International Journal of Communication 13(2019), 4000–4020 1932–8036/20190005

When the Journey Is as Important as the Goal: A Roadmap to Multilingual Dictionary Construction

FABIENNE LIND
JAKOB-MORITZ EBERL
TOBIAS HEIDENREICH
HAJO G. BOOMGAARDEN[1]
University of Vienna, Austria

Communication scientists have made rapid advances in the computer-assisted analysis of large quantities of media data, but research has focused on monolingual corpora and most often on English-language text. This study works toward the application of computer-assisted analysis in the framework of multilingual media content. Taking the measurement of migration frames in the news coverage in 7 languages as a case study, it systematically compares different strategies (i.e., keyword preselection, translation, evaluation) for the construction of a multilingual dictionary. Classification results are contrasted to each other and to results of English monolingual dictionaries that are applied to the translated text corpus version. Even though we do not yet achieve perfect agreement between manual coding and dictionary classification decisions, with the strategies compared here, we outline methodological techniques that may bring researchers closer to this goal.

Keywords: multilingual dictionaries, dictionary construction, text analysis, migration frames

With the increasing use of computer-assisted content analysis methods, dictionaries have become a decisive tool for concept measurement in digital texts. As a toolkit, a dictionary is a set of words or word-based indicators that is used to search texts (Neuendorf, 2002). Researchers have several options when selecting the best performing, most valid, or most appropriate dictionary to implement their measurement task. They apply open-source or commercially available dictionaries and modify and adjust these dictionaries to meet

Fabienne Lind: fabienne.lind@univie.ac.at
Jakob-Moritz Eberl: jakob-moritz.eberl@univie.ac.at
Tobias Heidenreich: tobias.heidenreich@univie.ac.at
Hajo G. Boomgaarden: hajo.boomgaarden@univie.ac.at
Date submitted: 2018–09–24

[1] This work was supported by the European Union's Horizon 2020 research and innovation program under Grant Agreement No. 727072. We would like to thank Petro Tolochko, Eva Luisa Gómez Montero, the editors, and our reviewers for their helpful feedback. We are grateful to Zehra Brakmic, Rachel Edie, Orsolya Fehér, Rubén Tamboleo García, Ewelina Marszowska, and Volina Șerban for their contribution to article coding.

their own needs. When researchers are interested in concepts that go beyond those that can be measured with available dictionaries, they must construct new, customized dictionaries (Grimmer & Stewart, 2013; Loughran & McDonald, 2011), applying deductive or inductive approaches, or a mix of both. The necessity of this task becomes particularly clear when analyzing a multilingual text corpus. Dictionaries in languages other than English are rare (Boumans & Trilling, 2016; Pang & Lee, 2008), and finding dictionaries that match specific research interests in other languages, which are also validly comparable across languages, is close to impossible.

Most of the literature on computer-assisted content analysis has dealt with English-language texts (Pang & Lee, 2008). This is not due to a lack of comparative research questions, which would, in fact, greatly benefit from analyses drawing on multilingual dictionaries. There are manifold reasons for the delayed attention to languages other than English and for the lack of simultaneous work with multiple languages. Among these reasons are the tremendous effort needed to build the necessary resources for "only" one language (Laver, Benoit, & Garry, 2003; Young & Soroka, 2012); the arguable dominance of English as the language of science (Ammon, 2001); the scarcity of multilingual resources and text-processing tools that assist text analysis; the high costs for hiring professional translators; and the immaturity of (affordable) machine translation technology.

At present, many of these factors are becoming less of an obstacle. For example, multilingual text corpora, multilingual text analysis resources (e.g., dictionaries, annotated corpora), as well as other helpful tools such as parallel corpora or multilingual thesauri are becoming increasingly available. As a consequence, work on "analysis strategies," which look at how to ideally use and combine resources and tools for multilingual dictionary construction, has become an active research field, and the first social science research teams constructed and applied dictionaries for multilingual text analysis (e.g., Baden & Stalpouskaya, 2015; Benoit, Schwarz, & Traber, 2012). Currently missing, however, is a systematic review and empirical comparison of different multilingual dictionary construction strategies and analysis approaches.

With a substantial interest in news media framing of human migration in seven European countries (Spain, UK, Germany, Sweden, Poland, Hungary, and Romania), this contribution therefore presents and systematically compares strategies for the construction of a multilingual dictionary. We here focus on keywords, which are the building units of a dictionary in its most basic form (Boumans & Trilling, 2016), and on three related central construction steps: keyword preselection (i.e., identification of potentially relevant keywords); keyword translation (an intermediate step where keywords available in one language are translated into another language); and keyword evaluation (i.e., reassessing the appropriateness of these keywords and their usefulness for the final keyword selection). The primary goal of this study is not to create the best performing dictionaries in each of the studied languages—although, as a side effect, of course, this is desirable—and thus to refine each dictionary to the limit. Rather, the goal is to evaluate the different intermediate steps and decisions that can be taken when constructing a multilingual dictionary.

Text Analysis With Dictionaries

Researchers using a dictionary in the analysis of a text corpus employ a top-down approach. The corpus (consisting of documents, text entities such as news articles, social media postings, or mere sentences)

is searched with the help of a predefined list of keywords, which may be words/word stems and phrases that represent the concept(s) of interest. The analysis strategy assumes that these dictionary keywords reflect the respective target concepts. Keyword counts (i.e., the frequency of keywords per text) and also, potentially, co-occurrences of keywords offer a reliable analysis of the occurrence of a given concept in a given text.

The construction of a new dictionary is relatively straightforward for clearly defined target concepts such as the occurrence of a specific actor in the news. For this purpose, a dictionary could simply include variations of that person's name or (in combination with) their position or the association they represent. However, this simple strategy reaches its limit quickly in the construction of a dictionary for less clear-cut concepts, such as frames, topics, or sentiment. The level of difficulty of the two abovementioned central construction steps, keyword preselection and keyword evaluation, increases considerably. In these cases, construction usually requires the preselection of more keywords to cover as much of the full concept as possible. More keywords, however, increases the danger of preselecting keywords that might represent not only one concept but two, or many more. For this reason, keyword evaluation is of particular importance here.

The advantages, challenges, and limitations of computer-assisted content analysis methods for social sciences, especially in contrast to manual content analysis, have been comprehensively discussed elsewhere (e.g., Boumans & Trilling, 2016; Grimmer & Stewart, 2013; Riffe, Lacy, & Fico, 2014; Young & Soroka, 2012). The main benefits of text analysis using a dictionary approach are the perfect reliability of the procedure and the capability to process large quantities of text. The main challenge is validity—that is, whether researchers are actually measuring the concepts in which they are interested. "Bag-of-words" text analysis approaches, such as the dictionary approach, process individual words regardless of order and context, and as a consequence wrongly assume "semantic independence" (Young & Soroka, 2012, p. 209). It is clear that computer-assisted methods that classify and process text are not equivalent to manual coding or human understanding of text. As Grimmer and Stewart (2013) note, "all quantitative models of language are wrong, but some are useful" (p. 269).

Overall, the construction of good-quality dictionaries has been described as "very difficult" (Young & Soroka, 2012, p. 208) and an "extensive effort" (Laver et al., 2003, p. 312) that requires time, money, and strong collaboration with human coders. Both keyword preselection and keyword evaluation are strongly impaired by one's "subjective conception" and "limited domain knowledge" (Burscher, Odijk, Vliegenthart, De Rijke, & De Vreese, 2014, p. 192). These outlined challenges and difficulties increase exponentially when going beyond the construction of monolingual dictionaries. Both the naturally limited language ability and domain knowledge of multiple national contexts, for instance, make it much more challenging to preselect and evaluate keywords. Strategies that may at least partly tackle these issues will become clearer within the next section, where we elaborate more on the particularities of multilingual dictionaries and outline the steps for constructing a multilingual dictionary.

Multilingual Dictionaries

A multilingual dictionary aims to hold keywords in different languages to provide valid measurement of the same concept. How similar the language-specific keyword lists should be to each other depends on the goal of the analysis.

In a noncomparative framework, a keyword list in one language often supports the creation of a keyword list in another language, as long as the studied concept is not understood diametrically differently in the different cultural contexts. This new and translated keyword list is subsequently adapted to the respective context. Ultimately, the two keyword lists can be detached from each other and used separately in the different (country) contexts. To date, this has been the predominant direction of development and usage of multilingual keyword lists in the social sciences (e.g., Duval & Pétry, 2016; Rauh, 2018; Sevenans, Albaugh, Shahaf, Soroka, & Walgrave, 2014). Although this approach can lead to the best possible measurement instruments (i.e., concept representation) for the one new context, the applications for comparative research are often limited.

In a comparative framework, the aim is usually to identify cross-country similarities and differences regarding one and the same concept. For this purpose, a multilingual dictionary includes several language-specific keyword lists that take context (i.e., country/region) into account, and at the same time validly map comparable concepts that comparatively describe the system-specific discourses. Here, in contrast to the noncomparative framework, it is crucial to search for ways to model new subconcepts (e.g., aspects of the topic to be measured) found during the refinement of the instrument in one context to the other examined contexts. Such endeavors—although highly beneficial for studying cross-national media discourses on diverse cross-national topics such as climate change, migration, or European integration—are still scarce. An exception is the INFOCORE project (Baden & Stalpouskaya, 2015), where the research group used a multistep mixed-methods strategy to construct concept keyword lists in nine languages. First, native speakers constructed language-specific lists with relevant concepts and related keywords based on their work with monolingual text samples. Concepts, and later keywords, were then compared, integrated, and revised across languages. Cross-checking across languages, thesauri, and word frequency analysis, as well as disambiguation strategies, further assisted to homogenize and improve the keyword lists. This approach included many feedback loops and a strong collaboration with native speakers. In what follows, we outline the basic steps needed to construct such a dictionary.

Steps for Multilingual Dictionary Construction

The first obvious task for the construction of a multilingual dictionary is defining the target concept. After specifying what the dictionary should actually measure, the subsequent steps include the preselection of dictionary keywords, their translation (if necessary) and evaluation, and, finally, the assessment of the overall dictionary performance. All these steps are performed in an iterative manner (i.e., insufficient keyword evaluation results and dictionary performance assessment may demand further specification of keywords, which in turn would have to be evaluated again) until the overall dictionary performance assessment yields satisfactory results.

Keyword Preselection

When it comes to the selection of keywords, researchers may choose among different techniques. For monolingual dictionary construction, social scientists have applied several strategies, including extracting seemingly relevant sentences, words, and phrases from text corpus samples (e.g., Vliegenthart & Roggeband, 2007); combining available dictionaries (e.g., Young & Soroka, 2012); consultation with human experts (e.g.,

Bengston & Xu, 1995); and making use of the "wisdom of the crowd" through crowd coding (e.g., Haselmayer & Jenny, 2017). Researchers also make use of available resources in other languages. They start off with a monolingual, mostly English-language template that is first translated word-by-word into the target language, and then enriched by working with various language-specific tools. Duval and Pétry (2016), for example, selected this strategy for their creation of the French Lexicoder Sentiment Dictionary. After a manual translation of the source dictionary, the English Lexicoder Sentiment Dictionary, they applied stemming, eliminated duplicates, added synonyms, and worked with Key Word in Context, as well as stop-word lists (Duval & Pétry, 2016). The Dutch-language Lexicoder Topic Dictionary was constructed following a similar approach, but for topics (Sevenans et al., 2014).

An alternative automated approach to creating a dictionary that fully emerges from the text corpus alone is based on principle component analysis or topic modeling (see Greussing & Boomgaarden, 2017; Heidenreich, Lind, Eberl, & Boomgaarden, forthcoming). Similarly, Lawlor and Tolley (2017) searched in their text corpus for the most frequently used words and phrases, then applied hierarchical clustering and examined whether terms that clustered together formed a logical frame. These terms were then used as keywords to construct English-language dictionaries for the automated measurement of frames.

Translation

Translation is the central preprocessing step for multilingual text analysis. Both human translation and machine translation have played a major role in multilingual dictionary construction. Machine translation technology, evolving from phrase-based to neural machine translation models, has matured significantly in the past decade.[2] Although it may not outperform manual translation, machine translation can complement work with multilingual text in meaningful ways (e.g., Balahur & Turchi, 2014). Studies comparing machine translation software (e.g., Hampshire & Salvia, 2010) recommended Google Translate for the machine translation of dictionaries and text corpora. Given optimization through automated translation procedures, it is an open question as to how beneficial the often costly (both in terms of time and financial resources) collaboration with native speakers still is.

Keyword Evaluation

Individual keywords are selected and evaluated based on their ability to, first, represent the target concept and, second, to match the vocabulary of the target text corpus. Related to these evaluation objectives is the dilemma of generalizability versus domain-specific knowledge discovery (demonstrated in Loughran & McDonald, 2011). Should the dictionary aim at measuring the concept in wide applications, or rather be customized for application to one specific text corpus? Although an answer to this question depends on the intended purpose and target concept, researchers should adapt dictionary keywords to the text domain to obtain meaningful results. The domain refers to the text type (e.g., language in legislative texts differs from news texts), and in a multilingual framework it also refers to the respective language- and

[2] Machine translation quality is different for each language pair (e.g., English ↔ Spanish) and each individual text, and depends on the direction of translation and the chosen machine translation technology, and will thus change with further improvement of these technologies (Koehn, 2009).

country-specific context. Different languages have a different diversity of words to express the same or similar meanings (i.e., the richness of a language). Languages also differ in terms of their morphologic complexity. For example, Hungarian, a Finno-Ugric language, is a highly inflective and agglutinative language, which requires special efforts (Pajzs et al., 2014), such as the application of customized preprocessing tools (e.g., lemmatizing).

In a comparative framework, keywords are ideally also assessed in terms of their consideration of country-specific contexts. Designed for the measurement of topics or frames from a comparative perspective across countries, a multilingual dictionary is ideally evaluated by its ability to, first, account for the individual national discourses and, second, to include elements that are part of a supranational discourse (i.e., general components likely to occur in any national context).

With this study, we want to focus on a keyword evaluation strategy technique that does not require the support of native speakers, which is, given the just-outlined evaluation aspect, the preferable, but often not available, option.

Assessment of Overall Dictionary Performance

There are different techniques and criteria to consider for the evaluation of the validity of a multilingual dictionary. A frequently applied technique to assess the empirical validity of a dictionary is to compare dictionary-coding decisions to manual coding decisions (e.g., Young & Soroka, 2012). This technique contrasts the final output of different coding processes (dictionary vs. manual coding) and the application of different dictionaries and quantifies their performance. Like conducting an intercoder reliability test for manual coders, one would compare the manual with the dictionary-based coding decisions, using Krippendorff's alpha (Krippendorff, 2004), for example, or recall, precision and F-score measures (see Sokolova & Lapalme, 2009). Applied to the text classification task examined here, recall is a measure of the conditional probability that a text is retrieved through the application of the dictionary, given that it is relevant (i.e., manually coded as such). Precision is a measure of the conditional probability that an article will be relevant (i.e., manually coded as such), given that it is retrieved by the dictionary. An F1 score is defined as the harmonic mean of recall and precision (i.e., how precise as well as how robust the applied research instrument is).[3] All three types of measures range from zero to one (best value). Manual human coders' classification decisions, are, after all, still referred to as the "most reasonable benchmark" (Rauh, 2018, p. 7), acknowledging the obvious differences between machines and humans in text-processing approaches. If reliability (agreement between dictionary and manual coding decisions) is high, then dictionaries are considered to provide an empirically valid measurement of concept.

Following this review of recent dictionary construction projects and central construction steps, one can conclude that the applied strategies are manifold and versatile. Indeed, the field lacks a systematic comparison of different strategies, especially regarding the steps of multilingual dictionary construction. We thus propose the following first research question.

---

[3] $\mathrm{F1} = \frac{2 * \mathrm{Recall} * \mathrm{Precision}}{\mathrm{Recall} + \mathrm{Precision}}$

RQ1: How do different strategies for keyword selection, keyword translation, and keyword evaluation contribute to the construction of high-quality multilingual dictionaries?

Analyzing a Multilingual Text Corpus: Two Approaches

Taking a step back, when social scientists decide to analyze a multilingual text corpus using an automated dictionary-based search, the application of a multilingual dictionary is just one possible methodological approach (Approach A). An alternative is the translation of the multilingual text corpus into the target language and the application of a dictionary in this language (Approach B).

Regarding Approach A, the construction of a multilingual dictionary has received the most attention, which was extensively discussed in the two previous sections. A different approach to analyzing the content of a multilingual text corpus (Approach B) is to first translate the entire corpus into one language and then apply tools designed for a single language. Lucas and colleagues (2015), for example, translated a Chinese–Arabic document–term matrix into English and applied structural topic modeling (Roberts, Stewart, & Airoldi, 2016), a topic model that can control for the original language. Working with a dictionary using this approach fully relies on monolingual, mostly English dictionaries, which are either adapted and refined or—in a "minimalist" approach—applied as they are. Benoit and associates (2012) used this method in their analysis of policy positions in legislative speeches (originally German, French, and Italian) in Switzerland.

Returning to our case study of migration news coverage, it is likely that Approach B misses out on important context-specific keywords. However, it is also an open question as to how important such context-specific keywords (e.g., "mojados" in Mexico; "boat people" in Spain) are in a comparative analysis compared with non-context-specific keywords (e.g., "refugee" or "migrant"). The great advantage of Approach B is that it bypasses the labor-intensive effort of selecting keywords for multiple languages (Approach A). But at what price?

Researchers usually decide in favor of one of the two approaches, preventing comparison between the two. In fact, at this point, we do not know how well either approach performs in direct comparison. We thus ask the following:

RQ2: How favorable is the construction of a multilingual dictionary in contrast to the translation of the target corpus to English and the subsequent application of an English-language dictionary?

Method

This study systematically evaluated the different steps in multilingual dictionary construction (RQ1) and compares Approach A (the application of a multilingual dictionary) with Approach B (the translation of the multilingual text corpus into a specific target language; RQ2). More precisely, we examined how strategies and approaches relate to dictionary performance (recall, precision, and F1 score), modeled as the dependent variable in subsequent analyses. To this end, we systematically created multiple keyword lists that differ in terms of the chosen preselection strategy, keyword translation, and keyword evaluation. The dictionary performance of each list was evaluated by contrasting it with human coding decisions.

The Case

The research questions were examined based on news coverage on migration. We thus worked toward the design of dictionaries for the measurement of four different migration frames in a multilingual news article text corpus. Studying media coverage on migration with computer-assisted methods, we understand frames as topics formed through reoccurring patterns of specific words that help us to categorize documents (Jacobi, van Atteveldt, & Welbers, 2016). Media frames on migration have been examined—mainly using manual content analysis—in numerous national contexts.[4] Frames relating to economic and budgetary, labor market, welfare, and security concerns have been identified as the most relevant in relation to migration coverage (Eberl et al., 2018) and were thus the focus of this project.

The multilingual text corpus consists of migration-related print and online news articles published between January 2000 and December 2017 in seven different countries and languages.[5] The languages belong to different language families and subfamilies, namely the Uralic (i.e., Hungarian = HU) and Indo-European language family, with the main subfamilies being Germanic languages (i.e., English = EN, German = DE, Swedish = SV), Romance languages (i.e., Spanish = ES, Romanian = RO), and Slavic languages (i.e., Polish = PL).

Article Subset for this Study

We randomly drew 1,000 articles from each of the seven text corpora. Native speakers decided first if an article dealt with migration (i.e., determined relevant). This was the case for the overall majority (UK: n = 977; ES: n = 996; DE: n = 950; SV: n = 925; PL: n = 978; HU: n = 932; RO: n = 932). To harmonize the size of the corpora, we further worked with 925[6] (randomly selected) news articles per corpus. In preparation for the analysis of RQ2, we machine translated the non-English article subsample (n = 925 articles per language) into English, article by article,[7] using the Google Translate API as well as the R package GoogleLanguageR.

Manual Article Classification

For their later use as a benchmark, a total of 6,475 articles (925 articles per coder) were then manually classified by the same seven native speakers. They manually evaluated the presence of each frame based on frame-specific coding instructions (see the codebook in Table A2 of the Online Appendix at https://osf.io/86nkx/?view_only=2ea2df83f8dd43e08bf2b0da4bc901fe). Intercoder reliability was

[4] For example, the UK (Caviedes, 2015) and Hungary (Vicsek, Keszi, & Márkus, 2008).

[5] Details about the seven corpora are available in Table A1 of the Online Appendix on the Open Science Framework (https://osf.io/86nkx/?view_only=2ea2df83f8dd43e08bf2b0da4bc901fe). Together they make up a large-scale multilingual corpus of about N = 1.5 million news articles that deal with migration.

[6] As 925 was the lowest common denominator of relevant classified articles per corpus.

[7] For better results, the translation of full documents is still to be preferred over document-term matrix translation (Reber, 2018).

assessed and deemed satisfactory.[8] Table 1 shows the relative frequencies for manual classification per corpus and per frame.

It is important to note that the manual coding decisions counted for and were thus connected to both the original-language and machine-translated versions of an article.

Table 1. Manual Article Classification Decisions per Corpus and Frame (Relative Frequencies).

	Corpus country context (language)							
	UK (EN)	Spain (ES)	Germany (DE)	Sweden (SV)	Poland (PL)	Hungary (HU)	Romania (RO)	Total *n* Articles
Articles (*n*)	925	925	925	925	925	925	925	6475
Frames (relative frequencies):								
Economy & budget	21	14	20	12	9	7	20	15
Labor market	38	15	36	30	41	14	22	28
Welfare	26	21	42	17	17	8	10	21
Security	39	19	24	22	27	44	35	30
At least one of the frames	75	51	70	59	67	59	63	63

Dictionary Creation

We relied on the following methods to construct the multilingual dictionaries.

Keyword Preselection Strategies

We followed two paths in building a basic stock of keywords for each migration frame. First, we selected keywords (*N* = 687, all English-language; see Table A4 in the Online Appendix at https://osf.io/86nkx/?view_only=2ea2df83f8dd43e08bf2b0da4bc901fe) from previously available keyword lists (here referred to as PAKL) with closely related categories (Albaugh, Sevenans, Soroka, & Loewen, 2013; Balaban,

[8] The intercoder reliability test for manual content analysis included two parts. The first, where all seven coders classified 70 English (original language) articles (Krippendorff's alphas: .71–.79). The second, where each native speaker manually coded 50 original-language articles. These coding decisions are then compared with the coding decisions of an English native speaker, who coded the machine-translated version of each of the 50 article sets (Krippendorff's alphas: .64–.92; details in Table A3 of the Online Appendix at https://osf.io/86nkx/?view_only=2ea2df83f8dd43e08bf2b0da4bc901fe).

Meza, & Vincze, 2018; Greussing & Boomgaarden, 2017; Lawlor & Tolley, 2017). These PAKL-based keywords represent words that were mostly compiled via topic modeling and related to the measurement of the four frames in news articles. An important consideration was that the keywords were in English, mostly originally gathered for other country contexts (e.g., immigration in Canada in Lawlor & Tolley, 2017). Largely for these reasons, we collected additional keywords from the Comparative Manifesto Project (here referred to as CMP; Volkens et al., 2015), which contains text in the relevant seven languages and country-specific contexts. The CMP has been recommended and tested for the development of issue-specific dictionaries in multiple languages (Merz, Regel, & Lewandowski, 2016). We selected—individually for each language—all sentences from the CMP database annotated by trained CMP coders with codes relating to our frames of interest. Based on these sentences, we extracted the most frequent keywords (*N* = 8,800, together for all frames and all seven languages; see Table A5 in the Online Appendix at https://osf.io/86nkx/?view_only=2ea2df83f8dd43e08bf2b0da4bc901fe) to be used as keywords for our dictionaries.

Keyword Translation

We further relied on machine translation to transfer each keyword from the basic keyword stock (*N* = 9,487) in each of the six non-English languages. English was here used as the pivot (bridge) language. Hence, all non-English keywords from the basic keyword stock (*N* = 7,200) were machine translated into English. All English keywords (*N* = 9,487) were machine translated into ES, DE, SV, PL, HU, and RO. We ended up with 66,409 keywords (*N* = 9,487 per language).

Keyword Evaluation

Two manual evaluation steps were integrated to improve the precision of research instruments. First, one researcher revised the English version of all the keywords (*N* = 9,487) and decided whether they were useful to search for a specific frame or not (see the codebook at https://osf.io/86nkx/?view_only=2ea2df83f8dd43e08bf2b0da4bc901fe, pp. 7–8). Intercoder reliability was then assessed and deemed satisfactory.[9] The decision made for the English version of a keyword was transferred to the equivalent keyword in all other languages. It was crucial to identify stop words. Other evaluation criteria were, for example, the "concept fit" and the "ambiguity of a keyword." Only 1,985 English keywords, and thus the equivalent 1,985 keywords in all other languages, were positively evaluated within this first evaluation round. This large number of negatively evaluated keywords is related to the high frequency of stop words, which were not captured by the imperfect multilingual stop-word lists used to clean the retrieved CMP keywords (see Silva & Ribeiro, 2003), and may also have originated in concerns about the coder reliability of the CMP (e.g., Mikhaylov, Laver, & Benoit, 2012). Given the poor quality of the many initially gathered keywords,[10] we decided to conduct all subsequent steps with only those keywords that passed this first researcher evaluation step.

[9] Another researcher coded a subsample of 500 keywords from these *N* = 9,487 keywords. Intercoder reliability for these 500 coding decisions, coded by the two researchers, was then assessed (Krippendorff's alpha = .83).

[10] We tested entirely unevaluated keyword lists and obtained very poor overall performance scores, as these unevaluated dictionaries produce a hit in almost every article.

These keywords were further evaluated by six native speakers in their respective native languages. Native speakers were asked to identify keywords that were likely to majorly disturb the analysis—for example, due to translation errors (see the codebook at https://osf.io/86nkx/?view_only=2ea2df83f8dd43e08bf2b0da4bc901fe, pp. 9–10). For example, the English keyword "unions" was not translated in a way that referred to "workers' unions" but to any other kind of union. Furthermore "patient" (e.g., in a hospital) was translated into the German adjective "*geduldig*" meaning "to be patient." However, at this stage, the number of keywords classified as potentially inflating precision of measurement was rather small and similar across languages (ES: 8%, DE: 3%; SV: 2%, PL: 7%; HU: 6%; RO: 2% of *N* = 1,985 keywords per language).

For a better overview, keyword preselection, keyword translation, and keyword evaluation are sketched in Figure A1 of the Online Appendix (https://osf.io/86nkx/?view_only=2ea2df83f8dd43e08bf2b0da4bc901fe). The procedure is repeated for each of the four frames.

Systematic Keyword List Creation

Every individual keyword compiled during the dictionary construction is categorized. The categories refer to different keyword characteristics such as keyword preselection strategy (i.e., whether it originated from PAKL or CMP) and the decision made during the keyword evaluation procedures. We also kept track of the current language, its original language, and whether the keyword was machine translated or not. Another essential category is the respective frame to which a keyword is assigned.

This keyword categorization is the basis for the systematic creation of various keyword lists (= dictionaries). Each keyword list represents another possible combination of the keyword preselection, keyword translation, and keyword evaluation steps for all seven different languages and four different frames. For example, one keyword list includes only keywords that are supposed to measure the welfare frame, originate from CMP, are in Spanish, untranslated, and have been additionally evaluated by a native speaker. Given the number of combinations, we ended up with 776 keyword lists[11] with distinct sets of keywords. As a final step in data cleaning, duplicate keywords were removed from each list.

Dictionary Article Classification

The previously created keyword lists were used to classify articles with regard to whether a frame was present (= 1) or not (= 0). As an important note, original language articles were classified applying all previously systematically created keyword lists that matched the respective language (i.e., the Spanish corpus was classified using each of the Spanish keyword lists individually); the machine translated English version of the articles were also classified using all English language keyword lists.

[11] That is, 194 per frame; see details in Table A6 of the Online Appendix (https://osf.io/86nkx/?view_only=2ea2df83f8dd43e08bf2b0da4bc901fe).

In preparation for the classification, news article words (original language version and machine-translated English version) and dictionary keywords were converted to lowercase and were lemmatized using the R package Udpipe (Wijffels, 2018). Relying on the R package Stringr (Wickham, 2018), we then acquired the number of matches between article words and dictionary keywords per article. This number of matches, which represents the number of keywords found in an article when a keyword list is used as a query, is referred to as a dictionary hit (Hashimoto & Kurohashi, 2007). It is a manual decision to define how many hits are required to classify an article as belonging to a frame (= 1) or not (= 0). Although one hit is often defined as sufficient (e.g., Caviedes, 2015; Vicsek et al., 2008), some projects have also used three hits (McLaren, Boomgaarden, & Vliegenthart, 2017). Here, we tested the implications of three scenarios (i.e., required dictionary hits: one, two, and three), and thus recoded the number of hits obtained for all dictionary applications for three different definition scenarios.

Evaluation of Dictionary Performance

The coding decisions (based on three different hit scenarios) resulting from the application of different keyword lists were systematically compared with each other by contrasting them to the previously introduced manual classification decisions of native speakers, which we defined as the best possible benchmark. The agreement between the classification decision of a keyword list and a human coder was compared and evaluated through three measures: recall, precision, and F1 scores. We thus calculated recall, precision, and F1 scores for each of the applied keyword list in three different required hit scenarios, which resulted in 3,336 values for recall, precision, and F1 scores, respectively. These performance measures were subsequently modeled as the dependent variable (recall: M = .80, SD = .26; precision: M = 0.32, SD = 0.19; F1 scores: M = 0.39, SD = 0.16).

Analysis Approach

For the analysis of RQ1, we exclusively included data points for multilingual keyword lists applied to the multilingual untranslated text corpus (all Approach A, N = 2,328). To examine the contribution of different multilingual dictionary creation techniques across languages (RQ1), we ran an OLS regression, with recall, precision, and F1 scores as dependent variables and the characteristics of the applied keyword lists as the independent variables, while including the required number of dictionary hits, the language corpus, and the frame as dummy variables.

For the analysis of RQ2, comparing Approach A (i.e., working with multilingual keyword lists and a multilingual untranslated text corpus) with Approach B (i.e., working with monolingual keyword lists and a machine-translated corpus), we exclusively examined dictionary performance (N = 3,168) for the untranslated versus machine-translated versions of the Spanish, German, Swedish, Polish, Hungarian, and Romanian news articles. The UK text corpus was not part of this analysis, because the machine translation of the English corpus into English would, of course, have been meaningless.

Results

Best Performing Dictionaries

The aim of this study was not necessarily to develop the best performing dictionaries in each of the studied languages or refine each dictionary to the limit. Rather, the goal was to evaluate different intermediate steps and decisions that could be taken when constructing a multilingual dictionary. To ground the reader's expectations about the final performance of our measures in this specific multilingual challenge, we show the F1 scores for the best performing dictionaries per news article corpus and frame (see Table 2).

Table 2. Best and Worst Performance Scores (F1 Scores) of Dictionaries per Corpus and Frame.

Frame	UK	Spain	Germany	Sweden	Poland	Hungary	Romania
Economy & budget							
	.52	.44	.59	.46	.31	.25	.63
Best	(.38; .82)	(.31; .75)	(.48; 78)	(.34; .68)	(.19; .80)	(.16; .53)	(.61; .65)
	.35	.24	.03	.21	.17	12	.33
Worst	(.21; 1)	(.14; 1)	(.20; 02)	(.12; .99)	(.09; .99)	(.06; .98)	(.20; 1)
Labor market							
	.69	.61	.70	.66	.71	.55	.67
Best	(.61; .78)	(.60; .62)	(.67; .74)	(.79; .57)	(.68; .75)	(.58; .52)	(.60; .76)
	.55	.26	.07	.27	.58	.24	.36
Worst	(.38; 1)	(.15; 1)	(.04; .93)	(.36; .22)	(.41; .99)	(1; .14)	(.22; 1)
Welfare							
	.57	.47	.72	.43	.46	.32	.46
Best	(.45; .77)	(.40; .59)	(.71; .73)	(.31; .72)	(.34; .72)	(.21; .67)	(.37; .59)
	.33	.14	.02	.10	.10	.05	.11
Worst	(.64; .22)	(.71; .08)	(.80; .01)	(.67; .05)	(.33; .06)	(.03; .22)	(.40; .07)
Security							
	.72	.48	.62	.53	.63	.78	.73
Best	(.66; .80)	(.36; .71)	(.53; .74)	(.43; .69)	(.53; .78)	(.66; .95)	(.64; .85)
	.56	.24	.05	.33	.43	.61	.24
Worst	(.39; .1)	(.27; .22)	(.67; .03)	(.32; .34)	(.27; .99)	(.44; .99)	(.70; .15)

Note. Precision and recall measures between brackets. All performance scores are based on the same number of articles—that is, 925 per language. Table A7 in the Online Appendix (https://osf.io/86nkx/?view_only=2ea2df83f8dd43e08bf2b0da4bc901fe) provides additional information on the best dictionary applications.

For example, the security frame in Hungarian migration news coverage is the frame for which we eventually obtained a dictionary with the highest overall performance (F1 = .78). Applying this dictionary, 95% of all articles that were manually classified as security frame are recalled, and 66% of all recalled articles were manually classified as security frame.

Generally, F1 scores greater than .5 were achieved when measuring the labor market and the security frame in almost all seven language corpora; the economy and budget frame in the UK, German, and Romanian corpora; and the welfare frame in at least the UK and German corpora. We acknowledge the low performance for some frames/languages of our tested dictionaries, which will be further addressed in the Discussion.

The Contribution of Different Construction Strategies

Turning to the results, overall performance (F1 scores) is significantly influenced by keyword evaluation, whereas keyword translation and preselection strategy did not greatly affect results (see Table 3).[12]

Native speakers' native-language keyword evaluation, in addition to researchers' keyword evaluation of the English keywords, turns out to be relevant to improve precision and overall dictionary performance (F1 scores).

Moreover, we found that the number of required dictionary hits matters. The expected patterns that an increasing number of required hits improves precision, but impairs recall, and vice versa, was observed for our case study across dictionaries for all frames and languages. Best overall performance (F1 scores) was achieved when one or two dictionary hits were defined as the threshold to classify whether an article belongs to a specific frame.

Finally, performance significantly depended on the language corpus and the frame (the concept to be measured with the respective instrument). Recall and overall performance were best for the English UK corpus, which is only natural given that the PAKL originally included only English-language keywords.

[12] We ran several robustness checks for the models. When additionally controlling for dictionary length (i.e., the number of keywords per dictionary), the main conclusions remained the same. However, due to endogeneity concerns between keyword preselection and dictionary length, we decided not to include this variable in the main models. Furthermore, all models where run without the UK news article corpus and also separately for English-language keyword lists applied to the UK corpus only. The main conclusions remained robust.

Table 3. Influence of Dictionary Construction Strategies on Recall,Precision, and F1 Scores.

	β		
	Model 1 Recall	Model 2 Precision	Model 3 F1 Scores
Keyword preselection (Reference: PAKL only)			
CMP only	0.33***	−0.23***	−0.03
PAKL + CMP	0.38***	−0.25***	−0.01
Keyword translation (Yes, applied)	0.05***	−0.06***	0.00
Keyword evaluation (Not only researcher evaluation of English keyword version but also additionally native speaker evaluation)	−0.62***	0.46***	0.16***
Required dictionary hits (Reference: 1 Hit)			
2 Hits	−0.15***	0.10***	−0.02
3 Hits	−0.25***	0.17***	−0.05**
Corpus (Reference: UK)			
Spain	−0.10***	−0.23***	−0.38***
Germany	−0.29***	0.09***	−0.35***
Sweden	−0.06**	−0.16***	−0.26***
Poland	−0.09***	−0.15***	−0.27***
Hungary	−0.01	−0.30***	−0.44***
Romania	−0.14***	−0.08***	−0.25***
Frame (Reference: Security frame)			
Economy & Budget	0.11***	−0.48***	−0.54***
Labor Market	−0.05**	0.03	−0.04*
Welfare	−0.27***	−0.19***	−0.49***
N	2,328	2,328	2,328
df	2,276	2,276	2,275
R^2	.63	.59	.44

Note. Estimates are standardized linear regression coefficients. The number of observations is composed via the different frames multiplied by the different languages (corpora), multiplied by preselection strategies, multiplied by keyword translation steps, multiplied by evaluation strategies, and multiplied by different required dictionary hit scenarios; see details in Table A6 of the Online Appendix (https://osf.io/86nkx/?view_only =2ea2df83f8dd43e08bf2b0da4bc901fe). *p < .05. **p < .01. ***p < .001.

Comparing Approach A and Approach B

We start by contrasting the applications of all obtained multilingual language lists (N = 2,160) to the untranslated corpus (Approach A) versus the usage of all obtained English-language keyword lists (N = 1,008) to the machine-translated English corpus (Approach B).

In comparison to Approach A, Approach B leads to better recall (β = .06, $p < .001$), and to marginally significant lower precision values (β = −.03, $p < .1$); overall, Approach B appears to lead to significantly better measurements (F1 scores) across the six language corpora and the four frames (β = .07, $p < .001$).[13] A possible explanation for this pattern is that the keywords included in a dictionary are more critical for valid concept measurement than the words in a text corpus, given their different role in the analysis. Machine translation errors have subsequently larger consequences if researchers translate the sensitive instrument (Approach A) rather than the corpus (Approach B).

These results are mirrored when the number of contrasting cases is limited to keyword lists from untranslated PAKL ($N = 144$) versus the machine-translated PAKL ($N = 144$). This finding may be informative for the frequently occurring situation in which an English-language keyword list is available, but there are no additional resources or language skills to preselect and refine additional keywords for multilingual applications. In this case, the results indicate that it is less beneficial to machine translate the available English dictionary into the language of the non-English text corpus (most simple version of Approach A). Rather, it is preferable to machine translate the text corpus to English, and subsequently apply the English-language dictionary (Approach B).

Discussion

We provided a comprehensive review of the state-of-the-art in dictionary construction for automated text analysis and presented a much-needed outline for a systematic approach to the methodological advancement of multilingual computer-assisted content analysis for comparative communication science.

We focused specifically on multilingual dictionary construction, where research is still scarce, because most studies tend to focus on English-language countries or refrain from comparative analyses altogether. We presented a review of key steps for automated dictionary construction and outlined strategies for keyword preselection, keyword translation, and keyword evaluation. We could show that the dictionary creation and application techniques presented here, led to the creation of instruments with an F1 score above .5 for the measurement of the labor market and security frames in six different language corpora, and for the measurement of the economy and budget and welfare frames at least for some corpora (the UK, German, and partly the Romanian corpus). This result was notably achieved by the means of comparatively simple and inexpensive construction steps that require, apart from the annotation of the validation article sets and a simple keyword evaluation step, no further involvement of native speakers. It is important to note, however, that the best performing dictionaries were constructed using different approaches, and the individual best practice strategy per language and frame may not be read as a direct decision guideline for other dictionary construction projects.

[13] We present standardized regression coefficients, standard errors, and significance level for the main independent variable of interest. The selection of additional independent variables (used as control variables in the multivariate regression analysis), their effect direction, and significance level were identical in the previously shown regression models (for reasons of space, refer to Table 3).

However, by contrasting the impact of different approaches, we can still formulate some cautionary suggestions that are aimed at helping other researchers on their journey toward multilingual dictionary construction. Taken together, it is beneficial for the performance of a dictionary to gather keywords from previously available dictionaries and/or from annotated multilingual text sources (here, CMP sentences) when available; to evaluate the English version of each keyword (i.e., all originally English-language keywords and the machine-translated equivalent of all original non-English keywords), a construction step easily applicable for any English language speaker; and to comparatively test different thresholds for the number of required hits.

Our results for the machine translation of keywords and corpora are in line with the assessments of other recent studies that perceive the combination of machine translation and automated text analysis as being useful for comparative research (e.g., de Vries, Schoonvelde, & Schumacher, 2018). We found that machine translations of keywords did not impair results and may thus be used for keyword translation on the "journey" toward creating a high-performing multilingual dictionary. Considering the findings for RQ2, if researchers lack the language skills or resources to further improve multilanguage dictionaries (e.g., additional efforts to account for country discourse-specific keywords, morphologic complexity), it is preferable to machine translate the corpus and apply English-language instruments (Approach B), compared with translating measurement instruments (dictionary) and applying the translation to the native corpus (Approach A).

Critically assessed, this study is limited to one topic and seven languages. Furthermore, the performance of many of the dictionaries tested here was less than ideal. This was true even for the most promising monolingual application case, the frame measurement in the English-language corpus with the arguably most perfect English keyword lists. Additional construction steps to further revise, refine, and extend keywords of multilingual dictionaries—and subsequently to improve performance measures—are thus needed and may include additional keyword preselection strategies (e.g., using JRC-EuroVoc, a multilingual thesaurus) and further collaboration with native speakers (via professional linguists or crowd coding). This will also provide the opportunity for improved comparisons between Approach A (i.e., working with multilingual keyword lists and a multilingual text corpus) and Approach B (i.e., working with monolingual keyword lists and a machine-translated corpus), which would involve adding more manual refinement/improvement steps for the multilingual dictionaries used in Approach A. Only then can the alleged benefits of a multilingual dictionary—such as its better consideration of language particularities and country-specific contexts—be fully developed and subjected to more thorough empirical tests.

To put the rather low performance scores for some frames into context, we would like to note that the dictionary's performance was assessed trough a comparison with human manual coding decisions, which represent a useful benchmark, but are themselves not free from error (e.g., visible in the acceptable, but still imperfect, intercoder reliability measures). We wish to take these results generally as an opportunity to repeatedly emphasize the importance of dictionary performance tests for valid automated concept measurement that is beyond chance. The low performance scores may also be read as a call to turn to other approaches for multilingual text analysis tasks. Although both supervised machine learning (Balahur & Turchi, 2014) and topic modeling (Lucas et al., 2015) are less transparent than dictionary approaches and

may require a sufficiently large annotated corpus (at least in the case of supervised machine learning), they could provide promising alternatives.

Generally, the constructed dictionaries are case specific and designed for an examination of the migration discourse in European media. Nevertheless, the strategies and techniques employed during the construction stage, the dictionary application approaches, evaluation procedures, and encountered challenges are not only informative but also draw attention to various issues and questions for other multilingual computer-assisted content analysis projects.

References

Albaugh, Q., Sevenans, J., Soroka, S., & Loewen, P. J. (2013, June). Lexicoder topic dictionaries (Versions June 2013) [Computer software]. Montreal, Canada: McGill University. Retrieved from http://www.lexicoder.com/

Ammon, U. (2001). *The dominance of English as a language of science: Effects on other languages and language communities*. Berlin, Germany: Walter de Gruyter.

Baden, C., & Stalpouskaya, K. (2015). *Common methodological framework: Content Analysis—A mixed-methods strategy for comparatively, diachronically analyzing conflict discourse* (INFOCORE Working Paper 2015/10). Retrieved from http://www.infocore.eu/wp-content/uploads/2016/02/Methodological-Paper-MWG-CA_final.pdf

Balaban, D., Meza, R., & Vincze, O. (2018, May). *The role of religion in Romanian news of the refugee crisis: A clusters-based frame analysis.* Paper presented at the preconference "Refugees, Religious Threats, and Political Radicalization: Theoretical and Empirical Perspectives" of the International Communication Association's 68th Annual Conference, Prague, Czech Republic.

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

Bengston, D. N., & Xu, Z. (1995). *Changing national forest values: A content analysis* (Research Paper NC-323). St. Paul, MN: United States Department of Agriculture, Forest Service, North Central Forest Experiment Station.

Benoit, K., Schwarz, D., & Traber, D. (2012, June). *The sincerity of political speech in parliamentary systems: A comparison of ideal points scaling using legislative speech and votes.* Paper presented at the Second Annual Conference of European Political Science Association, Berlin, Germany.

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.

Burscher, B., Odijk, D., Vliegenthart, R., De Rijke, M., & De Vreese, C. H. (2014). Teaching the computer to code frames in news: Comparing two supervised machine learning approaches to frame analysis. *Communication Methods and Measures, 8*(3), 190–206.

Caviedes, A. (2015). An emerging "European" news portrayal of immigration? *Journal of Ethnic and Migration Studies, 41*(6), 897–917.

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.

Duval, D., & Pétry, F. (2016). L'analyse automatisée du ton médiatique: construction et utilisation de la version française du Lexicoder Sentiment Dictionary [Automated media tone analysis: Construction and use of the French version of the Lexicoder Sentiment Dictionary]. *Revue Canadienne de Science, 49*(2), 197–220.

Eberl, J.-M., Meltzer, C. E., Heidenreich, T., Herrero, B., Theorin, N., Lind, F., . . . 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.

Greussing, E., & Boomgaarden, H. G. (2017). Shifting the refugee narrative? An automated frame analysis of Europe's 2015 refugee crisis. *Journal of Ethnic and Migration Studies, 43*(11), 1749–1774.

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.

Hampshire, S., & Salvia, C. P. (2010). Translation and the Internet: Evaluating the quality of free online machine translators. *Quaderns: Revista de Traducció, 17*, 197–209.

Haselmayer, M., & Jenny, M. (2017). Sentiment analysis of political communication: Combining a dictionary approach with crowdcoding. *Quality & Quantity, 51*(6), 1–24.

Hashimoto, C., & Kurohashi, S. (2007, June). Construction of domain dictionary for fundamental vocabulary. *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions,* 137–140. doi:10.3115/1557769.1557810

Heidenreich, T., Lind, F., Eberl, J.-M., & Boomgaarden, H. G. (forthcoming). Media framing dynamics of the "European refugee crisis": A comparative topic modelling approach. *Journal of Refugee Studies*. doi:10.1093/jrs/fez025

Jacobi, C., van Atteveldt, W., & Welbers, K. (2016). Quantitative analysis of large amounts of journalistic texts using topic modelling. *Digital Journalism, 4*(1), 89–106.

Koehn, P. (2009). *Statistical machine translation.* Cambridge, UK: Cambridge University Press.

Krippendorff, K. (2004). Reliability in content analysis. *Human Communication Research, 30*(3), 411–433.

Laver, M., Benoit, K., & Garry, J. (2003). Extracting policy positions from political texts using words as data. *American Political Science Review, 97*(2), 311–331.

Lawlor, A., & Tolley, E. (2017). Deciding who's legitimate: News media framing of immigrants and refugees. *International Journal of Communication, 11*, 967–991.

Loughran, T., & McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *The Journal of Finance, 66*(1), 35–65.

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.

McLaren, L., Boomgaarden, H., & Vliegenthart, R. (2017). News coverage and public concern about immigration in Britain. *International Journal of Public Opinion Research, 30*(2), 173–193.

Merz, N., Regel, S., & Lewandowski, J. (2016). The Manifesto Corpus: A new resource for research on political parties and quantitative text analysis. *Research & Politics, 3*(2), 1–8.

Mikhaylov, S., Laver, M., & Benoit, K. R. (2012). Coder reliability and misclassification in the human coding of party manifestos. *Political Analysis, 20*(1), 78–91.

Neuendorf, K. A. (2002). *The content analysis guidebook.* Thousand Oaks, CA: SAGE Publications.

Pajzs, J., Steinberger, R., Ehrmann, M., Ebrahim, M., Della Rocca, L., Simon, E., & Váradi, T. (2014, May). Media monitoring and information extraction for the highly inflected agglutinative language Hungarian. *Proceedings of the Ninth International Conference on Language Resources and Evaluation,* 2049–2056.

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

Rauh, C. (2018). Validating a sentiment dictionary for German political language—A workbench note. *Journal of Information Technology & Politics, 15*(4), 319–343.

Reber, U. (2018). Overcoming language barriers: Assessing the potential of machine translation and topic modeling for the comparative analysis of multilingual text corpora. *Communication Methods and Measures.* Advance online publication. doi:10.1080/19312458.2018.1555798

Riffe, D., Lacy, S., & Fico, F. (2014). *Analyzing media messages: Using quantitative content analysis in research.* New York, NY: Routledge.

Roberts, M. E., Stewart, B. M., & Airoldi, E. M. (2016). A model of text for experimentation in the social sciences. *Journal of the American Statistical Association, 111*(515), 988–1003.

Sevenans, J., Albaugh, Q., Shahaf, T., Soroka, S., & Walgrave, S. (2014, June). *The automated coding of policy agendas: A dictionary-based approach* (Version 2.0). Paper presented at the Seventh Annual Conference of the Comparative Agendas Project, Konstanz, Germany.

Silva, C., & Ribeiro, B. (2003, July). The importance of stop word removal on recall values in text categorization. *Proceedings of the International Joint Conference on Neural Networks, 3,* 1661–1666.

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management, 45*(4), 427–437.

Vicsek, L., Keszi, R., & Márkus, M. (2008). Representation of refugees, asylum-seekers and refugee affairs in Hungarian dailies. *Journal of Identity and Migration Studies, 2*(2), 87–107.

Vliegenthart, R., & Roggeband, C. (2007). Framing immigration and integration: Relationships between press and parliament in the Netherlands. *International Communication Gazette, 69*(3), 295–319.

Volkens, A., Lehmann, P., Matthieß, T., Merz, N., Regel, S., & Werner, A. (2015). The manifesto data collection: Manifesto project (MRG/CMP/MARPOR, Version 2015a) [Computer software]. Berlin, Germany: Wissenschaftszentrum Berlin für Sozialforschung.

Wickham, H. (2018). Stringr: Simple, consistent wrappers for common string operations (R Package Version 1.3.0) [Computer software]. Retrieved from http://stringr.tidyverse.org

Wijffels, J. (2018). Udpipe: Tokenization, parts of speech tagging, lemmatization and dependency parsing with the "UDPipe" "NLP" toolkit (R Package Version 0.6) [Computer software]. Retrieved from https://cran.r-project.org/web/packages/udpipe/index.html

Young, L., & Soroka, S. (2012). Affective news: The automated coding of sentiment in political texts. *Political Communication, 29*(2), 205–231.