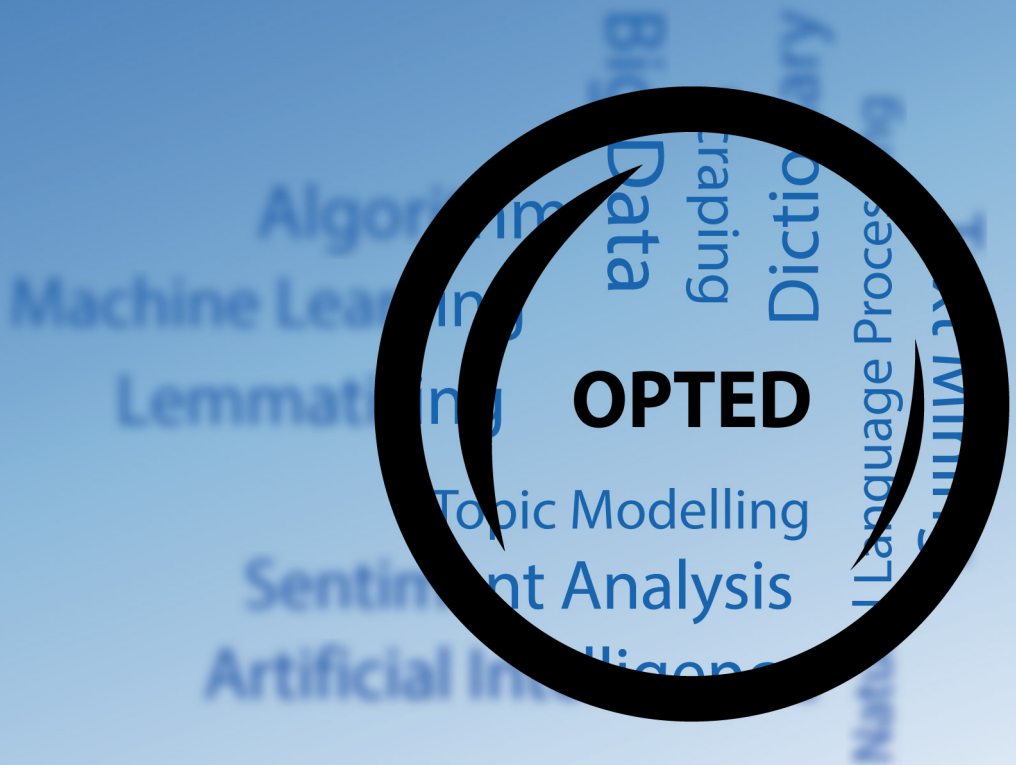
OPTED

Integrated standards and context-sensitive recommendations for the validation of multilingual computational text analysis

Christian Baden, Alona Dolinsky, Fabienne Lind, Christian Pipal, Martijn Schoonvelde, Guy Shababo, & Mariken A.C.G. van der Velden



Disclaimer

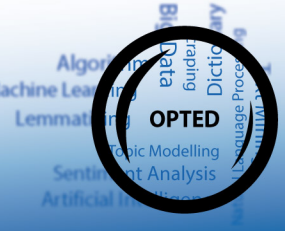
This project has received funding from the European Union's Horizon 2020 research & innovation programme under grant agreement No 951832. The document reflects only the authors' views. The European Union is not liable for any use that may be made of the information contained herein.

Dissemination level

Public

Type

Report



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

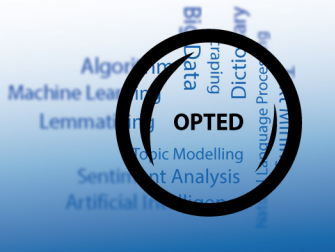
Integrated standards and context-sensitive recommendations for the validation of multilingual computational text analysis

Deliverable 6.2

Authors: Christian Baden[1], Alona Dolinsky[2], Fabienne Lind[3], Christian Pipal[4], Martijn Schoonvelde[5,2], Guy Shababo[1], & Mariken A.C.G. van der Velden[6]

[1] Hebrew University of Jerusalem
[2] University College Dublin
[3] University of Vienna
[4] University of Amsterdam
[5] University of Groningen
[6] Vrije Universiteit Amsterdam

Due date: September 2022



Executive Summary

The overall objective of WP6 is to identify and develop critical research infrastructures for the computational analysis of multilingual text. This work package takes stock of the state of the art at the cutting edge of the methodological research of computational social science and formulates a rigorous agenda for the validation and methodological evaluation of existing approaches.
As the second step to achieve this objective, D6.2 introduces a framework for the validation of computational text analysis methodology used to analyze multilingual textual data. Building on insights from a content analysis of published literature in the social sciences (which has been integrated with the multilingual hub that formed the core of Deliverable D6.1), and an expert survey with the respective authors, we first derive practical recommendations for the steps necessary to obtain valid measures of a construct that are comparable across languages and contexts. We then construct a framework that is sufficiently general to apply to very different forms of computational textual analysis and can thus offer guidance to researchers who want to ensure that they draw meaningful measurements from multilingual text data. This general formulation makes the framework applicable to various text types relevant to social scientists such as the text types curated in WP2 (citizen-produced texts), WP3 (journalistic texts), WP4 (texts by political organizations), and WP5 (parliamentary and government/ legal texts).

1 Introduction

Computational Text Analysis (CTA) has become an indispensable part of social science research. Fueled by the pioneering work of Lucas et al. (2015), CTA is also increasingly used for the analysis of text collections in different languages. Research teams have used multilingual text analysis to study various topics relevant to our understanding of democratic societies, for example, the tone of presidential candidates (Maurer & Diehl, 2020), civility on social media (Theocharis et al., 2016), the ideological positions of social media users (Temporão et al., 2018), the conflict orientation of news coverage (Baden & Tenenboim-Weinblatt, 2018) or cultural differences (Gutiérrez et al. 2016). Besides applied research, new methodological approaches are being proposed as well. Recent contributions include multilingual dictionaries (Lind et al., 2019; Proksch et al., 2019), multilingual supervised machine learning (Courtney et al., 2020; Lind et al. 2021), multilingual topic modeling (Chan et al., 2020; Lind et al., 2022; Maier et al., 2022), multilingual scaling (Watanabe, 2021), and multilingual word embeddings (De Vries, 2021; Licht, 2022). But even though these contributions all offer validations for their approaches, there is still little agreement on the appropriate strategies and sufficient criteria for validating multilingual text analysis tasks and methods. Also, in the methodological literature and reviews on the state of computational text analysis (e.g., van Atteveldt & Peng, 2018; Barberá et al., 2021; Gentzkow et al., 2019; Grimmer et al., 2022; Song et al., 2020), most discussions of validation strategies are limited to validating computational text analysis methods within the context of single-language applications, and do not specifically cover the validation of text analysis methods in a multilingual setting.

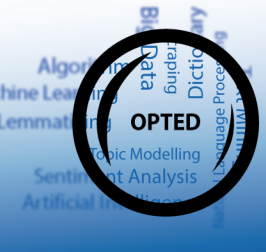
We argue that to stimulate the still reluctant use of multilingual automated text analysis methods in the social sciences (Baden et al., 2022), having a framework for validating multilingual uses of computational text analysis methods is instrumental. Adequate, accepted validation standards not only serve as a cornerstone of quality control to ascertain the meaningfulness and usefulness of research findings (e.g., Jacobs & Wallach, 2021), but also carry major implications in terms of communicating the viability and trustworthiness of computational text analysis methods for multilingual research. Validation not only serves as a justification for the obtained research findings, but also as a basis for theory development and interventions. Moreover, it builds experience and confidence among users of well-validated methods and can generate funding (Krippendorff, 2004, p. 313). Validation might even be more urgent if we consider the increasing popularity of analytical approaches that rely on third-party pretrained materials such as large language models. Research findings that have not been validated properly may have detrimental unintended consequences, ranging from erroneous and artifactual findings, to stereotypical information baked into available language models, to disastrous real-world implications such as algorithmic bias in hiring decisions (see e.g., Zhao et al. 2019; Bender & Friedman 2018; Bender et al. 2021). Validating text-based measurements is also important because a method that has worked well in the past does not necessarily have to perform equally well with similar but new text data or new tasks (Grimmer et al., 2022).

In computational text analysis in particular, validation is key given the numerous and often hidden decisions that a researcher needs to make towards deciding on the best possible representation and modeling of textual meanings to address the research question at hand – often based on only limited hands-on contact with the textual data. In fact, Grimmer et al. (2022, p.38) write that the 'best assurance that a text representation is working is extensive validation'. In a multilingual setting, alas, it is not 'just' the textual representation (i.e., the information in the text that is deemed relevant for the research question at hand) that needs to be validated, but also the performance and comparability of textual representations across languages and across cases. Whereas a researcher working in one language 'only' needs to ascertain that the textual data at hand meaningfully captures their concept of interest, a researcher working in more than one language needs to provide evidence that this link between the conceptual and empirical realm is comparable across cases and contexts. In other words, validation of multilingual textual analysis approaches extends well beyond what is required in a monolingual setting.

This contribution proposes a practical framework for the validation of multilingual computational text analysis. We start from the assumption that commensurability across languages and cases requires validation in several stages of the research process. In the first step, a researcher will need to ensure that the *documents* in a corpus can be compared across languages and cases (*data validation*). Data validation requires providing a (theory-informed) justification for the employed sampling strategy. Second, the researcher must screen and ideally establish that the *features* in an analysis can be considered equivalent across languages and cases (*input validation*). Input validation is demonstrated when the researcher offers a rationale for the preprocessing steps they apply to the corpus of documents towards obtaining a particular textual representation. The researcher will need to offer evidence that input alignment is achieved, depending on the comparability of the cases at hand, and the research goals (*etic or emic*) that the research question implies. For example, when conducting a multilingual sentiment analysis, the researcher may decide to focus on nouns or adjectives only (see e.g., van Atteveldt et al., 2022) when indeed these word functions are present and used in similar ways in all the languages analyzed. In the third step, the researcher will need to ensure that the *models or instruments* that are applied to the documents are capable of extracting equivalent information across languages (*process validation*). For example, when a researcher applies a multilingual dictionary to a set of documents, they will need to demonstrate that the ensuing categorization is as close as possible to being 'language-independent' (Bender, 2011) and thus comparable. This requires evidence that obtained categorisations rely on textual content with comparable meaning across languages and cases. Fourth, the researcher will need to assure that the *results* from a multilingual analysis can be meaningfully compared with each other in relation to the languages and cases from which they are generated (*output validation*). Either by comparing the outputs to manual codings or to divergent or convergent measures, the goal is to demonstrate *output alignmen*t across languages and cases (Lind et al., 2021).

Each of these four validation steps responds in distinctive ways to the characteristic challenges raised by multilingual text analysis. In what follows, we first define key concepts and discuss core challenges of validation in a multilingual setting. We then present the results of a content analysis of published literature in the social sciences we conducted to identify and benchmark existing validation strategies in the field. We also discuss evidence from an expert survey of published researchers in the social sciences that details their awareness and concerns about validation of computational text analysis methods in English, other languages, and cross-lingually. Building on the insights we draw from both data sources, we systematize the available validation strategies by differentiating between data, input, throughput, and output validation. We then provide context-sensitive recommendations, formulate integrated standards and generalized procedures suitable to engage in a systematic comparative validation of multilingual text analysis. We end with a discussion of capacities and limitations of existing validation techniques and outline paths to further advance validation practices.

This contribution offers a framework for validation of multilingual text analysis that distinguishes between validation at different stages of the research process. While the specific challenges and needs for validation remain of course sensitive to the specific setting and purpose of each research application, we believe that the framework presented here is sufficiently general to apply to very different forms of computational textual analysis and can thus offer guidance to researchers who want to ensure that they draw meaningful measurements from multilingual text data.



1.1 Concept definitions in multilingual and comparative context

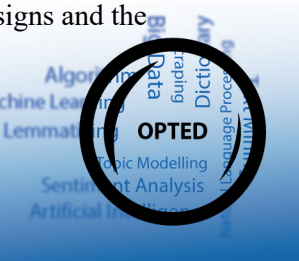
Before we introduce our framework, we first discuss the scope conditions of the types of research that our validation framework applies to. Within this scope we then define what goals validation seeks to achieve.
Our validation framework is designed for specific types of research. First, it is informative for research that seeks to analyze multilingual corpora (possibly from a case comparative perspective, or for joint analyses of data that includes more than one language) with CTA. Second, the framework is aimed at research interested in the measurement of universal or etic type constructs that builds on the assumption that information drawn from different cases or contexts (e.g., countries, regions) can correspond on a conceptual level (Goertz, 2006; Adcock & Collier, 2001). Following this logic, a construct is defined on an abstract level that intends to include all investigated cases. Third, the framework is general enough to apply to dictionary approaches, supervised, and unsupervised methods.

How exactly do we understand "validity" and "validation" within this scope of research? By *validity* we refer to measurement validity in an inclusive sense, i.e., the property that measurements obtained by means of a given (in our case computational text analysis) method indeed reflect meanings that these methods were intended to record. We distinguish such measurement validity from a broader notion of inferential validity, which requires that the theoretical inferences that a researcher may draw from such descriptive observations correspond to true causes, and thus extends beyond the realm of methodological validation. If concepts are compared across cases based on documents in different languages, a valid measurement implies that the correspondence between measurements and intended constructs holds for all of the languages and cases studied. *Validation*, in our use, refers then to any means employed by a researcher to ascertain the here defined validity of recorded measurements (Adcock & Collier, 2001). These means can relate to techniques used to reflect upon methodological choices, to empirically test the performance of used procedures, or to examine the correspondence between obtained measurements and conceptual demands.

This understanding of validity and validation in comparative and multilingual projects relies on a definition of "equivalence". More specifically, to make valid comparisons between cases, *equivalence* must be established and ideally demonstrated on the level of samples, inputs, procedures, and measurements (He & Van de Vijver, 2012). By *equivalence* we thus refer to a property of the collected samples, textual representations, selected procedures, and the recorded measurements. Equivalence does not suppose that these components of a CTA design are *identical* (which is impossible given that in a multilingual comparison meaning is expressed in different languages) but rather that they are selected to strengthen the comparability of the obtained measurements in support of the stated research objective. For example, measurements that are obtained from different data sets, cases and languages are *equivalent* if they help identify meaning from the underlying textual data that is comparable with regard to the research question at hand. This may mean that different measurements obtained from multiple cases may count as equivalent depending on the intended purpose. For instance, for an analysis interested in the use of incivility, references to semantically different slurs common in different cultural settings may be validly recorded as equivalent, while the same references would be non-equivalent – and thus, non-comparable – in a study focused on thematic patterns in the data. In this sense, "comparability" does not require that the data used to represent different cases or contexts is narrowly equivalent in all respects – obviously, one can very well compare apples to oranges; rather, comparability requires that any differences in obtained measurements are due to recognized differences between cases that are meaningful toward the analysis (e.g., that a French right-wing politician makes more use of a certain populist style than a Greek left-wing politician), while artifactual differences raised by unrecognized differences in the sampled data, constructed features, applied models or obtained measures can be ruled out with confidence.

1.2 Challenges of validation for multilingual text analysis methods

Given the paucity of systematic research into the implications of linguistic diversity for CTA, it is useful to first conceptualize the main ways in which multilingual applications may raise specific needs for targeted validation. The core challenge for social scientists when analyzing multiple languages comparatively is to infer comparable meanings from text despite important linguistic variation in how such meanings are encoded, along dimensions such as script (e.g, Latin, Cyrillic, Hebrew), morphology (the system through which words get formed) and syntax (how words are brought together to make sentences). Characteristically, what qualifies as "comparable meaning" additionally varies depending on the application and purpose of conducted research, focusing more on semantic (the literal meaning of language, centered upon the relation between signs and the



object they indicate) and pragmatic comparability (the social meaning of language, centered upon how present expressions are interpreted in context). These two linguistic dimensions are addressed in the text-as-data literature as two challenges related to a) the establishment of cross-lingual measurement equivalence and b) to the context-dependency of concepts (Licht & Lind, working paper).

The first of these challenges, which has also been referred to as the "Tower of Babel" problem (Chan et al., 2020, p.285), focuses on the semantic level and relates to the problem that comparable meanings are expressed using different vocabulary and phrases in different languages. As a result, those expressions recorded in different languages do not necessarily map neatly upon one another, with the result that numerical representations of textual contents are often not directly comparable. Some active bridging efforts are required to ensure that equivalent representations of texts in different languages indeed express comparable meanings. While humans achieve much of this bridging intuitively (e.g., when they translate sentences), machines often struggle with this task, which does not work out equally well for all concepts. Not only are comparable meanings often encoded in different ways, but also associated connotations and conceptual delimitations may vary (e.g., where one language distinguishes constructs that share one word in another), and surprisingly many words have no trivial translation in other languages (Sigismondi, 2018). In addition, it is possible that some information that is consequential for an intended analysis – such as information on the grammatical case, gender, or tense of expressed contents – may be linguistically available in some languages, but not in others. As a result, equivalent measurement strategies may need to operate with very different depths of information, potentially distorting results.

The second major challenge in multilingual text analysis focuses on pragmatic comparability. From a social science perspective, language can rarely be studied without considering the context in which language is produced. In comparative research designs that span multiple cases or contexts (e.g., countries, cultural settings, times), accordingly, there is often a need to consider the divergent meanings and connotations of semantically similar expressions, owing to the influence of historical, political, and social traditions on language use (Gurevitch & Blumler, 2003). For multilingual analysis to consider the context-dependency of concepts, measurement strategies must be adapted in ways that map concepts not merely at the level of semantic, denotative correspondence, but at the level of language uses and pragmatic implications.

To ensure that a computational text analysis method delivers valid, comparable results, both challenges raise distinct requirements that need to be considered in a validation framework. As a point of departure for formulating such a framework, we have designed two studies that jointly aim to capture current validation practices and validation-related concerns perceived within the research community.

2 Content analysis of content analysis literature

To understand current validation practices in multilingual research in the social sciences, we conducted a content analysis of all quantitative text-based research published in the top 20 highest ranked journals in the Web of Science categories of communication, political science, sociology and psychology, selected according to their 2019 SSCI 1-year impact factors. The analysis first inventorised all articles published in the selected journals between January 2016 and September 2020. Using a keyword search on the Web of Science, we then identified a total of 7,296 potentially relevant articles whose abstracts referred to some kind of textual contents or text analytic procedures (the search string we used to identify these papers can be found here). We then accessed the full text of these articles to determine whether the presented research included any form of quantitative textual analysis. This screening yielded a total of $N = 854$ articles. For further information on the sampling procedure, see the appendix of Baden *et al.* (2022) here.

Among the articles that relied on quantitative text analysis, we distinguished between manual and computational approaches. We defined manual approaches as all approaches where all classification decisions are made by a human following instructions. This includes both classic content analysis, various quantifying procedures embedded in otherwise qualitative methods, as well as crowd coding. Computational approaches referred to papers that had classification decisions taken by computational algorithms, including clustering techniques. Some papers included both manual and computational approaches, and we categorized those as mixed papers. We also distinguished between papers that relied on corpora that were only in English, in some other language, or that included multiple languages. The breakdown of papers across these categories can be found in Table 1.

A few things stand out. First, a clear majority of coded articles relies on English-language corpora ($N = 379$), with single-language studies in languages other than English a distant second ($N = 213$). Second, of the

multilingual papers – which in many cases report on comparisons between materials in English and one other language – a large majority ($N = 75$) relies on manual approaches for categorizing text. Researchers who make comparisons between languages rely less on computational methods and more on manual coding (for more information, please refer to the section below that reports details from an expert survey of the authors of these papers).

TABLE 1: PAPERS THAT RELY ON QUANTITATIVE TEXT ANALYSIS, ARRANGED BY ANALYTICAL APPROACH AND LANGUAGE.

	English	Another language	Multilingual	Row totals
Computational	165 (44%)	61 (29%)	28 (25%)	254
Manual	179 (47%)	135 (63%)	75 (68%)	389
Mixed	35 (9%)	17 (8%)	8 (7%)	60
Column totals	379 (100%)	213 (100%)	111 (100%)	703

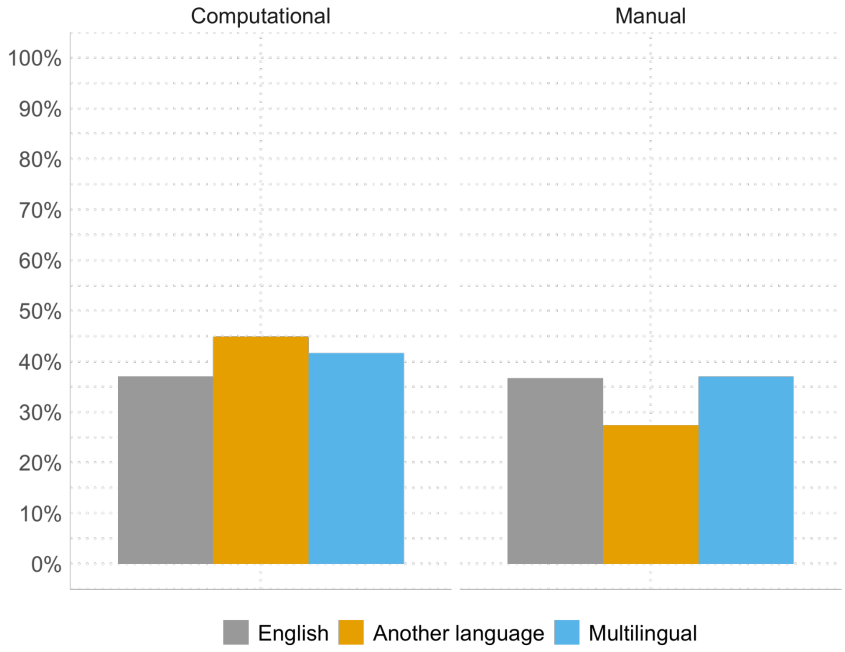
Note: The percentage in between brackets denote column percentages.

Further examining the validation efforts conducted in studies using computational methods (either alone or, rarely, together with manual approaches), we distinguished between 1) data validation, 2) input validation, 3) process / throughput validation and 4) output validation (the codebook we developed for this purpose can be found here). Data validation is demonstrated when an article reports the rationale for the sampling strategy it utilizes, arguing or demonstrating that the samples used enable a valid interpretation or comparison of findings about the pursued research question. Input validity refers to efforts to justify the preprocessing steps applied to the texts in order to ensure that subsequently employed measurement methods will recognize intended patterns in the data (e.g., its rationale for text unitization or its rationale for textual feature selection). For example, if an article offers a substantive rationale for including multiword expressions as opposed to individual tokens when developing a bag of words representation of a corpus this counts as a form of input validation. Process validation is demonstrated when an article explicitly ties the classification rules and algorithms it applies to a text to the dimensions of the concept it seeks to measure. If an article offers evidence that the classification criteria it imposes on the textual data are direct operationalizations of relevant conceptual categories, this is considered a form of throughput validation. Process validation, finally, concerns the 'adequacy of content' (Adcock & Collier, 2001) and as such is similar to content validation. Output validation aims to establish that those measurements yielded by a procedure validly reflect those meanings that the procedure aimed to measure, either by comparing model output to a (human-coded, and therefore presumed-valid) gold standard, or by assessing model output against other measures that are conceptually known to converge with, or diverge from, the measured variable. For example, a researcher who seeks to measure euroscepticism in parliamentary speeches may choose to compare the obtained euroscepticism scores against party expert surveys such as the Chapel Hill Expert Survey (Jolly et al. 2022), or manifesto-based databases such as the Manifesto Project (Volkens et al. 2021).

2.1 Data validation

Figure 1 shows the fraction of articles focusing on either kind of language setting and using either computational or manual methods that mentions data validity (justification for the corpus selection). We find that approximately 45% of articles that rely on computational methods and corpora in multiple languages report on data validation efforts, while about 35% of articles that rely on manual methods and corpora in multiple languages do the same. This indicates that while overall, less than half of the articles we surveyed reported on this form of validation, the numbers for computational and multilingual articles are not drastically different from other method-language pairs.

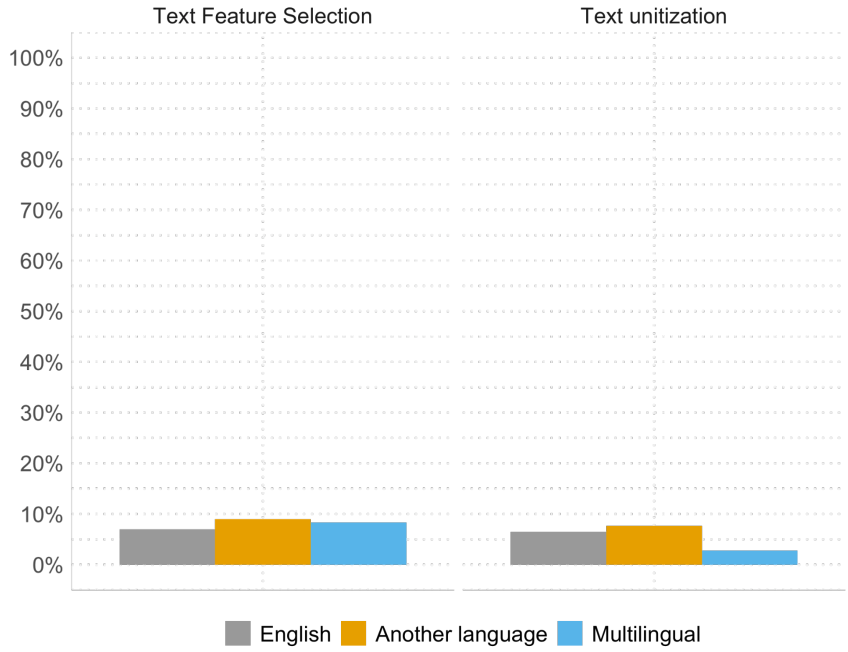
Figure 1 FRACTIONS OF ARTICLES THAT ENGAGE WITH DATA VALIDATION, BY METHOD AND LANGUAGE



2.2 Input validation

Figure 2 shows the fraction of articles focusing on either kind of language setting that mentions one of two types of input validation - *unitization* or *feature selection*. Unlike the above display, the following analyses focus solely on those papers using computational text analysis, as most preprocessing steps are not relevant for manual analyses. We see that compared to the fraction of papers addressing data validity, far fewer focus on input validity. Regardless of the languages analyzed, less than 10% of computational papers discuss a rationale for the preprocessing steps applied to their corpora, or the way in which features are selected for subsequent analysis. Practices are largely invariant across studied languages. Even where preprocessing is discussed, studies tend to cling to default steps set or suggested by subsequent modeling algorithms (often, Bag of Words representations), without considering whether other representations (e.g., bigrams retaining some word order information; stemmed words harmonizing or non-stemmed words distinguishing between related uses) might better capture constructs of interest. In political text corpora, bi- and trigrams such as "United Nations" or "International Monetary Fund" often play an important role for substantive analysis, however, papers rarely explain how they deal with such needs (if at all).
Similarly, unitization, or the level at which textual content is analyzed, is only rarely discussed in these papers. In many cases, articles stick to the default of taking given documents (e.g., newspaper articles, social media posts, entire parliamentary speeches) as the unit of observation, without regard to whether relevant meanings may be located in particular parts of these or might vary across such documents, possibly requiring unitization at a paragraph or sentence level, or permitting the exclusion of parts of the textual material.

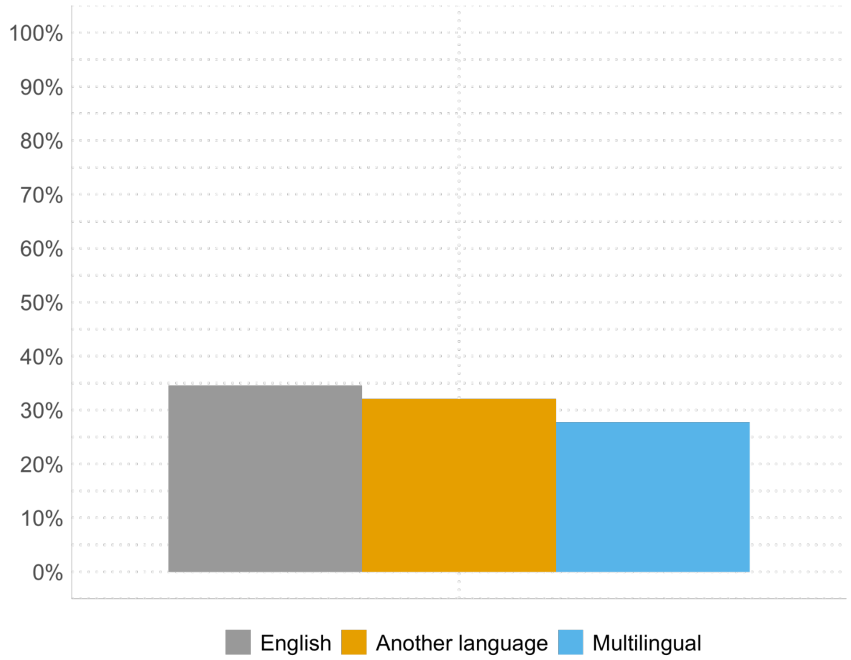
Figure 2 FRACTIONS OF ARTICLES USING COMPUTATIONAL METHODS THAT ENGAGE WITH INPUT VALIDATION, BY TYPE AND LANGUAGE



2.3 Process validation

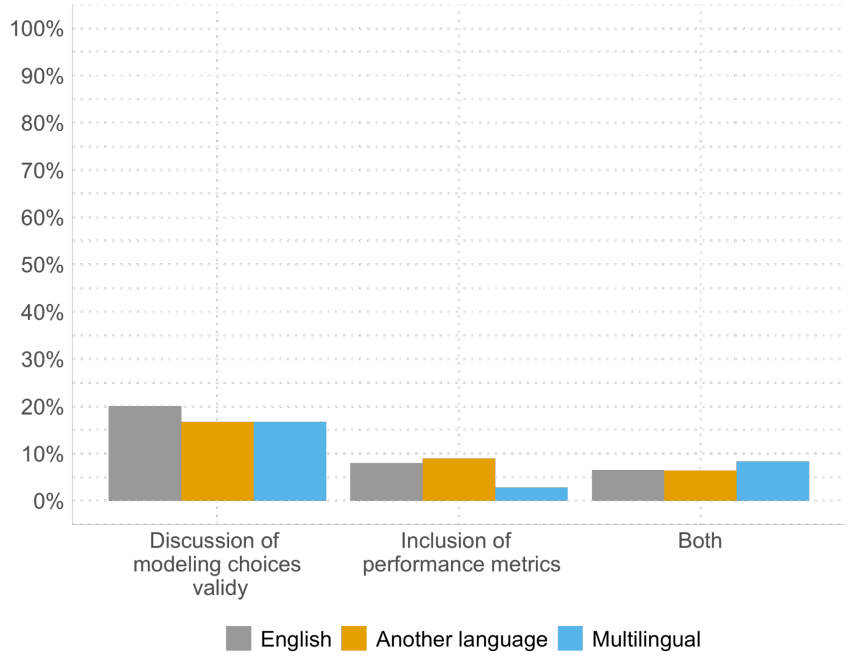
In the next step, we examine to what extent quantitative text articles discuss process validation. Process validation is demonstrated when an article explicitly ties the classification rules it applies to a text to the dimensions of the concept it seeks to measure. We coded these articles such that process validation can be reflected in either a discussion of modeling choices, a discussion of performance metrics or a combination of both. Modeling choices include both the express justification of selected algorithms (e.g., why a topic model with time as covariate is suitable to detect certain meanings, or why a dictionary can be expected to validly identify uses of a construct), but also any discussions of chosen parameters and hyperparameters within a selected model (e.g., expected or 'optimal' number of topics, word distance parameters, parametrization of machine learning tools), as long as these were discussed within the context of valid measurement. Performance metrics comprise numerical tests for the applicability of a particular measurement procedure (e.g., the coverage of the dictionary terms in annotated documents; analyses of the applicability of high-loading terms).
The results of this analysis are reported in Figure 3. For computational papers (which include papers that use mixed methods), approximately one in three contains a discussion of process validation regardless of the language of the corpus they analyze, although papers that use multilingual corpora show the lowest levels of concern for process validity (27%). Given the crucial question that process validation addresses – does the method we apply to our corpus give us a conceptually meaningful categorisation of documents that is comparable across cases? – these numbers are far too low.

Figure 3 FRACTIONS OF ARTICLES USING COMPUTATIONAL METHODS THAT ENGAGE WITH PROCESS VALIDATION, BY LANGUAGE



In the next step, we break down types of process validation for computational papers with Figure 4 displaying the results. Between 10 and 15% of papers discuss modeling choices one way or another, depending on the language. Regardless of the language under study, less than 10% of papers discuss performance metrics or include both types of process validation.

Figure 4 FRACTIONS OF ARTICLES USING COMPUTATIONAL METHODS THAT ENGAGE WITH PROCESS VALIDATION, BY TYPE AND LANGUAGE

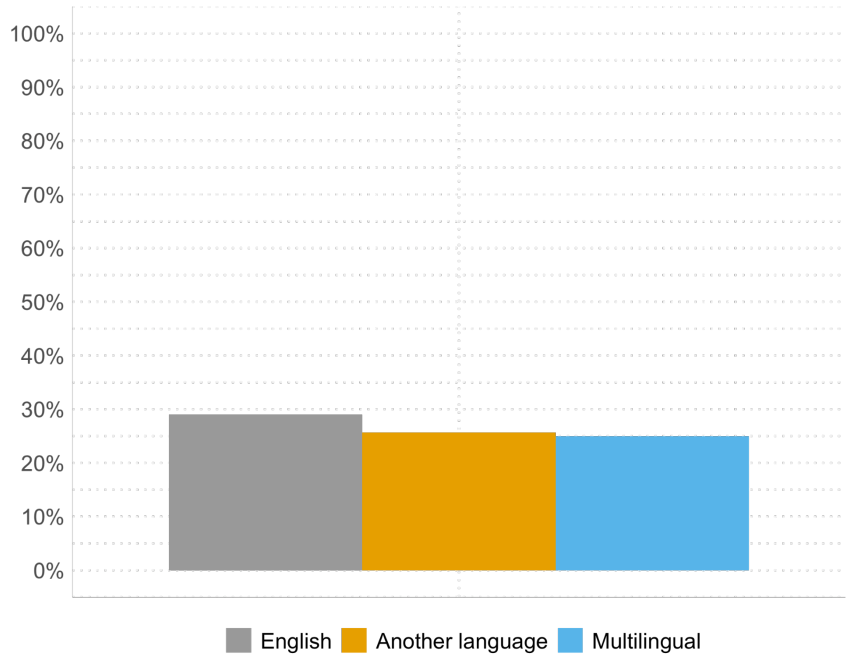


2.4 Output validation

In the final step of this analysis, we examined computational studies' engagement with output validation in these papers. In particular, we analyze what fraction of papers validates the outputs of the textual analysis against a human-coded gold standard (Figure 5), by assessing face validity or via the use of additional, divergent or convergent measures (Figure 6).

Figure 5 shows that less than one in three computational papers validates its measurements against a human coded benchmark, regardless of the language these papers are in.

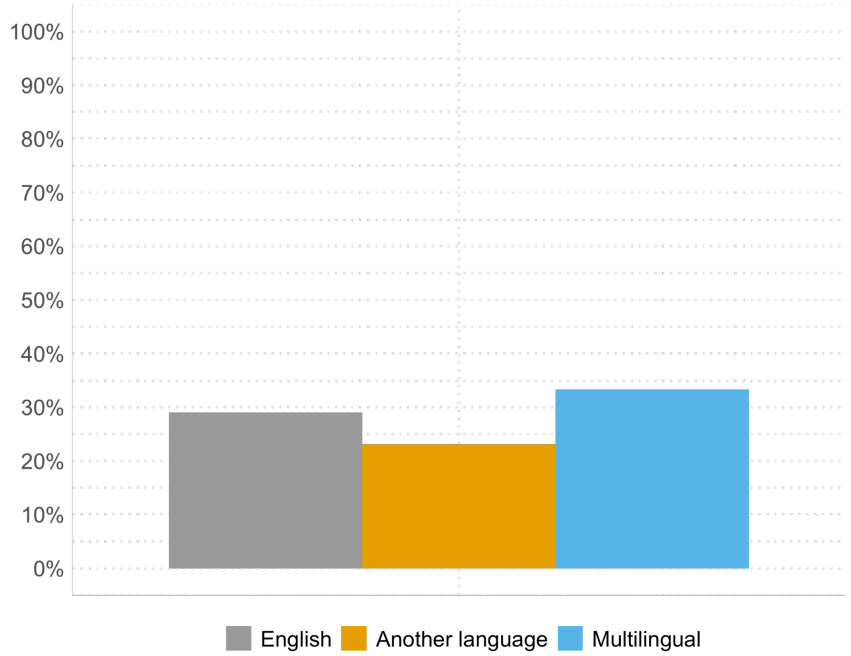
Figure 5 FRACTIONS OF ARTICLES USING COMPUTATIONAL METHODS THAT ENGAGE WITH OUTPUT VALIDATION USING A HUMAN-CODED GOLD STANDARD, BY LANGUAGE



Finally, in Figure 6 we display the fraction of papers that report either face validity, convergent validity or divergent validity. Again, approximately one in three articles mention either of these forms of output validation. For computational papers: 39 report face validity, and, out of the 46 papers that report convergent validity, 5 also report divergent validity. There are no papers that only report divergent validity.

The results of this analysis offer a few insights. Generally speaking, any form of output validation occurs in approximately one in three papers, which means that about two out of three papers do not benchmark their textual measures in any way. Considering the fact our analysis focused on studies published in top-ranking social science journals, it seems rather remarkable how few papers document an explicit effort to ensure the validity of measurement, be that by expressly justifying or testing operational and methodological choices, or by keeping 'humans in the loop' to validate the output created by a computational system. Second, we don't find dramatic language-specific differences. By and large, engagement with output validation occurs to a similar degree, regardless of whether the studied corpora are in English, in some other language, or multilingual.

Figure 6 FRACTIONS OF ARTICLES USING COMPUTATIONAL METHODS THAT ENGAGE WITH OUTPUT VALIDATION USING CONVERGENT OR DIVERGENT MEASURES, BY LANGUAGE



3 Survey with text analysis researchers

To gauge the perspective of users of quantitative textual analysis, next, we conducted a survey among all authors identified in the preceding content analysis. For each article, we identified the first three authors and looked up their contact emails. In this way, we identified a total of 1,668 unique authors, for 1,653 of whom we could identify functioning email-addresses. All of these authors were invited to fill in the questionnaire on March 4th of 2021, and received two reminders, each approximately a week after our last message (respectively on March 11th and March 16th of 2021). This yielded 433 responses, a response rate of approximately 25 per cent. This sample breaks down across gender, seniority, and disciplinary background as described in Table 2.

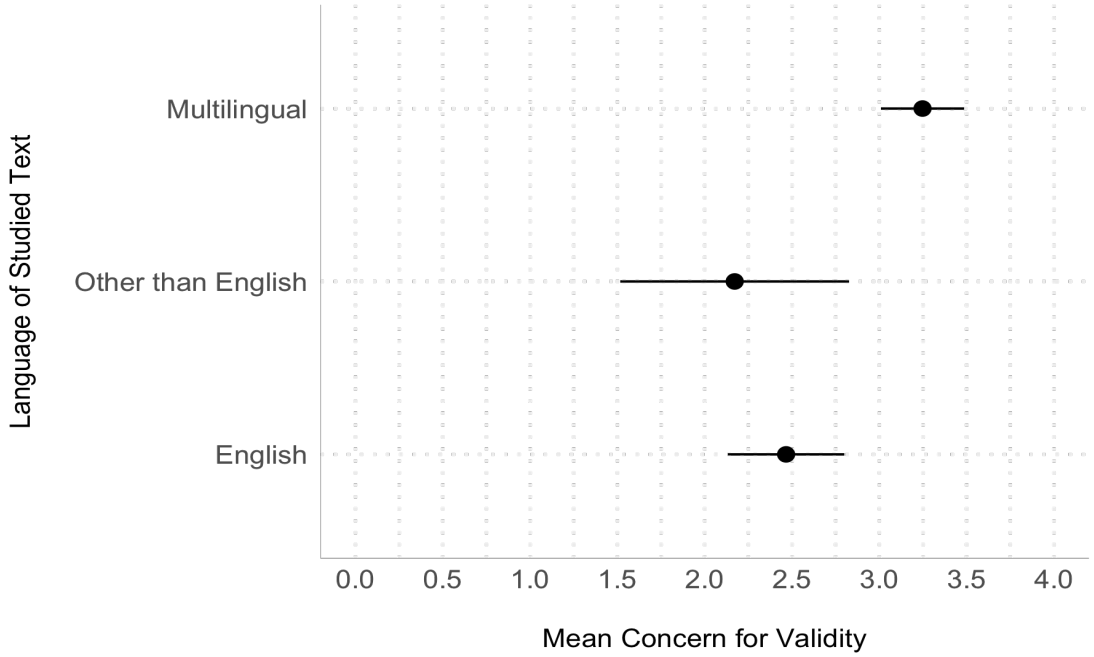
TABLE 2: SAMPLE CHARACTERISTICS

Variable	*Values*	*N*
Gender	Female	142
	Male	243
	Other	2
	No response	46
Seniority	PhD student	38
	Early-career researcher	110
	Mid-career researcher	158
	Senior researcher	80
	None of the above	4
	No response	43
Discipline	Communications	248
	Political Science	188
	Psychology	34
	Sociology	55
	Other	42

Note: For the Discipline variable, respondents were asked to indicate in what field they were active as a re-searcher by ticking all fields that apply. Since some respondents indicated more than one field, the number of responses in the table adds up to more than 433.

We first asked researchers about their concerns about the validity and availability of suitable computational text analysis methods. Concern for validity is measured using four items that capture concerns about the validity of CTAM methods more broadly and in a multilingual context more specifically. The results of the analysis are reported in Figure 7, where higher scores denote more validity concerns. Validity concerns are significantly more pronounced among researchers who work in more than one language, as opposed to those that do not. Working with corpora in more than one language makes researchers more conscious of the risks to validity when working with CTAM.

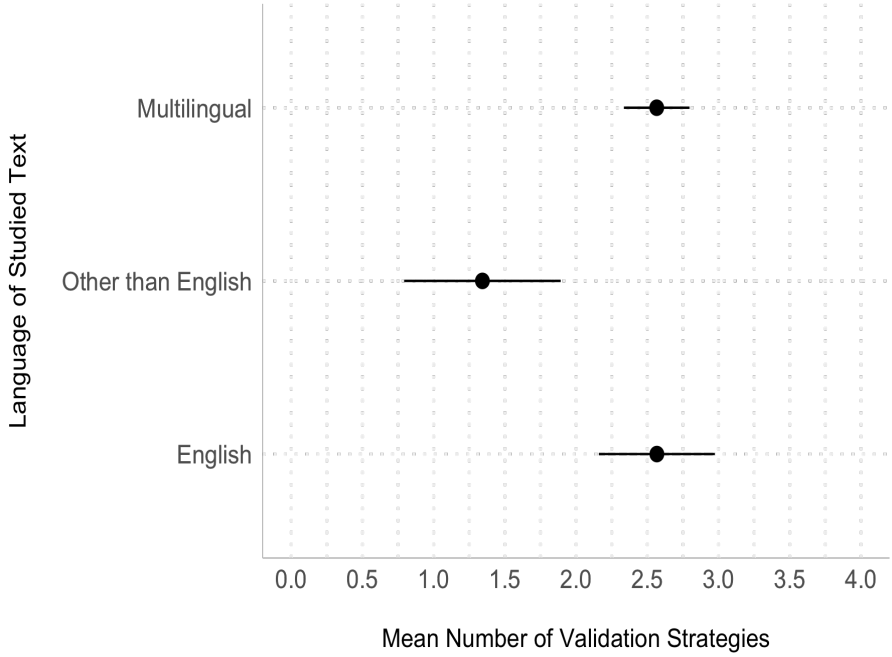
Figure 7 MEAN CONCERN FOR VALIDITY



Note: Mean concern for validity comparing researchers who study English-only materials, researchers who study materials in other languages, and researchers who study materials in more than one language, either in a comparative fashion or one at a time. Lines denote 95% confidence intervals. Lines denote 95% confidence intervals.

Crucially, however, these heightened concerns about validity suitable tools do not appear to result in researchers working with multilingual corpora using more extensive strategies to validate their computational approaches, as can be seen in Figure 8. Furthermore, researchers who work with texts in single languages other than English, who were similarly concerned about validation as researchers working with just English, use significantly fewer validation strategies in their computational research. These results are illustrative of a situation where researchers appear to struggle to address their concerns about the validity of computational analyses. While researchers working in English apparently still find several means for ascertaining validity, such strategies appear to be unavailable for other languages; and also among researchers working on multilingual texts, who are significantly more sensitive to the challenges of achieving validity, these researchers appear to be unable to identify additional strategies beyond what is commonly used and available for English language material.

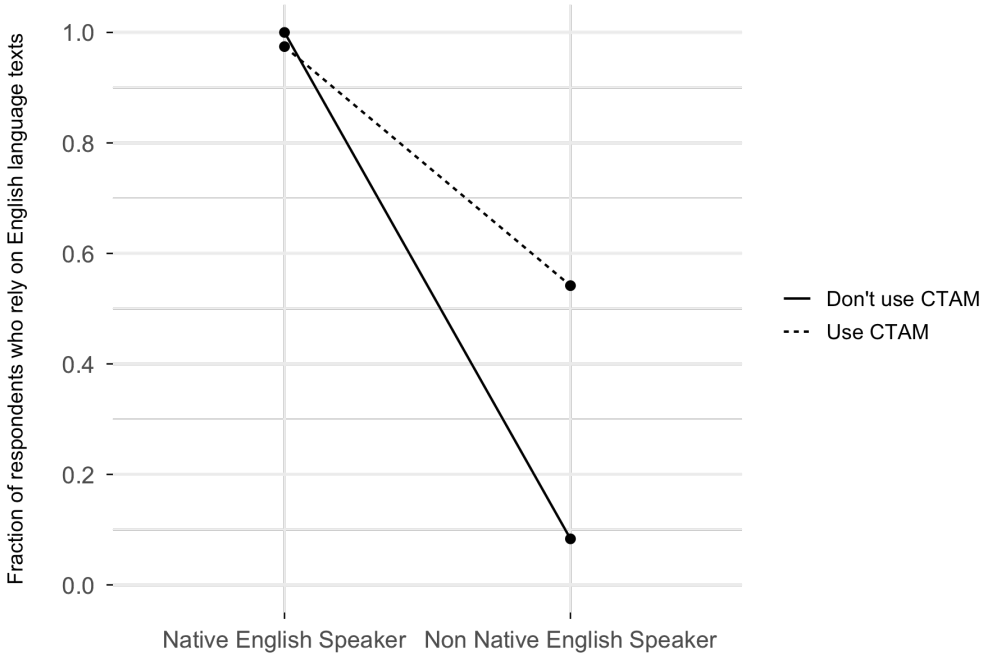
Figure 8 **MEAN NUMBER OF VALIDATION STRATEGIES**



Note: Mean number of strategies employed to validate use of computational methods comparing researchers who study English-only materials, researchers who study materials in other languages, and researchers conducting inter-lingual comparative research. Lines denote 95% confidence intervals.

In a final step, we combined researchers' native language and their use of CTAM to examine their shared relationship with the languages of the texts these researchers study. Owing to the dominance of English-language tools in CTAM, we expected that among native speakers of languages other than English, those who use CTAM are more likely to study English language materials than those who do not use CTAM. While researchers who are native English speakers almost always rely on English-language corpora, regardless of whether they work with CTAM or not, figure 10 shows that non-native speakers of English still default to English in more than half of all cases when they work with computational tools. This is compared to just over 10% of respondents who rely on manual forms of quantitative text analysis, who are thus far more likely to work in their own, native languages. Evidently, non-English native speaking researchers are faced with some dilemma when applying CTAM to other-language corpora, requiring them to choose between either opting against computational methods, or foregoing their native language setting in favor of English language materials.

Figure 9 FRACTION OF RESPONDENTS WHO RELY ON ENGLISH LANGUAGE TEXT, BY NATIVE LANGUAGE AND USE OF CTAM.



4 A framework for validating multilingual textual analysis

Building on the insights we draw from this content analysis and expert survey we propose a validation framework that is sufficiently general to apply to very different forms of computational textual analysis and can thus offer guidance to researchers who want to ensure that they draw meaningful measurements from multilingual text data. This framework is summarized in Figure 10.

Figure 10 A FRAMEWORK FOR VALIDATION OF MEASUREMENTS OBTAINED FROM MULTILINGUAL TEXTUAL DATA.

Steps to validate measures obtained from multilingual textual data				
	1. Data validation	2. Input validation	3. Process validation	4. Output validation
Key question	*Are the sampled documents comparable across cases and contexts?*	*Are the chosen textual representations equivalent across cases and contexts?*	*Are the chosen computational algorithms equally effective in all cases?*	*Are the obtained measurements of similar quality across cases and contexts?*
Objective	Mitigate potential sources of sampling bias in different cases and contexts.	Ensure relevance for the research question at hand and cross-lingual comparability of selected textual representation.	Ensure that the chosen algorithm categorizes texts in each case in a similar way.	Ensure that obtained measures are comparable across across cases and contexts.
Strategy	Confirm that textual data for all cases is similar quality and scope. Check corpus statistics for each case	Confirm that language-specific features correspond with dimensions of the construct of interest	Confirm that the obtained textual measures capture similar meaning in each language.	Examine precision and recall scores between model- and human-based categorizations in each language. Conduct language-specific error analysis

4.1 Data validation

When making comparisons across languages, a researcher needs to ascertain that the documents on which the analysis is based are indeed useful for answering their research question. This first requires a researcher to select equivalent document sources. When selecting a sample of documents (say, newspaper articles) we need to ensure that our source sampling procedure is as equivalent as possible across contexts, and be aware of any remaining discrepancies (e.g., when specific corpora, for which no suitable equivalent exists in other contexts, are included for substantive reasons) to address these later in the analysis. For a multilingual media analysis, for example, a common technique for selecting an equivalent sample of documents per case is to systematically choose the most widespread or most influential outlets (Rössler, 2012) and outlets of comparable types, such as two quality outlets and one boulevard outlet per case (Boomgaarden et al. 2013). When equivalent sources have been selected, if not the full population of documents is taken for each of them, document samples may have to be retrieved using equivalent time periods and potentially equivalent search terms.

Grimmer et al. (2022) also offer valuable insights on sampling validation in a comparative setting in their discussion of resource bias and retrieval bias in text samples. Resource bias occurs when we are comparing cases that have different resources at their disposal that may impact document availability. For example, richer countries with more effective central governments are also more likely to store government-produced documents in an accessible manner than governments with less resources at their disposal. Similarly, when comparing reporting on immigration in different media outlets, any observed differences in salience of immigration articles *may* be the result of different journalistic priorities - but they may also arise from other differences, such as the fact that some outlet includes columns and blogs in its archive while another doesn't; or simply different resources that enables one, but not another outlet to provide ongoing coverage also beyond acute crises. Retrieval bias occurs when the sampling strategy for identifying relevant documents may work differently in one case than in another case, in ways that may result in artifactual differences in subsequent measurement. For example, a search string for identifying immigration-related articles may have high precision and recall in one language, but the same, translated string may perform considerably worse in another language.[1] In other words, it cannot be assumed that the same sampling strategy, applied across contexts and languages, will reliably generate equivalent samples. While it is inevitable - and in fact often productive - that different cases or samples include some systematic differences, especially with regard to the pragmatics of discourse shaped by its respective context of use, it is essential that any differences that might affect findings are considered and acknowledged. For instance, a researcher may indeed be interested in how coverage of an issue is different if it is provided on a continuous basis or solely around major events; similarly, different etiquettes for public speech in compared contexts may either confound valid differences or constitute a substantive focus of interest, depending on the purpose of research. Data validation, accordingly, demands that any differences that are of substantial interest are explicated, while any sources of artifactual variation are controlled or kept as equivalent as is possible. In this way, the comparability of data across cases needs to be actively considered and justified, and ideally empirically validated, and if necessary adapted, on a case to case basis. This requires resources but since the quality of the analysis is crucially dependent on the quality of the sampling, these are resources well-spent.

4.1.1 Best practice data validation

While the overall rationale behind the sampling of data is still relatively frequently addressed in published research, only relatively few studies include a non-trivial discussion of equivalence of documents across languages. As a result, the inclusion of specific corpora may make plausible sense within the context of one case or language, but raise comparability issues in the overall analysis, as corresponding cases may be absent or unavailable in another case of language - a problem that is often found in comparative studies of news coverage (see https://meteor.opted.eu/ for a more extensive categorisation of news sources). An example of a thorough treatment of data validation can be found in Hager and Hilbig (2020), who offer an insightful discussion of

[1] For example, if important keywords that indicate relevant subdimensions of the immigration concept in this case are missing in the translated version of the search string. Such biases may even arise with seemingly trivial sampling criteria - for instance, the keyword "Trump" may work reasonably well in non-US media for identifying coverage of the former US president, but may generate precision problems in US media (which may also cover other members of the family to some extent), and it will completely miss relevant transcripts of "The Late Show", since its host Colbert systematically replaces that name with slanderous nicknames.

how their text corpus came about, and how the specific selection of texts enables valid comparison (even though their study focuses solely on German texts), allowing them to draw convincing causal inferences. Although similarly precise mappings may not be viable for many research projects - since in many cases in computational text analysis corpora are `found data` (Salganik, 2018, p.82-83) for which the exact generative process is unknown - the quality of the inferences drawn from a comparative textual analysis crucially depend on the equivalence of the documents under study.

4.2 Input validation

Speaking generally, the purpose of text preprocessing is that the information rendered available to the computational model is neatly focused on the variation that is responsible for the textual expression of studied meanings, with as much irrelevant variation as is possible removed from the data. In the multilingual study of textual data, preprocessing additionally fulfills the need to ensure that the information entered into the computational model is as equivalent across languages as possible – a challenge that almost always requires ensuring semantic equivalence, and often additionally touches upon pragmatic equivalence (i.e., that corresponding expressions not only denote the same, but carry equivalent significance in their respective contexts of use). This can of course happen in many ways: Reber (2019) has argued that machine translation can offer a suitable way for aligning the semantic representations of text in different languages; Lind et al. (2022) rely on comparable training documents, Chan et al.'s (2020) multilingual topic modeling tool relies on word embeddings to inform the computational model about equivalent contents; and various pre-trained language models are available for supervised applications that purport to bridge the gap. Also, simple pre-processing steps such as stopword removal or lemmatization could contribute to achieving equivalence across languages. Under any circumstances, however, it is imperative that such steps be taken consciously, for explicated reasons, and validated with regard to their effectiveness at achieving equivalent representations on a semantic and ideally also on a pragmatic level.

For machine-translation, all documents or features from a multilingual corpus are translated into one language prior to analysis (Lucas et al., 2015), which allows the presentation of monolingual material to a selected algorithm. If machine-translation is used, it must be carefully aligned with other pre-processing steps. For example, for better translation quality and if resources permit, it is advised to machine translate first and only tokenize or lemmatize afterwards.[2] Machine translation with state of the art neural translation models like DeepL or Google Translate can provide high quality translation for many languages (de Vries et al., 2018), however, the quality of translation varies between different pairs of languages, and even well-performing tools may fall for language peculiarities in ways that affect the analysis.[3] At the same time, even bad machine translation may suffice for some purposes (e.g., as long as those key terms needed for measurement are dependably translated). Accordingly, it is necessary to validate and ascertain that translation quality suffices for a given analytic purpose. To validate machine translation, researchers can, for example, calculate automated metrics based on (human) reference translation or via human judgements (Chatzikoumi, 2019). In a similar vein, also the transformation of documents into multilingual embeddings with (pretrained) multilingual word embedding models is used to provide language agnostic inputs to an algorithm (Chan et al. 2020; Licht, 2022). Multilingual word embeddings have shown great promise in downstream prediction tasks, even if their lack of interpretability can be a drawback when the research goal concerns measurement of latent constructs (as is often the case in social science research). That said, well-validated embeddings are often unavailable especially in less well-resourced language communities (Wu & Dredze, 2020; Joshi et al. 2020), while the validation of embeddings as an input for downstream research objectives can be highly challenging (Rodman, 2020; Rodriguez & Spirling, 2022). It should also be noted that such models tend to "overrepresent hegemonic viewpoints and encode biases potentially damaging to marginalized populations" (Bender et al., 2021, p. 610).

With regard to the redaction of irrelevant information, the first key focus of validation concerns unitization, i.e., the determination on what level of text relevant meanings are expressed and to what units of text they then pertain. Consequently, it is necessary to represent studied corpora to the modeling algorithm in units that contain all of, and only, the required textual information – which may mean that the coding unit (to which a

[2] For a comparison between full-text translation vs. token translation see, for example, Reber (2019); Van der Veen (2022)

[3] For instance, in a project that one of the authors participated in, names of the same key political actors were sometimes transliterated and sometimes translated during machine translation, requiring careful post-hoc re-homogenization.

code is attached) differs from the text required to determine the correct classification. While the removal of required information (e.g., adjacent words/sentences, prior turns, headlines - depending on the focus of the analysis) can result in systematic misclassification, the use of too inclusive units adds noise to the process and can result in considerable measurement uncertainty (e.g., in the topical classification of multi-thematic documents). While this validation stage is relatively[4] language-independent, it is heavily dependent on both the nature of the measured meaning, and the generic conventions of the studied text.

Second, validation efforts need to consider at what level, and in what form, relevant meanings are encoded in the text. For instance, if the text relies on emojis to express relevant meaning, these need to be preserved (and likely separated)[5], whereas the same can arguably be discarded for topical analyses or the study of official reports. Similarly, syntactic information and punctuation may often be uninformative, but where it is informative, strategies must be chosen that retain the relevant variation in some form (e.g., by treating headlines differently from the main text; by tagging contents based on the kinds of clauses they belong to). If certain grammatical information such as subject/object differentiation or tense is important for the anticipated modeling task, it needs to be preserved if present (e.g., keeping certain stopwords or morphological distinctions) or explicated if absent in the text (e.g., using POS taggers). In the same vein, researchers need to homogenize any variation in the textual representation of relevant tokens that are of no consequence for the analysis – for instance, by using available stemming and lemmatizing algorithms to remove case, number and/or gender information where it is deemed uninformative, or by segmenting or tokenizing words to create identical tokens for identical meanings. In order to ensure that equivalent information is retained in multiple languages, it may thus be required to use different preprocessing strategies for different languages.[6] To determine what strategies are valid, some extent of linguistic familiarity – and to the extent that pragmatic equivalence is required, cultural familiarity – may be necessary, so researchers are advised to consult with native speakers before preprocessing languages that they are unfamiliar with.

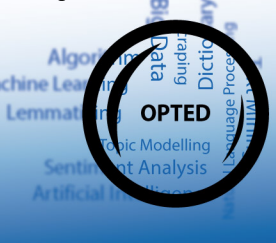
Third, researchers are well-advised to anticipate and, if necessary, disambiguate expressions that convey different relevant information. Depending on the focus of the analysis, this can mean different things. Under most circumstances (with the exception perhaps of contextualized word embeddings), disambiguation is worth considering for important homonyms, where the same spelling is used to express different meanings (e.g., 'rock', which can be a mass of stone, a musical genre, a movement, and more; some languages contain many more homonyms than others, so especially for Semitic languages, POS tagging may be necessary to efficiently distinguish the many possible meanings of identical character sequences). Another type of disambiguation helps distinguish different roles of words, for instance, if an analysis is supposed to distinguish between countries as places and countries as actors, between actors as perpetrators or as victims. In many languages, moreover, words may appear by themselves or as part of standing expressions, whose meaning may not be validly captured by the sum of constituent words (e.g., red tape, long shot) and thus might need to be separated from these (entity recognition tools may be useful for this task). Especially in dictionaries, this may often mean that auxiliary criteria are necessary to determine whether a given token expresses the intended meaning (Baden & Stalpouskaya, 2015). For example, a useful auxiliary rule can be to only count the occurrence of a certain keyword in a text if it appears together with another set of keywords in the same sentence – a strategy that can also serve to distinguish pragmatically different uses of semantically equivalent expressions. The consideration of language-specific disambiguation and collaboration with linguistic experts can hardly be avoided if this validation step is to be implemented in a careful manner.

Fourth, care must be given to select a textual representation format that is adequate for an intended analysis. For a valid analysis, the information required to recognize relevant meanings may be fully contained in single words or phrases, require larger textual units or entire documents, or even extend beyond these. Where

[4] But it is not fully language-independent. For instance, if the coding unit is the word or expression, language use different numbers of words to encode the same meaning - compare "health insurance certificate" vs. "Krankenversicherungsschein" - so equivalent strategies may need to segment words in one or concatenate words in other languages, either in the representation of the textual data or in measurement tools (e.g., dictionaries - the English keyword "insurance" corresponds to the German keyword "*versicherung*", with * denoting truncation)

[5] Interestingly, emojis were found to be a European language-independent resource for automated sentiment analysis (Kralj Novak et al., 2015), a result that makes the application of the same pre-processing strategies for emojis (in studies with European language data sets) seem appropriate.

[6] It should be noted that some computational techniques such as self-supervised learning do not require any preprocessing; however, it is unclear whether such models are indeed capable of autonomously focusing on equivalent information in the data or whether they perpetuate biases rooted in the linguistic encoding.



word order is irrelevant, Bag-of-Words representations (possibly after concatenating standing expressions, named entities, multi-word constructs or whatever may be needed for analysis) may be suitable, while different strategies (using higher n-grams or sequential representations) might be required if word order matters. That said, for cross-lingual analysis, we need to additionally recognize different languages' different strategies for word formation, which may affect the choice of equivalent representations. In agglutinative languages (e.g., German, Russian), complex constructs are expressed by concatenating multiple lemmas, with the effect that one token contains considerably more (and more specific) information than it would in a more analytic language (e.g., English), where each lemma tends to constitute a separate word (e.g., the English 3-gram "health insurance certificate" corresponds semantically to the German 1-gram "Krankenversicherungsschein"). Pretrained multilingual models such mBert, XLM and XLM-RoBERTa, which are trained on as much as 100 languages show great promise towards addressing issues of comparability but their applicability for obtaining valid measurements across languages are still open for debate.

Under all circumstances, it is instrumental that every preprocessing step is documented and justified by reference to some operational need or consideration. Since preprocessing is highly consequential for the analysis (see e.g., Denny & Spirling, 2018, for unsupervised models), selected procedures should never be applied by default, but may be adequate or not depending on the measured contents, processed genres, and other considerations (see also Grimmer et al. (2022) on what they call 'avoiding the default'). Accordingly, any employed NLP technologies should be validated not only with regard to their correct performance (e.g., the share of POS tags correctly applied), but also with regard to their operational purpose (i.e., what they achieve for focusing the analysis on the required information). The key test of input validity is that whatever meaning is studied can still be easily recognized from a preprocessed text, while all variation that is lost is immaterial to the pursued research question. In addition, it is desirable that in the representation that will be passed to the modeling algorithm, there is as little redundancy in the relevant tokens as is possible (e.g., there should not be many tokens that express the same information but differ in ways that are uninformative for the analysis, such as pre- or suffices, capitalization, etc.). One strategy that should help with detecting validity issues and biases in multilingual analysis at the level of pre-processing is to translate a few texts and subject them to the respective preprocessing stages, comparing the achieved representation of relevant contents across languages. The key test of cross-lingual validity is then that the representations obtained by preprocessing map well upon one another, such that the way in which relevant information is expressed (e.g., by single tokens, by token collocation patterns, or by sequentially arranged tokens) is similar between languages.

4.2.1 Best practice input validation

While many computational studies at least document the preprocessing steps taken to prepare the textual data, very little research discusses or offers evidence for how selected preprocessing steps are suitable for achieving representations of the textual data in a multilingual corpus that are comparable across languages. As one example for a study that does this well, Segev (2019) discusses and argues for the selection of certain word types (i.e., country names) and the exclusion of others (i.e., nationalities) to increase the cross-national and cross-linguistic validity of the measurement instrument. Furthermore, Denny and Spirling (2018) offer a useful, language-independent workflow for how to identify the impact of certain pre-processing steps on subsequent unsupervised models.

4.3 Throughout validation

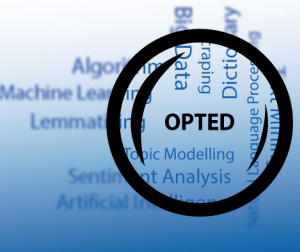
Once it is assured that the textual data is represented in a fashion that focuses on all of, and only all of the relevant variation in the raw text, an algorithm can be chosen for the intended analysis. The primary consideration governing the validity of this choice is whether the meaning intended to be measured is indeed expressed in those ways that are recognized by the algorithm. For some algorithms, that is fairly straightforward. For instance, if single words or expressions are sufficient to express a relevant meaning, simple dictionaries should be able to recognize these, provided that all relevant indicators are adequately included and disambiguated.[7] However, whenever more than one indicator is required for an analysis, additional modeling considerations are required and demand justification. Practically speaking, this primarily concerns the weighting of available information, as well as the modeling of association in the text.

[7] This relates primarily to manifest variables that are easily recognizable (e.g., counting how often a specific organization is mentioned).

Regarding weighting, the key question is whether all information should be regarded equally, or whether there are any deductive assumptions about what information matters more. For instance, for topical analyses, headlines often convey key information, while other textual components matter less. Many sentiment analyses assume that some terms are more powerful indicators of tone than others, or weigh terms by where they appear in the text (e.g., by proximity to other constructs of interest). Likewise, many algorithms include rules about the presence of competing indicators. For instance, a text that contains two indicators of class A, thereof one in the headline, and three indicators of class B in the main text, could be classified as A (privileging position), B (privileging frequency), both (privileging presence) or neither (privileging exclusivity), or along any range of additional rules that require express consideration and justification. One very common assumption in textual analysis is that tokens are the more important the more frequent they are, and the more they discriminate between different documents (i.e., terms that are frequent but occur only in some documents are important; terms that occur everywhere all of the time are less important; and terms that are rare are relatively unimportant; an assumption that is implemented in the widespread tendency to weight vocabularies by term frequency and inverse document frequency, tf-idf). All of these assumptions may or may not be compelling for any given analysis, and cause distinct challenges for multilingual analysis. For instance, agglutinative languages (that concatenate lemmas into long words) have a much steeper long tail distribution of word frequencies than analytic languages, such that discounting "infrequent" terms eliminates much information that would still be retained in analytic languages (e.g., "Krankenversicherungsschein" has a presence of only three tokens per million in the DeReKo Corpus, while "health", "insurance" and "certificate" account for 246, 71, and 29 tokens per million respectively in the British National Corpus). Inversely, synthetic languages (that express grammar by means of affixes or morphology) contain much fewer highly common words than English, such that discounting the most common terms mostly eliminates articles, prepositions, conjunctions etc. in English, but reaches well into common nouns used for construct formation in other languages (e.g., "beit" ["house"] in Hebrew, which appears as part of very many construct state nouns: "beit sefer" ["school"], "beit mishpat" ["court"], "beit knesset" ["synagogue"]). While theoretical and operational reasoning should normally suffice to justify which weighting assumptions make sense for within-language analyses, for multilingual analysis it seems thus desirable to ascertain that equivalent weighting choices indeed emphasize and discount equivalent information. The weighting questions are therefore, just as previously discussed steps, at best taken together with linguistically skilled collaborators.

With regard to the modeling of association, most rule-based algorithms permit considerable researcher control and thus facilitate the justification of modeling choices. For instance, researchers can define an association between two recognized terms as present when it is grammatically expressed, or the terms are collocated within the same sentence, paragraph or floating window, or based on the frequency of their co-presence within or across numerous documents. Matters are more complicated for any algorithms that rely on the extraction and possibly comparison of patterns among tokens – i.e., unsupervised and supervised algorithms, where association measures are typically hard-coded into the algorithm. Researchers need to consider whether document-level word co-occurrences (as assumed, for instance, in most classic machine learning systems) are an adequate representation of the expressed meaning, or whether more restrictive measures need to be obtained (e.g., by designing specific features that focus estimation on specific patterns in the data, or by implementing different algorithms for obtaining the covariance matrices upon which all supervised and unsupervised tools depend). Again, just as for input validation, the easiest way of ascertaining validity seems to subject a few texts to the algorithm's transformation and determine whether the represented information (e.g., weighted scores, associations) reflects meaningful features of the processed text – in case of multilingual analysis, whether it reflects the same meaningful features in equivalent ways.

Next, modeling algorithms tend to make assumptions about the manifest structure of indicative patterns. For instance, many clustering algorithms such as PCA or KNN assume that every token belongs to exactly one cluster, while factor analytic or topic modeling algorithms permit that the same token may contribute to more than one pattern. Especially in languages with many homonyms or words that appear as part of different multiword expressions, therefore, such clustering solutions are inappropriate for certain analyses, as the same token cannot be constrained to one role exactly. In the same vein, topic models assume that most documents are dominated by only a few topics. Depending on the nature of the studied texts and pursued measurements, these assumptions may be plausible or not: While topic modeling tweets may be statistically challenging (due to data sparsity), for instance, it is substantively plausible that most tweets are monothematic but re-use selected words that may feature across multiple topics. By contrast, topic modeling parliamentary minutes at the doc-



ument level might raise validity issues if the debate covers many topics and might thus require some segmentation of the document prior to analysis. This challenge also relates to the study of multilingual corpora to the extent that different generic conventions govern texts sampled from different languages - for instance, where one language tends to use a much more narrative style than another, regularly progressing from topic to topic within the span of one document. Agglutinative languages may cause issues as key terms in the analysis are too infrequent to constitute recognized patterns (e.g., because "health insurance", "insurance industry", "health policy" etc. each constitute independent tokens, depressing term overlap: "Krankenversicherung", "Versicherungsindustrie", "Gesundheitspolitik"). Substantive consideration should also be given to the choice of relevant hyperparameters: For instance, there should be explicated reasons beyond statistical criteria for why one expects a corpus subjected to topic modeling to contain few or many topics (is it plausible that an entire news debate can be reduced to only a dozen topics?).

Such considerations often also require some deeper understanding of the algorithmic design - such as topic models' tendency to constitute *any* systematic pattern as a topic. This tendency commonly results in a range of "garbage" topics that result from non-topical variations in language use that survived the preprocessing stage. Importantly, however, different languages tend to generate different kinds of additional systematic patterns beside texts' topical structure (e.g., owing to stylistic devices or register markers)[8], which may need to be considered prior to, and validated during the modeling stage. Especially for supervised algorithms, providing substantive justifications for the choice of specific algorithms or hyperparameters often presents a major challenge, as very little is known how exactly modeling co-variation in a hierarchical fashion (e.g., decision tree algorithms), as a vector space (e.g., SVM), or in a neural network might affect classification. While it is evident (and plausible) that such choices are consequential, virtually no theory is available for justifying specific choices.

4.3.1 Best practice process validation

With regard to research that offers a transparent discussion of how a chosen computational instrument or model links up with the concept of interest and how it was produced, best practices inevitably vary considerably depending on the specific algorithms chosen for analysis. As one outstanding example, Maurer and Diehl (2020) study populism in French and English tweets, using a multilingual dictionary as a primary tool. In so doing, they first distinguish between valence on the one hand, references to the people and references to the elites on the other (in line with a popular ideational approach of understanding populism). They then go in important detail on how they constructed and validated their French and English dictionaries against crowdsourced benchmarks, forcefully convincing their readers of the quality of their instrument. As an example of a use of unsupervised models, Baden et al. (2020) go to quite some length discussing what exactly the extraction of regularities in their data via inductive topic models can, and cannot be interpreted to be, before relegating the procedure to an auxiliary role rather than a standalone analytic procedure.

4.4 Output validation

The final validation stage is concerned with the quality of obtained measures and their equivalence across languages and across cases. Just as in monolingual or single case studies, the validity of the computationally obtained estimates can be assessed through a comparison with an established benchmark, which is often manually coded material. Such a test can indicate to what extent the automated measurements can mimic human understanding of text (e.g., Song et al., 2020). Recall and precision are typically calculated metrics to express the degree of compliance of automated measurement with the gold standard.

In multilingual applications, output validation needs to be considered for each included language. As there are numerous sources of bias that arise from linguistic differences, it cannot be assumed that a tool that performs well in one language does the same in another one. Ideally, researchers work with a high-quality gold standard for each language and for each case in their dataset. In order to establish a gold standard that captures comparable meanings in different languages and contexts, the codebooks, human coder training, and intercoder reliability tests must be designed accordingly. More practically, the definitions, rules, and examples described in the codebook should be indicative for all languages and cases involved. Just as for monolingual projects, it

[8] For instance, the differentiation between formal and informal register is likely to generate systematic patterns in Korean that may inflate pattern counts and result in garbage topics, while the same differentiation is unavailable in other languages.

is advantageous to train all involved coders in joint (online) sessions and to clarify issues or adjust the codebook collaboratively (Rössler, 2012). If such joint sessions are impractical, for example, in cases of crowd-coding, instructions and test questions used to filter out potential spammers should be carefully selected (Lind et al., 2017). Regardless of the coding mode (more classic manual expert coding or crowd-coding), intercoder reliability tests should cover reliability across languages/cases as well as within languages/cases (Peter & Lauf, 2002). If all coders are skilled in one language, they can code the same material to establish intercoder reliability (Hopmann et al., 2016). Care should be taken to ensure that the material is representative for all languages and cases. To achieve this, a random subset per language and case can be translated into a common language (Courtney et al., 2020). As additional tests and to better assess intercoder reliability within language/case at least two coders code original language material.

Apart from self-created baselines, the obtained measures can also be compared with variables known to measure the same concept (convergent validation) or with variables known to measure concepts that differ (discriminant validation). However, in such cases, context-dependent differences between such third variables need to be considered. For instance, a convergent measure may constitute a close equivalent in one language or context but exhibit important differences or flaws in another. External data may differ in quality from context to context, such that identical measurement performance may not result in equivalent convergent or discriminant patterns. For instance, party membership may offer a fair reference variable for political ideology in one country's party system, but not in another, which is marked by a more multi-dimensional set of cleavages; it may be comparatively weaker as an indicator in context where party cohesion is low, or where there is considerable fluidity in parties and party memberships.

For the evaluation of obtained validation results, furthermore, it is often useful to consider substantive differences between compared cases and languages: If a measured meaning is far more dominant, or largely absent, in one compared case, performance scores may need careful interpretation, as accuracy measures can reach very high values without valid discrimination (e.g., if a meaning occurs only in 3% of cases, never coding it at all classifies 97% of cases correctly; chance- and category frequency-corrected measures may be necessary to obtain an informative measure of measurement validity). A mere appraisal of overall performance across languages may be grossly misleading.

One key strategy for output validation, therefore, is to examine recall and precision as well as the corresponding misclassifications. While overall performance metrics may offer a decent first sense of classification performance, even good performance leaves plenty of room for systematic biases. For instance, translated dictionary indicators may work well overall, but if some minority group in one language uses a different term to express the same, all such references will be missed, while accuracy can still be high. To detect such biases, error analysis offers valuable insights. Systematic issues in precision are easily detected by pulling up any coded but irrelevant cases and examining whether these share any relevant characteristics. Systematic recall issues are harder to detect but can be found by drawing a random sample of texts for manual classification, before examining which cases are missed by the method. If a proper gold standard is available, examining both false positives and negatives with an eye toward detecting patterns of misclassification can offer not only valuable information about possible biases, but also prepares the ground for fixing any detected issues and thus improving the performance of employed tools.

4.4.1 Best practice output validation

By far the most seen validation strategy found in computational text analysis research is one that benchmarks obtained measures against a (combination of) human-coded benchmarks or against convergent or divergent measures. For example, Temporão et al., (2018) assess convergent and predictive validity of their computationally obtained ideological scaling measures of social media users. Lo et al. (2016) assess convergent validity of their scaling measures of party manifestos by comparing the obtained estimates with alternative existent measures.

5 Discussion and Conclusion

There are currently 7,151 spoken languages in the world, with 23 languages accounting for the mother tongues of half the world's population.[9] If social scientists want to learn from texts produced in these languages, the need to engage with multilingual text analysis methods is obvious. Besides a reliance on manual analyses

[9] See https://www.ethnologue.com/guides/how-many-languages [accessed August 2022].

and multilingual coders, especially the use of automated methods for dealing with large text corpora raises important challenges for validation (Grimmer & Stewart, 2023; van Atteveldt & Peng, 2018). With the increasing popularity of CTA methods in general but not for multilingual projects in particular (Baden et al., 2022) we here focused on validation best practices for multilingual projects. As a contribution to ensuring quality social science, in this document, we proposed a validation framework to be of use for applied research. Building on insights from a content analysis of published literature in the social sciences, and an expert survey with the respective authors, we derived practical recommendations for the central methodological steps necessary to obtain valid observations for a concept. In detail, the framework outlines validation techniques for the selection of equivalent data sources and samples, the preparation of equivalent input features, the selection of equivalence data processing techniques, and finally equivalent measurements.

In a nutshell, our review has shown that validation remains a major challenge across computational text analysis research. Crucially, neither existing methodological discourse, nor applied research practices reflect even remotely as thoroughly upon validation as would be needed to confidently trust in many machine-made measurements. To date, most validation efforts focus on documenting the relevance of used materials, as well as the demonstration of valid measurement outputs – data validation and output validation, respectively. And indeed, without data appropriate to the research objective, much effort invested in subsequent validation steps may be worthless. Likewise, comparing input, throughput and output validation, output validation is especially indicative of a successful concept measurement. However, whichever biases or validity concerns may be detected during output validation are most likely derived from validity issues that could have been detected and need to be addressed at the stages of validating input and throughput – two stages that are to date largely neglected in the published literature. One of the key insights that arises from our investigation thus is that the need for validation far exceeds demonstrating predictive performance relative to a given benchmark, but that data, input and processes constitute key considerations, which hold far-reaching implications for measurement validity, and should be taken more strongly into account.

Of course, the observed paucity of attention to validation – especially, but not only, with regard to validating input and throughput – only reflects what has been documented in the surveyed publications. It is possible, and indeed likely, that additional validation steps have been taken in some of these studies, but were not reported, be that owing to a perception that these were less important, a concern about alerting reviewers to possible issues, or merely due to restrictions in space. In light of the key importance of validation, a first takeaway from this intervention might thus be the acute need to document and justify methodological choices and report any validation efforts taken. Not only should authors dedicate some space to enabling readers to follow taken validation steps and thus build confidence in the measurement, but reviewers need to demand and scrutinize presented validation efforts (keeping in mind that a presented, less than perfect validation is still more valuable than none), and editors need to set aside the requisite space for the purpose. Beyond enabling a more informed appraisal of achieved measurement validity in computational text analysis, such transparency is also required as a foundation for any more systematic methodological discourse on validation strategies and a key source of information and insights into possible improvements in the development of CTA technologies.

In light of this acute need to attend systematically to issues of validation in CTA, questions about the practical implementation of such validation immediately arise. What concrete steps are necessary to validate the suitability of specific designs, algorithmic strategies and methods, and what are the adequate benchmarks and standards for successful validation? In order to advance validation practices further, proposing a comprehensive validation framework is merely setting the stage, raising broad questions that need to be asked, but that may lead to different practices and strategies depending on the specific nature of intended measurement. In order to execute the validation requirements outlined in the framework, it is clear that the corresponding resources are needed. Many of these steps can be supported by better research infrastructures such as OPTED. Services like the curated inventories for data sources (e.g., D3.2, D4.2, D5.1) and tools that support the input and processing stage (e.g., D3.3) help research teams across Europe to save resources. The community-promoting character of such infrastructures is another driving force. After all, cooperation partners and thus often experts for a diverse set of languages and cases can be found through them, who are essential for all presented validation stages.

References

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

Baden, C., Kligler-Vilenchik, N. and Yarchi, M., (2020). Hybrid content analysis: Toward a strategy for the theory-driven, computer-assisted classification of large text corpora. *Communication Methods and Measures, 14*(3), 165–183. https://doi.org/10.1080/19312458.2020.1803247

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

Baden, C. & Stalpouskaya, K. (2015). Common methodological framework: Content analysis. A mixed-methods strategy for comparatively, diachronically analyzing conflict discourse. *INFOCORE Working Paper 2015/10*. https://www.infocore.eu/results/

Baden, C., & Tenenboim-Weinblatt, K. (2018). The search for common ground in conflict news research: Comparing the coverage of six current conflicts in domestic and international media over time. *Media, War & Conflict, 11*(1), 22-45. https://doi.org/10.1177/1750635217702071

Barberá, P., Boydstun, A. E., Linn, S., McMahon, R., & Nagler, J. (2021). Automated text classification of news articles: A practical guide. *Political Analysis*, *29*(1), 19–42. https://doi.org/10.1017/pan.2020.8

Bender, E.M. (2011). On achieving and evaluating language-independence in NLP. *Linguistic Issues in Language Technology*, *6(3)*. https://doi.org/10.33011/lilt.v6i.1239

Bender, E.M. & Friedman, B. (2018). Data statements for natural language processing: Toward mitigating system bias and enabling better science. In Lee, L., Johnson, M., Toutanova, K., Roark, B. (Eds.), *Transactions of the Association for Computational Linguistics, Volume 6*, (pp.587-604). MIT Press.

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

Boomgaarden, H. G., De Vreese, C. H., Schuck, A. R., Azrout, R., Elenbaas, M., Van Spanje, J. H., & Vliegenthart, R. (2013). Across time and space: Explaining variation in news coverage of the E uropean U nion. European Journal of Political Research, 52(5), 608–629.

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. https://doi.org/10.1080/19312458.2020.1812555

Chatzikoumi, E. (2020). How to evaluate machine translation: A review of automated and human metrics. *Natural Language Engineering, 26*(2), 137–161. https://doi.org/10.1017/S1351324919000469

Courtney, M., Breen, M., McMenamin, I., & McNulty, G. (2020). Automatic translation, context, and supervised learning in comparative politics. *Journal of Information Technology & Politics*, *17*(3), 208–217. https://doi.org/10.1080/19331681.2020.1731245

de Vries, E. (2021). The Sentiment is in the details: A Language-agnostic approach to dictionary expansion and sentence-level sentiment analysis in news media. *Computational Communication Research*. https://doi.org/10.31235/osf.io/8y3jq

de Vries, E., Schoonvelde, M., & Schumacher, G. (2018). No longer lost in translation: Evidence that Google Translate works for comparative bag-of-words text applications. *Political Analysis*, *26*(4), 417–430.

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.

Gentzkow, M., Kelly, B., & Taddy, M. (2019). Text as data. *Journal of Economic Literature*, *57*(3), 535–74.

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

Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political analysis*, *21(3)*, 267–297. https://doi.org/10.1093/pan/mps028

Gurevitch, M., & Blumler, J. G. (2003). Der Stand der vergleichenden politischen Kommunikationsforschung: Ein eigenständiges Feld formiert sich [The state of comparative political communication research: An independent field is forming]. In Esser, F. & Pfetsch, B. (Eds.), *Politische Kommunikation im internationalen Vergleich [Political communication in international comparison]* (pp. 371-392). VS Verlag für Sozialwissenschaften.

Gutiérrez, E.D., Shutova, E., Lichtenstein, P., de Melo, G. & Gilardi, L., (2016). Detecting cross-cultural differences using a multilingual topic model. In Lee, L., Johnson, M., & Toutanova, K. (Eds.), *Transactions of the Association for Computational Linguistics, Volume 4*, (pp.47-60). MIT Press.

Hager, A. & Hilbig, H. (2020). Does public opinion affect political speech? *American Journal of Political Science, 64*(4), 921–937. https://doi.org/10.1111/ajps.12516

He, J., & van de Vijver, F. (2012). Bias and equivalence in cross-cultural research. *Online Readings in Psychology and Culture, 2*(2), 2307–0919. http://dx.doi.org/10.9707/2307-0919.1111

Hopmann, D. N., Esser, F., de Vreese, C. H., Aalberg, T., van Aelst, P., Berganza, R., ... & Strömbäck, J. (2016). How we did it: approach and methods. In C. de Vreese, F. Esser, & D. N. Hopmann (Eds.), *Comparing political journalism* (pp. 10-21). Routledge.

Jacobs, A.Z. & Wallach, H., (2021). Measurement and fairness. In Proceedings of the 2021 ACM conference on fairness, accountability, and transparency (pp. 375-385). https://doi.org/10.1145/3442188.3445901

Jolly, S., Bakker, R., Hooghe, L., Marks, G., Polk, J., Rovny, J., Steenbergen, M. and Vachudova, M.A., 2022. Chapel Hill expert survey trend file, 1999–2019. *Electoral Studies, 75. https://doi.org/10.1016/j.electstud.2021.102420*

Joshi, P., Santy, S., Budhiraja, A:, Bali, K., & Choudhury, M. (2020). The state and fate of linguistic diversity and inclusion in the NLP world. In Jurafski, D., Chai, J., Schluter, N., & Tetreault, J. (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp.6282–6293). Association for Computational Linguistics. https://aclanthology.org/2020.acl-main.560.pdf

Kralj Novak, P., Smailović, J., Sluban, B., & Mozetič, I. (2015). Sentiment of emojis. *PloS one*, *10*(12), https://doi.org/10.1371/journal.pone.0144296

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

Licht, H. (2022). Cross-Lingual classification of political texts using multilingual sentence embeddings. Working paper. URL:https://osf.io/384wr/?view_only=abcfb31cada64dbca8f7b43a59b1e696

Licht, H. & Lind, F. (2022). Going cross-lingual: A guide to multilingual text analysis. Working paper.

Lind, F., Eberl, J. M., Eisele, O., Heidenreich, T., Galyga, S., & Boomgaarden, H. G. (2022). Building the Bridge: Topic modeling for comparative research. *Communication Methods and Measures*, 16(2).https://doi.org/10.1080/19312458.2021.1965973

Lind, F., Eberl, J. M., Heidenreich, T., & Boomgaarden, H. G. (2019). Computational communication science| when the journey is as important as the goal: A roadmap to multilingual dictionary construction. *International Journal of Communication*, *13*, 4000–4020.

Lind, F., Gruber, M., & Boomgaarden, H. G. (2017). Content analysis by the crowd: Assessing the usability of crowdsourcing for coding latent constructs. *Communication methods and measures*, *11*(3), 191-209. https://doi.org/10.1080/19312458.2017.1317338

Lind, F., Heidenreich, T., Kralj, C., & Boomgaarden, H. (2021). Greasing the wheels for comparative communication research: supervised text classification for multilingual corpora. *Computational Communication Research*, *3*(3), 1–30. https://computationalcommunication.org/ccr/article/view/109

Lo, J., Proksch, S. O., & Slapin, J. B. (2016). Ideological clarity in multiparty competition: A new measure and test using election manifestos. *British Journal of Political Science*, *46*(3), 591–610.

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. https://doi.org/10.1080/19312458.2021.1955845

Maurer, P., & Diehl, T. (2020). What kind of populism? Tone and targets in the Twitter discourse of French and American presidential candidates. *European Journal of Communication*, *35*(5), 453–468. https://doi.org/10.1177/0267323120909288

Peter, J., & Lauf, E. (2002). Reliability in cross-national content analysis. *Journalism & Mass Communication Quarterly, 79*(4), 815–832. https://doi.org/10.1177/107769900207900404

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. https://doi.org/10.1111/lsq.12218

Reber, U. (2019). Overcoming language barriers: Assessing the potential of machine translation and topic modeling for the comparative analysis of multilingual text corpora. *Communication methods and measures*, *13*(2), 102–125. https://doi.org/10.1080/19312458.2018.1555798

Rodman, E. (2020). A timely intervention: Tracking the changing meanings of political concepts with word vectors. *Political Analysis, 28*(1), 87–111.

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

Rössler, P. (2012). Comparative content analysis. In F. Esser & T. Hanitzsch (Eds.), *The handbook of comparative communication research* (pp. 481-490). Routledge.

Salganik, M.J. (2019). *Bit by bit: Social research in the digital age*. Princeton University Press.

Segev, E. (2019). From where does the world look flatter? A comparative analysis of foreign coverage in world news. *Journalism*, *20*(7), 924-942. https://doi.org/10.1177/1464884916688292

Sigismondi, P. (2018). Exploring translation gaps: The untranslatability and global diffusion of "cool". *Communication Theory, 28*(3), 292–310. https://doi.org/10.1093/ct/qtx007

Song, H., Tolochko, P., Eberl, J. M., Eisele, O., Greussing, E., Heidenreich, T., Lind, F., Galyga, S. & Boomgaarden, H. G. (2020). In validations we trust? The impact of imperfect human annotations as a gold standard on the quality of validation of automated content analysis. *Political Communication*, *37*(4), 550–572. https://doi.org/10.1080/10584609.2020.1723752

Temporão, M., Kerckhove, C. V., van der Linden, C., Dufresne, Y., & Hendrickx, J. M. (2018). Ideological scaling of social media users: a dynamic lexicon approach. *Political Analysis*, *26*(4), 457–473.

Theocharis, Y., Barberá, P., Fazekas, Z., Popa, S. A., & Parnet, O. (2016). A bad workman blames his tweets: the consequences of citizens' uncivil Twitter use when interacting with party candidates. *Journal of Communication*, *66*(6), 1007–1031. https://doi.org/10.1111/jcom.12259

van Atteveldt, W., & Peng, T. Q. (2018). When communication meets computation: Opportunities, challenges, and pitfalls in computational communication science. *Communication Methods and Measures*, *12*(2-3), 81–92. https://doi.org/10.1080/19312458.2018.1458084

van Atteveldt, W., Trilling, D. and Calderon, C.A., (2022). *Computational Analysis of Communication*. John Wiley & Sons.

Van der Veen, M. (2022). *Machine translation for the rest of us*. Working Paper. https://github.com/amaurits/translation4tru

Volkens, A., Burst, T, Krause, W., Lehmann, P., Matthieß, T., Regel, S., Weßels, B., & Zehnter, L. (2021): The Manifesto Data Collection. Manifesto Project (MRG/CMP/MARPOR). Version 2021a. Berlin: Wissenschaftszentrum Berlin für Sozialforschung (WZB). https://doi.org/10.25522/manifesto.mpds.2021a

Watanabe, K. (2021). Latent semantic scaling: A semi-supervised text analysis technique for new domains and languages. *Communication Methods and Measures*, *15*(2), 81–102.

Wu, S., & Dredze, M. (2020). Are all languages created equal in multilingual BERT?. In Gella, S., Welbl, J., Rei, M., Petroni, F., Lewis, P.,.. Hajishirzi, H. (Eds.). *Proceedings of the 5th Workshop on Representation Learning for NLP* (pp.120-130). https://aclanthology.org/2020.repl4nlp-1.16.pdf

Zhao, J., Wang, T., Yatskar, M., Cotterell, R., Ordonez, V. and Chang, K.W. (2019). Gender bias in contextualized word embeddings. *Computation and Language.* arXiv preprint arXiv:1904.03310.