the circle corresponding to the prevalence of the topic in the whole corpus. Put plainly, the larger the area of a circle the higher the number of documents the represented topic appears in. The second feature, the difference of topics, is visualized by the proximity and distance two circles exhibit in the two-dimensional space of the diagram.

While of little substantive use in this case, the LDA algorithm itself actually did exactly what it was supposed to. As LDA relies on co-occurrences of words, and words in different languages generally do not tend to occur together in the same document, the model will recognize languages as topics. The model cannot recognize that the German word “partei” and the English “party” are semantically the same. While this simple approach is therefore an effective method to sort documents in a multilingual corpus by language, most questions in comparative social science require smarter strategies to incorporate the more basic topic models such as vanilla LDA or STM, or alternatively command the consideration of more advanced models with additional language-related specifications.

Basic Topic Modelling Approaches for Cross-Country Research

What methodological techniques have been used so far to apply more basic topic modelling approaches like LDA and STM for comparative research? We conducted a systematic literature review to answer this question.

Aiming to identify studies using topic modelling for country comparative research, we started by assessing the 20 studies examined in a recent systematic review by Maier and colleagues (2018: 115-116), a set representing communication science articles published up to May 2016 that apply LDA topic modelling. Interestingly, all of these studies are based on monolingual data and none of them are aimed at cross-country comparisons. In fact, language or cross-country aspects are only briefly mentioned in two studies. (1) Koltsova and Koltcov (2013) applied LDA topic modelling to blog posts from a Russian blog platform. The analysed material was foremost in Russian, but a few blog posts were in Ukrainian and English, which resulted in Ukrainian and English terms being clustered together in “language topics” (p. 218) and could not further be considered for any substantive interpretation. (2) Elgesem, Feinerer, and Steskal (2016), who base their analyses on English text produced by bloggers from several countries, critically discussed that even under the assumption of a transnational blogosphere, it may be beneficial to include local political contexts and thus a country/system comparative component (p. 188). Is sum, it seems that although topic modelling has found widespread use in communication science, the approach has not often been used for country comparative research questions. Given this finding, we extended our search to other disciplines and conducted a systematic review based on a keyword search on Web of Science.[6] We retrieved *N* = 75 possibly relevant studies. Only retaining studies that describe and explore differences and similarities between countries via topic modelling approaches[7] left us with *N* = 10 studies. We then coded these studies regarding the a) type of text documents used in their analysis, b) the language(s) of text documents, c) the topic model used, d) the purpose of the analysis, and e) the number of countries studied. See Table A1 (Appendix) for the detailed results of our study annotation.

In sum, while keeping language issues to a minimum and still using largely basic topic model algorithms, the reviewed studies apply three different strategies to compare topics across countries. Studies either (a) use corpora combining monolingual texts stemming from different countries, (b) compute one topic model per language, or (c) translate documents into one target language.

A: Monolingual Text Selection

[6] Web of Science: All categories (= scientific disciplines/subfields); search in title, abstract, author keywords, and Keywords Plus; timespan: All years, all document types; Search string: TS = (("topic model" OR "topic modelling") AND ("comparative" OR "countr*" OR "cross-national*" OR "cross national")); The search was performed on July 19, 2019.

[7] Most studies were excluded because "compare" in the abstract did not refer to a country comparison. Some studies applied topic modelling to an English corpus from authors with different nationalities (i.e. scientific publications), but because the interest in analysis subsequently did not refer to country comparisons, they were also excluded.

The first set of studies (*n* = 6) approaches this issue by naturally holding the language of documents constant. Here one option is to select only countries that share the same official language (e.g. English tweets from the United States, Canada, Britain, and South Africa, Abdelwahab, Robles, Chiru, & Rebedea 2014). The second option is to work with text types that are naturally published in the same language across several countries with different official languages. Examples for such text types are scientific publications – written by authors from different countries but all in English – (Hassan & Haddawy 2015; Kim, Hong, & Jung 2019), international student reports – a dataset by students from 167 countries but filtered for exclusively English language documents (Perez-Encinas & Rodriguez-Pomeda 2019), English blog posts written by tourists from different countries (Rahmani, Gnoth, & Mather 2018) or English news articles with audiences in various countries, such as international wire services, news outlets like *The New York Times International*, *Le Monde International* or national English newspapers such as the *Shanghai Daily* in China (see: Jiang et al. 2017).

B: One Topic Model per Language

A second set of studies (*n* = 3) runs several topic models – one per country/language – to conduct country comparative research. Here, the multilingual corpus is first divided into country/language specific sup-corpora. Each sub-corpus is then modelled with its own topic model. The main challenge of this approach is the subsequent comparison of topic model outputs, since the resulting models may not generate the same number of topics, and topics within the different sub-corpora may not necessarily overlap conceptually. To put it differently, “when learning two models independently, we cannot guarantee that the topic representations will be comparable” (De Smet & Moens 2009: 57). The reviewed studies have different strategies to deal with this issue. Zheng and colleagues (2014), for example, first apply an LDA topic model for Japanese blog posts and another for Chinese blog posts. They then manually label the produced topics separately (thus not knowing the output of the other model, respectively). Afterwards, the labelled topics are then classified in three categories (topics found only in Japanese blog posts, topics found only in Chinese blog posts, and topics found in both). Only topics that are in that third category are directly compared to each other. See also Chen, Liu, Wei, Yan, Hao, and Ding (2017) and Sakamoto and Takikawa (2017) for other strategies.

Another study that chose this approach for the analysis of REMINDER media data is the content analysis by Heidenreich and colleagues (forthcoming). LDA models were run separately for Spanish, English, German, Swedish, and Hungarian news articles all dealing with refugee and asylum discourses. First, topics within individual models were labelled with the help of native speakers and country experts. Then, labels were harmonized whenever topics appeared to refer to similar topics across languages/models. See results of these topic models in Figure 3 below.

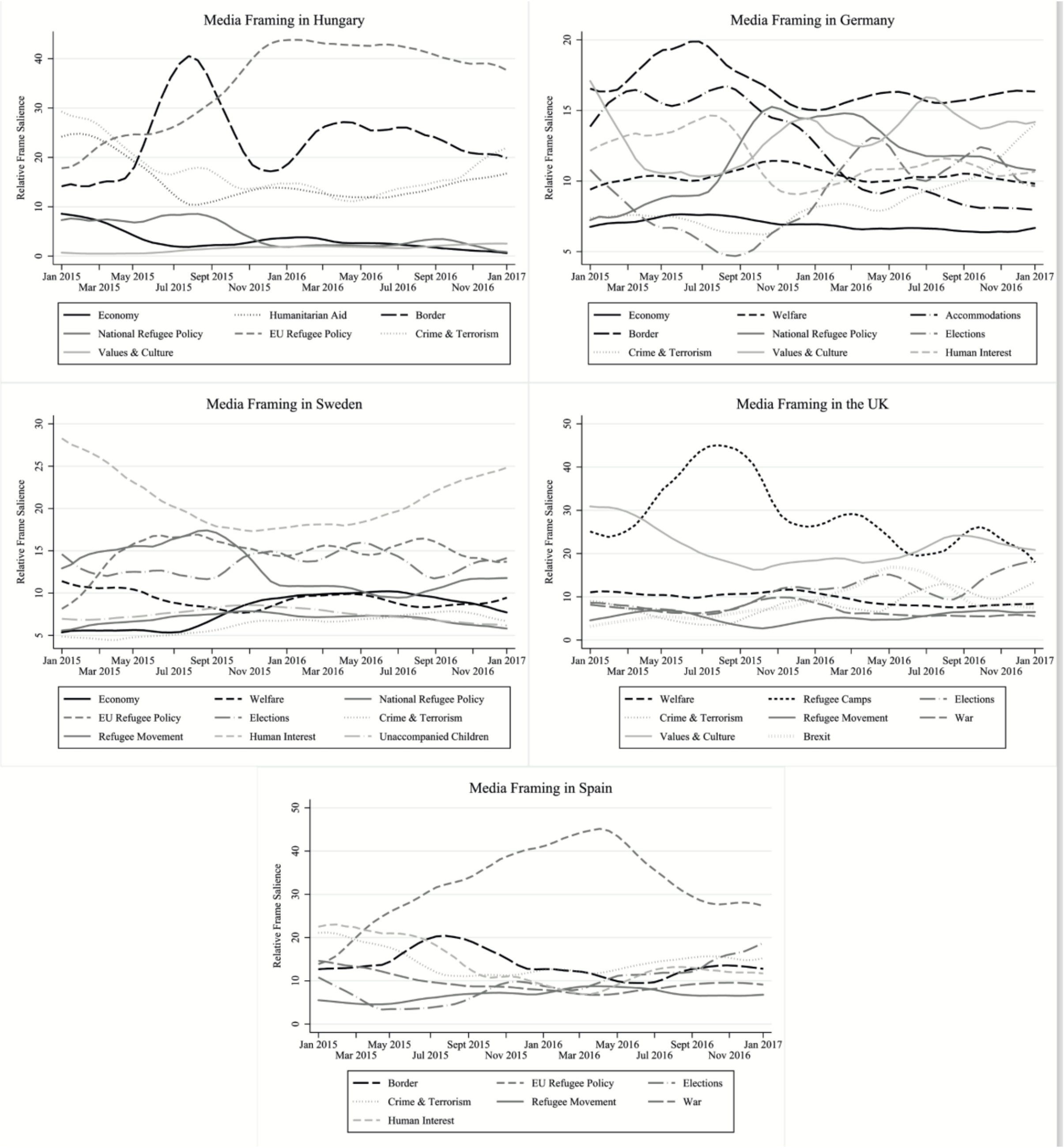


Figure 3. Heidenreich and colleagues (forthcoming) Figure 2 entitled Dynamics of Refugees Framing in Europe

C: Document Translation

The third strategy (applied by only one study in our sample) is to machine translate the multilingual text documents into a common (target) language, and to use the original language of a document as a covariate for Structural Topic Modelling (STM) (Roberts et al. 2014). Lucas and colleagues (2015) introduce this strategy for comparative political text analysis and demonstrate its application for Chinese and Arabic social media posts. They first translate the Chinese and Arabic texts with a machine translation software into English. Note that English (a third language) is chosen, among other reasons, to expose each text corpus to the same level of machine translation (p. 269). The next step is then to fit one STM topic model to the resulting English language corpus. STM is used as it allows the introduction of covariates at the document level. Such a covariate (i.e. content covariate related to the original language of a document) can account for the circumstance that a machine translation software may make different mistakes for each language.

On a practical note, we would like to point to some recent findings comparing term-document-matrices and LDA topic model results for human-translated documents and machine-translated documents. In fact, De Vries, Schoonvelde, and Schumacher (2018) recommend machine translation. Furthermore, while the machine translation software itself (comparing Google Translate vs. Deepl) plays a subordinate role, full document translation is preferred over the mere document-term matrix translation (Reber 2019).

Advanced Topic Modelling Approaches for Analysis Across Languages

A more advanced set of algorithms to model topics in multilingual text corpora are referred to as “Multilingual Probabilistic Topic Models” (MuPTM-s) (Vulić et al. 2015). Currently, MuPTM-s are not common for comparative research in the social sciences. However, we believe that they have great potential for future application, and therefore will use this chapter to introduce them. For this purpose, we now provide an overview of different types of MuPTM-s and explain how they work. Table A2 (Appendix) lists all mentioned models.

We then focus on application cases and refer to some recent work in computer science with the potential to push forward its applicability for social science research more generally and comparative social science in particular.

The Special Features of Multilingual Probabilistic Topic Models

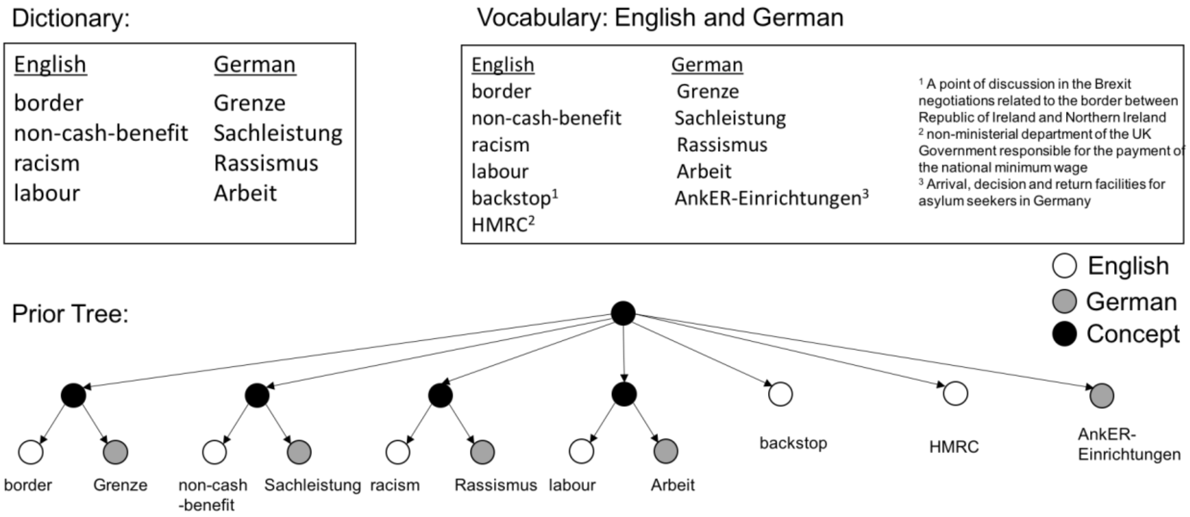
In brief, MuPTM-s are “trained on the individual documents in different languages, and their output are joint latent cross-lingual topics in an aligned latent cross-lingual topical space” (Vulić et al. 2015: 123). In contrast to previously presented models, MuPTM-s need no document translation. They also apply only one topic modelling algorithm to model the multilingual corpus. In fact – and this is what makes them special – MuPTM-s manage to represent documents in different languages “within the same vector space” (Ni, Su, Hu, & Chen 2009: 1155), irrespective of the fact that they may originate from different language corpora. This allows them to identify cross-language latent topics within the texts. Extracting common topics shared in multiple languages – without the need for manual post-hoc matching of topics by researchers – is probably their most promising feature (Zhang Mei, & Zhai 2010: 1128).

Bridging the Chasm Between Languages

MuPTM-s employ different strategies to “bridge the chasm between languages” (Hu, Zhai, Eidelman, & Boyd-Graber 2014: 1166). Bridging refers to the strategy used to “tie the languages together” (Boyd-Graber & Blei 2009: 75). In the following, we introduce two different strategies for such bridging. The strategies rely in fact on two different types of alignment information: (1) pre-existing lexical resources and (2) a pre-selection of topically comparable documents. MuPTM-s usually use one of the two kinds of alignment information.

Tying languages together via lexical resources

The first class of models relies on word-level information such as lexical resources (e.g. bilingual dictionaries) to connect different languages. Dictionaries serve as input for the model, inducing the 'clues' for alignment on the vocabulary level and guiding the topics. Hu and colleagues (2014), who propose the "Polylingual Tree-Based Topic Model" (ptLDA), give an easily-understandable technical explanation of the role of bilingual dictionaries in tree-based topic models. We adopted their graphical example (p. 1169) and transferred it to an example for English and German migration-related texts (Figure 4). Note that semantically equivalent words – as defined by a dictionary – (e.g., EN: racism, DE: Rassismus) are grouped into one concept. All concepts are then connected to one root node. Words that are not included in the dictionary (e.g. Figure 4: "DE: AnkER-Einrichtungen") have also a direct link to the root node. This resulting structure is called prior tree, and serves as prior for topic models. Words that are grouped in one concept will share the probability of being in a topic.



Dictionary:

English	German
border	Grenze
non-cash-benefit	Sachleistung
racism	Rassismus
labour	Arbeit

Vocabulary: English and German

English	German
border	Grenze
non-cash-benefit	Sachleistung
racism	Rassismus
labour	Arbeit
backstop[1]	AnkER-Einrichtungen[3]
HMRC[2]	

[1] A point of discussion in the Brexit negotiations related to the border between Republic of Ireland and Northern Ireland
[2] non-ministerial department of the UK Government responsible for the payment of the national minimum wage
[3] Arrival, decision and return facilities for asylum seekers in Germany

Figure 4: Adopted by Figure 8.1 from Boyd-Graber, Hu, & Mimno (2017: 96), here transferred to an example for migration-related texts

There are at least three other models in addition to the ptLDA (Hu et al. 2014) that rely on bilingual dictionaries to bridge the languages: (1) The "Multilingual Topic Model" (MUTO), which operates with word matchings induced among others by dictionaries and orthographic features (Boyd-Graber & Blei 2009). (2) The "JointLDA" model (Jagarlamundi & Daumé III 2010), which – in contrast to other models – allows one word to be grouped with multiple words in the other language. (3) The "Probabilistic Cross-Lingual Latent Semantic Analysis" (PCLSA) model (Zhang et al. 2010). Unlike to the MUTO model and JointLDA, the PCLSA model is not an extension of LDA but of the "Probabilistic Latent Semantic Analysis" (PLSA).

Tying languages together via topically-comparable text documents

This second class of models ties the languages by making use of pre-selected training documents in different languages that are topically comparable. An important task is therefore to obtain such topically comparable corpora as an input in the model. For example, such inputs can be direct translations of documents (e.g., the Europarl corpus, a collection of European Parliament proceedings in 11 languages; Koehn 2005). Other possibilities are to obtain comparable documents in different languages through shared named entities (e.g. Montalvo et al. 2007; Vulić et al. 2015) or shared time stamps (e.g. Wang, Zhai, Hu, & Sproat 2007). The expectation is that sharing named entities (i.e. persons, organizations, locations) or time stamps (e.g. news articles' publication date) increases the probability that documents have content in common or are somehow related. Another frequently-used topically-comparable corpus is the Wikipedia collection. Wikipedia can be seen as a database that has a large set of pre-selected articles on the same subjects in multiple languages. For example, while the English and Spanish Wikipedia articles about the European migrant crisis are not an exact translation of each other, they share a thematic similarity – and are thus topically-comparable text documents.

Two similar but independently developed models that require topically-comparable text collections are the "Polylingual Topic Model" (PLTM) (Mimno et al. 2009) and the

"Multilingual LDA" model (ML-LDA) (Ni, Su, Hu, & Chen 2009, 2011).[8] Both models are extensions of LDA (Blei et al. 2003). Relying on the model presentation by Mimno and colleagues (2009: 881-882), we here briefly introduce some of the more technical backgrounds of PLTM.

In the PLTM framework, the model input – again the topically-comparable text documents – are referred to as "tuples". One tuple holds one document per language (in most cases), and thus basically pairs sets of vocabulary in different languages. A basic model assumption is that all documents in a tuple share the same tuple-specific distribution over topics. Also, each topic is assigned a distinct topic-word distribution, one per language. By requiring documents within the same tuple to share a distribution over topics, the model is able to identify the words associated with a given topic across languages. The two basic output sets of PLTM are probability distributions. The first is a set of word distributions in all languages that depict the shared topics across languages. One multilingual topic is then basically a set of word lists, where each word list represents the topic version for a different language. The second output set are per-document topic distributions, which may be used to examine the relative salience of identified shared topics in each language. When applying PLTM, the main steps are to first train a topic model with training documents, second to evaluate the topic model with test documents, and third to infer topics for new documents.

[8] On a side note, the bilingual LDA (BiLDA) model shares the basic structure and assumptions with PLTM but is designed for documents in two languages (for a comprehensive discussion, see Vulić et al. 2015).

Use Cases

Thus far, MuPTM-s have been employed for various tasks. In computer science or library science, these models are widely used for improving machine translation (Koehn 2009; Boyd-Graber, Hu, & Mimno 2017; Krstovski & Smith 2013), search engine optimization (Jiang, Tong, & Song 2016), bilingual dictionary extraction (Liu, Duh, & Matsumoto 2015), or cross-lingual information retrieval (Vulić, De Smet, & Moens 2013).

The tasks that appear most promising for a wide range of applications in the social sciences are cross-lingual document exploration, event detection, and document classification. Document exploration refers, for example, to facilitated categorizing and summarizing of large news collections available in multiple languages (e.g. Ni, Sun, Hu, & Chen 2011). The goals of event detection might be to organize a multilingual corpus based on common topics/events (Jagarlamundi & Daumé III 2010) or to link stories across languages that report on the same event. (e.g. see De Smet & Moens 2009).

Conclusion

None of the approaches described offers the perfect strategy. Both the basic and more advanced approaches have weaknesses. When following a topic modelling approach in comparative social science, the use of the more basic approaches has so far been preferred. Lucas and colleagues (2015), briefly mentioning MuPTM-s, provide reasons for why these models have not yet gained a foothold in the social sciences. While we cannot eliminate these concerns entirely, we will now discuss them, and ultimately emphasise how useful MuPTM-s may be regardless of their limitations.

For a comparative assessment, let us first briefly explain the downsides of the three more basic approaches described above. The strategy (A) to select only monolingual text, is very much limited by its dependency on such monolingual but cross-country or international text resources. Research questions are restricted to countries with a shared language. Moreover, such very specific and often elite international text resources may strongly restrict the generalizability of results. When relying on approach (B), running one topic model per language, comparisons between languages are deemed possible, but they are made only very cautiously and to only a limited extent. As Heidenreich and colleagues (forthcoming) note, “direct comparability of topics across countries is limited, as some topics may be similar but not the same” (p. 8) across languages. Finally, one decisive disadvantage of strategy (C), document translation, retains the costs of proper machine translation, especially when working with large corpora. In sum, although they are used with some regularity, the more basic approaches are all strongly limited either in restricted data selection options, possible country comparisons and the degree of comparability of models, or by possible errors and imprecisions due to document translation.

Which limitations are characteristic of the more advanced methods? To start with MuPTM-s that rely on lexical resources, their main limitation is the imperfection of these resources. Dictionaries will never include all possible translations and will always only consider words that have a counterpart in another language. They may miss more technical vocabulary, may not be available for low-resource languages, and usually pair two words and not whole sets of vocabulary. In addition, dictionaries are usually created specifically for two languages. The use of two such dictionaries for the joint analysis of three languages, for example, therefore brings with it problems of comparability. One characteristic of the second type of MuPTM-s, is that the required text corpora (only comparable corpora) already need to be “relatively well understood or annotated” (Boyd-Graber & Blei 2009: 75). Therefore, with models like PLTM the big advantage of clustering unknown text documents using unsupervised methods cannot be fully exploited.

More generally, both types have in common that the induced alignment information (lexical information or topical document comparability) requires validation. The same applies to the model output, “the user needs to verify that the topic word distributions are comparable across languages “(Lucas et al. 2015: 262). Another weakness of MuPTM-s is their inability to include additional document metadata such as the specific media outlet in which a document has been published (Lucas et al. 2015: 262). As a last point, it is true that MuPTM-s hardly come with a publically-available model estimation software. Nothing has changed since Lucas and colleagues' note this in 2015. It continues to be the case that, from the previously mentioned models, only the PLTM has a publicly-accessible Java implementation in Mallet (McCallum 2002).

Ultimately, we would argue that none of these concerns makes it impossible to use these models; instead, we should view them as methodological challenges. While it seems that giving up on document-level covariates is currently a prerequisite for the application of MuPTM-s, the fact that there is one public software implementation for PLTM is in fact a great opportunity.

Because this entry hurdle has already been removed, it is easy to deal directly with other methodological questions concerning the applicability of PLTM in comparative social science research. We agree that the evaluation of model input and model output are no trivial tasks, but pointing to recent work on model evaluation strategies (Hao, Boyd-Graber, & Paul 2018; Pruss et al. 2019), we are optimistic about the likelihood of finding solutions soon. Following the encompassing methodological experiments by Pruss and colleagues (2019), in which the authors compared model performances for different types and amount of training data, the call for extended methodical knowledge on model implementation and specification is wide open. How do different types of training data affect model output and model evaluation? What are best-case strategies to identify topically-comparable news articles across countries? What is the model’s applicability for text across domains (new media coverage, social media, political speeches)?

In order to justify the increased effort, we would like to point out a special benefit of PLTM for comparative research. PLTM allows the characterization of differences in topic prevalence, not only at the document level but also at the language level. With PLTM’s focus on linguistic details, it is possible to identify differences and similarities in topic emphasis between languages, and to answer questions like whether different languages have different (or similar) perspectives on a thematically-similar article. The ability to allow for close inspection of linguistic peculiarities is a unique selling point of PLTM compared to the basic approaches used in the social sciences so far.

Also quite encouraging is a recently-published study in which PLTM was used in a convincing way for a country comparative analysis. Pruss and colleagues (2019) extract the key topics of discussion across a multilingual Twitter dataset in English, Spanish, and Portuguese. They are able to identify, for example, the top three topics with the highest average topic probabilities in each location, and show figures on the volume and distribution of topics over time. Other applications for substantive cross-country comparative research are pending.

In sum, we consider PLTM a promising approach and argue that it has to be tested and applied further to flesh out its potential for comparative research. The media corpus that was gathered in WP8 consists of migration-related media texts in seven European languages (Spanish, English, German, Swedish, Polish, Hungarian and Romanian), and thus provides an excellent basis for investigating the discussed methodological questions further. We hope that the answers will not only lead to an extended understanding of this collection of texts (i.e. what and how European migration topics are emphasised similarly or differently in different languages and countries), but also to suggestions and guidance for many other comparative social science projects.

Literature

Abdelwahab, Ahmed, Jose Robles, Costin-Gabriel Chiru, and Traian Rebedea. "Tweets Topic Modelling Across Different Countries." *eLearning & Software for Education,* no. 1 (2014): 134–141.

Blei, David M. “Probabilistic topic models.” *Communications of the ACM* 55, no. 4 (2012): 77–84.

Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation." *Journal of Machine Learning Research* 3, no. Jan (2003): 993–1022.

Blei, David, Lawrence Carin, and David Dunson. "Probabilistic Topic Models: A focus on graphical model design and applications to document and image analysis." *IEEE Signal Processing Magazine* 27, no. 6 (2010): 55–65.

Boumans, Jelle W., and Damian Trilling. "Taking stock of the toolkit: An overview of relevant automated content analysis approaches and techniques for digital journalism scholars." *Digital Journalism* 4, no. 1 (2016): 8–23.

Boyd-Graber, Jordan, and David M. Blei. "Multilingual topic models for unaligned text." In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pp. 75–82. AUAI Press, 2009.

Boyd-Graber, Jordan, Yuening Hu, and David Mimno. "Applications of topic models." *Foundations and Trends® in Information Retrieval* 11, no. 2-3 (2017): 143–296.

Brown, Gillian, and George Yule. *Discourse Analysis*. Cambridge University Press, 1983.

Chen, Xieling, Ziqing Liu, Li Wei, Jun Yan, Tianyong Hao, and Ruoyao Ding. "A comparative quantitative study of utilizing artificial intelligence on electronic health records in the

USA and China during 2008–2017." *BMC medical informatics and decision making* 18, no. 5 (2018): 55–69.

De Vries, Erik, Martijn Schoonvelde, and Gijs Schumacher. "No longer lost in translation: Evidence that Google Translate works for comparative bag-of-words text applications." *Political Analysis* 26, no. 4 (2018): 417–430.

De Smet, Wim, and Marie-Francine Moens. "Cross-language linking of news stories on the web using interlingual topic modelling." In *Proceedings of the 2nd ACM workshop on Social web search and mining*, pp. 57–64. ACM, 2009.

Directorate-General for Research and Innovation. *Research and Innovation performance in the EU. Innovation Union progress at country level*. European Commission. 2014.

Eberl, Jakob-Moritz, Christine E. Meltzer, Tobias Heidenreich, Beatrice Herrero, Nora Theorin, Fabienne Lind, Rosa Berganza, Hajo G. Boomgaarden, Christian Schemer, and Jesper Strömbäck. "The European media discourse on immigration and its effects: a literature review." *Annals of the International Communication Association* 42, no. 3 (2018): 207–223.

Eberl, Jakob-Moritz, Sebastian Galyga, Fabienne Lind, Tobias Heidenreich, Hajo G. Boomgaarden, Beatriz Herrero Jiménez Montero, Eva Luisa Gómez, Rosa Berganza. “European Media Migration Report: How media cover migration and intra-EU mobility in terms of salience, sentiment and framing.” Working paper prepared as part of the REMINDER project (2019).

Elgesem, Dag, Ingo Feinerer, and Lubos Steskal. "Bloggers’ responses to the Snowden affair: Combining automated and manual methods in the analysis of news blogging." *Computer Supported Cooperative Work (CSCW)* 25, no. 2-3 (2016): 167–191.

Greene, Derek, and James P. Cross. "Exploring the political agenda of the European parliament using a dynamic topic modelling approach." *Political Analysis* 25, no. 1 (2017): 77–94.

Grimmer, Justin. "A Bayesian hierarchical topic model for political texts: Measuring expressed agendas in Senate press releases." *Political Analysis* 18, no. 1 (2010): 1–35.

Günther, Elisabeth, and Emese Domahidi. "What communication scholars write about: An analysis of 80 years of research in high-impact journals." *International Journal of Communication* 11 (2017): 3051–3071.

Hao, Shudong, Jordan Boyd-Graber, and Michael J. Paul. "Lessons from the Bible on Modern Topics: Low-Resource Multilingual Topic Model Evaluation." *arXiv preprint arXiv:1804.10184* (2018).

Hassan, Saeed-Ul, and Peter Haddawy. "Analyzing knowledge flows of scientific literature through semantic links: a case study in the field of energy." *Scientometrics* 103, no. 1 (2015): 33–46.

Heidenreich, Tobias, Fabienne Lind, Jakob-Moritz Eberl, and Hajo G. Boomgaarden. "Media Framing Dynamics of the 'European Refugee Crisis' A Comparative Topic Modelling Approach." *Journal of Refugee Stud*ies (forthcoming).

Hofmann, Thomas. "Probabilistic latent semantic analysis." In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*, pp. 289–296. Morgan Kaufmann Publishers Inc., 1999.

Hu, Yuening, Ke Zhai, Vladimir Eidelman, and Jordan Boyd-Graber. "Polylingual tree-based topic models for translation domain adaptation." In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1166–1176. 2014.

Jacobi, Carina, Wouter Van Atteveldt, and Kasper Welbers. "Quantitative analysis of large amounts of journalistic texts using topic modelling." *Digital Journalism* 4, no. 1 (2016): 89–106.

Jagarlamudi, Jagadeesh, and Hal Daumé III. "Extracting multilingual topics from unaligned comparable corpora." In *European Conference on Information Retrieval*, pp. 444–456. Springer, Berlin, Heidelberg, 2010.

Jiang, Di, Yongxin Tong, and Yuanfeng Song. "Cross-lingual topic discovery from multilingual search engine query log." *ACM Transactions on Information Systems (TOIS)* 35, no. 2, article 9 (2016).

Jiang, Hanchen, Maoshan Qiang, Peng Lin, Qi Wen, Bingqing Xia, and Nan An. "Framing the Brahmaputra river hydropower development: different concerns in riparian and international media reporting." *Water Policy* 19, no. 3 (2017): 496–512.

Kim, Hyunuk, Inho Hong, and Woo-Sung Jung. "Measuring national capability over big science's multidisciplinarity: A case study of nuclear fusion research." *PloS one* 14, no. 2 (2019).

Koehn, Philipp. *Statistical machine translation*. Cambridge University Press, 2009.

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

Koltsova, Olessia, and Sergei Koltcov. "Mapping the public agenda with topic modeling: The case of the Russian livejournal." *Policy & Internet* 5, no. 2 (2013): 207–227.

Krestel, Ralf, and Bhaskar Mehta. "Learning the importance of latent topics to discover highly influential news items." In *KI 2010: Advances in Artificial Intelligence*, edited by Rüdiger Dillmann, Jürgen Beyrer, Uwe D. Hanebeck and Tanja Schultz, pp. 211–218. Berlin, Germany: Springer, 2010.

Krstovski, Kriste, and David A. Smith. "Online polylingual topic models for fast document translation detection." In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pp. 252–261. 2013.

Lau, Jey Han, David Newman, and Timothy Baldwin. "Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality." In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 530–539. 2014.

Lee, Ju-Sung, and Adina Nerghes. "Refugee or migrant crisis? Labels, perceived agency, and sentiment polarity in online discussions." *Social Media+ Society* 4, no. 3 (2018). Advance online publication. https://doi.org/10.1177/2056305118785638

Liu, Xiaodong, Kevin Duh, and Yuji Matsumoto. "Multilingual Topic Models for Bilingual Dictionary Extraction." *ACM Transactions on Asian and Low-resource Language Information Processing* 14, no. 3 (2015): Article 11.

Lind, Fabienne, Jakob-Moritz Eberl, Tobias Heidenreich, and Hajo G. Boomgaarden. "When the Journey Is as Important as the Goal: A Roadmap to Multilingual Dictionary Construction." In "Computational Methods for Communication Science: Towards A Strategic Roadmap" ed. JungHwan Yang, Leonard Reinecke, and Emese Domahidi, special issue, *International Journal of Commu*nication 13, (forthcoming).

Lucas, Christopher, Richard A. Nielsen, Margaret E. Roberts, Brandon M. Stewart, Alex Storer, and Dustin Tingley. "Computer-assisted text analysis for comparative politics." *Political Analysis* 23, no. 2 (2015): 254–277.

Maier, Daniel, Annie Waldherr, Peter Miltner, Patrick Jähnichen, and Barbara Pfetsch. "Exploring issues in a networked public sphere: Combining hyperlink network analysis and topic modeling." Social Science Computer Review 36, no. 1 (2018): 3–20.

Maier, Daniel, Annie Waldherr, Peter Miltner, Gregor Wiedemann, Andreas Niekler, Alexa Keinert, Barbara Pfetsch et al. "Applying LDA topic modeling in communication research: Toward a valid and reliable methodology." *Communication Methods and Measures* 12, no. 2-3 (2018): 93–118.

Mann, Gideon S., David Mimno, and Andrew McCallum. "Bibliometric impact measures leveraging topic analysis." In *Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries*, pp. 65–74. ACM, 2006.

McCallum, Andrew Kachites. “Mallet: A machine learning for language toolkit.” Available at http://mallet.cs.umass.edu. 2002.

Merrill, Samuel, and Mathilda Åkerlund. "Standing Up for Sweden? The Racist Discourses, Architectures and Affordances of an Anti-Immigration Facebook Group." *Journal of Computer-Mediated Communication* 23, no. 6 (2018): 332–353.

Mimno, David, Hanna M. Wallach, Jason Naradowsky, David A. Smith, and Andrew McCallum. "Polylingual topic models." In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2-Volume 2*, pp. 880–889. Association for Computational Linguistics, 2009.

Montalvo, Soto, Raquel Martínez, Arantza Casillas, and Víctor Fresno. "Bilingual news clustering using named entities and fuzzy similarity." In *International Conference on Text, Speech and Dialogue*, pp. 107–114. Springer, Berlin, Heidelberg, 2007.

Nguyen, Thien Hai, and Kiyoaki Shirai. "Topic modeling based sentiment analysis on social media for stock market prediction." In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 1354–1364. 2015.

Ni, Xiaochuan, Jian-Tao Sun, Jian Hu, and Zheng Chen. "Cross lingual text classification by mining multilingual topics from Wikipedia." In *Proceedings of the fourth ACM international conference on Web search and data mining*, pp. 375–384. ACM, 2011.

Ni, Xiaochuan, Jian-Tao Sun, Jian Hu, and Zheng Chen. "Mining multilingual topics from Wikipedia." In *Proceedings of the 18th international conference on World Wide Web*, pp. 1155–1156. ACM, 2009.

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval, 2*(1/2), 1–135.

Paul, Michael J., and Mark Dredze. "Discovering health topics in social media using topic models." *PloS one* 9, no. 8 (2014): e103408.

Pennacchiotti, Marco, and Siva Gurumurthy. "Investigating topic models for social media user recommendation." In *Proceedings of the 20th international conference companion on World Wide Web*, pp. 101–102. ACM, 2011.

Perez-Encinas, Adriana, and Jesus Rodriguez-Pomeda. "Geographies and Cultures of International Student Experiences in Higher Education." *Journal of International Students* 9, no. 2 (2019): 412–431.

Proksch, Sven-Oliver, Will Lowe, Jens Wäckerle, and Stuart Soroka. "Multilingual sentiment analysis: A new approach to measuring conflict in legislative speeches." *Legislative Studies Quarterly* 44, no. 1 (2019): 97–131.

Pruss, Dasha, Yoshinari Fujinuma, Ashlynn R. Daughton, Michael J. Paul, Brad Arnot, Danielle Albers Szafir, and Jordan Boyd-Graber. "Zika discourse in the Americas: A multilingual topic analysis of Twitter." *PloS one* 14, no. 5 (2019): e0216922.

Rahmani, Kamal, Juergen Gnoth, and Damien Mather. "Hedonic and eudaimonic well-being: A psycholinguistic view." *Tourism Management* 69 (2018): 155–166.

Reber, Ueli. "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, no. 2 (2019): 102–125.

Roberts, Margaret E., Brandon M. Stewart, and Dustin Tingley. "stm: R package for structural topic models." *Journal of Statistical Software* 10, no. 2 (2014): 1–40.

Sakamoto, Takuto, and Hiroki Takikawa. "Cross-national measurement of polarization in political discourse: Analyzing floor debate in the US the Japanese legislatures." In 2017 IEEE International Conference on Big Data (Big Data), pp. 3104–3110. IEEE, 2017.

Sievert, Carson, and Kenneth Shirley. "LDAvis: A method for visualizing and interpreting topics." In *Proceedings of the workshop on interactive language learning, visualization, and interfaces*, pp. 63–70. 2014.

Utsuro, Takehito, Takashi Horiuchi, Yasunobu Chiba, and Takeshi Hamamoto. "Semi-automatic compilation of bilingual lexicon entries from cross-lingually relevant news articles on WWW news sites." In *Conference of the Association for Machine Translation in the Americas*, pp. 165–176. Springer, Berlin, Heidelberg, 2002.

Vulić, Ivan, Wim De Smet, and Marie-Francine Moens. "Cross-language information retrieval models based on latent topic models trained with document-aligned comparable corpora." *Information Retrieval* 16, no. 3 (2013): 331–368.

Vulić, Ivan, Wim De Smet, Jie Tang, and Marie-Francine Moens. "Probabilistic topic modeling in multilingual settings: An overview of its methodology and applications." *Information Processing & Management* 51, no. 1 (2015): 111–147.

Wang, Xuanhui, ChengXiang Zhai, Xiao Hu, and Richard Sproat. "Mining correlated bursty topic patterns from coordinated text streams." In *Proceedings of the 13th ACM*

SIGKDD international conference on knowledge discovery and data mining, pp. 784–793. ACM, 2007.

Wijffels, Jan. 2018. Udpipe: Tokenization, parts of speech tagging, lemmatization and dependency parsing with the “UDPipe” “NLP” toolkit (R Package Version 0.6). Computer software. Available at https://cran.r-project.org/web/packages/udpipe/index.html

Zhang, Duo, Qiaozhu Mei, and ChengXiang Zhai. "Cross-lingual latent topic extraction." In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pp. 1128–1137. Association for Computational Linguistics, 2010.

Zheng, Liyi, Tian Nie, Ichiro Moriya, Yusuke Inoue, Takakazu Imada, Takehito Utsuro, Yasuhide Kawada, and Noriko Kando. "Comparative topic analysis of japanese and chinese bloggers." In *2014 28th International Conference on Advanced Information Networking and Applications Workshops*, pp. 664–669. IEEE, 2014.

Appendix

Table A1: Systematic review of studies that use topic models for cross country comparisons

Author	Type of Text Documents	Language of Text Documents	Model	Purpose of Analysis	Countries
Monolingual Data Type Selection					
Abdelwahab et al. (2014)	Tweets	English	LDA	Topic extraction and comparison between countries	4
Hassan & Haddawy (2015)	Scientific publications	English	LDA with distance matrix	Compare topics in Japanese and Chinese papers that cite scientific literature produced by researchers from the United States	2
Jiang et al. (2017)	International media reporting	English	STM	Topic extraction and comparison (similarities and differences) between countries	34
Kim, Hong, & Jung (2019)	Scientific publications	Not specified; probably English	Dynamic LDA	Country's research capability in nuclear fusion research	14
Perez-Encinas & Rodriguez-Pomeda (2019)	International student reports	English	LDA	Shared experiences of international students from different countries	21
Rahmani, Gnoth, & Mather (2018)	Tourists' blog posts	English	LDA	Compare France's and New Zealand's well-being positioning in contrast to a global tourist destination baseline	2
One separate Topic Models per language/country					
Chen et al. (2018)	Scientific publications	Not specified; probably English	LDA	Topic extraction and comparison (similarities and differences) between countries	2
Sakamoto & Takikawa (2017)	Legislative debates	English, Japanese	LDA	Cross-country differences in collective articulation of public agendas among relevant political actors	2
Zheng et al. (2014)	Blog posts	Chinese, Japanese	LDA	Compare Chinese and Japanese bloggers' concerns, opinions, and cultures	2
Document Translation					
Lucas et al. (2015)	Social media posts	Arabic, Chinese	STM	Comparison of social media reaction of citizens in China and the Middle East about Snowden	2

Table A2: Multilingual Probabilistic Topic Models: An overview

Authors	Application Demonstrations	Corpora Type	Model	Input/Bridge	Public Software Implementation
Zhang et al. (2010)	Extract multilingual topics from an unaligned corpus	Unaligned	Probabilistic Cross-Lingual Latent Semantic Analysis Model (PCLSA)	Bilingual dictionary	-
Boyd-Graber & Blei (2009)	Pair related documents across languages	Unaligned	Multilingual Topic Model (MuTo)	Bilingual dictionaries	-
Jagarlamudi & Daumé III (2010)	Extract multilingual topics from an unaligned corpus	Unaligned	JointLDA	Bilingual dictionaries	-
Hu et al. (2014)	Domain adoption for statistical machine translation improvement	Not specified	Polylingual tree-based Topic Model (ptLDA)	External dictionaries + word alignments from aligned sentences in a parallel corpus	-
Vulić et al. (2015)	Cross-lingual event-centered news clustering; cross-lingual document classification; cross-lingual semantic similarity; cross-lingual information retrieval	Comparable	Bilingual LDA (BiLDA)	Tuples composed of comparable documents in each language of the corpus	-
Mimno et al. (2009)	compare topic emphasize across languages, link topics in non-comparable documents, enhancing lexicons by aligning topic-specific vocabulary, adapt machine translation systems to new domains	Comparable	Polylingual Topic Model (PLTM)	Tuples composed of comparable documents in each language of the corpus	Mallet (McCallum 2002)
Ni et al. (2009, 2011)	Cross-lingual text classification, cross-lingual document recommendation	Comparable	Multilingual LDA (ML-LDA)	Tuples composed of comparable documents in each language of the corpus	-

REMINDER

ROLE OF EUROPEAN MOBILITY AND ITS IMPACTS IN NARRATIVES, DEBATES AND EU REFORMS

The REMINDER project is exploring the economic, social, institutional and policy factors that have shaped the impacts of free movement in the EU and public debates about it.

The project is coordinated from COMPAS and includes participation from 12 consortium partners in 8 countries across Europe

UNIVERSITY OF OXFORD

This project has received funding from the European Union's Horizon 2020 research & innovation programme under grant agreement no 727072

COMPAS

CENTRE ON MIGRATION · POLICY & SOCIETY