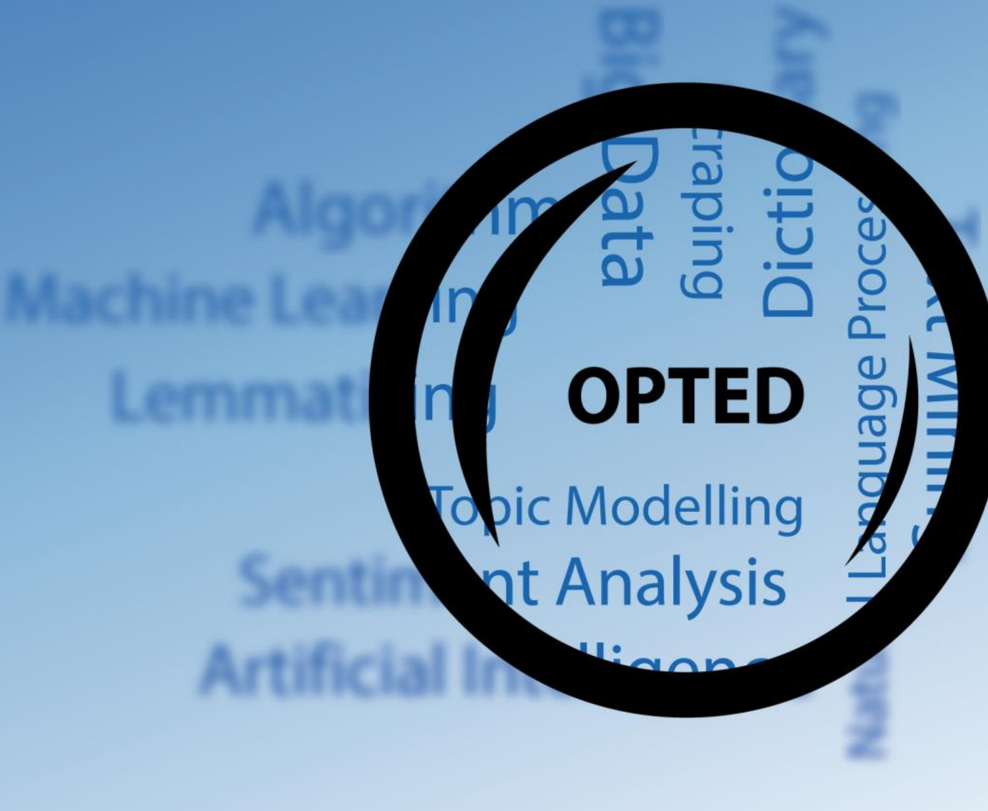
OPTED

Assessing Cross-Lingual Differences in the Performance of Computational Text Analysis Methods: Challenges and Initial Findings

Christian Baden, Alona Dolinsky, Martijn Schoonvelde, Guy Shababo,

Anna Smoliarova, Mariken van der Velden, & Avital Zalik



Disclaimer

This project has received funding from the European Union's Horizon 2020 research & innovation programme under grant agreement No 951832. The document reflects only the authors' views. The European Union is not liable for any use that may be made of the information contained herein.

Dissemination level

Public

Type

Report

OPTED
Observatory for Political Texts in European Democracies:
A European research infrastructure

Assessing Cross-Lingual Differences in the Performance of Computational Text Analysis Methods: Challenges and Initial Findings

Deliverable 6.4

**Authors: Christian Baden[1], Alona Dolinsky[2], Martijn Schoonvelde[2,4],
Guy Shababo[1], Anna Smoliarova[1], Mariken van der Velden[3], & Avital Zalik[1]**

1 *The Hebrew University of Jerusalem*

2 *University College Dublin*

3 *Vrije University Amsterdam*

4 *University of Groningen*

Due date: September 2023

Executive Summary

In this Deliverable, we examine a key challenge in the cross-lingual application of computational text analysis methods (CTAM): Driven by the demand that applied measures be 1) valid within their respective linguistic and cultural context, and 2) equivalent with regard to their capacity to measure key constructs of interest, we argue that validation in cross-lingual CTAM cannot content itself with demonstrating acceptable performance of selected tools on all relevant languages. Instead, we postulate that a systematic effort at validation needs to document that CTAM are capable of yielding commensurable, valid measurement on equivalent textual data in different languages, and that they do so by demonstrably relying on equivalent, valid variation in the linguistic data. To demonstrate validity, a key role falls to parallel corpora as a benchmark for validation, which sustain the expectation that applied measures must yield highly similar, unbiased measurements in each language. In addition, we emphasize the importance of grounding cross-lingual computational measurement in a conceptual understanding of linguistic differences relevant to a given measurement task, ideally supported by a qualitative investigation both of potentially meaningful differences in how equivalent meanings are expressed in different languages, and how the same computational measures respond to these relevant variations in the texture of the data.

To substantiate our arguments, we present five experiments that illustrate key avenues for the cross-lingual validation of CTAM, and can serve both as foundation stones for the much-needed accumulation of experience and conceptual knowledge, and as templates for designing future validation efforts. In our first experiment, we demonstrate that languages not only differ considerably in what information is redacted in the process of stopword removal, but also respond quite differently to available strategies for defining lists of stopwords that can be removed from the data. The second experiment investigates the implications of grammatical gender differentiation, which is pervasive in some and nearly absent in other languages, upon the study of gender-related biases in discourse, pointing out worrisome artifacts raised by the comparative analysis. Experiment three demonstrates that some languages' tendency to form long words comprised of multiple concatenated lemmas profoundly alters the way in which inductive pattern extraction algorithms can access key thematic information in the text. Experiment four, which is still in progress, examines the effects of word order differences – due to grammatical word order or passive mood – upon the information available via low-order n-grams, a key building block in supervised text analysis methods. In the fifth experiment, finally, we investigate to what extent broad linguistic similarities and differences can be leveraged to account for documented losses in the performance of CTAM across different languages.

Besides raising several substantive insights into the complex, at times pervasive differences that arise from applying CTAM to textual data in different languages, a key contribution of this deliverable consists in documenting important design challenges, resource issues and other bottlenecks in the validation and application of cross-lingual CTAM. In this respect, we argue that a central, open research infrastructure can 1) provide a critical link that facilitates conceptual development, needed to bridge present mismatches between available theory and the need for operational knowledge; 2) address the pressing insufficiency of critical resources – notably, parallel corpora and cross-lingually validated tools; 3) foster the urgently needed documentation and exchange of experiences and knowledge in the field; and 4) help establish qualitative, human oversight within the computational loop as an indispensable quality control mechanism needed to ascertain measurement validity within the present, rapid advance of black-boxed CTAM.

1 Introduction

As the field of computational textual research in the social sciences is undergoing its comparative turn (van der Velden et al., 2023), evidence is quickly accumulating that the performance of computational text analysis methods (CTAM) varies across different languages. What is more, available research, as well as recent conceptual inquiries into the challenges presented by multilingual textual research suggest that there are systematic differences in what CTAM do well or badly in different languages, raising concern about consequential measurement biases that threaten the validity of comparative multilingual computational research (e.g., Ho & Chan, 2023; Laurer et al., 2023; Mate et al., 2023; Steimel et al., 2019; van der Veen, 2023; Windsor et al., 2019). In this Deliverable, we address these concerns by presenting concrete strategies for empirically evaluating the performance and measurement bias that arises from linguistic differences.

To do so, we build upon Deliverable D6.3 (Shababo et al., 2023), in which we have identified the main sources and suspected impacts of linguistic differences upon computational measurement, as well as Deliverable D6.2 (Baden et al., 2022), which has proposed a validation framework for cross-lingual computational text analysis. In the present deliverable, we apply these two conceptual contributions to develop and test concrete designs suitable for validating specific uses of computational methods on multi-lingual text. We identify key challenges in the design and implementation of suitable validation strategies, and present initial findings on the nature and scope of language-dependent measurement biases. Specifically, we aim to make two main contributions:

1. We offer practical guidance to researchers aiming to validate computational measures in multilingual research and review key bottlenecks and challenges in the process

2. We gauge the extent of potential biases and measurement problems that arise from different kinds of linguistic differences and applications to help researchers identify critical and less critical concerns in the validation of their computational methods.

2 Objectives and foci of cross-lingual validation

In computational textual research, validation generally refers to the effort spent to ascertain that computational measures pick up on expressions within text whose meanings correspond to those conceptual variables that they are supposed to measure. In multilingual textual research, this means that validation is governed by two overarching objectives: First, to ascertain that measures record expressions that are valid within their respective linguistic and cultural context; and second, to ascertain that measures record expressions in different languages whose meanings are equivalent to one another in conceptual terms (Licht & Lind, 2023). While the former objective is well-familiar from monolingual textual analysis, and merely needs to be ascertained for each language again, the latter objective is specifically relevant to cross-lingual work, and familiar mostly from comparative research.

As we have argued in Deliverable 6.2 (Baden et al., 2022), validation in computational textual research can be conceived of as part of a multi-stage process.

1) **Data validation** ascertains that the textual materials used for analysis are meaningfully comparable with regard to a pursued research objective (see also Palicki et al., 2023);

2) **Input validation** ascertains that the text is pre-processed and represented in a way that focuses subsequent analyses on equivalent, relevant variation (see also Shababo et al., 2023);

3) **Process validation** ascertains that the presented input is processed in equivalent ways by the algorithmic procedure of measurement;

4) **Output validation**, finally, ascertains that the algorithmic procedure yields valid, equivalent measurement performance in each language (see also De Vries et al., 2018; Maier et al., 2021).

Depending on the nature of the measurement task, this validation process can be fairly straightforward or rather complex: For instance, validation is relatively simple if corresponding corpora are easily available in different languages; if no or limited preprocessing is required; if computational algorithms can be trusted to process different-language data in equivalent ways; and if corresponding gold standards are available for each languages, and/or researchers' adequate understanding of the material in each language enables them to manually ascertain measurement validity. However, at each stage, problems are likely to arise: Sampled materials may be comparable only up to some degree; Different preprocessing routines may be necessary to create equivalent representations; The same algorithm may pick up specific contents better in one language than in another; And overall measurement performance may differ, or extract different kinds of patterns, distorting subsequent analyses. For our purpose, it makes sense to group these concerns into two broad kinds of challenges that must be considered when validating algorithmic tools' cross-lingual performance: On the one hand, it is useful to consider the differences in linguistic structure (which shapes the demands on preprocessing; *Input validation*) jointly with the ways in which algorithms access and process textual contents (*Process validation*) – that is, to focus on the way in which a computational method is enabled to recognize specific contents or patterns in textual data (Shababo et al., 2023). On the other hand, we can group the concern over comparable textual materials (*Data validation*) with the concern over comparable measurement outputs (*Output validation*; Tufis & Ion, 2007) – that is, we can focus on the extent to which differential measurement can be confidently attributed to meaningful differences in the analyzed contents, or may arise from irrelevant or artifactual variation in the language of the analyzed material or the performance of applied procedures.

2.1 Do different measurements reflect meaningful differences?

To ascertain that different measurements reflect meaningful differences in the textual material, the most opportune strategy is to rely on textual materials in different languages that are known to contain equivalent information with regard to the research question. 'Parallel' corpora – that is, corpora that contain the same documents translated into multiple languages – are key to this endeavor because they sustain the expectation that equivalent, valid measures should result in identical (or very similar) measurement regardless of the language of the analyzed material (De Vries et al., 2018; Öhman, 2016). Most parallel corpora are derived via human translation. Most famously, the European Parliament translates all of its minutes into all member languages of the EU (Koehn, 2005), as do several multilingual news organizations; but also canonical texts such as the bible, as well as literary texts have been found a useful point of departure for constructing parallel corpora (Christodouloupoulos & Steedman, 2015). Where suitable parallel corpora are unavailable, researchers may also try to create parallel benchmark data themselves (e.g., by using machine translation).

Especially where parallel corpora permit a mapping of contents in fine grain (e.g., at the level of paragraphs or even sentences), researchers can gain confidence that their CTAM indeed yields equivalent measurements. Using parallel corpora, researchers can not only compute agreement measures that offer a standardized assessment of the extent and distribution of differences in measurement performance; by analyzing confusion matrices, researchers can rapidly identify what contents are responsible for differential measurement performance, informing efforts at improving available measures. Similar metrics can even be devised for inductive methods, which measure not classification agreement, but determine the extent to which a tool extracts similar patterns from data in different languages (e.g., Correlation Matrix Distance measures [CMD] can be used to determine whether different topic models adjoin and separate the same sets of documents in a parallel corpus; Maier et al., 2022).

At the same time, researchers still need to additionally ascertain that equivalently recorded differences are indeed valid, i.e., they correspond to meaningful differences in the textual data that can be verified by human readers (Adcock & Collier, 2001; Grimmer et al., 2022; Lind et al., 2017). Similar to monolingual research, this is most easily achieved by obtaining or creating manually annotated gold standards, wherein a part of the material is analyzed manually by human readers. Measurement validity is then ascertained by comparing the outputs generated by CTAM to the human analysis, using precision, recall and F1 metrics to assess the accuracy of classification performance, or other suitable strategies for determining the validity of inductively identified patterns. The main advantage of using parallel corpora at this stage is that the human gold standard needs to be constructed only once, as it applies equally to all included languages.

2.2 Does the computational procedure respond to equivalent information in the text?

To ascertain that computational procedures respond to variation in the textual data that validly express the intended meaning, and do so equivalently across languages, it is necessary to consider how exactly measured meanings are expressed in different languages (Lucas et al., 2015; Shababo et al., 2023). Depending on the measurement task, identifying exactly what information needs to be retained or redacted during preprocessing, and how subsequent algorithmic procedures need to appraise the retained content may be fairly straightforward or quite complex, as discussed below.

Input and process validation are most straightforward whenever relevant meanings are expressed in predictable and fairly limited ways in the textual data. For instance, if the measurement task concerns the extraction of specific named entities or fixed expressions, a dictionary typically suffices to list the relevant expressions that an algorithm needs to record, while any other variation can be safely disregarded. To construct such multilingual dictionaries (for valuable guides, see e.g., Baden et al., 2018; Lind et al., 2019), researchers need to a) obtain all translations and transliterations of all relevant names or expressions in each language (e.g., to record references to the Russian opposition politician Alexei Navalny, one needs to consider variegated case inflections in Russian: Навальный, Навального, etc.; while his English name remains uninflected but may come in various spelling variants); b) adjust truncation rules to accommodate different languages' tendencies to inflect words or affix additional characters (e.g., the Arabic prefix و meaning 'and') ; and c) check for possible conflicts with homographs of different meaning, which may require disambiguation (e.g., efforts to measure references to peace are likely to yield many false positives in Hebrew, where the word for peace, 'שלום', doubles as greeting). Things get more complex for concept references, latent constructs or pragmatic behaviors that permit many paraphrases and indirect references (e.g., there are many more ways of making a promise other than saying 'I promise…'; Zalik & Baden, 2023; evaluative expressions are both variegated and exceedingly difficult to disambiguate, as evidenced by the ongoing debate over the – limited – validity of common sentiment measures; Boukes et al., 2020; Overbeck et al., 2023; van Atteveldt et al., 2021). This is true especially where different cultural norms govern the use of direct vs. indirect references (e.g., recognizing references to homosexuality is likely more straightforward in U.S. English than in Egyptian Arabic; measuring disagreement should be harder in Asian languages where cultural norms sanction the use of direct contradiction or challenging of expressed positions). However, by and large, methods that rely on specific, listable expressions to recognize specific contents can typically be validated relatively easily by identifying relevant passages in random samples of documents in different languages, adjusting keywords and rules, and iterating until all relevant passages are recognized with equivalent quality.

By contrast, input and process validation are complex and may require considerable effort and thought whenever measurement relies on an extraction of broader patterns from the textual material – that is, supervised and unsupervised machine learning procedures (van Atteveldt et al., 2022; Shababo et al., 2023). Unlike the validation of equivalent output, the primary concern here is not whether a supervised algorithm classifies instances correctly, or whether patterns extracted by unsupervised algorithms are interpretable to human analysts (Maier et al., 2021); rather, the question is whether the textual variation that is leveraged by the algorithm for constructing patterns and classifying instances is meaningful and equivalent with regard to the intended measurement task (Akçakir et al., 2023; Chan et al., 2020; Hirst et al., 2014). This is important especially if measurement performance is less than perfect: In order to improve measurement performance, but also to identify possible biases or validity issues in imperfect measurement, it is critical to understand what exactly an algorithm draws upon, and how textual variation is modeled in the process of computational measurement (Baden et al., 2021). Even if output validation on a parallel corpus indicates decent performance, major biases may persist to the extent that this performance is achieved by reliance on non-commensurable information in different languages (e.g., if an algorithm primarily relies on ideological vocabulary to classify texts' partisan leaning in one language, but relies on markers of left- and right-wing parties' executive and opposition roles in another, as was documented in Hirst et al., [2014], measurement will be biased whenever these two are not aligned).

Unfortunately, the way in which both unsupervised and supervised methods extract patterns from textual data responds – at least potentially – to just about any variation in the presented data. In Deliverable 6.3, we have reviewed a wide range of linguistic differences that may create non-commensurable variation in the data: Besides differences in script, languages differ in their morphology, forming, concatenating and inflecting words in very different ways, as well as their word order and various other properties (Shababo et al., 2023). Denny and Spirling (2018) have impressively demonstrated how even minor differences hold the potential to profoundly alter the process of pattern extraction. To list just a few ways in which seemingly minor differences between languages may be responsible for consequential differences down the line: scripts that create many homographs merge expressions of different meanings in one, but not in other languages; morphological variants may create multiple tokens where other languages use only one (Lucas et al., 2015), tilting the frequency distribution of unique character sequences, and often creating differences in what (e.g., grammatical, gender, number, tense) information is available and distinguished for subsequent analysis; word order differences affect the construction of n-grams, and the list continues. To justify the expectation that a method will extract patterns with roughly equivalent meaning from text in different languages, elaborate efforts at preprocessing and harmonization may thus need to be in order (Bender, 2011). To name just a few possible avenues, one strategy suitable specifically for thematic analysis may be to redact all content except for nouns, verbs, and adjectives, which are thoroughly lemmatized to eliminate any grammatical information that might be encoded differently; or one may directly model equivalent tokens, segmenting concatenated words in synthetic languages (e.g., 'Kranke|(n)|versicherung') to match the same expressions in analytic ones (here, 'health insurance'; Lucas et al., 2015); using POS-annotation to augment grammatical differentiation absent in one language, or stemming to remove such differentiation everywhere. Thus, ensuring that pattern extraction algorithms have access to equivalent, relevant variation in each language,[1] while all non-equivalent and irrelevant variation is redacted, we optimize the chance that identical algorithms will indeed construct equivalent patterns despite the underlying language differences - and thereby minimize the risk that differential measurements arise from irrelevant, language-specific variation in the underlying data.

3 Gauging the impact of cross-lingual differences: Five instructive experiments

Based on the above discussion, the *cross-lingual validation of CTAM can be broadly understood as an effort to explicitly model textual representations and algorithms in such a way that the method treats equivalent meanings in equivalent ways, even if these are expressed differently in different languages*. As a key test of such validation, CTAM need to obtain identical measurements from parallel documents in different languages. Accordingly, cross-lingual validation critically depends on a) the availability of parallel corpora, b) a solid understanding of both linguistic and cultural variations that affect the expression of any meanings relevant to the analysis, and c) an explicit grasp of the underlying computational algorithms and procedures. Fostering the creation of these resources and rendering these available to the community of computational researchers in the social sciences and humanities constitutes a key desiderate for a research infrastructure such as has been designed by OPTED.

In order to start assessing the relative impact of specific linguistic differences upon critical uses of computational methods, we designed and implemented five exemplary experiments with three main criteria in mind: First, we selected experiments that relate to common research practices and applications, hoping to identify validation concerns that are directly relevant to ongoing work. Second, we constructed experiments to maximize expected differences, selecting languages with contrasting linguistic properties suitable to profoundly affect a given type of analysis. Third, we designed our experiments to exemplify the bandwidth of possible approaches, prioritizing strategies that can be readily adapted to other applications and linguistic differences. By doing so, we highlight strategies and challenges that are widely applicable in the validation of multilingual computational text analysis, while obtaining a glance at a possible 'worst case scenario' of language-specific biases and measurement implications. While our first four experiments target four specific linguistic features – the encoding of grammar, the distinction of grammatical gender, the formation of compound words from multiple morphemes (or lemmas in computational diction), and prevalent word order – the final experiment looks at linguistic difference holistically, attempting to predict overall performance based on a broad range of relevant linguistic features. In order to inform the construction of an open research infrastructure for computational text researchers, each experiment is intended to offer a transparent workflow that is reproducible and can be adapted to the study of similar multilingual validation challenges, or the replication and analysis of other key multilingual CTAM papers in social science research. Moreover, these experiments are intended to support the accumulation of theoretical knowledge about how exactly and why CTAM perform specific tasks in different ways on different languages, and how such biases can be counteracted.

[1] Equivalence here refers to equivalence in meaning with regard to the intended measurement, which is not necessarily the same as equivalence in terms of grammatical form or function; in fact, to include equivalent variation in the textual representation, it may sometimes be necessary to deliberately preserve some differences in the textual representation, as the same meaning may be expressed in different ways in different languages.

3.1 Experiment 1: Do stopwords matter?

Corpus:
European Parliament Proceedings Parallel Corpus (Parallel Corpus): https://www.statmt.org/europarl/

Languages:
English, Finnish, German, Lithuanian, Polish, Spanish

Linguistic Focus:
Grammatical and Syntactic Function Words

Key measures:
Parts-of-Speech, Topical/thematic patterns extracted by topic modeling algorithms

Rationale:

Different languages express grammatical functions and syntax in different ways, using different balances of separate function words (e.g., articles, conjunctions, prepositions) and morphological variations (e.g., case inflections). As a consequence, so-called ‘stopwords’ - function words that are typically removed prior to analysis, as they are highly frequent but carry low information value (e.g., Lucas et al., 2015; Maier et al., 2021) - vary considerably across languages. Nevertheless, one common approach to the creation of stopword lists is to identify a fixed percentage of most frequent tokens, or to translate from stopword lists in other languages (typically, English). Another, arguably superior but more laborious strategy is to identify all words that belong to certain grammatical categories - notably, articles, pronouns, prepositions, conjunctions (Zheng & Gaowa, 2010). In this experiment, we evaluate what information is redacted by such stopword removal, identifying differences in what is included and excluded in language-specific stopword lists, and what else may get removed by following common pruning approaches that eliminate a fixed percentage of most common words. The selection of languages covers a broad range of language families (Germanic, Romance, Slavic, Finno-Ugric, Baltic) and systematically varies languages’ reliance on morphology to mark grammatical and syntactic functions (from English, which is morphologically poor, to Finnish, which is morphologically rich).

Expectations:

Strongly analytic languages such as English, which allocate most grammatical and syntactic functions to separate words, should result in the removal of notably more information during stopword removal than is the case for languages that rely more heavily on morphological variation to express similar information (Shababo et al., 2023). Depending on a language’s strategy for expressing such information, stopword removal should remove very different information from the text. Moreover, languages that make use of more elaborate case- or gender-differentiation should contain relatively more unique words to perform the same functions. As a consequence, identical pruning thresholds should result in the removal of much more substantive information in languages that use relatively few unique function words than in those that use many, and different thresholds are adequate in different languages to redact most function words while preserving a maximum of substantial information.

Method:

For this experiment, we rely on a combination of word frequency statistics and Part-of-Speech tagging to evaluate different strategies for redacting stopwords from the text. Using word frequency statistics, we identified the top 1% most frequent tokens in each language. For all tokens, we then used the Part-of-Speech tagger included in the *spaCy* Python package to determine the grammatical type of each word. Based on the ordered list of most frequent tokens, we then determined the ratio of valid stopwords, as opposed to substantively informative words (verbs, nouns, adjectives), achieved by higher or lower pruning thresholds. We also determined the composition of the list of valid stopwords from different types of function words. To determine the impact of different pruning strategies, we applied a) a flat frequency threshold; b) a language-specific, optimal frequency threshold; or c) a curated list of valid stopwords in each language. The resulting text was then subjected to topic modeling, using the LDA algorithm available in the *gensim* Python library. Resulting topic solutions were systematically compared with regard to their tendency to group documents in similar ways, expressed by pairwise correlations of documents’ topic composition in the estimated theta matrix, using a

Correlation Matrix Distance (CMD) measure (Maier et al., 2022). The CMD result is a number between 0 and 1, with 0 denoting full identity and 1 denoting no similarity.

Key insights from the process:

There is very limited documentation that comes with stopword lists embedded in common text preprocessing tools and packages. As a consequence, it is often hard to determine how a given stopword list had been compiled - e.g., whether it is primarily based on term frequencies and/or document frequencies, grammatical functions, based on translation, a manual collection effort or a heuristic statistical procedure, and whether validation has taken place. Available stopword lists differ considerably in inclusiveness also within the same language, as one may miss important stopwords when translating from English, or accidentally include non-stopwords when following a frequency-based approach.

Main Findings:

Different strategies for constructing stopword lists lead to markedly different sets of excluded tokens. Frequency-based approaches are considerably more error-prone than those that regard grammatical or syntactic functions. In our experiment, a pruning threshold of 6 - 7‰ was found to be optimal across most languages, but at this cutoff, the pruned tokens still contain a non-trivial amount of semantically meaningful tokens. Based on an analysis of grammatical functions, we arrive at very differently-sized lists of true stopwords (i.e., words with no additional semantic value, such as count words, syntactic or grammatical function words) in different languages. In our experiment, Finnish - and thus contrary to expectations, a morphologically rich language - yielded the most stop words, owing in part to the fact that the same function word typically brought along multiple morphological variants. The shortest list of stopwords was found for Polish, a Slavic language (most Slavic languages don't use articles and mark much grammar morphologically).

Different selections of stopwords measurably affect the output of subsequently deployed pattern-extraction algorithms. Similarly to what Denny and Spirling (2017) reported, different stopword removal routines applied to the same corpus alter the topic solutions found by standard topic modeling algorithms. However, the CMD measure shows that languages are affected differently, with Finnish showing the least impact, while English and Spanish are more profoundly affected.

Implications:

Not all stop word lists are created equal, and the mode of creation matters. Following our investigation, it appears that translating from English is probably the least suitable way to obtain a valid list; relative pruning can be effective if a sufficiently narrow threshold is selected, but for a properly validated list, considering grammatical functions is inevitable.

The impact of stopword removal is rather large in English, and somewhat more limited in other languages, confirming the important role of languages' differential strategies for encoding grammatical information.

When applying stopword removal as part of preprocessing, we need to consider each language separately, selecting an optimal strategy for each language. To this end, it may be necessary to think beyond stopword removal and consider also other ways of removing grammatical information from morphologically richer languages (notably, stemming, lemmatizing). It may be wise to evaluate the robustness of findings given different stopword removal strategies.

3.2 Experiment 2: Does grammatical gender matter?

Corpus:

PressEurope news coverage (Parallel Corpus), obtained from the Intercorp collection at the Czech National Corpus: https://korpus.cz/clarin

Languages:

English, French, German, Polish, Spanish

Key measures:

Topical/thematic patterns extracted by topic modeling algorithms

Rationale:

If individuals of different genders are commonly discussed differently in the news, an inductive thematic analysis should be capable of unearthing gendered patterns in the material – i.e., themes that apply primarily to men, or to women, or to both and neither. It is amply documented that one type of actors that is subject to such gendered practices is politicians, and that news discourse engages in gendered discussions of political personnel (e.g., van der Pas & Aaldering, 2020). We focus here on a subcorpus of PressEurope coverage – namely, news commentary, editorials and op-eds – that should be more expressive than reports in its reliance on gendered role stereotypes and differential standards. To the extent that gendered differences exist in how male and female politicians are discussed in news commentary, such patterns should be notably easier to detect if gender is abundantly marked in the textual material (Shababo et al., 2023). Accordingly, applying an inductive, unsupervised procedure to the selected sub-corpus should result in differential measurement performance when applied to languages that strongly, mildly, or barely encode grammatical gender (Prewitt-Freilino et al., 2012).

Expectations:

Topics should be most gender-differentiated in languages with pervasive marking of grammatical gender – notably, Polish, where nouns, verbs, adjectives and many other words are gender-differentiated; and least so in languages with little gender differentiation – notably, English; German, French, and Spanish should lie in between. This differentiation should also result in a need for more (gender-differentiated) topics in gender-differentiated languages, while less gender-differentiated languages should fail to distinguish some of these.

Method:

For each language, we created three versions of the parallel corpus: One 'raw', unchanged version; one 'minimally preprocessed' version that retains any gender differentiation marked morphologically or by the use of different pronouns, prepositions, etc.; and one 'maximally preprocessed' version that harmonizes these differentiations by means of removing also gendered stopwords and lemmatizing gender-differentiated variants of the same word into a common token (the expectation is that the lemmatized version of most languages will yield results similar to those obtained from a non-gender differentiated language, in our case, English; this lemmatized version serves to separate the impact of grammatical gender (which is preserved in the minimally preprocessed version, and redacted in the maximally preprocessed one) from other sources of difference (which may still distinguish between the maximally preprocessed versions of each language). The raw version serves to test whether neural topic modeling algorithms such as TopicBERT, which ingest raw text, are affected similarly. The corpus was passed to the topic modeling procedure disaggregated into sentences. For each language, we identified optimal K parameters by assessing coherence, exclusiveness, heldout and residual statistics (conveniently offered by the STM R package; Roberts et al., 2019) for both the minimally and maximally preprocessed data. To check robustness, five models were run per language and preprocessing type (using K-2, K-1, K, K+1, and K+2, while keeping other hyperparameters at default value). Resultant topics were compared in three ways: First, we obtained aggregate-level similarity metrics by determining the extent to which different models sorted the same documents in similar ways, using a CMD metric that ranges from 0 (if two topic models result in identical theta matrices) to 1 (if theta matrices are uncorrelated; Maier et al., 2022). Second, we performed a keyword-based classification of topics, rating each topic in each model based on a) its tendency to expressly refer to men or women, and b) its overall sentiment score, using the translated sentiment dictionaries available on Kaggle (Chen & Skiena, 2014).[2] Based on these metrics, we determined whether some models tend to differentiate more between men and women than others, and whether they present women and men using different evaluative tendencies. Finally, any topics that were identified as informative using the latter strategy were analyzed qualitatively to interpret relevant differences and draw conclusions about the underlying biases in the topic modeling procedure.

[2] These dictionaries were developed with the objective to achieve some degree of comparability across languages, validated on Wikipedia data; they show good polarity agreement, but only modest to low correlations of sentiment scores across languages; still, they are among the better (more comparable) dictionaries available.

Key insights from the process:

During validation, we identified numerous non-parallelisms in the parallel corpus: As the PressEurope corpus is created such that coverage is initially written in one language and then translated into all others by the media organization, this translation process involved numerous edits and transformations that reduce the parallelism of the data (see also below). Moreover, we found numerous missing or incomplete entries, likely an outcome of data handling. We had to exclude all non-parallel lines from the study, and recommend careful scrutiny before using parallel corpora even for published and renowned data sources. Moreover, some non-commensurabilities arise from the availability of preprocessing algorithms: Stopword lists are not necessarily equivalent (neither in the sense of including the same types of terms, nor being similarly inclusive), so we relied on our own optimal selection of stopwords validated in the previous experiment instead. Lemmatizers likewise vary in performance. During analysis, it turned out that the few top tokens typically considered for labeling topics are insufficient for detecting meaningful differences in gender focus and sentiment. Instead, we relied on the entire beta matrix to weigh recognized references to men or women, positive or negative sentiment by relevant tokens' probability within a given topic. During analysis, it became clear that grammatical gender has a complex relationship to the gender of discussed actors, mostly because all languages that differentiate grammatical gender do so not only for individuals, but also for organizations, constructs, and other nouns (e.g., in German, the EU is female, the EU Council is male). Since grammatical gender was thought to help differentiate discussions of individuals, we decided to try and focus the analysis on discussions of natural persons only. To this end, we conducted a crowd coding task to decide whether a sentence discussed a female or male individual, multiple individuals (to be included) or organizations/institutions, policies or any other object (to be excluded; a final category applied when there was no recognizable object of discussion). The crowd coding procedure identified a minority of cases as relevant for the narrowed-down analysis, the results of which are still pending.

Main Findings:

Across all languages, if any gender differentiation is identified, it is that men receive slightly more negative treatment than women (who are also discussed notably less frequently). Initial results suggest that retaining grammatical gender differentiation indeed helps identify differential discussions of women and men in those languages when such differentiation is available. The minimally preprocessed variants of the data, which retained such differentiation, overall generated more topics that include explicit gender focus, reflecting differential evaluative tendencies as indicated by the topic-specific sentiment scores. By contrast, very few topics in English and in the fully preprocessed versions explicitly focus on one gender, and also evaluative differences are diminished.

In the English data, removing what little grammatical gender information there is during preprocessing barely alters the analysis. In French and German, the redaction of grammatical gender information left the overall tendencies in the findings unchanged, but several gender-differentiated topics disappeared. Differences are more pronounced in Spanish and Polish. In Spanish, removing grammatical gender information changed the overall findings, from generally negative discussions of topics focused on either gender, with a slightly more negative sentiment for topics focused on men, to almost balanced presentations of both genders. In Polish, we found the opposite: before the removal of grammatical gender information, we did not observe any systematic association between topics' gender focus and their sentiment, but after removal, both genders were associated with positive sentiment. Comparing findings across languages, the analysis of the English data (both versions) yielded the most negative results for both genders, Polish was most positive, with German and French closer to English, and Spanish closer to Polish - differences that may have to do with non-commensurabilities in the performance of the translated sentiment dictionaries. Under all circumstances, the data versions with lesser grammatical gender differentiation (English, maximally preprocessed versions of the other languages) showed less pronounced gender differences in sentiment than those that contained grammatical gender information. Curiously, we found more gender-differentiated topics in Polish after removing grammatical gender information.

Implications:

When studying gender-related differences in text comparatively between different languages, it is essential to consider that different languages express gender differently. Where grammatical gender is encoded, this information changes (and arguably improves) the analysis. This of course may create a dilemma between prioritizing equivalent data representations by removing information available only in some languages, but not others, and attaining the most informative analysis available in each language. One strategy may be to focus on languages that offer similar differentiation with regard to grammatical gender (which may mean dropping English from the analysis); but if findings are compared across languages that encode different amounts of grammatical gender, linguistic differences may account for part of the findings, which need to be separated from substantive differences by means of manual validation.

3.3 Experiment 3: Do polymorphemes matter?

3.3.1 Corpus*:*

European Parliament Proceedings Parallel Corpus (Parallel Corpus): https://www.statmt.org/europarl/

3.3.2 Languages*:*

English, German

3.3.3 *Key* measures*:*

Topical/thematic patterns extracted by topic modeling algorithms

3.3.4 Rationale*:*

Topic models tend to rely on tokens that are frequent but whose use is distributed unevenly over documents (high tf-idf value). They seek to allocate tokens into topics based on their tendency to be co-used. One type of terms that shows such behavior are the components of compound nouns in so-called “analytic” languages - that is, languages that tend to keep different morphemes as separate tokens - such as English: “health” and “insurance”, for instance, are both fairly common but appear, and often jointly, only in certain documents. By contrast, “synthetic” languages tend to concatenate compound nouns into longer words, such that the constitutive morphemes no longer appear as separate tokens. In German, for instance, the same meaning is expressed by the single token “Krankenversicherung”, which is much rarer in natural language use than either “health” or “insurance”, and thus much less likely to be informative toward topic formation (Lucas et al., 2015; Shababo et al., 2023). Moreover, while “health insurance” shares a token with many thematically related nouns - e.g., health policy, mental health, insurance policy - “Krankenversicherung” does not. As a result, it seems plausible that topic modeling will assign central roles to tokens that play important roles in compound nouns in English, which are relatively high in abstraction and obtain their specific meaning only through the compounding. As a consequence, the algorithm is enabled to construct many topics around broad thematic domains, whose key tokens (such as “health”) appear frequently and in stable collocations with other relevant tokens. In German, by contrast, compound nouns are both rare and lack the same stable collocation structure, likely resulting in their discounting during topic formation. We focus on the European parliament corpus, which is characterized by a relatively technical kind of discourse that is rich in compound nouns. Following our rationale, we expect that many of the expressions that are key to the construction of topics in English are unavailable or relatively uninformative in German, such that the algorithm will be unable to form the same topics.

3.3.5 Expectations*:*

Most topics in the analytic language should be arranged around nouns that are commonly used in compound nouns, referring to the broad policy domains, types of activities and policies that the European Parliament debates. The same should be less common in German, as the compounded names of most policies are too rare to play a key role in topics, and the constitutive morphemes appear rarely as independent tokens. Instead, we expect that the German topics tend to be organized around the actors and institutions, as well as the verbs used to describe parliamentary activities, which are not affected by the synthetic language structure.

3.3.6 Method*:*

We use the first 300,000 sentences of the German and English document sets of the Europarl corpus, applying standard preprocessing (stopword removal, stemming, using the built-in preprocessing routines offered by the STM R package [Roberts et al., 2019]; more elaborate lemmatization appears unnecessary, as neither language is morphologically rich). Contents are passed to the topic modeling procedure at the sentence level. We use the STM R package for estimation, identifying optimal K parameters using the same procedure as in the previous experiment. As a robustness check, we again run models with K-2, K-1, K, K+1, and K+2, keeping all other hyperparameters at default value. After estimation, topics were labeled and examined for their reliance on nouns (and specifically, nouns that refer to policy domains and outcomes), names (of institutions, organizations, groups), and verbs. Topic solutions were compared based on these metrics, as well as qualitatively.

3.3.7 Key insights from the process:

Despite its popularity, the Europarl data set is reliably parallel mostly at the level of sentences (with some differences, as some sentences in one language became two in another, which is shown in the data set by adding an empty line before the joined sentence). We initially had intended to use paragraphs or turns, but identified several inconsistencies in paragraph segmentation in the document-level data set.

Another insight is that the in-built stopword redaction and stemming routines in STM do not process German text remotely as effectively as English text: Several stopwords (e.g., "dass") and especially plural forms survived the preprocessing. Better performance should be achievable by checking and adjusting inbuilt word lists and using a dedicated stemmer.

3.3.8 Main Findings:

In line with expectations, we find dramatic differences in what kind of information structures both topic models. In the English data, depending on the model, between half and a third of the estimated topics are indeed topical in the sense that they refer to policy areas, issues and topics of political debate; another third to quarter of the remaining topics concern procedural aspects (negotiation procedures, adoption procedures, implementation, etc.), and the rest collects various patterns of interactive speech (e.g., greetings, evaluations). In the German data, by contrast, only a small minority of topics can be connected uniquely to specific policy fields or issues; far more topics are dedicated to procedural and interactive aspects, or collect groups of qualifications that vaguely generic political terminology, references to actors and institutions, and other aspects of procedural and interactive commentary. What is more, the German model occasionally adjoins quite distinct policy areas that share specific verbs (e.g., poverty, immigration, and terrorism both appear jointly in a topic organized by "to fight"; industry, research and environment are joined by "to support") or other terminology (e.g., a topic that combines women, children, and victims). Even where policy areas are recognizable in the topics, key labels, which would often be long words in German, are mostly absent.

Implications*:*

One key implication of the presented findings is that unsupervised algorithms are highly sensitive to morphological differences; while we tested here the impact of concatenated, long polymorphemes, it stands to reason that also morphological variation, unless treated, holds the power to completely refocus algorithmic pattern extraction. Comparative research should be more than wary of differences detected between patterns extracted from morphologically different languages. Even between closely related languages - English and German are both Germanic languages that share numerous similarities - profound differences can arise during the analysis. At the same time, there is also a potential in this algorithmic sensitivity: Once it is understood how exactly algorithms such as topic models respond to specific variations in how information is parsed into tokens, it may be possible to deliberately aim them at specific kinds of patterns in the data, foregrounding thematic, behavioral, or other information.

3.4 Experiment 4: Do passive mood and word order matter?

3.4.1 Corpus:

Scientific abstracts from selected journals (Parallel corpus, self-created)

3.4.2 Languages:

English, Russian, Korean

3.4.3 Key measures:

Part-of-Speech, frequency distribution of n-grams, SML performance on detecting actors, actor-specific sentiment, etc.

3.4.4 Rationale:

In English, sentences typically begin with the grammatical subject, which often is the agent of a sentence, and the verb, indicating the relevant activity. Many algorithms depend on this characteristic front position and adjacency of agent and activity. Both word order and passive mood potentially undermine this adjacency: Different word orders may adjoin subject and object, which are harder to tell apart than subject and verb, or locate the subject far away from the predicate, complicating its confident identification. The same is true for passive mood, which additionally may result in the agent being redacted altogether in a sentence. Accordingly, algorithms that rely on an n-gram representation of the text to extract key information about actors and their activities and attributes (e.g., for studying media bias, influence, representation, or blame attribution) should perform notably worse whenever languages don’t follow SVO word order (such as Korean, which follows a dominant SOV word order) or rely pervasively on passive mood (as do some registers in Russian, such as academic discourse; Boginskaya, 2022; Shababo et al., 2023). Likewise, freer word orders, which typically mark grammatical roles morphologically and not via position, as in English, can derail such algorithms. As it is much less likely that subject and verb, or subject and attributes appear within the same n-gram, both efforts at identifying the agent and attempts to correctly attribute expressed activities, sentiments or other qualities should suffer recognizable losses in performance.

Expectations:

For our analysis, we expect that a classic, morphologically rich SVO-ordered language such as English facilitates the identification of agent-action and agent-attribute patterns, as these are expressed nearby, likely found in the same n-gram, and uninflected. For Korean, subject and verb are placed far apart, likely eroding performance. In Russian, SVO word order is less dominant in English, while especially in technical and formal registers - such as our case of scientific abstracts - passive mood is very common (Stokes, 2020). As passive mood tends to separate subject and verb, if the subject isn’t redacted altogether, and intersperse additional auxiliary verbs (Sepehri et al., 2020), longer n-grams are needed to capture the required information. As a result, we expect a marked deterioration in performance. In addition, the reversal of word positions in passive mood - foregrounding the object - might lead to a systematic misattribution of activities, sentiments and other qualities to the object, rather than the agent of a sentence.

Method:

We created two parallel corpora, one for English and Russian, and one for English and Korean, by scraping the abstracts published in well-established social science and humanities journals in Russia and South Korea, over the past years. For each corpus, we created several different sizes of n-grams, using up-to-2-gram, up-to-3-gram, and up-to-4-gram representations of the text. For Russian, we also created a version with and one without stopword removal prior to tokenization, to examine the impact of stopword removal upon the ambiguity of passive constructions. For a random subset of documents, we use a POS-tagger to identify the grammatical functions of words prior to the creation of the n-gram representation, enabling us to compare the relative frequency of n-grams that contain specific kinds of information, along with a manual examination of the most common n-grams in each language. Subsequently, conduct several classic machine classification tasks - identifying the agent, sentiment classification, etc. - to compare algorithm performance for each language pair and n-gram representation.

3.4.5 Key insights from the process:

It is rather difficult to identify suitable parallel corpora that reach beyond European languages. We initially had hoped to rely on academic abstracts translated into multiple languages, however, such practices are too rare and unsystematic to yield a suitable base for comparison. Bilingual parallel texts - typically including a domestic language and English - are still reasonably common, but already trilingual materials are rare.

While there is quite a bit of linguistic literature and theory on word order differences between languages, the same is not true for differences between registers, or the prevalence of stylistic differences, such as the use of passive mood in formal registers. There is a need for more systematic investigation and theorization on consequential linguistic variation below the level of languages, concerning different language practices that depend on genre, register, etc.

We also know too little about the process of creating translations such as the translated abstracts to evaluate possible sources of error (e.g., due to machine translation, which is prone to mistranslating technical jargons). For instance, is the convention to translate more or less literally, or is there an effort to adapt foreign-language abstracts' English translations to 'Western' scientific communication styles and conventions? More transparency about the mode of translation would be helpful

3.4.6 Main Findings:

This experiment is still in progress. Findings will be reported once they are verified.

3.4.7 Implications:

This experiment is still in progress. Implications will be reported once they are verified.

3.5 Experiment 5: Can we predict measurement performance differences between languages?

3.5.1 Corpus:

Debate records of the European Parliament 1996-2011 (Parallel corpus, created by Proksch et al., 2019)

3.5.2 Languages:

Bulgarian, Czech, Danish, Dutch, English, Estonian, Finnish, French, German, Greek, Hungarian, Italian, Latvian, Lithuanian, Polish, Portuguese, Romanian, Slovakian, Slovenian, Spanish, Swedish

3.5.3 Key measures:

IVs: typological indicators from existing databases of structural properties of languages; DVs: performance metrics of multilingual sentiment analysis tool replicated from Proksch et al. 2019

3.5.4 Rationale:

With the comparative turn in computational text analysis, there is a growing body of work that reports differential performance of computational tools on multilingual corpora. For example, Proksch et al. (2019) use machine-translated sentiment dictionaries to capture conflict in legislative speech in all 24 official EU languages. Licht (2023) relies on multilingual sentence embeddings to determine how well a topic classifier trained on one set of source languages can predict topics in a different target language. However, we still know little about when and why performance varies across languages and between specific language pairs. Therefore, we ask to what extent structural linguistic differences - notably, regarding their syntax and morphology - can account for observable differences in the cross-lingual comparability and performance of CTAM in the social sciences (Shababo et al., 2023).

3.5.5 Expectations:

Broad syntactic features that govern the content of a word's context (such as word order) should affect context-sensitive measures, such as word embeddings; but also sentiment dictionaries that consider negated terms should perform better if word order requires negators to be adjacent. Depending on the nature of compared languages and the specific measure in question, different expectations apply.

3.5.6 Method:

We replicate Proksch et al., 2019's construction of a multilingual sentiment dictionary, followed by integration with data on linguistic features. Based on this replication, we examine the correlation between automated translations of dictionaries and hard-coded gold-standards, as well as automated translations of dictionaries and the English dictionary across languages. At this point, we have focused on a comparison between analytic (e.g., English) and synthetic (e.g., Hungarian) languages and across different language families. Future work will include additional, and more specific language features as predictors.

3.5.7 Key insights from the process:

Significant effort was necessary to replicate the construction of the multilingual sentiment dictionary constructed in Proksch et al. 2019, mostly due to recent changes in relevant code packages and functions, before it was ready to be deployed for analysis. It should also be noted that we were able to reproduce the findings from the original paper chiefly thanks to the authors' extensive work to make the reproduction materials publicly available.

Substantial and substantive knowledge of linguistics is needed to meaningfully cluster languages by relevant features. Syntactical and morphological differences between languages are far from uni-dimensional but involve numerous features, whose complex differences complicate efforts to cluster languages into distinct categories that are sufficiently cleanly delimited to inform a meaningful analysis. Much of the rich available linguistic information[3] requires considerable further work to be translated and transferred toward meaningfully informing social scientific textual research.

3.5.8 Main Findings:

Clustering languages by high-level linguistic features turns out to be too crude a way to gain insight into possible patterns of relationships across languages in our replication of a key paper in political science. Further analysis at the level of more specific linguistic differences is needed to tease out which language features matter and why. However, this endeavor is further complicated by the large number of potentially relevant features, given the limited range of languages in which suitable materials exist to analyze.

3.5.9 Implications:

While linguistic theory identifies many language features that affect the encoding of meanings targeted in social scientific textual analysis, it is not obvious at what level of abstraction these features impact CTAM measurements. At this stage, work in social science that relies on multilingual CTAM would benefit from further theoretical development and a strengthened dialogue between social science, computer science and (computational) linguistics.

4 Lessons for cross-lingual validation

As the above experiments show, cross-lingual validation may include a wide range of facets: Establishing the validity and equivalence of measures may take place in abstract, studying the differential effects of language features on specific algorithms in general, or already in relation to specific envisaged measurement tasks; it may target specific gold standards in deductive classification, or investigate the validity of inductive treatments of the materials; and it may focus on specific linguistic differences, or examine performance across a wide range of potentially relevant variations in language. The above experiments illustrate some of these avenues, but of course remain far from exhausting the many directions in which cross-lingual validation will need to develop. Rather they should be understood in the context of OPTED's effort to design an open research infrastructure for the accumulation of knowledge and exchange of resources in computational textual research,

[3] E.g., World Atlas of Language Structures (https://github.com/cldf-datasets/wals/releases); AUTOTYP (https://github.com/autotyp/autotyp-data#dataset-overview)

and are intended primarily as examples that can inspire further research and the conceptual integration of consistent findings.

In the rest of this deliverable, we briefly discuss several issues that emerged from our validation efforts and appear valuable for guiding future work. In the following section, we discuss practical issues, resource concerns and bottlenecks that complicate the conduct of cross-lingual validation. The subsequent section then draws upon our preliminary findings to discuss initial insights into the scope and pertinence of cross-lingual validity issues in CTAM. Without any claim to finality or completeness, we believe that the items mentioned hereafter should form part of a research agenda aimed to supply key resources and capacities for computational textual research (Baden et al., 2021), and will be valuable to consider before designing cross-lingual research applications in the social sciences.

4.1 Notes from the kitchen: Bottlenecks in cross-lingual validation

Despite their different foci and design strategies, all presented validation experiments critically depend on a range of key resources that enable a confident evaluation of CTAM's comparative performance across different languages. Chiefly, each design departs from a necessary conceptual understanding of how recognizable linguistic differences might affect intended measures (Baden et al., 2022; Bender, 2011); other key resources include suitable parallel corpora, as well as well-documented and -validated preprocessing tools tailored to the specific needs of different languages. Based on our experience, however, all three of these resources remain in short supply, and riddled with issues and challenges that present themselves often only during the validation process.

4.1.1 Linguistic theory is not well-matched to social-scientific validation needs.

Cross-lingual differences are far too many and complex to permit open-ended strategies for cross-lingual validation: Not only are there typically several interrelated differences between languages that might account for important performance differences or biases, such that observed differences are difficult to attribute to specific causes - a necessary prerequisite for being able to mitigate or account for resulting validity concerns; but even the mere process of identifying meaningful differences in the performance of CTAM on different-language data is constantly at risk of getting lost in entropy unless a clear conceptual focus is available to guide the evaluation.

At the surface, and thankfully, there exists a wealth of linguistic theory available to formulate expectations and design targeted validation experiments (Payne, 2017; for an invaluable collection, see Haspelmath et al., 2005's World Atlas of Linguistic Structures WALS). However, at a closer glance, such linguistic theories are often difficult to operationalize with regard to their implications for social scientific text analysis and its validation. There are several aspects to this. One is that much linguistic theory is highly specific to very particular variations, each of which is too rare to systematically bias textual measurements - even if their joint effect may be significant. At the same time, inversely, known, broad differences between languages (e.g., at the level of dominant word order or morphological principles) do not necessarily characterize actual language use consistently, and numerous important variations may persist: For instance, theory would suggest that both English and Russian are SVO-ordered languages - but in specific formal registers, Russian's dominant word order is overlaid by its preference for passive mood, with the result that the theoretical classification does not actually apply very well for informing hypotheses during the validation process. Moreover, different language features often interact, undermining efforts to gauge their isolated effects: For instance, languages not only mark grammatical gender more or less pervasively, they also differ in how they conceptualize grammatical gender in relation to the social phenomena described (e.g., German uses neuter gender for certain biologically female referents; in Hebrew, many institutions are grammatically female, while organizations tend to be male). Several broad assumptions about the general impact of specific linguistic features do not hold up to scrutiny if examined in relation to a specific measurement task (e.g., many stopwords in gendered languages still carry valuable gender-related information). As a consequence, seemingly clear linguistic differences often translate into rather messy language uses relevant to a given measurement task.

4.1.1.1 What can be done?

To better bridge the gulf that separates linguistic theory from social scientific validation needs, a first desiderate concerns the conceptual translation of theoretical knowledge: Much remains to be said about how linguistic features affect language use in specific textual genres, registers, and contexts relevant to social scientific textual research, and distort the measurement of those semantic and pragmatic meanings that social scientists tend to be interested in (Shababo et al., 2023). Such an effort at theory building would need to actively consider the ways in which different linguistic features interact to shape the way in which commensurable meanings are expressed in different languages, and cannot take the form of linear translations and transfers. A central hub where relevant conceptual knowledge can be accessed and linked to empirical efforts at validation should therefore form a valuable part of a desirable research infrastructure for cross-lingual textual analysis. Second, there is also a place for a careful, qualitative appraisal of the specific textual phenomena under consideration prior to the design of suitable validation procedures. Many times, the specific discursive practices relevant to a given measurement actualize only selected linguistic differences, while others may still be true in general, but less relevant for a given measurement task. Such a qualitative investigation may not only help researchers to formulate more specific, more accurate expectations, but also help target validation experiments as well as measurement tools to zoom in on those passages and variations in text that actually matter (for instance, in our second experiment, we decided that measurement performance could be improved by excluding references to commonly discussed organizations and institutions prior to classification, focusing recorded gender differences on natural persons). Also after validation, a careful, manual analysis of identified errors can reveal important insights into possible causes and remedies (Ho & Chan, 2023). Similarly to the efforts at integrating theory, also such empirical efforts at pinning down consequential cross-linguistic differences should be collected and rendered accessible through a central knowledge hub – especially since many valuable observations in the course of cross-linguistic research practice are likely too small to be publishable in their own right. Third, it is useful to devise quantitative strategies for detecting expected issues and biases especially in the output generated by inductive pattern-finding algorithms. Eyeballing a few top tokens per topic, for instance, is unnecessarily effortful, unsystematic, and limited in its appraisal of relevant information (Maier et al., 2021). Based on a conceptually founded expectation, however, it is relatively easy to compare the salience of specific groups of terms within the entire vocabulary associated with each topic (e.g., if one expects systematic differences between languages with regard to evaluative tendencies, even a quick-and-dirty sentiment scoring of topics may suffice to either substantiate or dismiss this expectation). An open research infrastructure will be valuable for further facilitating the construction and exchange of tools for facilitating quantitative validation.

4.1.2 Parallel corpora are rare and often less parallel than assumed.

Multilingual comparative textual research does not normally use parallel corpora, for the simple reason that identical measurements could be expected for each language. For the very same reason, however, parallel corpora are a key resource in cross-lingual validation (e.g., De Vries et al., 2018; Öhman et al., 2016). However, one key insight from our validation experiments is that suitable corpora are often less ‘parallel’ than required to sustain the confident expectation that differential measurements indicate biases in the measurement procedure. Even when using well-established and published parallel corpora, important language- and context-specific differences in the textual material continued to add noise to our comparative validation. One rather obvious, but annoyingly persistent source of noise is the continued presence of technical errors and glitches: Especially machine translation occasionally commits translation errors, or fails to translate (parts of) sentences or documents (Maier et al., 2022). Also automated scraping and data-preparation procedures are error-prone, leaving (for instance) html tags or boilerplate content behind that may subsequently derail machine translation and sentence alignment. Segmentation routines are likewise often inconsistent, with the effect that one sentence or paragraph in one language may map upon multiple in another - and not all such instances are necessarily caught during data curation. Where alignment issues propagate (e.g., if units are assigned running numbers, the first misaligned document ruins alignment for everything that follows), such issues can do pervasive damage; however, on the whole, our impression was that most such errors primarily added noise and depressed performance, but did not result in strong systematic biases.

The same is, however, not true for another kind of difference, which arises from the fact that corpora can be ‘parallel’ in somewhat different senses of the word. While machine-translated corpora are more or less parallel in a literal sense (save for translation errors), most human translations simultaneously adjust translated texts to their new cultural and generic context, which may result in a range of idiosyncratic adaptations (Bellos, 2011). To begin, translators are aware that what has been “the government” in English may need to become “the British government” in French, and vice versa, that “Macron” will suffice where the original had referred to “French President Emmanuel Macron” (see also Baden et al., 2018). Second, translations regularly adjust the amount of detail presented to match their relevance to foreign audiences. Such domestication practices may result not only in more succinct or elaborate paraphrases, but also in the omission of entire sentences deemed uninteresting to a specific audience, or the addition of detail and context suitable to link the translated content to domestically salient debates and narratives. Third, adjustments may also respond to differential generic conventions, such as a more or less evaluative and subjective style in different journalistic cultures, or different etiquettes in parliamentary debates (Hemppanen et al., 2012). Translators may prefer lexical choices than would be appropriate in a country’s cultural context over a closer translation deemed culturally inappropriate; or they may adjust stylistic practices to generic conventions, switching between active and passive mood, or adding or redacting ornamental speech, etc. Such practices of cultural domestication differed notably between different kinds of parallel corpora, reflecting the purpose of translation (Bellos, 2011): Parallel corpora created with the intention of cross-lingual validation tend to maintain rather narrow equivalence at a lexical level; legal translations such as those created by the European parliament’s translation service are faithful at the level of semantic content while adjusting phrasings to culturally appropriate styles; and translations created for the purpose of rendering texts available to foreign audiences (e.g., multilingual press coverage; translated abstracts) regularly take considerable license to adjust contents to the different cultural setting and debate.

4.1.2.1 What can be done?

To manage such limits to the parallelism of multilingual corpora, a first, general advice is to carefully check the alignment of documents, both to avoid pervasive misalignment and to get a sense of the nature and extent of cultural domestication practices and variation in the data. A central research infrastructure can contribute to this effort not only by documenting known issues and offering background information about key linguistic resources, but also by facilitating access to high-quality resources or even offering its own, curated and validated reference corpora. Second, it is helpful to consider at what level of meaning an intended validation needs to establish equivalence: If a measure is intended to capture equivalent pragmatic language behavior, it may indeed be fortunate when compared documents are parallel at a pragmatic level, but not necessarily literally, thus reflecting more accurately what people would actually say and write in each language; for other validation purposes, literal equivalence may be important, and machine-translated corpora may be preferable accordingly. It is thus generally advisable to consider not only the type of discourse and contents present in a parallel corpus, but also its mode of translation, and resulting sense of parallelism. If a corpus’ type of parallelism does not adequately match the intended validation, it may still be possible to identify or construct subcorpora that come closer to the required kind of parallelism. A central research infrastructure will be instrumental for fostering awareness of the consequences of relying on different kinds of resources, and enabling researchers to make informed choices when designing their own validation strategies.

4.1.3 Tools with identical labels sometimes do different things.

During the text preprocessing stage, one common challenge is to find tools and language resources that perform specific preprocessing tasks not only with high accuracy, but also in ways that do not add new differences to the existing linguistic variation. Of course, one may sometimes exactly need different tools to achieve similar outcomes - e.g., when the grammatical information expressed by stopwords in one language is encoded by morphological variation in another language (Shababo et al., 2023). However, even when different languages require the same preprocessing treatment, problems may arise. High-quality resources may simply be unavailable for many languages, especially for more demanding tasks such as lemmatization or POS annotation (Baden et al., 2021; Bender, 2011). But also when tools exist, important differences may remain between these that add, rather than help remove, differences to the data in different languages. For example, as we have shown in our first experiment, stopword lists that are translated from English remove systematically different information from the text than those created based on a relative pruning approach, or through an investigation of grammatical functions. Sentiment tools are notoriously hard to compare across languages: Translations may faithfully maintain innate construction logics and avoid major differences in inclusiveness, but disregard that translated terms may not the most common ways of expressing a sentiment in another language (Proksch et al., 2019; see also Lind et al., 2019; Cheng & Zhang, 2023); especially machine translations might reduce linguistic variability if they translate distinct words in one language into the same word in another. Separately constructed tools (e.g., most pretrained tools) may be well-validated in their original language, but may pick up and miss different meanings across languages (Araújo et al., 2020). Inversely, tools that do strictly the same thing (e.g., stemmers) may have drastically different effects when applied to different languages (e.g., some languages are easy to stem, as suffices are regular, as in Turkish, or relatively rare, as in English; but may mostly create a mess in morphologically more complex languages).

4.1.3.1 What can be done?

To control the performance of tools, it is wise not only to consider the respective linguistic features in each language that a tool is supposed to process (e.g., what does it mean substantively to redact stopwords in Hungarian, in Dutch or in Korean?), but also to carefully examine the provenance and construction of the respective tools: How were embedded word lists constructed, or what exactly were tools pre-trained on? How does the algorithm itself function - and is there reason to think that it will treat a given language adequately? Unfortunately, such efforts are undermined by the often absent or haphazard documentation of preprocessing tools – a key lacuna that a research infrastructure may help address. Given preprocessing's continuing (but mistaken! See Denny & Spirling, 2018) treatment as a relatively trivial interventions, many text processing packages offer built-in preprocessing options for many languages that are applied in a 'one-size-fits-all' fashion (Bender, 2011), without permitting researchers to interrogate or tweak the algorithm. A key desiderate in this respect is therefore the systematic documentation and comparative validation of preprocessing routines, ideally in connection with a conceptual debate over the known implications of preprocessing interventions and criteria that govern their applicability.

4.2 Notes from the analysis: The differential impact of cross-lingual variation

At the present time, the analysis of all of the above experiments is still underway, and whatever findings we can draw upon remains preliminary. What appears safe to say at this point already is that the implications of cross-lingual variation upon computational measurement are complex and intensely dependent on the nature of intended measurement. We have so far found no support for common assumptions that more similar (e.g., related) languages generally raise fewer issues in comparative treatment, or that certain languages generally 'work well' or worse with CTAM (see also Paridon & Thompson, 2021). While the availability of suitable language resources still presents an obstacle for the analysis of less well-resourced languages (Bender, 2011; Maté et al., 2023; van der Veen, 2023), linguistic variation is too complex to permit simple generalizations about the implications of language-specific structural features upon CTAM. To take just an example, German shares many qualities with English, to which it is closely related, however, experiment three documents pervasive differences that arise from one key difference in word formation; yet, there is little reason to assume that other types of algorithms are necessarily affected in similar ways. Even our own assumption that most CTAM work better on English text (Baden et al., 2021) seems to be challenged by the finding that English language text is disproportionately sensitive to stopword removal routines.

What we can say is that many differences appear to be overall minor, likely as a result of simultaneous, opposite effects and complex interactions between these. Where different distortions cancel one another out, in the luckiest of circumstances, we may obtain minor biases and mostly random noise added to the analysis. At the same time, at least some of the noted differences are pronounced, and of a kind capable of derailing valid comparative analysis. For instance, our third experiment suggests that a comparative analysis of German and English parliamentary records will likely invite a researcher to conclude that German politicians are way more obsessed with procedure than they in fact are, while the English debate will appear much more focused on actual issues - even if the underlying data is substantively identical. Likewise, one might conclude that gender plays a lesser role in English than in Polish news coverage, when an unknown proportion of identified differences are primarily reflections of the much more pronounced marking of grammatical gender in Polish discourse. Without being able to separate methodological artifacts created by linguistic differences from valid differences in the analyzed materials, there is a real danger of mistaking spurious differences for meaningful information, or missing valid effects amid the methodological noise.

For a more thorough analysis and understanding of cross-lingual validity issues in the use of CTAM, it appears inevitable to adopt a much more nuanced approach capable of regarding the variegated interactions and contingencies in how linguistic features influence computational measurement. One key prerequisite for such an endeavor is the systematic accumulation of experiences gleaned during the research process, including both applied uses and validation efforts of CTAM in various languages (Baden et al., 2021; Bender, 2011). It is arguably here where a European research infrastructure will make its most profound impact. Beyond this, there is an acute need for explicit theorizing. On the one hand, cross-disciplinary collaboration between researchers in the social sciences and linguistics will be instrumental for transferring linguistic knowledge to the domain of social scientific text research, both in relation to CTAM and beyond. On the other hand, qualitative and inductive discourse research holds ample potential for informing the much-needed theorizing effort, as it unites social scientific research concerns with an acute awareness of language and textual nuance (e.g., Danziger, 2023; Hellinger & Ammon, 1996; Stubbs, 1997). Finally, much additional validation work is needed all across the computational social sciences. Importantly, validation efforts need to be minutely documented, also (especially!) when they fail or yield less than satisfactory results, collected and synthesized to offer an empirical check on the concomitant development of methodological theory. One key contribution of a research infrastructure will thus be to create a space where also minor, unsuccessful, and highly specific validation experiences can be collected that would otherwise remain undocumented or hidden away in appendices and footnotes. The described effort will require much collaboration across and beyond the social sciences, and benefit from several consequential changes in how computational text analysis research is conducted: From researchers engaging cross-lingual validation; to reviewers requesting validation efforts beyond summative output validation, to shed light on the implications of preprocessing routines and computational algorithms; to editors facilitating (or mandating) the documentation of validation results, along with open data and materials suitable to enable collaborative investigations; to developers opening up and documenting their code and embedding explicit choices into their packages; to the research community as a whole, which cannot content itself with blindly assuming that the latest tools offer somehow adequate, comparable performance, when every evidence suggests otherwise. The OPTED research infrastructure is intended to build a key venue where the much-needed accumulation of theoretical, methodological and practical knowledge on multilingual CTAM can commence, connecting experiences, findings, researchers, and research agendas. We hope that with our work, we have helped build a foundation for this effort.

References

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

Akçakir, G., Jiang, Y., Luo, J. & Noh, S. (2023). Validating a mixed-method approach for multilingual news framing analysis: A case study of COVID-19. *Computational Communication Research.*

Araújo, M., Pereira, A., & Benevenuto, F. (2020). A comparative study of machine translation for multilingual sentence-level sentiment analysis. *Information Sciences, 512*, 1078-1102.

Baden, C., Jungblut, M., Micevski, I., Stalpouskaya, K., Tenenboim-Weinblatt, K. Berganza Conde, R., Dimitrakopoulou, D., & Fröhlich, R. (2018). *The INFOCORE Dictionary*. https://osf.io/f5u8h/

Baden, C., Dolinsky, A., Lind, F., Pipal, C., Schoonvelde, M., Shababo, G., & van der Velden, M. A. C. G. (2022). Integrated standards and context-sensitive recommendations for the validation of multilingual computational text analysis. *OPTED Deliverable D6.2*. https://www.opted.eu/results/project-reports/

Baden, C., Pipal, C., Schoonvelde, M., & van der Velden, M. A. C. G. (2021). Three gaps in computational text analysis methods for social sciences: A research agenda. *Communication Methods & Measures, 16*(1), 1-18.

Bellos, D. (2011). *Is that a fish in your ear? Translation and the meaning of everything*. Penguin.

Bender, E. M. (2011). On achieving and evaluating language-Independence in NLP. *Linguistic Issues in Language Technology, 6*(3), 1–26.

Boginskaya, O. A. (2022). Functional categories of hedges: A diachronic study of Russian research article abstracts. *Russian Journal of Linguistics, 26*(3), 645-667.

Boukes, M., van de Velde, B., Araujo, T., & Vliegenthart, R. (2020). What's the tone? Easy doesn't do it: Analyzing performance and agreement between off-the-shelf sentiment analysis tools. *Communication Methods & Measures, 14*(2), 83-104.

Chan, C. H., Zeng, J., Wessler, H., Jungblut, M., Welbers, K., Bajjalieh, J. W., & Althaus, S. L. (2020). Reproducible extraction of cross-lingual topics (rectr). *Communication Methods and Measures, 14*(4), 285-305.

Chen, Y., & Skiena, S. (2014). Building sentiment lexicons for all major languages. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, 2,* 383-389.

Cheng, C. Y. & Zhang, W. (2023). C-MFD 2.0: Developing a Chinese moral foundation dictionary. *Computational Communication Research.*

Christodouloupoulos, C. & Steedman, M. (2015). A massively parallel corpus: The Bible in 100 languages. *Language Resources and Evaluation, 49*, 375-395.

Danziger, R. & Kampf, Z. (2020). Interpretive Constructs in Contrast: The Case of Flattery in Hebrew and in Palestinian Arabic. *Contrastive Pragmatics 2*(2), 137–167.

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

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

Haspelmath, M., Dryer, M., Gil, D., & Comrie, B. (Eds., 2005) *The World Atlas of Language Structures.* Oxford University Press.

Hellinger, M. & Ammon, U. (Eds., 1996). *Contrastive sociolinguistics.* Mouton de Gruyter.

Hirst, G., Riabinin, Y., Graham, J., Boizot-Roche, M., & Morris, C. (2014). Text to ideology or text to party status? In B. Kaal, I. Maks, & A. van Elfrinkhof (Eds.), *From text to political positions: Text analysis across disciplines* (pp. 93–115). John Benjamins.

Ho, J. & Chan, C.-H. (2023). Evaluating transferability in multilingual text analysis. *Computational Communication Research, 5(2).*

Kemppanen, H., Jänis, M., & Belikova, A. (Eds., 2012). *Domestication and foreignization in translation studies.* Frank & Timme.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. *MT Summit.*

Laurer, M., van Atteveldt, W., Casas, A. & Welbers, K. (2023). Lowering the language barrier: Investigating deep transfer learning and machine translation for multilingual analyses of political texts. *Computational Communication Research, 5(2).*

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

Licht, H. & Lind, F. (2023). Going cross-lingual: A guide to multilingual text analysis. *Computational Communication Research.*

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

Lind, F., Gruber, M., & Boomgaarden, H. G. (2017). Content analysis by the crowd: Assessing the usability of crowdsourcing for coding latent constructs. *Communication Methods & Measures, 11*(3), 191-209.

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

Maier, D., Baden, C., Stoltenberg, D., De Vries-Kedem, M., & Waldherr, A. (2022). Machine translation vs. multilingual dictionaries assessing two strategies for the topic modeling of multilingual text collections. *Communication Methods and Measures, 16*(1), 19–38.

Maier, D., Waldherr, A., Miltner, P., Wiedemann, G., Niekler, A., Keinert, A., Pfetsch, B., Heyer, G., Reber, U., Häussler, T., Schmid-Petri, H. & Adam, S. (2021). Applying LDA topic modeling in communication research: Toward a valid and reliable methodology. In W. van Atteveldt & T.-Q. Peng (Eds.), *Computational Methods for Communication Science* (pp. 13-38). Routledge.

Maté, A., Sebők, M., Wordliczek, Ł., Stolicki, D. & Feldmann, A. (2023). Machine translation as an underrated ingredient? Solving classification tasks with large language models for comparative research. *Computational Communication Research, 5(2).*

Öhman, E., Honkela, T., & Tiedemann, J. (2016). The challenges of multi-dimensional sentiment analysis across languages. *Proceedings of the Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media,* 138-142.

Overbeck, M., Baden, C., Aharoni, T., Amit-Danhi, E. R. & Tenenboim-Weinblatt, K. (2023). Beyond sentiment: An algorithmic strategy for identifying evaluations within large text corpora. *Communication Methods & Measures.*

Palicki, S. K., Walter, S., van Atteveldt, W., Beazer, A. & Bravo, I. (2023). Selecting relevant documents for multilingual content analysis: An evaluation of keyword and semantic similarity search approaches. *Computational Communication Research.*

Payne, T. E. (2017). Morphological typology. In A. Y. Aikhenvald & R. M. W. Dixon (Eds.), *The Cambridge Handbook of Linguistic Typology* (p. 7894). Cambridge University Press.

Prewitt-Freilino, J. L., Caswell, T. A., & Laakso, E. K. (2012). The gendering of language: A comparison of gender equality in countries with gendered, natural gender, and genderless languages. *Sex Roles, 66*(3), 268-281.

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

Roberts, M. E., Stewart, B. M., & Tingley, D. (2019). Stm: An R package for structural topic models. *Journal of Statistical Software, 91*(2), 1–40.

Sepehri, A., Mirshafiee, M. S. & Markowitz, D. M. (2022). PassivePy: A tool to automatically identify passive voice in big text data. *Journal of Consumer Psychology.*

Shababo, G., Baden, C., Dolinsky, A., Lind, F., Pipal, C., Schoonvelde, M., Smoliarova, A., van der Velden, M. A. C. G., & Zalik, A. (2023). How do linguistic differences impact computational text analysis methods? A road map for future validation and integrated strategy development. *OPTED Deliverable D6.3*. https://www.opted.eu/results/project-reports/

Steimel, K., Dakota, D., Chen, Y., & Kübler, S. (2019). Investigating multilingual abusive language detection: A cautionary tale. *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP)*, 1151–1160.

Stokes, S. R., 2020. *Political efficacy and the Russian language: The role of language in politics*. PhD Thesis. Texas A&M University.

Stubbs, M. (1997). Whorf's children: Critical comments on critical discourse analysis (CDA). *British Studies in Applied Linguistics, 12*, 100-116.

Tufis, D. & Ion, R. (2007). Parallel corpora, alignment technologies and further prospects in multilingual resources and technology infrastructure. *Proceedings of the 4th International Conference on Speech and Dialogue Systems*, 183–195.

Van Atteveldt, W., Trilling, D. C., & Arcila Calderon, C. (2022). *Computational analysis of communication*. Wiley.

van Atteveldt, W., van der Velden, M. A. C. G., & Boukes, M. (2021). The validity of sentiment analysis: comparing manual annotation, crowd-coding, dictionary approaches, and machine learning algorithms. *Communication Methods and Measures,* 1–20.

Van der Pas, D. J. & Aaldering, L. (2020). Gender differences in political media coverage. A meta-analysis. *Journal of Communication, 70*(1). 114-143.

Van der Veen, M. (2023). Word-level machine translation for bag-of-words text analysis: Cheap, fast, and surprisingly good. *Computational Communication Research, 5(2).*

Van der Velden, M. A. C. G., Schoonvelde, M., & Baden, C. (2023). Introduction to the Special Issue on Multilingual Computational Text Analysis. *Computational Communication Research.*

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

Zalik, A. & Baden, C. (2023). *The future is (ever) promising: Elected representatives' direct and indirect action promises in routine parliamentary discourse*. Presented at the 73rd ICA Annual Conference, Toronto, Canada.

Zheng, G. & Gaowa, G. (2010). The selection of Mongolian stop words. *IEEE International Conference on Intelligent Computing and Intelligent Systems*.