

## Three Gaps in Computational Text Analysis Methods for Social Sciences: A Research Agenda

Christian Baden, Christian Pipal, Martijn Schoonvelde & Mariken A. C. G van der Velden

To cite this article: Christian Baden, Christian Pipal, Martijn Schoonvelde & Mariken A. C. G van der Velden (2022) Three Gaps in Computational Text Analysis Methods for Social Sciences: A Research Agenda, Communication Methods and Measures, 16:1, 1-18, DOI: [10.1080/19312458.2021.2015574](https://doi.org/10.1080/19312458.2021.2015574)

To link to this article: <https://doi.org/10.1080/19312458.2021.2015574>



© 2021 The Author(s). Published with license by Taylor & Francis Group, LLC.



Published online: 27 Dec 2021.



Submit your article to this journal [↗](#)



Article views: 4974



View related articles [↗](#)







View Crossmark data [↗](#)



Citing articles: 1 View citing articles [↗](#)

# Three Gaps in Computational Text Analysis Methods for Social Sciences: A Research Agenda

Christian Baden , Christian Pipal , Martijn Schoonvelde ,  
and Mariken A. C. G van der Velden 



The Hebrew University of Jerusalem Mount Scopus; University of Amsterdam; University College Dublin; Vrije Universiteit Amsterdam

## ABSTRACT

We identify three gaps that limit the utility and obstruct the progress of computational text analysis methods (CTAM) for social science research. First, we contend that CTAM development has prioritized technological over validity concerns, giving limited attention to the operationalization of social scientific measurements. Second, we identify a mismatch between CTAMs' focus on extracting specific contents and document-level patterns, and social science researchers' need for measuring multiple, often complex contents in the text. Third, we argue that the dominance of English language tools depresses comparative research and inclusivity toward scholarly communities examining languages other than English. We substantiate our claims by drawing upon a broad review of methodological work in the computational social sciences, as well as an inventory of leading research publications using quantitative textual analysis. Subsequently, we discuss implications of these three gaps for social scientists' uneven uptake of CTAM, as well as the field of computational social science text research as a whole. Finally, we propose a research agenda intended to bridge the identified gaps and improve the validity, utility, and inclusiveness of CTAM.

Over the past decade, computational methods for the analysis of digital texts have experienced an unprecedented boom across the social sciences (e.g., see overview articles of Brady, 2019; Hilbert et al., 2019; Lazer & Radford, 2017; van Atteveldt & Peng, 2018; Van Atteveldt et al., 2019). In step with the rapid expansion of available data, the accessibility and capabilities of analytic software have also advanced rapidly. Not only were software and ideas from the computational sciences introduced into social science research, but also social scientists' own efforts at developing computational text analysis tools have regained momentum. We have seen the emergence of computational social science research centers, the establishment of (social) data science degree programs, as well as new divisions, journals, networks and research infrastructures dedicated to computational social science research. Clearly, computational text analysis methods (CTAM) are here to stay.

Reflecting the increasing importance of CTAM in cutting-edge social science research, computational methods are used in a growing share of studies published in leading journals, with several recent special issues specifically dedicated to CTAM in social research.<sup>1</sup> Yet, available tools are taken up unevenly. While some algorithms – such as SentiStrength (Thelwall et al., 2010) or topic models (Blei & Lafferty, 2006) – are widely adopted across social science scholarship (Günther & Quandt, 2016), many – especially, high-powered algorithms, such as Neural Network classifiers (e.g., Choi et al., 2021) – remain a rare sight. In the existing scholarship, this uneven uptake of CTAM is typically explained by

**CONTACT** Christian Baden  [c.baden@mail.huji.ac.il](mailto:c.baden@mail.huji.ac.il)  Department of Communication & Journalism, The Hebrew University of Jerusalem Mount Scopus, Jerusalem, 9190501, Israel

<sup>1</sup>e.g., Roberts (2016); Theocharis and Jungherr (2020); van Atteveldt and Peng (2018).

© 2021 The Author(s). Published with license by Taylor & Francis Group, LLC.

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives License (<http://creativecommons.org/licenses/by-nc-nd/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited, and is not altered, transformed, or built upon in any way.

reference to the rapid pace of development in the computational sciences (Boumans & Trilling, 2016), as well as social scientists' limited computational literacy (Domahidi et al., 2019). While we agree that important potentials remain to be unlocked by training social scientists in the use of CTAM, we argue in this paper that additionally, social scientists often have good reason to forgo available computational solutions and to prefer manual strategies, despite the (often considerable) required effort.

In particular, we delineate three major gaps that hamper the utility and attractiveness of CTAM for many social science applications. First, we identify a mismatch between CTAM developers' emphasis on technological and statistical properties, and social scientists' primary concern for operational demands and measurement validity (see, for example, Nicholls & Culpepper, 2020; van Atteveldt et al., 2021). Where CTAM capabilities fail to map onto the specific needs of valid measurement, researchers may prefer methods that afford them a greater degree of manual control and transparency. Second, we identify a mismatch between CTAMs' tendency to focus exactly one kind of information and social scientists' need for the simultaneous measurement of multiple, often internally complex textual contents (e.g., object-specific evaluations, frames; Liu & Zhang, 2012; Walter & Ophir, 2019). As a consequence, researchers often need to combine or concatenate different tools, leveling if not reversing the advantages of computational processing. Finally, we identify a mismatch between the growing linguistic diversity and orientation toward comparative research in the social sciences, and the continued, heavy dominance of English in CTAM (Bender, 2011). Especially for the growing community of European and non-Western scholars, CTAM rarely offer adequate capabilities for their application in cutting-edge research (e.g., Amram et al., 2018).

We substantiate our claims through a broad review of the state of the methodological literature, supported by data from two recent efforts at taking stock of the state of CTAM in the social sciences, both conducted within the context of the OPTED project (<http://www.opted.eu>), of which the authors are a part. On the one hand, we draw upon a content analysis of research articles published in the top 20 journals in communication science, political science, sociology and psychology<sup>2</sup> between 2016 and 2020. Identifying all 854 articles that involved some form of quantitative textual analysis, this analysis investigated the prevalence and uses of either manual or computational text analysis methods in the social sciences (for details about the data collection, see Online Appendix A: <https://osf.io/s7h8b>). On the other hand, we drew upon a survey administered to all authors of quantitative text analytic research identified via said content analysis, which inquired about researchers' considerations and concerns in the application of computational and manual text analytic strategies (for details about the data collection, see Online Appendix B: <https://osf.io/s7h8b>). Seen against the present state of the field reflected in both data sources, the selective uptake of CTAM not only illustrates the widespread impact of the three gaps that we identify, but points to important implications for the future development of social science textual research. Following our discussion of each gap, we conclude by proposing a research agenda for addressing these challenges.

## A broad typology of CTAM

Computational text analysis methods (CTAM) are an umbrella term for many different methods (Boumans & Trilling, 2016), from tools for extracting specific contents using simple keywords or formatting rules (e.g., hashtags) to statistically complex software solutions (e.g., BERT or other large-scale language models; Devlin et al., 2019). They include generalized tools suitable to process just about any type of textual data, as well as knowledge-heavy, highly-specialized packages. Methods require variable degrees of human supervision – from a few parametric choices to expansive training sets, databases or reference corpora. Depending on the type of CTAM, supervision may occur in the

<sup>2</sup>Web of Science categories *Communication*, *Political Science*, *Sociology*, *Social Science: Mathematical Models*, *Psychology: Applied* and *Psychology: Mathematical*; for the list of included journals, please refer to Online Appendix A: <https://osf.io/s7h8b>.

form of human-created models and classification rules, pre-trained tools, ex-post validation efforts, or human supervision kept in the loop continuously (e.g., Baden et al., 2020; Munro, 2020; Ramage et al., 2009).

Even within the same family of methods, there exists an impressive variety of available solutions. For example, dictionaries have been used to classify broad themes or textual sentiments, recognize complex theory-informed constructs, and even to extract the semantic organization of complex debates (e.g., Lind et al., 2019; Tenenboim-Weinblatt & Baden, 2021; van Atteveldt et al., 2021). Numerous clustering techniques have been proposed, ranging from recently popular topic models to strategies capable of organizing entire document collections into discrete events, ongoing news stories or chains of reused materials (e.g., Maier et al., 2018; Papacharissi & de Fatima Oliveira, 2008; Welbers et al., 2020). Given the immense scholarly creativity and rapid expansion of development both inside and beyond academia, any attempt to organize CTAM remains necessarily limited.

Accordingly, in this paper, we cast a wide net, distinguishing broadly between rule-based, supervised, and unsupervised approaches. Rather than a precise distinction of tools, this categorization recognizes three modes of thinking that underlie these approaches. Where CTAM takes a *rule-based* approach, they assume that rules needed to classify text in conceptual categories are known and can be fully specified. Such rules may take the form of a) exhaustively listing all relevant forms of a content (e.g., keyword-based strategies, dictionaries); b) specifying formal rules for recognizing relevant contents (e.g., link extraction; shallow parsers), or c) any combination of these (e.g., dependency parsers). By comparison, *supervised* approaches assume that classification rules are *not* fully specified, but can be inferred from observed classifications based on shared regularities in the text. *Unsupervised* approaches, finally, drop the assumption that a given variable of interest corresponds with a pre-defined class and instead classify instances inductively based on observed regularities in the text. Despite their internal heterogeneity, therefore, each group of CTAM is characterized by a specific set of demands and assumptions, which are consequential for their utility and performance in textual research.

Given their immense diversity, CTAM's use in social scientific research obviously depends on a wide range of factors. Ready-to-use off-the-shelf software solutions may be easier to apply than tools that require extensive tweaking and a more advanced computational skill set. Long-established approaches supported by rich experience may inspire more confidence in researchers than the most recent, still unfamiliar tools. Additionally, well-documented tools that transparently communicate embedded assumptions may be more appealing than complex, black-boxed algorithms. Doubtlessly, the need for computational literacy presents one major obstacle for social scientists attempting to make use of available CTAM. However, in the following, we will argue that there are also several sound reasons that may lead (*especially* computationally literate!) social scientists to opt against using CTAM in their research.

## Gap I: Technology before validity

The first gap concerns the disconnect between social scientific methodological discourses, and those methodological discourses that accompany the development of CTAM. In social science research, much textual measurement focuses on latent and abstracted constructs. As these can be referenced in natural discourse in myriads of ways (Kantner & Overbeck, 2020; Nicholls & Culpepper, 2020), there is rarely a straightforward way to operationalize them. Hence, measurement validity and the operationalization of complex constructs are of particular concern in empirical social science text analysis. Aiming to improve upon coders' less than precise intuitive grasp of measured constructs, content analytic research relies on conceptual definitions and often detailed operational rules, engaging in careful coder training to ensure a shared understanding of coded constructs (Krippendorff, 2004). Despite remaining difficulties at obtaining high levels of reliability (Weber et al., 2018), considerable resources are dedicated to ensuring the validity of textual measurements. By comparison, CTAM cannot build upon an intuitive understanding of textual meaning, and thus rely entirely on human

supervision to ensure measurement validity. Yet, only the development of rule-based approaches is sometimes accompanied by similar efforts at concept-driven operationalization and the development of substantive validity criteria (e.g., in the construction of dictionaries, word banks and rule sets). For supervised applications, validation efforts remain largely limited to the manual creation of training data sets, and are rarely reported. Subsequently, operationalization is replaced by powerful algorithms trained to identify any patterns and indicators that correlate with provided annotations, effectively supplanting validity with predictive performance (see also Theocharis & Jungherr, 2020). In their effort to match human annotations or given ground truths, algorithmic classifiers show little interest in separating valid variation in the material from accidental, meaningless regularities and confounding patterns. Relying on salient, correlated patterns identified in the data, these tools may still frequently guess correctly, while potentially introducing systematic biases into the analysis. As a drastic example, Hirst et al. (2014) demonstrate how a machine classifier trained to recognize political ideology ended up classifying incumbency instead, which happened to correlate with ideology in their data but was easier to recognize for the machine. Things are worse still for unsupervised CTAM, whose inductive outlook prevents their evaluation based on their predictive performance. Instead, the validity of findings from unsupervised procedures is often evaluated no more scrupulously than by checking their interpretability (hardly a high standard, considering humans' capacity to perceive meaning in random patterns; Shermer, 2000) or consistency with expectations (not a tough standard either, given human confirmation bias; Nickerson, 1998). Where fit metrics are offered to assist model evaluation, these primarily serve to rate qualities that have little bearing upon measurement validity. Owing to a narrow understanding of validity in these CTAM, operational validity is supplanted by a cursory, parametric and usually post-hoc check whether the achieved measurement appears plausible, while systematic biases may remain undetected.

Between social scientists' focus on operational validity and CTAM developments' emphasis on predictive performance, this separation of methodological perspectives is reflected in at least three key disconnects, which diminish the utility of CTAM for social science research. First, social scientists rarely find their long-established knowledge – about language and discourse, genres and styles, constructs and measures – reflected in CTAM development. Computational tools' alignment with social scientific and linguistic knowledge is rarely included as an evaluation criterion or objective throughout the development process. For example, the vast majority of CTAM processes textual contents without regarding their position within a document. Thereby, it ignores decades of research documenting the systematic ordering of most textual genres – be that journalism's "inverted pyramid" style of front-loading key information (e.g., van Dijk, 1985), the narrative organization of political speech (e.g., Pipal et al., 2021), or the heavily relational organization of interactive (online and offline) discourse (e.g., Baden et al., 2020). Instead of following a knowledge-based operational logic (for a rare example, see Mulder et al., 2021), CTAM, especially the supervised and unsupervised versions, tend to hand large amounts of (relevant and irrelevant) data to the machine, hoping that the algorithm will identify and rely on the "right," valid patterns (Conway, 2006; Hirst et al., 2014). At the same time, we know hardly anything about how different textual genres affect the performance of available CTAM.

Neither does social scientists' knowledge about the operational demands of measured constructs inform the development of CTAM. For instance, despite topic models' (Blei & Lafferty, 2006) ostensible reference to topicality, neither the original introduction, nor subsequent developments refer to social scientific knowledge about the topicality of texts (e.g., van Dijk, 1985). Instead, their output is simply equated with the vaguely related construct of topicality (Günther, 2020). When novel topic modeling algorithms were introduced to treat social media data, development was driven not by social scientific insights about the topical organization of interactive social media discourse, but by data sparsity problems created by the need to process very short documents (Mehrtz et al., 2013). This incommensurability between validity criteria and CTAM development is even evident in many computational tools developed within the social sciences themselves. For example, few tools amid the wild growth of CTAM to measure frames elucidate how proposed algorithms *validly* operationalize the construct (Baden, 2010; David et al.,

2011; Nicholls & Culpepper, 2020; Papacharissi & de Fatima Oliveira, 2008; Walter & Ophir, 2019). Not even the broadest methodological concerns debated in the literature on social science quantitative text analysis (e.g., regarding unitization and the role of context; Baden, 2018; linguistic variability and polysemy; Boxman-Shabtai, 2020; van Atteveldt et al., 2021; or statistical classification biases; Geiss, 2021; Krippendorff, 2004) are well-reflected in computational tool development. Yet, all of these debates are directly relevant to computational text classification. Similar arguments have been made with regard to available knowledge in linguistics research, which likewise holds immense potential for informing more valid computational text classification (Bender, 2011).

Second, consequently, the validation of CTAM is typically externalized from the development process, and left to the stage of application (Budnitsky & Hirst, 2006). Only at this stage, researchers' knowledge of the investigated texts and meanings informs the augmentation of dictionaries (e.g., Muddiman et al., 2018); shapes the pre-processing of textual materials (Denny & Spirling, 2018); or serves as a benchmark for selecting one model among multiple that have been estimated (Nicholls & Culpepper, 2020). In this way, computational social scientists are slowly accumulating experience. They observe, for instance, that off-the-shelf sentiment dictionaries rarely work well beyond the text genres they were developed from (van Atteveldt et al., 2021); that lemmatizing tends to help focus topic models on textual qualities that more closely resemble valid topics (Günther, 2020), or that SVM classifiers usually outperform Naïve Bayes or Random Forest algorithms for English language classification (Stalpouskaya, 2020). However, any such considerations are applied on a case-by-case and on an a-theoretical basis, and do not feed back into the development of computational methods. Inversely, those scores that do, to some extent, feed back into development – notably, precision and recall – tend to neglect the role of confounding patterns, chance, and other issues that are well-understood in social scientific text analysis (e.g., Krippendorff, 2004) and offer little information about systematic classification biases that threaten the validity of measurement. Other metrics employed when no ground truth is available to benchmark against – e.g., agreement with other tools, robustness scores or classifier performance metrics such as AUC or the distinctness of topics – say little about the validity of machine-made decisions. By dissociating predictive performance from operational validity, these indicators fail to communicate researchers' concerns about measurement validity and systematic error and provide little incentive to CTAM developers to engage with existing text analytic knowledge in the social sciences.

Third, owing to the far-reaching absence of operational considerations in the development and documentation of CTAM, social scientists trying to make an informed choice about their use of computational methods frequently fail to find relevant guidance in the existing methodological literature (Grimmer & Stewart, 2013). Reflecting the described mode of development, documentation regularly omits reference to what known linguistic, discourse-practical or conceptual properties an algorithm attempts to model (for exceptions, see Baden, 2010; Liu & Zhang, 2012). There is little knowledge about what preprocessing stages are suitable for analyzing different kinds of discourse or tuning computational methods toward the detection of specific textual properties, or which tools are better suited for specific measurement tasks, and why (e.g., van Atteveldt et al., 2021). At the same time, seemingly arbitrary decisions in the data pre-processing stage can alter results in dramatic ways, as demonstrated by Denny and Spirling (2018). However, in place of discussions tying methodological choices to operational concerns, existing methodological debates offer mostly statistical reasoning (e.g., concatenate short documents, so as to avoid zero inflation; Mehrtra et al., 2013), metrics without transparent theoretical meaning (e.g., “lift” or “exclusivity” scores; Roberts et al., 2019), and unspecified rules of thumb (e.g., it usually helps to remove prepositions). Consequently, studies applying CTAM routinely try an entire range of procedures and model specifications, selected without much operational justification based on prior experience or most recent technological advances. Findings are reported for whichever model produced superior predictive accuracy or plausible interpretability, abusing the “researcher degrees of freedom” (Simmons et al., 2011) left behind by the underdeveloped



literature, in hope that such findings are the result of superior model fit and not of chance variation (Bakker et al., 2021). Even if specific methods have been shown to perform well in the past, this practice instills little confidence that the same choice is appropriate for a given project.

Taken together, these three disconnects create a situation where social science researchers may be well-advised to opt against CTAM for conducting their textual research. Especially where operational knowledge is important for achieving valid measurement, manual or rule-based computational tools permit a controlled modeling of key classification criteria (e.g., adding rules for context-based disambiguation; ascertaining grammatical relatedness). Alternatively, researchers may take a leap of faith that supervised or unsupervised CTAM will independently arrive at valid classifications, while suitable strategies for ex-post validation remain underdeveloped.

A quick glance at the present state of quantitative social science text analysis documents the pertinence of the laid-out concerns. Among those text-analytic studies published in top social science journals (see Online Appendix A: <https://osf.io/s7h8b> for details), manual and rule-based, researcher-controlled approaches by-far outstripped the popularity of more machine-reliant CTAM. Of all surveyed authors (see Online Appendix B: <https://osf.io/s7h8b> for details), 40% identified validity as a major concern, and another 35% as a minor concern, in the application of CTAM. Studies that relied on less researcher-controlled applications of CTAM reported redoubled efforts to demonstrate validity. Classification accuracy was assessed for virtually all uses of supervised CTAM. Even for the notoriously hard-to-validate unsupervised tools, validation efforts were reported in two thirds of all papers. At the same time, while validation efforts concerning the use of manual or rule-based classification strategies focused primarily on the operational construction of instruments, users of supervised and unsupervised CTAM almost exclusively relied on post-hoc validity demonstrations. For rule-based operational procedures, validation was frequently supported by references to prior validation efforts, experiential knowledge and methodological discussions. By contrast, preprocessing and modeling choices in (un)supervised CTAM were rarely discussed, and hardly ever justified by reference to prior methodological knowledge. Absent a methodological discourse that links algorithmic choices in CTAM to substantive operational needs and validity criteria, researchers are left to try out different model specifications without reporting or overseeing the consequences of the particular selected configurations. Taken together, we put forward that this first gap constitutes a major obstacle not only to the application of CTAM in social science research, but for the development of computational social science as such.

## Gap II: Specialization before integration

The second gap concerns the highly specialized methodological trajectories of development in CTAM. At present, most tools focus on the extraction of exactly one kind of textual meaning. For example, sentiment tools score overall the sentiment of texts (van Atteveldt et al., 2021), and dependency parsers classify specific phrases based on their grammatical roles (van Atteveldt et al., 2017). While some computational methods are sufficiently versatile to measure multiple textual contents of the same kind (e.g., references to different countries; Segev, 2014), few are capable of measuring meanings that are expressed in different textual form (e.g., references to named entities and framing, or document-level topicality and sequential patterns). Moreover, most available tools either focus on classifying entire documents (e.g., scoring their sentiment or ideological leaning) or recognize specific, localized contents within the text (e.g., identifying pronouns or references pre-defined issues). While focusing tools on specific data levels, representations and patterns makes sense from a CTAM development perspective, this raises important complications for the use of CTAM in social science text research. There, most textual measurement is aimed at reconstructing embedded meanings that arise from the specific arrangements of textual contents (Baden, 2018), with the information required to address specific research questions typically scattered unevenly across the text. It may involve specific localized expressions (e.g., Stalpouskaya, 2020) as well as longer narratives (e.g., Welbers et al., 2020), selective linguistic patterns (e.g., Tenenboim-Weinblatt & Baden, 2021) and broad, global themes (e.g.,

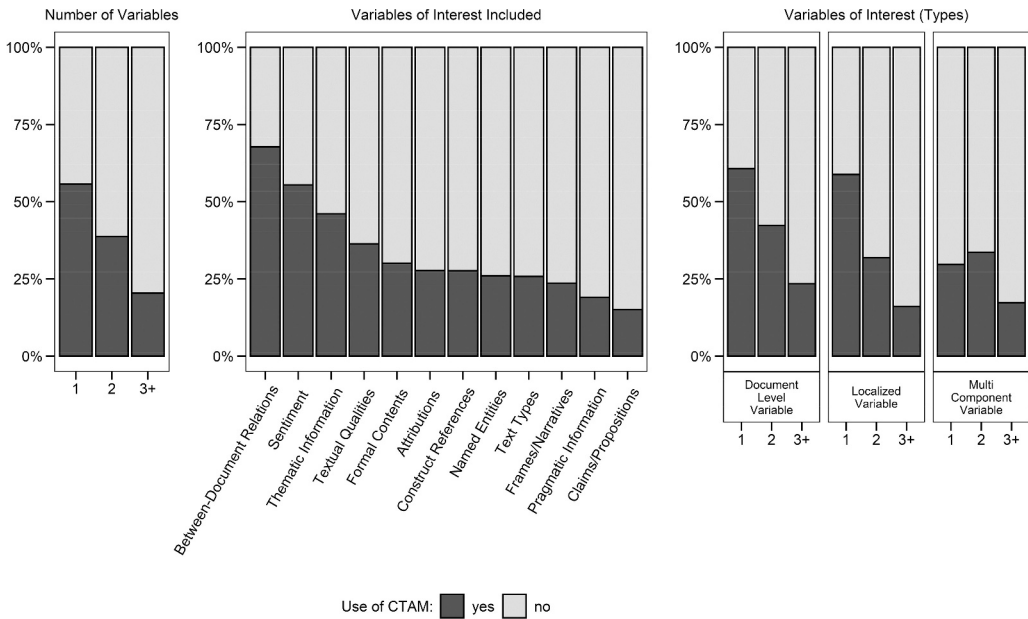
Burscher et al., 2014; Nelson, 2017), or any combination of these. For instance, social scientists are typically less interested in the “tone” of a text as a whole than in measuring the evaluative sentiment that is expressed by *specific* sources, with regard to *specific* topics (e.g., the economy), constructs (e.g., a policy) or entities (e.g., a country) referenced in the text (Liu & Zhang, 2012). Studies rarely stop at measuring whether an actor is mentioned, but continue to ask whether mentioned actors are blamed for certain things, how they are characterized, or whether they appear in certain roles (Fogel-Dror et al., 2018). In addition, many social scientific constructs are themselves modular, in the sense that they are composed of multiple components that are related to one another in specific ways (e.g., frames; Matthes & Kohring, 2008). On top of these challenges, many analyses seek to establish relations between different textual contents (Geiss, 2021). This not only necessitates the measurement of multiple constructs that may be expressed in quite different ways in the text (for a classic example, consider frames and object-specific evaluations; e.g., Lind et al., 2019). It may additionally require the extraction of specific semantic relationships expressed in the text (e.g., if blame for specific acts is attributed to different actors; van Atteveldt et al., 2017). Accordingly, the same social scientific text analysis frequently requires a multitude of different measurements, which may be easily combined in manual content analysis, but which for CTAM almost inevitably require the concatenation of different measurement procedures (Schoonvelde et al., 2019).

Yet, many available computational methods are conceptualized and built as standalone tools, using different coding languages and data standards and offering specialized interfaces that do not support concatenation within more complex pipelines (Liu & Zhang, 2012; Tenney et al., 2019). Accordingly, measured contents can often be put into relation with one another only *post hoc*, based on the observed correlation of occurrences within broad textual units (Geiss, 2021): Texts that mention a certain construct also tend to show positive sentiment. To extract specific relations expressed in the text, or operationalize complex, modular constructs, this strategy leaves much to be desired. Another strategy is to feed the entire, complex constellation of interrelated constructs coded by manual coders (e.g., of frame-embedded object evaluations) to powerful machine classifiers, hoping that these somehow learn to recognize relevant cases as holistic patterns. However, rising construct complexity (and the immense variability of indicative linguistic patterns) tends to severely diminish the accuracy of supervised classification, not to mention the increasing demands on large and carefully composed training sets.

Even where diverse algorithms are available on the same platform (as is increasingly the case for Python or R, such as the *quanteda* R package; Benoit et al., 2018), important challenges remain. Tools frequently make incommensurable assumptions regarding the processed material and require specific preprocessing steps, which, in turn, ask for different skill sets to implement and validate. More consequentially, tools rarely combine well. Neither the algorithms themselves, nor their accompanying methodological debates anticipate the possibility of interactions and concatenations between different analyses. One issue concerns the problem of multiplying errors – an issue that is particularly vexing as most errors incurred by CTAM are not random but systematic, degrading accuracy and amplifying possible biases. If we know little about how specific preprocessing algorithms affect the performance of subsequent topic models or supervised classifiers (Denny & Spirling, 2018), we know even less about the concatenation of analytic algorithms, such as the construction of semantic networks from topic models (e.g., Walter & Ophir, 2019), or the use of dictionary-identified features or PoS tags in machine classification (e.g., Stalpouskaya, 2020).

Owing to this second gap, the utility of CTAM for social science text research depends critically on the number and kinds of constructs to be measured in a corpus. Computational strategies have much to offer for the measurement of single, relatively well-defined constructs. Especially where contents can be classified by either searching for specific expressions in a text or by appraising entire documents, CTAM can add considerable value to the social science text analyst’s tool kit. To illustrate, Figure 1 shows that among those articles published in top social science journals, according to the OPTED content analysis, 56% of studies that measured exactly one kind of textual contents used some form of computational measurement. Figures are even higher for the measurement of simple constructs such





**Figure 1.** Use of CTAM by variables of interest. Note: Shares relative to all articles published in top-20 journals in communication science and political science that use quantitative text within each category. N varies considerably, from N = 300 (Thematic Information) to 27 (Inter-Document Relations).

as names or entities that can be recognized as localized instances in a text (59%, mostly using dictionaries or keyword-based analyses), and for the measurement of constructs that are reflected in the overall distribution of words over a document, such as sentiment or document types (61%, with topic models dominating thematic classification, while dictionaries and supervised CTAM are used for most other document-level classification tasks). However, the utility of CTAM diminishes rapidly as the complexity of textual measurements increases. For constructs that are neither localized in a text nor easily recognized at the document level – such as attributions, propositions, frames and narratives – CTAM use drops to 33%. Studies that aimed to measure two different kinds of textual content used CTAM only in 39% of all cases, and the share drops further to 20% for three or more kinds. Especially for the measurement of the most complex constructs such as frames (24%), pragmatic contents (19%) or propositions (15%), studies rarely resorted to CTAM. Contrary to claims that computational methods help researchers to reduce manual effort, researchers systematically opt against CTAM for more complicated, effortful classifications.

An anomaly, the still rather common use of CTAM – especially supervised and unsupervised approaches – for the measurement of frames points to another insight: The gap between specialized CTAM and the need for measuring complex constructs also raises the temptation to sacrifice measurement validity, so as to enable computational measurement without elaborate measurement pipelines. Especially in communication science, researchers' enthusiasm for computationally measuring frames while neglecting their constitutive internal structuring risks rendering frames operationally indistinct from themes (Baden, 2018; Günther, 2020). In a similar vein, also the widespread tendency to rely on textual sentiment to measure much more demanding constructs such as evaluative tone or even object-specific evaluations threatens not only the validity of obtained findings, but also erodes important conceptual distinctions (Boukes et al., 2020). Especially in conjunction with the first gap, the absence of an integrated methodological discourse on measurement validity in CTAM, computational methods' tendency to focus on exactly one kind of textual pattern at a time bears a tangible risk of prioritizing measurability over validity. In consequence, this second gap between complex operational demands and relatively simple operational capabilities presents a major limitation to the

application of CTAM in social scientific text research. Beyond the rapidly diminishing utility of CTAM whenever any required constructs cannot convincingly be measured algorithmically, the temptation of avoiding laborious manual analysis by resorting to crude computational approximations may hinder the conceptual development of social science textual research.

### Gap III: English before everything

The third gap concerns CTAMs' heavy focus on English language, and Germanic languages more generally, with researchers seeking to study other languages frequently finding the required tools lacking or non-existent. Over the past two decades, social science research has experienced a rapid internationalization. As the participation of non-anglophone researchers in international social science research has multiplied (Wilson & Knutsen, 2020), an increasing share of research published in leading journals investigates phenomena located in countries in which English is not the main language. Following the boom in internationally comparative survey research (e.g., Hanitzsch et al., 2019), moreover, also social scientific text analysis is increasingly conducted in an internationally, inter-lingually comparative fashion (e.g., Lind et al., 2019).

This rapid internationalization of research, however, is barely reflected in the development of CTAM. Owing to the considerable head start of English language computational development – both due to early U.S. dominance in computer technology, and due to the special role of English as scientific *lingua franca* – many resources, tools and experiences required for cutting-edge CTAM development are available only, or in far better quality, for English-language text. The same development has facilitated the growth of linguistic knowledge about natural discourse especially in English (Bender, 2011). Aiming to incrementally advance the state of the art, researchers – even in many non-anglophone countries – default to English in order to exploit the much richer knowledge, linguistic resources and tool box (e.g., van Atteveldt et al., 2017): Despite some recent efforts to translate or expand tools to include other languages (e.g., LIWC; NRC Emotion Lexicon) and the initiation of dedicated NLP development efforts in many language communities, English-language text analysis methods continues to advance at a pace that is unmatched by any other language. As the availability, sophistication and performance of tools in other languages continues to lag far behind available English-language tools, adequate algorithms are either unavailable or severely deficient for many analyses of non-English textual material.

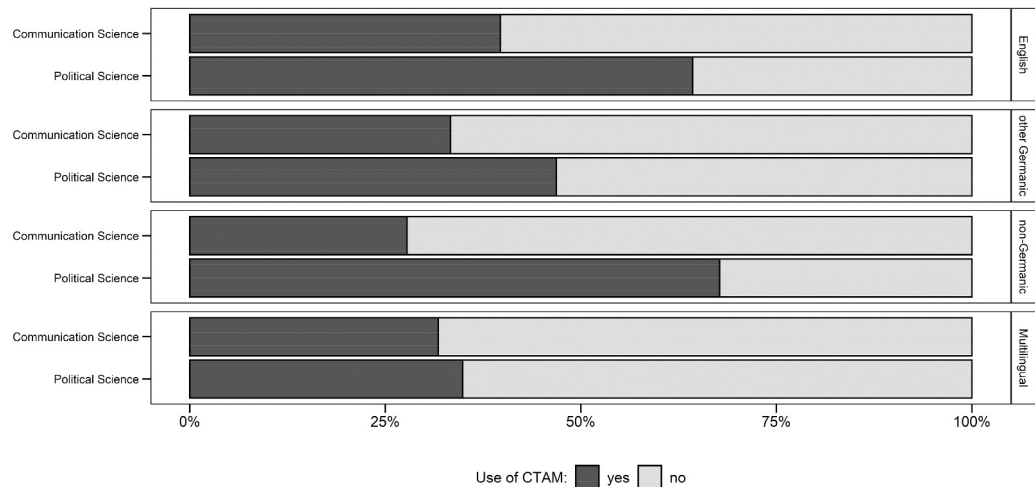
Even where adequate tools exist in multiple languages, furthermore, their measurement is rarely comparable across languages (Chan et al., 2020; Reber, 2019; however, see Proksch et al., 2019). In CTAM development, little attention is given to the consequential differences between languages (Bender, 2011). While many tools technically support their application to multiple languages, doing so often leads to incommensurable results owing to the hidden impact of language-specific differences (Chan et al., 2020; Maier et al., 2021). Moreover, as English serves as a global standard for computational tool development, its particular properties have in many ways become hard-coded into computational linguistic thinking, development and technology (Amram et al., 2018; Bender, 2011). The simple morphology of English verbs and nouns, as well as its tendency to allocate most grammatical functions to separate words have largely become naturalized in computational text analysis, and are hard to remove from existing technologies. Yet, the ubiquitous focus on space-delimited tokens as carriers of meaning raises important questions for morphologically richer languages (Goldberg & Elhadad, 2013). Similarly, English word order informs the perception that bi- and trigrams can serve to capture multi-token names and expressions, and present a reasonable approximation of word order. However, many languages follow looser word orderings or concatenate expressions in ways other than by adjacency. The more unlike English a language is, the less convincing are many assumptions that inform computational tool development.

Things get still more complicated yet where different scripts are involved. For example, *abjads* (vowel-less scripts; e.g., Arabic, Hebrew) typically include numerous homonyms that are disambiguated only by context, confusing token-based algorithms (Tsarfaty et al., 2013). Likewise, the use of

identical signs for syllables and words in logo-syllabic scripts (e.g., Chinese) violates common assumptions about the uniqueness and separation of linguistic tokens. As a consequence, computational methods designed for English often suffer severe performance losses or require major adjustments as they become translated. Trying to force different languages into the corset of English-like language structure, researchers confront many open questions and preprocessing needs (e.g., artificially tokenizing morphologically rich languages; Goldberg & Elhadad, 2013) that diminish the appeal of existing CTAM (Amram et al., 2018).

Owing to the vast and still widening gap between CTAM development in English and most other languages, accordingly, researchers studying resource-poorer languages face considerable difficulties matching the fast-evolving standards expected for (English-language) cutting-edge research. This imbalance appears, first of all, in a starkly asymmetric uptake of CTAM depending on the language studied in our analysis of published articles. If 64% of studies processing English language materials in political science, and 40% in communications, resort to computational methods, this share drops to 47% and 33%, respectively, whenever another Germanic language other than English was studied, as is shown in Figure 2. In communication, the share drops further to 28% for non-Germanic languages.<sup>3</sup> In political science, reliance on CTAM almost halves to 35% for studies processing multiple languages. Moreover, to the extent that languages unlike English are studied, the content analysis of leading publications shows that researchers primarily apply less sophisticated CTAM solutions – notably, tools for extracting formal contents such as links or hashtags, as well as simple keyword searches. Also seemingly agnostic, unsupervised tools such as topic models are still comparatively widespread in the study of non-English textual material, but mostly in research focused on one language only. Whenever adequate, comparable tools are unavailable for any included language, computational methods cease to offer a viable solution, and researchers default to manual classification.

Beyond the marked drop in reliance on CTAM for non-English textual analysis, the OPTED survey reveals another striking implication of the asymmetric development of computational methods. Even among the researchers who are not based in anglophone countries or do not have English as a native language, English-language textual materials dominate selected research sites and



**Figure 2.** Use of CTAM by language. Note: Shares relative to all articles using quantitative text within each category. N varies considerably (Communication Science: English: N = 282; Other Germanic: N = 102; Non-Germanic: N = 90; Multilingual: N = 87; Political Science: English: N = 140; Other Germanic: N = 32; Non-Germanic: N = 30; Multilingual: N = 44).

<sup>3</sup>One exception is Chinese, where a community of computational social science researchers is developing solutions that work for Mandarin. The high share of CTAM use on non-Germanic languages in political science also reflects the small number of studies falling into this category.

foci. To capitalize on the availability of superior CTAM technologies, researchers regularly decide against investigating other, less studied language communities, defaulting to the same few research sites studied already by their anglophone colleagues. Consequently, English language materials are heavily dominant in top ranking publications – 69% of all textual studies in political science, and still 59% in communication, study English language only. Studies on German and Dutch text are still common, and frequently rely on CTAM, benefitting from the relatively easy transferability of English-language tools into other Germanic languages. Other than these, some non-Germanic language communities maintain a small but steady presence in top-ranking social science journals – mostly studying text in Hebrew, Chinese, and Spanish – predominantly relying on manual methods. For most other language families, textual studies were rare, and computational studies were nearly or entirely absent.

The state of CTAM, thus, exacerbates existing imbalances in favor of English-speaking research, which already benefits from the easier accessibility of its research sites to global scientific communities; native language skills for scientific publishing; many leading journals' origin as U.S. or U.K. national journals; and many other factors. Similarly, the state of CTAM acts as a disincentive against multilingually comparative studies, which make up less than 16% of all top ranking studies on communication, and 18% in political science. Among the top-ranked journal publications, the inclusion of a second language decreased the use of computational methods from 36% to 28% in communication science, and from 64% to 41% in political science. Accordingly, the overwhelming focus on English-language CTAM development severely limits their utility for the study of most other languages, as well as the study of multilingual and inter-lingually comparative text corpora. Researchers are far more likely to find and apply suitable CTAM for studying English texts, privileging anglophone researchers and anglophone research sites. Where researchers intend to study textual material in other languages, by contrast, CTAM often require considerable efforts at adaptation, fall short in performance or are entirely unavailable, with painful consequences for the diversification and de-Westernization of social science research (Henrich et al., 2010).

## Discussion

In light of the presented three gaps, we argue that social science textual researchers' selective uptake of CTAM is not necessarily an outcome of limited computational literacy or related inhibitions, but may in many cases be motivated by sound reasoning. To the extent that available CTAM fail to adequately address pertinent concerns about operational validity, the measurement of multiple and internally complex constructs, and the needs of languages other than English, social science researchers may be well-advised to eschew the promises of computational tools and invest instead into carefully researcher-controlled, limited-scale manual studies. Arguably, these gaps are of concern primarily for scholars in communication and political science, who rely most heavily on quantitative text analysis.

At the same time, we have argued that existing CTAM have their applications, and are indeed taken up eagerly, as long as the available tools satisfy the needs of the analyst. As we have shown, CTAM has penetrated somewhat farther into political science text research than into communication. The price paid for this – or the enabling condition – appears to be a much heavier focus on English-only content. In communication science, CTAM is applied to a notably wider variety of measured variables and languages. Yet, in both disciplines, CTAM is applied selectively, weighing the availability and validity of suitable tools, regularly leading researchers to prefer manual classification for more complex and demanding measurement tasks. For studies that aim to measure few constructs, preferably of one kind, for whose recognition in English language text available CTAM offers plausible strategies, CTAM is widely applied. CTAM is even widely applied for measurements where at least some of these assumptions are in question – as is documented by the ready uptake of topic models despite their somewhat sketchy claim to validity (Günther, 2020), the enthusiastic development of sentiment or frame measurement tools that attempt to recognize complex contents based on much simpler patterns

(Baden, 2018; Overbeck et al., 2021), or the numerous cases where English-optimized CTAM were adapted or simply transferred to process materials in other languages and scripts (e.g., Chan et al., 2020; Tsarfaty et al., 2013). However, the farther an application strays from ideal conditions, the more sound objections can be raised that should – and as the OPTED content analysis and survey show, do – give social science researchers a pause when considering the use of CTAM.

Of course, we do not mean to suggest that computational illiteracy or skepticism play no role in the selective uptake of CTAM. Given the much stronger and longer-standing tradition of CTAM development in anglophone countries, researchers focusing on English language texts might be overall more computationally literate; the tools required to measure more complex textual contents tend to be more advanced, and thus less well-understood (Domahidi et al., 2019); and the most closely researcher-controlled computational tools tend to be also more established and thus familiar (Boumans & Trilling, 2016). Yet, each gap demonstrably contributes to explaining where and when social science researchers tend to use or eschew CTAM, and how these tools are applied when they are chosen. Researchers' concern about validity, as well as their elaborate efforts at somehow constructing a narrative that inspires confidence in the validity of achieved measurements, underscores the pertinence of our first gap as well as the need for an integrated, operationalization-focused methodological discourse on CTAM. Researchers clearly recognize CTAMs' superior ability to measure simple, localized or document-level variables. By comparison, they struggle visibly when aiming to extract more complex constructs (e.g., Liu & Zhang, 2012) – a fact also documented by the variegated efforts to enable CTAM to measure frame-like patterns, typically at the cost of cutting one or another corner (Baden, 2010; David et al., 2011; Walter & Ophir, 2019).

In addition, researchers' tendency to either default to English-language research sites or to retreat to manual classification demonstrates the need for more language-aware, language-inclusive, and interlinguistically comparable computational measures (Bender, 2011). Of course, the disproportionate emphasis on English-language text may also have to do with the legacy of many leading journals, which have developed out of domestic journals in the U.S. and U.K. and still double as preferred outlets for domestically-oriented research. Research on other language communities may face additional hurdles breaking into the English- and Western-dominated flagship journals, and may be more commonly published in domestically oriented journals elsewhere (Wilson & Knutsen, 2020). Yet, English-language materials dominate even beyond the share of authors based in anglophone countries, revealing researchers' need to focus on English-language cases if they wish to benefit from available computational tools.

Crucially, selective uses of CTAM consistent with our three gaps were evident even in top-ranking published research and a survey of leading authors in the field. Arguably, these cases are least likely to be explained by a lack of the resources, knowledge and computational skills required to implement any textual research methods deemed optimal for a given purpose. In addition, these authors' methodological choices had to pass the strictest peer-review processes, and thus deserve to be considered current best practices in the field. In this view, it seems to be that especially computationally literate researchers regularly decide against using CTAM for reasons related to our three gaps.

At the same time, at least some of the described challenges require some degree of computational literacy to become at all apparent. Especially the concerns raised in our first gap presume a certain awareness of algorithmic modeling procedures and computational logic, and might be less relevant among less computationally literate researchers. That said, the scarcity of methodological guidance presents additional hurdles for researchers seeking to acquire the requisite skills. Likewise, the absence of suitable tools for measuring many constructs or processing most languages not only explains computationally literate researchers' hesitation at applying CTAM in many cases, but also presents a potential disincentive especially for non-Western researchers entering the field of computational social science. On the one hand, thus, the data used to illustrate the impacts of our three gaps are not necessarily representative of the field as a whole, owing to a range of publication biases that not only reward superior (access to) skills and resources, but also tend to privilege research on well-established

subjects, Western research sites and English-language materials. Yet, on the other hand, even beyond the present state of the art of social science quantitative text research, each gap raises concerns that obstruct or threaten the development of the field as a whole.

### ***Implications for social science text research***

To the extent that cutting-edge computational tools are becoming normalized in social scientific text research, our three gaps point at several important imbalances in the suitability and performance of CTAM. While CTAM may perform reasonably well under specific conditions – notably, for measuring single, operationally well-defined constructs, which can be recognized in contiguous expressions or as document-level properties, in English-language text – any other uses raise important challenges. Notably, the unequal applicability to different languages puts non-anglophone researchers at a structural disadvantage. The comparative inadequacy of non-English language tools diminishes the chances of research on other languages at scoring top-ranking journal publications and adds to the existing biases responsible for the continuing under-representation especially of non-Western scholars (Wasserman, 2020). Moreover, the superior toolbox available for studying Germanic-language contents disincentivizes researchers from other language communities from entering the realm of computational social sciences, actively counteracting present efforts at diversifying textual research beyond the present dominance of WEIRD (Western, Educated, Industrialized, Rich & Developed) nations (Henrich et al., 2010).

Similarly, the disproportionate efforts and unavailable methodological knowledge required to computationally operationalize social scientific constructs incentivizes textual researchers to dilute conceptual standards. To the extent that textual sentiments are passed off as evaluations, or frames are operationalized in ways indistinct from themes, topics, or issues, the progressing application of CTAM threatens to blur important theoretical distinctions (Baden, 2018; Overbeck et al., 2021). Moreover, development efforts and methodological knowledge are distributed unevenly over different computational approaches and types of measurement: Some methods are widely applied and tested for specific uses in a given field or discipline, while other avenues remain underexplored. Researchers may thus be incentivized to rely on well-documented, popular tools where different, but less accessible approaches might offer superior measurement. The prevalent mode of CTAM development outside the social sciences, without much concern for researchers' operational demands, privileges the study of phenomena for simply being measurable, compromising text researchers' control of the research agenda. Without the capacity to translate available algorithms and tools into intelligible models of textual meanings and operational capabilities, social scientists face considerable hurdles for making optimal use of available computational technologies (Domahidi et al., 2019). Inversely, without a recognition of the knowledge generated by decades of social scientific, linguistic and other text-based research, computational tool developers are likely to miss or inadequately model important textual properties, and are bound to laboriously reinvent the wheel through trial and error. The persistent disconnect between social scientific, validity-focused methodological debates, and the prevalent mode of technologically driven development – especially, but not solely in the computational sciences – deprives both developers and users of CTAM of valuable and urgently needed synergies.

### ***Research Agenda***

That said, there are also several encouraging insights to be gleaned from this appraisal of the present state of computational text analysis in the social sciences. For one, building on researchers' growing experience from application-based validation efforts, we perceive a redoubled commitment to systematic validation. Social science methodologists increasingly debate the operational implications of various computational procedures in light of existing methodological knowledge in textual research (e.g., Maier et al., 2020; van Atteveldt et al., 2021). Research teams in both social science and AI are beginning to concatenate tools and methods in order to boost the validity and nuance of algorithmic



extraction (e.g., Human-In-The-Loop approaches/Active learning Munro, 2020); and there are several notable efforts at developing language-specific computational tools (e.g., Tsarfaty et al., 2013), as well as a capacity for inter-lingual comparative validation and research (e.g., Chan et al., 2020). In the following, we wish to sketch several ideas for a future research agenda, which is suitable to narrow the presented gaps and address their problematic implications for social scientific research. While we discuss possible responses to each gap in turn, our points are closely interrelated, and each step toward bridging one gap also provides important building blocks for addressing the other ones.

In order to address the disconnect between methodological discourses in the social sciences and CTAM development, one obvious desiderate concerns the intensification of existing, mutual communication efforts (van Attevelde & Peng, 2018; Mulder et al., 2021). However, to communicate effectively, some shifts in perspective may be useful. One much needed shift concerns the widespread tendency (especially in the computational sciences) to view social scientists primarily as *users* of computational tools. This perception overlooks social scientists' (and linguists') rich knowledge about textual discourse and language use, including numerous methodological challenges in text analysis. Social scientists can – and should – teach their computational colleagues many a thing about issues that can (and should) be considered – and if possible, modeled – in the design of CTAM: Pertinent examples range from the topical organization of discursive genres (Günther, 2020), to the nonrandom use of evasive or figurative speech (Muddiman et al., 2018), to the challenges of polysemy, embedded assumptions and implicatures (Baden, 2018), to matters of measurement bias and the blind spots of accuracy metrics in validation (Krippendorff, 2004). By relying on existing experience in social scientific text research, developers may not only create much more nuanced, valid and useful tools, which anticipate users' need to tweak algorithms to accommodate the requirements of different measured constructs or textual genres – they can also skip a few detours that social scientists have long explored. Inversely, social scientists would do well to abandon their perspective as CTAM users themselves and participate, if not in CTAM development, then at least in the accumulation of methodological knowledge and guidance, as well as the strategic improvement of available tools (Chan et al., 2020). One key step here is to shift emphasis in validation from demonstrating *that* a computational method adequately performs certain tasks (predictive accuracy), to discussing exactly *how* and *how convincingly* it operationalizes relevant conceptual properties (measurement validity; Liu & Zhang, 2012; Stalpouskaya, 2020). In the same vein, social science methodologists are well-placed to scrutinize the differential implications of computational preprocessing and modeling choices (Denny & Spirling, 2018; Günther, 2020; Schoonvelde et al., 2019). Analyzing misclassification patterns or identifying the sources of observed classification biases, social scientists can generate valuable knowledge to inform the improvement of available CTAM.

In order to address the disconnect between the fragmented development of computational tools, and social scientists' need for modular, interoperable and integrated measurement instruments, much credit goes to the handful existing efforts at gathering diverse capabilities in unified text analysis platforms and packages (notably, Benoit et al., 2018; Trilling et al., 2018). Also the “re-discovery” of algorithmic pipelines in CTAM is doubtlessly a step in the right direction (Liu & Zhang, 2012; Tenney et al., 2019). However, considerable additional work is needed to understand the complex interdependencies that arise from the sequential concatenation, parallel combination and hierarchical nesting of different computational tools. Beyond the obvious question how different pre-processing steps alter subsequent computational procedures, especially the algorithmic modeling of textual meaning, as well as the specific interactions and incompatibilities that arise from embedded assumptions deserve attention (Baden et al., 2020). In addition, offering explicit conceptual definitions and operational models of textual meanings and semantic relatedness, and initiating a systematic exchange between social sciences researchers and CTAM developers over the operational requirements of more demanding measurement tasks can contribute valuably to this research agenda. Especially research in linguistics and discourse studies offers a rich vocabulary for adding precision to operational choices, enabling their better translation into algorithmic procedures.

In order to address the disconnect between English-focused CTAM development and the progressing internationalization and comparative orientation of social scientific research, ongoing developments in the field of computational linguistics appear to lead the way toward a possible response. Beyond the reinforcement of existing efforts at developing tools for less well-resourced languages (e.g., the dedication of specific fora to resource-poorer languages), one key step concerns raising the visibility of non-English computational text research and the challenges that arise from such undertakings. Following an important intervention by Emily Bender (2011), raising awareness especially among English-language tool developers that different languages work differently may be instrumental to facilitating cross-lingual cooperation and comparison. Social scientists and computational tool developers alike need to reflect upon those properties of languages under investigation (e.g., morphology, word order; Moravcsik, 2013) that align or deviate from tools' (typically implicit) modeling assumptions. To mitigate the inherent publication bias in favor of English language textual research, which benefits from superior computational tools and resources, editors, reviewers and also authors need to acknowledge the value and specific challenges of advancing and applying CTAM in other languages. Especially in comparative textual research, explicitly addressing linguistic differences that impact the performance of computational tools will be instrumental not only for discriminating between meaningful and artifactual differences in the analysis (Chan et al., 2020; Lind et al., 2019), but also to systematically expose and address these issues in future development (Bender, 2011). In close collaboration with ongoing efforts in computational linguistics, both social scientists and computational tool developers can work toward a next generation of CTAM that are transparent, tweakable, and sensitive to language-specific differences, so as to enable valid comparative research across different genres and languages.

## Disclosure statement

No potential conflict of interest was reported by the author(s).

## Funding

This work was supported by the Horizon 2020 Framework Programme [951832].

## ORCID

Christian Baden  <http://orcid.org/0000-0002-3771-3413>

Christian Pipal  <http://orcid.org/0000-0002-5395-2035>

Martijn Schoonvelde  <http://orcid.org/0000-0003-4370-2654>

Mariken A. C. G van der Velden  <http://orcid.org/0000-0003-0227-9183>

## References

- Amram, A., Ben David, A., & Tsarfaty, R. (2018). Representations and architectures in neural sentiment analysis for morphologically rich languages: A case study from modern Hebrew. *Proceedings of the 27th International Conference on Computational Linguistics*, Santa Fe, NM, 2242–2252. ACL. <https://aclanthology.org/C18-1190>
- Baden, C., Kligler-Vilenchik, N., & 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. (2010). *Communication, contextualization, & cognition: Patterns & processes of frames' influence on people's interpretations of the EU constitution*. Eburon.
- Baden, C. (2018). Reconstructing frames from intertextual news discourse: A semantic network approach to news framing analysis. In P. D'Angelo (Ed.), *Doing news framing analysis ii: Empirical and theoretical perspectives* (pp. 3–26). Routledge.

- Bakker, B. N., Jaidka, K., Dörr, T., Fasching, N., & Lelkes, Y. (2021). Questionable and open research practices: Attitudes and perceptions among quantitative communication researchers. *Journal of Communication*, 71(5), 715–738. <https://doi.org/10.1093/joc/jqab031>
- Bender, E. M. (2011). On achieving and evaluating language-Independence in NLP. *Linguistic Issues in Language Technology*, 6(3), 1–26. <https://doi.org/10.33011/lilt.v6i.1239>
- Benoit, K., Watanabe, K., Wang, H., Nulty, P., Obeng, A., Müller, S., & Matsuo, A. (2018). Quanteda: An R package for the quantitative analysis of textual data. *Journal of Open Source Software*, 3(30), 774. <https://doi.org/10.21105/joss.00774>
- Blei, D. M., & Lafferty, J. D. (2006). Dynamic topic models. *Proceedings of the 23rd International Conference on Machine Learning*, Pittsburgh, PA, 113–120. ACM. <https://doi.org/10.1145/1143844.1143859>
- Boukes, M., van de Velde, B., Araujo, T., & Vliegthart, R. (2020). What's the tone? easy doesn't do it: Analyzing performance and agreement between off-the-shelf sentiment analysis tools. *Communication Methods and Measures*, 14(2), 83–104. <https://doi.org/10.1080/19312458.2019.1671966>
- Boumans, J. W., & Trilling, D. (2016). Taking stock of the toolkit. *Digital Journalism*, 4(1), 8–23. <https://doi.org/10.1080/21670811.2015.1096598>
- Boxman-Shabtai, L. (2020). Meaning multiplicity across communication subfields: Bridging the gaps. *Journal of Communication*, 70(3), 401–423. <https://doi.org/10.1093/joc/jqaa008>
- Brady, H. E. (2019). The challenge of big data and data science. *Annual Review of Political Science*, 22(1), 297–323. <https://doi.org/10.1146/annurev-polisci-090216-023229>
- Budanitsky, A., & Hirst, G. (2006). Evaluating Wordnet-based measures of lexical semantic relatedness. *Computational Linguistics*, 32(1), 13–47. <https://doi.org/10.1162/coli.2006.32.1.13>
- Burscher, B., Odijk, D., Vliegthart, R., de Rijke, M., & de Vreese, C. H. (2014). Teaching the computer to code frames in the news: Comparing two supervised machine learning approaches to frame analysis. *Communication Methods and Measures*, 8(3), 190–206. <https://doi.org/10.1080/19312458.2014.937527>
- Chan, C.-H., Zeng, J., Wessler, H., Jungblut, M., Welbers, K., Bajjalieh, J. W., van Atteveldt, W., & Althaus, S. L. (2020). Reproducible extraction of cross-lingual topics (rectr). *Communication Methods and Measures*, 14(4), 285–305. <https://doi.org/10.1080/19312458.2020.1812555>
- Choi, S., Shin, H., & Kang, -S.-S. (2021). Predicting audience-rated news quality: Using survey, text mining, and neural network methods. *Digital Journalism*, 9(1), 84–105. <https://doi.org/10.1080/21670811.2020.1842777>
- Conway, M. (2006). The subjective precision of computers: A methodological comparison with human coding in content analysis. *Journalism & Mass Communication Quarterly*, 83(1), 186–200. <https://doi.org/10.1177/107769900608300112>
- David, C. C., Atun, J. M., Fille, E., & Monterola, C. (2011). Finding frames: Comparing two methods of frame analysis. *Communication Methods and Measures*, 5(4), 329–351. <https://doi.org/10.1080/19312458.2011.624873>
- 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. <https://doi.org/10.1017/pan.2017.44>
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of NAACL-HLT*, Minneapolis, MN, 4171–4186. dblp. <https://doi.org/10.18653/v1/n19-1423>
- Domahidi, E., Yang, J., Niemann-Lenz, J., & Reinecke, L. (2019). Outlining the way ahead in computational communication science: An introduction to the IJOC special section on “Computational methods for communication science: Toward a strategic roadmap.” *International Journal of Communication*, 13, 1–9. <https://ijoc.org/index.php/ijoc/article/view/10533>
- Fogel-Dror, Y., Shenhav, S. R., Sheaffer, T., & van Atteveldt, W. (2018). Role-based association of verbs, actions, and sentiments with entities in political discourse. *Communication Methods and Measures*, 13(2), 69–82. <https://doi.org/10.1080/19312458.2018.1536973>
- Geiss, S. (2021). Statistical power in content analysis designs: How effect size, sample size and coding accuracy jointly affect hypothesis testing – A Monte Carlo simulation approach. *Computational Communication Research*, 3(1), 61–89. <https://doi.org/10.5117/CCR2021.1.003.GEIS>
- Goldberg, Y., & Elhadad, M. (2013). Word segmentation, unknown-word resolution, and morphological agreement in a Hebrew parsing system. *Computational Linguistics*, 39(1), 121–160. [https://doi.org/10.1162/COLI\\_a\\_00137](https://doi.org/10.1162/COLI_a_00137)
- 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>
- Günther, E., & Quandt, T. (2016). Word counts and topic models. *Digital Journalism*, 4(1), 75–88. <https://doi.org/10.1080/21670811.2015.1093270>
- Günther, E. (2020). *Topic Modeling: Theoretische Einordnung algorithmischer Themenkonzepte in Gegenstand und Methodik der Kommunikationswissenschaft* [Unpublished doctoral dissertation]. WWU Münster.
- Hanitzsch, T., Hanusch, F., Ramaprasad, J., & de Beer, A. S. (2019). *Worlds of journalism: Journalistic cultures around the globe*. Columbia University Press.
- Henrich, J., Heine, S. J., & Norenzayan, A. (2010). Most people are not weird. *Nature*, 466(7302), 29–29. <https://doi.org/10.1038/466029a>

- Hilbert, M., Barnett, G., Blumenstock, J., Contractor, N., Diesner, J., Frey, S., González-Bailón, S., Lamberson, P. J., Pan, J., Peng, T.-Q., Shen, C., Smaldino, P. E., van Atteveldt, W., Walldherr, A., Zhang, J., Zhu, J. H. et al. (2019). Computational communication science: A methodological catalyzer for a maturing discipline. *International Journal of Communication*, 13, 1–23. <https://ijoc.org/index.php/ijoc/article/view/10675>
- Hirst, G., Riabinin, Y., Graham, J., Boizot-Roche, M., & Morris, C. (2014). Text to ideology or text to party status? In B. Kaal, I. Maks, & A. van Elfrinkhof (Eds.), *From text to political positions: Text analysis across disciplines* (pp. 93–115). John Benjamins.
- Kantner, C., & Overbeck, M. (2020). Exploring soft concepts with hard corpus-analytic methods. In N. Reiter, A. Pichler, & J. Kuhn (Eds.), *Reflektierte algorithmische Textanalyse. Interdisziplinäre(s) Arbeiten in der CRETA-Werkstatt* (pp. 169–189). De Gruyter.
- Krippendorff, K. (2004). *Content analysis: An introduction to its methodology*. Sage.
- Lazer, D., & Radford, J. (2017). Data ex machina: Introduction to big data. *Annual Review of Sociology*, 43(1), 19–39. <https://doi.org/10.1146/annurev-soc-060116-053457>
- Lind, F., Eberl, J.-M., Heidenreich, T., & Boomgaarden, H. G. (2019). When the journey is as important as the goal: A roadmap to multilingual dictionary construction. *International Journal of Communication*, 13. <https://ijoc.org/index.php/ijoc/article/view/10578>
- Liu, B., & Zhang, L. (2012). A survey of opinion mining and sentiment analysis. In C. C. Aggarwal & C.-X. Zhai (Eds.), *Mining text data* (pp. 415–463). Springer.
- Maier, D., Baden, C., Stoltenberg, D., De Vries, M., & Walldherr, A. (2021). Machine translation v. multilingual dictionaries: Assessing two strategies for the topic modeling of multilingual text collections. *Communication Methods and Measures* [Advance Online Publication], 1–20. <https://doi.org/10.1080/19312458.2021.1955845>
- Maier, D., Niekler, A., Wiedemann, G., & Stoltenberg, D. (2020). How document sampling and vocabulary pruning affect the results of topic models. *Computational Communication Research*, 2(2), 139–152. <https://doi.org/10.5117/CCR2020.2.001.MAIE>
- Maier, D., Walldherr, A., Miltner, P., Wiedemann, G., Niekler, A., Keinert, A., Pfetsch, B., Heyer, G., Reber, U., Häussler, T., Schmid-Petri, H., & Adam, S. (2018). Applying LDA topic modeling in communication research: Toward a valid and reliable methodology. *Communication Methods and Measures*, 12(2–3), 93–118. <https://doi.org/10.1080/19312458.2018.1430754>
- Matthes, J., & Kohring, M. (2008). The content analysis of media frames: Toward improving reliability and validity. *Journal of Communication*, 58(2), 258–279. <https://doi.org/10.1111/j.1460-2466.2008.00384.x>
- Mehrtra, R., Sanner, S., Buntine, W., & Xie, L. (2013). Improving LDA topic models for microblogs via tweet pooling and automatic labeling. *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, Dublin, Ireland, 889–892. ACM. <https://doi.org/10.1145/2484028.2484166>
- Moravcsik, E. A. (2013). *Introducing language typology*. Cambridge University Press.
- Muddiman, A., McGregor, S., & Stroud, N. J. (2018). (Re)claiming our expertise: Parsing large text corpora with manually validated and organic dictionaries. *Political Communication*, 36(2), 214–226. <https://doi.org/10.1080/10584609.2018.1517843>
- Mulder, M., Inel, O., Oosterman, J., & Tintarev, N. (2021). Operationalizing framing to support multiperspective recommendations of opinion pieces. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency Canada* [Virtual Event], 478–489. ACL. <https://doi.org/10.1145/3442188.3445911>
- Munro, R. (2020). *Human-in-the-loop machine learning: Active learning, annotation, and human-computer interaction*. Manning.
- Nelson, L. K. (2017). Computational grounded theory: A methodological framework. *Sociological Methods & Research*, 49(1), 3–42. <https://doi.org/10.1177/0049124117729703>
- Nicholls, T., & Culpepper, P. D. (2020). Computational identification of media frames: Strengths, weaknesses, and opportunities. *Political Communication*, 38(1), 159–181. <https://doi.org/10.1080/10584609.2020.1812777>
- Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, 2(2), 175–220. <https://doi.org/10.1037/1089-2680.2.2.175>
- Overbeck, M., Baden, C., Aharoni, T., & Tenenboim-Weinblatt, K. (2021). Beyond sentiment: An algorithmic strategy for identifying evaluations within large text corpora [Conference Paper]. *71st ICA Annual Conference*, Denver, CO [Conference Paper]. ICA.
- Papacharissi, Z., & de Fatima Oliveira, M. (2008). News frames terrorism: A comparative analysis of frames employed in terrorism coverage in US and UK newspapers. *The International Journal of Press/Politics*, 13(1), 52–74. <https://doi.org/10.1177/1940161207312676>
- Pipal, C., Schoonvelde, M., & Schumacher, G. (2021). *Taking context seriously: Joint estimation of sentiment and topics in textual data*. OSF Preprints. <https://doi.org/10.31219/osf.io/mt3cs>
- 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>

- Ramage, D., Hall, D., Nallapati, R. M., & Manning, C. (2009). Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, Singapore, 248–256. ACL. <https://aclanthology.org/D09-1026>
- 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>
- Roberts, M. E., Stewart, B. M., & Tingley, D. (2019). Stm: An R package for structural topic models. *Journal of Statistical Software*, 91(2), 1–40. <https://doi.org/10.18637/jss.v091.i02>
- Roberts, M. E. (2016). Introduction to the virtual issue: Recent innovations in text analysis for social science. *Political Analysis*, 24(V10), 1–5. <https://doi.org/10.1017/S1047198700014418>
- Schoonvelde, M., Schumacher, G., & Bakker, B. N. (2019). Friends with text as data benefits: Assessing and extending the use of automated text analysis in political science and political psychology. *Journal of Social & Political Psychology*, 7(1), 124–143. <https://doi.org/10.5964/jspp.v7i1.964>
- Segev, E. (2014). Visible and invisible countries: News flow theory revised. *Journalism*, 16(3), 412–428. <https://doi.org/10.1177/1464884914521579>
- Shermer, M. (2000). *How we believe: Science, skepticism, and the search for God*. Henry Holt.
- Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, 22(11), 1359–1366. <https://doi.org/10.1177/0956797611417632>
- Stalpouskaya, K. (2020). *Automatic extraction of agendas for action from news coverage of violent conflict* [Doctoral dissertation], Ludwig Maximilian University Munich. <https://edoc.ub.uni-muenchen.de/25807/>
- Tenenboim-Weinblatt, K., & Baden, C. (2021). Gendered communication styles in the news: An algorithmic comparative study of conflict coverage. *Communication Research*, 48(2), 233–256. <https://doi.org/10.1177/0093650218815383>
- Tenney, I., Das, D., & Pavlick, E. (2019). Bert rediscovers the classical NLP pipeline. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy, 4593–4601. ACL.
- Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., & Kappas, A. (2010). Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology*, 61(12), 2544–2558. <https://doi.org/10.1002/asi.21416>
- Theocharis, Y., & Jungherr, A. (2020). Computational social science and the study of political communication. *Political Communication*, 38(1–2), 1–22. <https://doi.org/10.1080/10584609.2020.1833121>
- Trilling, D., Van De Velde, B., Kroon, A. C., Löcherbach, F., Araujo, T., Strycharz, J., Raats, T., De Klerk, L., Jonkman, J. G. F. et al. (2018). INCA: Infrastructure for content analysis. *2018 IEEE 14th International Conference on e-Science*, Amsterdam, Netherlands(, 329–330. IEEE. <https://doi.org/10.1109/eScience.2018.00078>
- Tsarfaty, R., Seddah, D., Kübler, S., & Nivre, J. (2013). Parsing morphologically rich languages: Introduction to the special issue. *Computational Linguistics*, 39(1), 15–22. [https://doi.org/10.1162/COLI\\_a\\_00133](https://doi.org/10.1162/COLI_a_00133)
- van Atteveldt, W., & Peng, T.-Q. (2018). When communication meets computation: Opportunities, challenges, and pitfalls in computational communication science. *Communication Methods & Measures*, 12(2–3), 81–92. <https://doi.org/10.1080/19312458.2018.1458084>
- van Atteveldt, W., Sheaffer, T., Shenhav, S., & Fogel-Dror, Y. (2017). Clause analysis: Using syntactic information to automatically extract source, subject, and predicate from texts with an application to the 2008–2009 Gaza war. *Political Analysis*, 25(2), 207–222. <https://doi.org/10.1017/pan.2016.12>
- van Atteveldt, W., van der Velden, M. A., & Boukes, M. (2021). The validity of sentiment analysis: Comparing manual annotation, crowd-coding, dictionary approaches, and machine learning algorithms. *Communication Methods and Measures*, 15(2), 121–140. <https://doi.org/10.1080/19312458.2020.1869198>
- van Atteveldt, W., Welbers, K., & van der Velden, M. (2019). Studying political decision making with automatic text analysis. *Oxford Research Encyclopedia of Politics*, 1–11. <https://doi.org/10.1093/acrefore/9780190228637.013.957>
- van Dijk, T. A. (1985). Structures of news in the press. In T. A. van Dijk (Ed.), *Discourse and communication* (pp. 69–93). De Gruyter.
- Walter, D., & Ophir, Y. (2019). News frame analysis: An inductive mixed-method computational approach. *Communication Methods and Measures*, 13(4), 248–266. <https://doi.org/10.1080/19312458.2019.1639145>
- Wasserman, H. (2020). Moving from diversity to transformation in communication scholarship. *Annals of the International Communication Association*, 44(1), 1–3. <https://doi.org/10.1080/23808985.2019.1706429>
- Weber, R., Mangus, J. M., Huskey, R., Hopp, F. R., Amir, O., Swanson, R., Gordon, A., Khooshabeh, P., Hahn, L., & Tamborini, R. (2018). Extracting latent moral information from text narratives: Relevance, challenges, and solutions. *Communication Methods and Measures*, 12(2–3), 119–139. <https://doi.org/10.1080/19312458.2018.1447656>
- Welbers, K., van Atteveldt, W., Althaus, S., Wessler, H., Bajjalieh, J., Chan, C.-H., & Jungblut, M. (2020). *Media portrayal of terrorist events: Using computational text analysis to link news items to the global terrorism database*. 70th ICA Annual Conference, Gold Coast, Australia [Virtual Event]. ICA.
- Wilson, M. C., & Knutsen, C. H. (2020). Geographical coverage in political science research. *Perspectives on Politics* [Advance Online Publication], 1–16. <https://doi.org/10.1017/S1537592720002509>