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

Abstract

We identify three gaps that hamper the utility and progress of computational text analysis methods (CTAM) for social science research. First, CTAM development has given insufficient attention to social scientists' concerns about measurement validity. Second, we identify a mismatch between the focus of many computational tools and that of social science research. Third, we argue that the dominance of English language tools depresses comparative research and inclusivity towards scholarly communities examining languages other than English. To substantiate our claims, we draw upon a content analysis of all research published in the top ranked journals in communications, political science, sociology, and psychology. Identifying a total of 854 articles between 2016 and 2020 that use quantitative text analysis, we examined studies' reliance on CTAM, what variables were measured, what validation efforts were undertaken, and what languages were present in the studied materials. We show that each gap contributes to explaining the uneven uptake of CTAM, and discuss how each gap has implications for research practice in the social sciences. In order to address these gaps, we propose a research agenda for CTAM development.

Keywords: Computational Methods, Validation, Integration, Multilingualism, Research Agenda

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

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, Welbers, & van der Velden, 2019). One key driver of this computational turn has to do with the increased availability of data, evidenced by digitally available repositories for textual data (Possler, Bruns, & Niemann-Lenz, 2019), such as ParlSpeech (Rauh & Schwalbach, 2020) or EUSpeech (Schumacher et al., 2020). These large scale databases allow researchers to examine their research questions using high-resolution quantitative evidence, provided that there are adequate computational tools available to process the data.

In step with the expansion of available data, also the accessibility and capabilities of analytic software have 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 considerable momentum. Examples range from versatile text analysis platforms such as AmCAT (van Atteveldt, Ruigrok, Takens, & Jacobi, 2014) or INCA (Trilling et al., 2018), to dedicated tool collections such as the `quanteda` package (Benoit et al., 2018) or the `NLTK` modules (Bird, Klein, & Loper, 2009). 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 leading journals, with several recent special issues specifically dedicated to CTAM in social research[1]. Yet,

[1] e.g., M. E. Roberts (2016); Theocharis and Jungherr (2020); van Atteveldt and Peng (2018)

available tools are taken up unevenly. While some algorithms – such as SentiStrength (Thelwall, Buckley, Paltoglou, Cai, & Kappas, 2010) or topic models (Blei & Lafferty, 2006) – are rapidly and widely adopted across social science scholarship (Günther & Quandt, 2016), many – especially, high-powered algorithms, such as Neural Network classifiers (e.g., Choi, Shin, & Kang, 2021) – remain a rare sight. In the existing scholarship, this uneven uptake of CTAM is typically explained by reference to the rapid pace of development in the computational sciences (Boumans & Trilling, 2016), as well as social scientists' often limited computational literacy (Domahidi, Yang, Niemann-Lenz, & Reinecke, 2019). In this view, training social scientists in the use of CTAM should enable them to take full advantage of the available tool box. We agree that important potentials remain to be unlocked by advancing social scientists' computational literacy. However, in this paper, we maintain that social scientists often have good reason to forego available computational solutions and to prefer manual strategies, despite their often considerable required effort.

In particular, we argue that there are 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 (Nicholls & Culpepper, 2020; Yarchi, Baden, & Kligler-Vilenchik, 2020) and measurement validity (see, for example Boukes, van de Velde, Araujo, & Vliegenthart, 2020). Where researchers identify a mismatch between CTAM capabilities and the specific needs of valid measurement, they may be well-advised to prefer methods and tools 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 – typically extracted at the document level (e.g., topicality, sentiment) or located in specific, contiguous expressions (e.g., named entities, incivil expressions) – and social scientists' need for the simultaneous measurement of multiple, often internally complex textual contents (e.g., object-specific evaluations, frames; Liu & Zhang, 2012; Ophir, Walter, & Marchant, 2020). To measure multiple textual

contents expressed using different kinds of linguistic patterns, it is typically necessary 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 resources and capabilities as would be required for their application in cutting-edge research (e.g., Amram, Ben David, & Tsarfaty, 2018).

To substantiate our claims and gauge their implications for computational social science textual research, we draw upon a content analysis of all research published in the top 20 journals in communication science, political science, sociology and psychology[2] between 2016 and 2020. Identifying a total of 854 published articles that include some form of quantitative textual analysis, we coded 1) whether studies employed any form of CTAM, 2) what efforts were made to validate the computational measurement, 3) what variables were measured in the text, and 4) what languages were examined. We show that social science researchers do not generally eschew CTAM, but apply them selectively, in ways that are broadly consistent with the identified gaps. Following a brief discussion these gaps' detrimental implications for the advancement of social science research, we conclude by proposing a research agenda for overcoming these challenges.

A broad typology of CTAM

Computational text analysis methods (CTAM) are an umbrella term for many different methods and tools (Boumans & Trilling, 2016). CTAM range from tools for extracting specific contents using simple keywords or formatting rules (e.g., hashtags) to statistically

[2] 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 Supportive Information (SI) A

complex software solutions (e.g., BERT, large-scale language models Devlin, Chang, Lee, & Toutanova, 2019). They include largely agnostic 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, which may be limited to few parametric choices or require expansive training sets, data bases or reference corpora to extrapolate from. Moreover, supervision may take place in the form of pre-trained tools and validated feature sets, as a training phase prior to application, or humans may be kept in the loop continuously to improve and evolve algorithmic classification (e.g., Baden, Kligler-Vilenchik, & Yarchi, 2020; Munro, 2020; Ramage, Hall, Nallapati, & Manning, 2009).

Even within the same family of methods, there exists an impressive variety of available solutions. For example, dictionaries have been used to score textual sentiments, measure broad themes, recognize complex, theory-informed constructs, and even to extract the semantic organization of complex debates (e.g., Boukes et al., 2020; Lind, Eberl, Heidenreich, & Boomgaarden, 2019; Tenenboim-Weinblatt & Baden, 2018). 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 re-used materials (e.g., Maier et al., 2018; Papacharissi & de Fatima Oliveira, 2008; Welbers et al., 2020). Among technologies available for scaling up human classification decisions, algorithms range from reasonably transparent Bayesian or decision tree classifiers to neural network/deep learning tools whose classification choices are, at least for now, beyond the full comprehension of computer scientists and mathematicians (Burscher, Odijk, Vliegenthart, & de Rijke, 2014; Tenney, Das, & Pavlick, 2019). 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. More than a precise distinction of

tools, this categorization recognizes three modes of thinking that underlie each of the three approaches. Where CTAM take 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 (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 already contains substantively predefined classes and instead classify instances inductively based on observed regularities in the text.

Of course, there are still many methods that defy unique classification. For instance, most tools designed to identify bot-like communication patterns on social media rely on a combination of specified rules, inductively determined interaction patterns and sometimes also a corpus of reference cases (Hilbert & Darmon, 2020). Bibliometric tools are primarily unsupervised in that they detect referencing patterns in an inductive manner, but they also include pre-set rules that enable them to distinguish different types of sources and source attributes (Thor, Bornmann, Haunschild, & Leydesdorff, 2020). Word embeddings are derived from an unsupervised procedure, but are often used to augment supervised classification tasks (Chan et al., 2020; Rudkowsky et al., 2018). Yet, while it is sometimes impossible to uniquely classify a tool *per se*, its underlying algorithms and present applications can usually be located within one of the above three broad approaches.

Given their immense diversity, CTAMs' 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;

Well-documented tools that transparently communicate embedded assumptions may be more appealing that complex methods that involve obscure 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, well-understood and substantive 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 empirically social science text analysis. Yet, the knowledge accumulated by social science text research remains largely absent from ongoing development in computational methods (see also Theocharis & Jungherr, 2020). Neither the validity criteria driving social scientists' operational choices, nor those insights gained from social scientists' efforts at validating available tools are effectively communicated toward computational sciences' tool development. Instead, debates concerning CTAM development primarily revolve around algorithms' statistical properties and limitations, whose bearings upon social scientific measurement concerns remain unclear. This gap between social scientific validity criteria and methodological development comprises at least three major disconnects that diminish the utility of computational tools 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. Not even the many methodological debates about quantitative text analysis in the social sciences – which have discussed classification biases (e.g., Geiss & Monzer, 2020), the structuring of texts (Baden, 2018; Pipal, Schoonvelde, & Schumacher, 2019), and many more issues that are directly relevant to computational measurement (e.g., van Atteveldt, van der Velden, & Boukes, 2021) – are well-reflected in computational tool development. Similar arguments have been made with regard to available knowledge in linguistics research (Bender, 2011).

In the same vein, computational tools' alignment with social scientific and linguistic knowledge rarely constitutes a relevant evaluation criterion or objective throughout their development process. For example, topic models have been developed to evaluate corpora of text that are prohibitively large to read by any one person (Blei, 2012; Boumans & Trilling, 2016). Yet, despite the ostensible reference to topicality, neither the original introduction, nor subsequent developments refer to social scientific knowledge about the topicality of texts. 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 (Mehrtra, Sanner, Buntine, & Xie, 2013). As a consequence, we know hardly anything about how the very different organization of different genres of discourse - be that social media interactions, hierarchically structured news discourse, or parliamentary debates - affect the performance of available topic models. This neglect of measurement validity is even evident in many tools developed within the social sciences themselves. As a particularly pertinent example, few tools amid the ongoing wild growth of computational methods for the measurement of frames elucidate how proposed algorithms *validly* operationalize the construct (Baden, 2010; David, Atun, Fille, & Monterola, 2011; Nicholls & Culpepper, 2020; Papacharissi & de Fatima Oliveira, 2008; Walter & Ophir, 2019).

Second, and consequently, social scientists trying to make an informed choice about their use of computational methods frequently fail to find relevant information in the existing methodological literature (Grimmer & Stewart, 2013). Reflecting the described mode of development, available documentation regularly omits reference to what known linguistic, discourse-practical or conceptual properties an algorithm attempts to model (for notable exceptions, see Baden, 2010; Liu & Zhang, 2012). Moreover, there is little knowledge about what preprocessing stages are suitable for analyzing different kinds of discourse, or for tuning computational methods toward the detection of specific textual properties. In many cases, multiple similar tools exist, but there is limited methodological guidance on which tools are better suited for specific measurement tasks, and why (e.g., van Atteveldt et al., 2021). In place of discussions tying methodological choices to operational concerns, the debate offers 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; M. Roberts et al., 2020), and rules of thumb (e.g., it usually helps to remove prepositions). Even if specific methods have been shown to perform well in the past, these recommendations instill little confidence that a choice is appropriate for a given project.

Third, the validation of CTAM is typically externalized from the development process, and left instead to the stage of application (Budanitsky & Hirst, 2006). Since very little is known about how different kinds of measured constructs, different forms of discourse or other operational variations might affect the performance of a tool, it is largely left to the user to ascertain whether a tool is suitable for a given measurement task. Only at this stage, then, researchers' knowledge of the investigated texts and meanings informs the augmentation of dictionaries (e.g., Muddiman, McGregor, & Stroud, 2018); shapes the pre-processing of textual materials (Denny & Spirling, 2018); or serves as a benchmark for selecting a 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 lemmatizing seems to help focus topic models on certain textual qualities (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 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 – chiefly, precision and recall scores – respond primarily to the most common expressions in the textual material. Neglecting non-standard expressions, chance, and other sources of classification bias that are well-understood in social scientific text analysis (e.g., Krippendorff, 2004), these indicators fail to communicate researchers concerns about measurement validity and systematic error.

Taken together, these three disconnects undermine the emergence of a methodological debate that connects the development of CTAM to existing experience in social scientific textual analysis. Social scientists' operational knowledge, needs and experience in textual research are barely communicated toward computational tool development, and developers' modeling assumptions are rarely reflected and rendered meaningful in social scientific methodological discourse. Not only does this gap thus deprive developers of the knowledge and incentives needed to further improve the utility of existing tools for application in social science research (Grimmer & Stewart, 2013); but it also denies social scientists the information needed to make confident, informed operational choices in the application of computational methods.

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. Sentiment tools score overall the sentiment of texts (Boukes et al., 2020). Topic models extract systematic token co-occurrence patterns (Maier et al., 2018). Dictionaries

identify references to specific entities or constructs within the text (Lind et al., 2019). Dependency parsers classify specific phrases based on their grammatical roles (van Atteveldt, Sheafer, Shenhav, & Fogel-Dror, 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; document-level topicality and sequential patterns). Moreover, most available tools focus either 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).

By contrast, much of social scientific text analysis is aimed at reconstructing embedded meanings that arise from the specific arrangements of textual contents (Baden, 2018). Consequently, the information required to address social scientific research questions is 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 (Tenenboim-Weinblatt & Baden, 2018) and broad, global themes (Burscher et al., 2014; Nelson, 2017), or any combination of these. For instance, social scientists are rarely interested in the “tone" of a text as a whole (for an exception, see Proksch, Lowe, Wäckerle, & Soroka, 2019), or whether a specific actor is mentioned at all. Much more commonly, they wish to access the specific evaluative sentiment that is expressed *by specific sources* (Overbeck, Baden, Aharoni, & Tenenboim-Weinblatt, 2021) and *with regard to* some topic (e.g., the economy), construct (e.g., a policy) or entity (e.g., a country) referenced in the text (Liu & Zhang, 2012). They study whether mentioned actors are blamed for certain things, how they are characterized, or whether they appear in certain roles. 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; Entman, 1993; Matthes & Kohring, 2008). On top of these

challenges, many analyses seek to establish relations between different textual contents (Geiss & Monzer, 2020), which may require the extraction of specific semantic relationship expressed in the text (van Atteveldt et al., 2017). Accordingly, the same social scientific text analysis frequently involves the recognition of multiple discrete meanings and relations between these, which almost inevitably require different measurement procedures (Schoonvelde, Schumacher, & Bakker, 2019).

Yet, given the structure of most existing computational tools, measured contents can often only be put into relation with one another *post-hoc*, based on the observed correlation of occurrences within broad textual units: Texts that mention a certain construct also tend to show overall mostly positive sentiment; passive voice is more common in texts that also contain a certain topic. To extract specific relations expressed in the text, or operationalize complex, modular constructs, this strategy leaves much to be desired.

In fact, 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). Only recently, with the emergence of R and Python as the two dominant coding environments for computational text analysis software, have there been some systematic efforts to gather the capabilities needed for different kinds of analysis in a unified environment (notably, the quanteda R package; Benoit et al., 2018). Moreover, consequential challenges remain even if diverse algorithms are increasingly available on the same platform, using compatible data formats. Often, different tools make different assumptions regarding the processed material and require different preprocessing steps, which require different skill sets to implement and validate. Applying multiple algorithms side-by-side, researchers need to repeat the process of methodological justification, modeling and analysis multiple times, wasting time, space and resources (see also Windsor, 2020, for a related point).

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 important as most errors incurred by CTAM are not random but systematic. Using multiple algorithms in sequence, errors incurred early early on in the pipeline are carried forward and exacerbated by subsequent steps, degrading accuracy and amplifying possible biases. For example, Denny and Spirling (2018) show that the results of unsupervised topic models and dimensional scaling are very much dependent on seemingly arbitrary decisions in the data pre-processing stage. If we know little about how specific preprocessing algorithms affect the performance of subsequent topic models or supervised classifiers, 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). How can we incorporate measurement uncertainties during one classification step to inform subsequent stages (Baden et al., 2020)?

Taken together, the described tensions between fragmented CTAM development and social scientists' need for integrated, modular measures and pipelines severely constrains the utility of computational tools for social science text research. Whenever at least one of multiple required constructs cannot convincingly be measured algorithmically, CTAM cede their utility. In the choice between crude approximations (e.g., evaluations measured as sentiment, frames measured as topic models, Nicholls & Culpepper, 2020), dropping constructs for which there are no tools available, and performing a manual analysis on a suitable sub-sample, the latter may often be the most convincing choice.

Gap III: English before everything

The third gap concerns CTAMs' heavy focus on English language, and Germanic languages more generally. While a variegated development efforts and resources rapidly advance the

capabilities of computational tools developed for a small handful of languages, researchers seeking to study other languages frequently find available 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, Hanusch, Ramaprasad, & de Beer, 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). As a consequence, English-language text analysis methods continues to advance a pace unmatched that is unmatched by any other language.

Only recently has an increasing number of tools been translated or expanded to include also other languages (e.g., LIWC; NRC Emotion Lexicon). Many language communities by now sustain their own NLP and computational tool development efforts. Still, the availability, sophistication and performance of tools in other languages continues to lag far behind available English-language tools. Accordingly, for many analyses focusing

or including non-English textual material, adequate algorithms are either unavailable or severely deficient.

Even where adequate tools exist in multiple languages, furthermore, their measurement is rarely comparable across languages (Chan et al., 2020; Reber, 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 such as different syntax, but also the limited translatability of equivalent denotations and connotations (Chan et al., 2020; Maier, Baden, Stoltenberg, De Vries, & Waldherr, 2020).

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, which appears reasonable for English, 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 presents 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 – or more widely, Germanic languages – 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, Seddah, Kübler, & Nivre, 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 require major adjustments as they become translated. Try to force different languages into the corset of English-like language structure, researchers need to device additional, often complex preprocessing steps (e.g., artificially tokenizing morphologically rich languages Goldberg & Elhadad, 2013), confronting important open questions that diminish the appeal of existing computational methods (Amram et al., 2018).

As a consequence of the vast and still widening gap between CTAM development in English and most other languages, researchers applying computational tools for the study of resource-poorer languages face considerable difficulties matching the fast-evolving standards expected for (English-language) cutting-edge research. Structural differences between both tools and languages impose severe limitations upon the comparative study of texts across different languages. And whenever adequate, comparable tools are unavailable for any included language, computational methods cease to offer a viable solution.

Expectations

Based on the three gaps discussed above, we thus expect that social scientists' limited use of CTAM can be explained by a) researchers' concerns about measurement validity, b) difficulties in the operationalization of more demanding textual contents, and c) the inadequacy or absence of tools in languages other than English – on top of possible hesitations rooted in limited computational literacy.

To the extent that validity concerns influence researchers' methodological choices, we expect that social scientists exhibit an overall preference for relatively closely researcher-controlled computational tools – notably, dictionaries and other rule-based methods. In addition, we expect that social scientists will explicitly discuss validation

efforts in their work.

To the extent that the fragmentation of CTAMs impedes social scientists' efforts to operationalize more demanding constructs, we expect that researchers tend to opt against using CTAM as the diversity and complexity of measured textual contents increases. Specifically, researchers should default to manual classification for the measurement of internally complex constructs (e.g., attributions, propositions, frames) and whenever different variables require measurement at different levels of abstraction (e.g., document-level and localized measurements).

To the extent that the quality and availability of CTAMs in different languages presents a challenge, finally, we expect that researchers tend to rely less on computational tools when studying languages other than English, and especially, languages that are very different from English. In addition, computational tools should be much more limited for work that includes more than one language, and requires the comparable measurement of textual contents in different languages.

Of course, the above expectations overlap in part with what one might expect also as an outcome of researchers' limited computational literacy: 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; the most closely researcher-controlled computational tools tend to be also more established and thus familiar. To evaluate such alternative explanations, we interpret authors' methodological choices against the specific demands of measured constructs, processed materials and chosen tools. The extent to which authors engage in extensive tweaking and targeted validation efforts, prefer established or more recent tools, or are based at major anglophone institutions should offer valuable insights into the relative weight of literacy-related reasons.

By examining on the use of CTAM specifically in the top tier of research

publications, we furthermore focus on a population of researchers that should be least likely to suffer from computational illiteracy and most enabled to acquire any required computational skills. In addition, these leading authors' methodological choices had to pass the strictest peer-review processes, and thus deserve to be considered current best practices in the field. At the same time, the practices observed in these leading publications are not necessarily representative of the field as a whole, both due to authors' likely above-average skill set, and a range of well-known publication biases, which arguably privilege research on well-established subjects, Western research sites and English-language materials.

Data & Methods

In the following, we seek to assess the impact of these gaps for empirical text analysis in cutting-edge social science research. Specifically, we conducted a content analysis of all quantitative text-based research published over the past five years in the top 20 highest ranked journals in the Web of Science categories of communication, political science, sociology (including mathematical models) and psychology (multidisciplinary and mathematical), selected according to their 2019 SSCI 1-year impact factors.

Sampling. Our analysis departed from an inventory of all articles 45,437 published in the selected journals (see Appendix A) 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. We then accessed the full text of these articles to determine whether the presented research included any form of quantitative textual analysis. Quantitative textual analysis was defined broadly to include any form of processing natural language that identified specific kinds of textual contents with the purpose of classification and quantitative analysis. Analyses that relied solely on metadata or pre-existing classifications were excluded, as were investigations accessing only formal properties of the sampled texts (e.g., length). We included analyses of multi-modal media (e.g., posters,

television) as long as textual contents were informative toward classification. Purely methodological contributions discussing specific potentials or limitations of available methods were excluded, unless they included applied demonstrations wherein actual textual data was processed. Articles were considered relevant as soon as they used any form of quantitative textual analysis, even if it was used merely in an auxiliary capacity (e.g., a content analysis to identify frames to be used in an experiment; sentiment analyses of open-ended survey responses). This screening yielded a total of $N = 854$ articles, which form the basis for the following analysis.

Analysis. For all relevant articles included in the analysis, we manually coded four main variables. First, we determined whether manual or computational strategies were chosen and–in case that CTAM were selected–what specific methods were used. For all studies using computational methods, we furthermore coded whether the article discussed specific strategies to validate the computational measurement. Next, we determined what variables of interest were coded in the textual data. Finally, we recognized the language or languages of analyzed textual materials. In the following, we briefly explain the coding process for each variable and introduce the specific coded categories.

Used Methods. For the classification of used methods, all classifications were based on the actual use of textual material and quantitative text analysis, and not based on authors' own labeling. Many articles did not expressly label their methodological approach, or used labels in ways that did not match our criteria (e.g., "content analysis" as a label for methods that did not involve any text-based classification or quantification, or "machine learning" as umbrella term for any use of computational tools).[3]

For the purpose of this study, we defined *Manual approaches* as all approaches where all classification decisions are made by a human following instructions. This includes both

[3] Notably, most "qualitative" content analyses are in fact quantitative analyses and thus relevant; the "qualitative" component usually referred to a preceding, inductive stage of category development (which was not considered for this study).

classic content analysis, various quantifying procedures embedded in otherwise qualitative methods, as well as crowd coding.

Rule-based approaches were defined as all methods that classify textual contents based on specific pre-defined properties or expressions. Among the content-focused variants, this includes keyword-based classification and various forms of dictionaries, which we further differentiated into researcher-constructed and -controlled dictionaries, and dictionaries embedded in existing software packages or otherwise available for reuse (e.g., LIWC). Property-focused variants include any tools that extract contents based on their formatting (e.g., hashtags, links) as well as natural language processing technologies ranging from simple word frequency extraction to grammar tagging and parsing tools.

We defined *Unsupervised approaches* as any methods that extract patterns from the textual co-occurrences of unknown tokens, based on some mathematical model. These approaches included various forms of topic modeling and related clustering algorithms, semantic network analysis, as well as a small group of tools measuring the overall similarity between documents (akin to plagiarism detection tools) in an inductive fashion.

Supervised approaches, finally, were defined as any methods that used algorithmic strategies to extrapolate underlying rules from (usually manually) pre-classified textual data and apply these to automatically classify additional cases. This class includes supervised machine learning approaches (), document scaling tools, as well as hybrid content analysis (Baden et al., 2020), which classifies large text corpora based on their use of manually classified topics.

Validation. For all studies that included any use of CTAM, we determined whether the article mentioned specific efforts to evaluate the validity of algorithmic text classification decisions. Owing to the diversity of tools and measurements, we included a wide range of approaches to validation, from comparing coded outcomes against gold standard data, to the auditing of classification rules, to the establishment of convergent validity based on additional measures or external knowledge and data. Such discussions of

measurement validity were distinguished from mere interpretations of presented data, which assume measurement validity as given (e.g., labeling topics). Likewise, we excluded references to prior validation based on different data (e.g., others' validation studies; software documentation). We also excluded formal robustness checks, which are indifferent to measurement validity, as well as validations of the statistical analyses mounted upon classified textual data.

In addition, we recognized whether any data or details were provided about the outcomes of such validation efforts, demonstrating measurement validity (e.g., demonstrations of face validity; precision, recall, and F1 scores; confusion matrices or error analyses; correlations for convergent validity). Reflecting journals' different standards for presenting methodological documentation, we considered any demonstrations of measurement validity presented either in the main manuscript or in attached or online appendices.

Variables of interest. For the classification of the variables of interest coded in the textual materials, we distinguished between four broad classes and various subclasses.

A first class focuses on information obtained from *entire documents* as the unit of observation. Within this class, we distinguished between analyses interested in the topical focus of documents (including their classification by issue, theme or covered event); analyses interested in the classification of document types (e.g., by genre, authorship, etc.); analyses interested in the tone or textual sentiment of documents; and analyses interested in other qualities present in these documents (e.g., stylistic properties, offensiveness, or embedded news values).

A second class focused on the identification of localized instances *within a text*. Within this class, we distinguished between analyses interested in texts' inclusion of formal contents (e.g., the use of links, visual elements, or bylines); analyses interested in references to specific named entities (e.g., people, firms, countries); and analyses interested in references to theoretically constituted, abstracted constructs (e.g., value references,

metaphors, political conflict).

The third class includes any *pre-structured information* to be obtained from the analysis that is non-reducible to document properties or localized instances. Within this class, we distinguished analyses interested in the presentation or evaluation of specific objects or actors, including the attribution of specific qualities; analyses interested in specific propositional contents (e.g., claims, arguments) in the text; and analyses interested in more complex, multipartite systems of claims and ideas, such as frames, narratives, and similar constructs.

Finally, we considered a somewhat heterogeneous class of analyses that studied text in order to look *beyond the text* itself. Within this class, we considered analyses interested in specific relationships between texts (e.g., one being referenced, derived from or plagiarized by another); analyses interested in the behavior reflected or expressed in the text (e.g., speech acts, interactive communication behavior); and a final set of studies using textual data to reconstruct social networks.

Included languages. With regard to the classification of languages, very few articles explicitly identified the languages present in the analyzed textual contents (Bender, 2011). Explicit references were found mostly 1) in studies focusing on multilingual countries (notably, Belgium, Switzerland, India), 2) when studies relied on non-native language materials (e.g., studying English language news media from non-English-speaking countries) or 3) when authors acknowledged the removal of contents in languages other than the one(s) in focus from the corpus (typically, removing tweets in languages other than English).

In most cases, therefore, language had to be inferred from the description of the analyzed material. For formal publications, this is straightforward: The Times of India publishes in English, and El País in Spanish. Matters are less straightforward for informal and social media communication. While it makes sense to assume English language to be prevalent (and the focus of analysis) in US politicians' campaign tweets, there is good

reason to expect also some tweets in other languages (notably, Spanish). Likewise, if an article analyzed transcripts from interviews with students at an US-based university, or with policy makers in a global forum on internet policy, we may assume that transcripts were probably in English.

The largest uncertainty arose for social media samples composed by following specific hashtags. While the likely most common language contained in a sample can usually be inferred from the selected case and setting (e.g., most uses of `#blacklivesmatter` are likely in English language tweets), it is reasonable to assume that also other languages occur – especially when certain debates transgress national boundaries (e.g., `#metoo` was also used in various European Twitter debates) or when certain hashtags are spelled identically in multiple languages (e.g., `#stopnazis` works in several languages). While it is thus likely that many such corpora contained material in multiple languages, few studies documented awareness of this possibility, and fewer still proposed strategies to address it.

For the purpose of this study, therefore, we assume that such foreign-language contents were not considered in the analysis and show up as noise in the reported findings, unless a different treatment was mentioned. Accordingly, we generally recorded the dominant domestic language relevant to the presented corpus, unless the article expressly acknowledged whether contents in other languages were treated as well. Finally, a very small number of studies was explicitly agnostic to language and thus included contents in any language.

For each case, we recorded all quantitative text analysis methods used in a study, all languages represented in the analyzed material, and all types of information extracted for a study by means of textual classification. Validation efforts were coded as binary variables (present/absent). The full codebook is available in Supportive Information (SI) B.

Coding reliability. Each article selected for manual screening and content analysis was read and assessed by one coder. To ensure the reliability of our coding, a second coder repeated both the screening procedure and the coding process, using a

random sample of articles classified at each stage. We then assessed the agreement between coders using Cohen's kappa. For the initial relevance screening of articles based on their use of quantitative text analysis, this sample consisted of 200 randomly selected articles from each discipline, sampled from the results of the keyword search. Inter-coder reliability for this stage was very high (communication science: $\kappa = 0.96$, political science: $\kappa = 0.97$, sociology: $\kappa = 0.94$, other social sciences: $\kappa = 0.93$). For the subsequent classification of relevant articles based on their used methods, validation procedures variables of interest, and included languages, we sampled 10% from all articles selected for the content analysis (n = 85). Again, this procedure yielded very high inter-coder reliability, with $\kappa > 0.9$ for all variables.

Findings

Prevalence of quantitative text analysis. Overall, quantitative text analysis was found to be much more prominent in communication science than in political science and the other considered social sciences. In communication journals, 9% of all published articles made at least some use of quantitative text analysis, whereof 34% (3% of all articles) employ any computational methods. Quantitative text analysis was especially common in journals dedicated to the study of political communication and journalism (17% of published articles). By comparison, only 4% of articles in political science journals,[4] 1% of articles in sociology[5] and less than half a percent in psychology studied text in a quantitative perspective. Despite a higher relative prominence of computational text analysis methods in these fields, communication journals published more computational text analytic studies, and far more quantitative text analyses, than the

[4] About 16% of these were published in *Political Communication*, which is listed in both communication and political science.

[5] Almost half of these were published in *Information, Communication & Society*, which is listed in both communication and sociology.

other fields combined. Owing to the low number of relevant studies beyond communication and political science, we will in the following focus on these two fields only.

Use of CTAM. In communication science, 64% of all studies relied on manual content analysis as the sole means of textual classification. The remaining 36% involved at least some computational tools. Proportions were reversed in political science, with 43% using manual classification only, and 57% including some use of computational tools. About 27% of studies in communication science, and 49% in political science relied solely on CTAM.[6]

Among computational methods, rule-based approaches dominated (24-28% of all studies): Most commonly, studies relied on researcher-created dictionaries (7-13%) and natural language processing software (6-7%). Unsupervised approaches – usually, topic modeling – were more common in political science (19%) than in communication (9%), and the same is true for supervised approaches (14%, 6%). Where studies combined multiple methods, most augmented manual classification with rule-based technologies (6%). In communication, also combinations of unsupervised with rule-based or manual approaches were found regularly (3% each).

Validation. Across all fields and CTAM, a majority of publications that use computational tools expressly addressed validation as an important concern. In political science, 71% of relevant articles mentioned efforts at validating applied methods. In communication, only 57% did the same. The validity of textual measurements was demonstrated in a majority of CTAM applications in political science (54%), but only a minority of articles in communication (32%).

That said, validation efforts varied systematically between different CTAM. For supervised approaches, validation efforts were addressed almost universally (90%) and validation scores (typically, precision, recall and F1-scores) were presented in 71-79% of

[6] This includes uses of supervised machine learning, which often involves the manual pre-classification of part of the material.

cases. Validity was less dependably addressed and demonstrated for unsupervised methods, whose inductive nature bars validation by reference to a given gold standard. Still, a majority of uses discussed efforts to validate inductive patterns by assessing convergent validity with known phenomena or establishing face validity (around 80%).[7] Among rule-based methods, dictionaries underwent documented validation efforts in 73-74% of cases, but were rarely accompanied by a demonstration of validity (42-43%). For pre-existing software solutions and, less so, NLP applications, measurement validity was still frequently discussed, but only demonstrated in a minority of cases. For relatively simple tasks, such as link extraction and keyword-based counts, less than half of all applications argued that such procedures validly operationalized some relevant construct, and most skipped explicit validation.

Variables of interest. In all fields, thematic information constituted the most common focus of analysis (37-38%). In communication science, but not in political science, also specific constructs (32%, as opposed to 6%) and named entities (29%, 17%) were similarly common. Of all studies, 11-13% were interested in textual sentiments, and 22% in other textual qualities. More complex constructs such as object-specific evaluations and attributions (16-17%), frames and narratives (11-15%) were less common, but still relevant. Propositional contents (11%), as well as pragmatic information (16%) were studied occasionally in communication science, but rarely in political science.

Included Languages. 84% of text analytic studies in communication, and 82% of those in political science, relied on material in one language only. English was strongly dominant in political science (69%), and less so in communication (59%). German ranked second in both fields (10%), in communication closely followed by Dutch (9%). No other single language accounted for more than 5% of studies in either field, and only German, Dutch, French, Spanish, Chinese, and Hebrew contents were studied in more 1% of

[7] Validation efforts were less common for topic models and semantic networks, especially in communication science, where only one in two uses discussed, and one in four demonstrated validity.

single-language studies. Looking at wider language families, Germanic languages made out 81-85% of all single-language studies in each field, followed by Romance languages (5-7%), Semitic languages (2-3%) and a group of East-Asian language families (Sino-Tibetan, Japonic, Kra-Dai, Austroasiatic and Austronesian; 4-5%). Slavic and other Indo-European languages, as well as other Eurasian language families (e.g., Uralic, Turkic) account for less than 3% each. We found only one language beyond these families, that is, one study of computer language. African and American indigenous languages were completely absent.

Among those studies that treated material in multiple languages, English was included in 79-84% of all cases, followed by German (52-61%), French (43-45%), Spanish (25-39%), Dutch (18-36%), Chinese (9-10%), and Hebrew (5-9%). No other language was included in more than 6% of coded multilingual studies. Germanic languages were included in virtually every multilingual study in political science (95%), and 86% of studies in communication. Around 62-64% of all multilingual studies included at least one Romance language, and around 18-22% included at least one Slavic language. Four combinations of languages were commonly observed: 1) studies focusing on lead authors' (or their institutions') domestic languages that included another English-language case for comparison; 2) various combinations of major European languages (German, French, Spanish, English, Italian), often augmented by one or two smaller languages; 3) broad global comparative studies that included English, some other European and some Asian languages; and 4) smaller, case-driven comparative studies, usually between neighboring languages.

Only Germanic languages were commonly included in both single-language and multilingual studies; Romance languages were commonly included in multilingual studies but rarely appeared as sole focus (as an extreme case, French occurred in 43-45% of multilingual, but <1% of single-language studies); Chinese maintained a marginal presence (9-10% of multilingual, 2-3% of single-language studies); and everything else was rare.

Use of CTAM by variable of interest. With regard to the use of computational methods to study different kinds of textual contents, a differentiated picture emerges. To extract thematic information, but also to determine the relations between documents, rule-based (22%) or unsupervised tools (16%) were commonly used. Sentiment classification relied heavily on rule-based (46%) and, less commonly, supervised (12%) methods. Formal contents, named entities, and constructs were still measured occasionally using computational methods (31%), while most complex constructs (notably, propositions, frames, pragmatic contents) were overwhelmingly studied manually, and occasionally using rule-based CTAM (<20%). Only for the measurement of frames in communication research did studies regularly resort also to unsupervised tools (10%). Manual classification was used in 44-63% of studies interested in sentiment, thematic information, or document relations; 67-80% of studies interested in specific instances of entities, constructs or formal contents; and 70-88% of studies focused on more complex textual meanings.

Most strikingly, computational methods were much more common in studies focused on extracting exactly one kind of information from the text: 29% of such studies involved rule-based tools, 17% unsupervised methods, and 11% use supervised ones. For studies measuring two different kinds of textual information, shares dropped to 27%, 9%, and 6% respectively, and further to 19%, 6%, and 5% for studies recording three or more kinds. Reliance on manual classification, inversely, rises from 49% for single-focus studies, to 71% and 84% for the measurement of a second and third kind of information. Paradoxically, increasing complexity of measurement thus *decreased* the number of methods included in the analysis, as researchers defaulted to manual classification.

Use of CTAM by language. Computational tools played a notably bigger role in single-language studies than in the study of multilingual contents. The inclusion of a second language decreased the use of computational methods from 35% to 28% in communication science, and from 65% to 41% in political science (the drop was even more dramatic if one excludes tools that rely on language-independent textual properties, such

as the extraction of links or hashtags). In communication science, the use of NLP tools dropped from 8% to 2%, and the use of software receded from 5% to 2% if multiple languages were considered. In political science, mostly the use of dictionaries decreased (from 14% to 5%). In both fields, multilingual studies were more likely than single-language studies to algorithmically extract formal contents (links, hashtags) and to rely on simple keyword-based analyses. Multilingual studies were about half as likely to use unsupervised methods as are single-language studies. By contrast, supervised methods were roughly similarly common in multilingual and single-language studies in both fields. In political science, but not in communication, also manual classification was more common in multilingual (66%) than in single-language studies (48%).

Within single-language studies, computational methods dropped from 40% to 31% (communication science) and from 64% to 57% (political science) if the studied language was not English. In communication, but not in political science, it dropped further if the language was not Germanic (28%). Especially off-the-shelf software packages (e.g., sentiment tools, LIWC), dictionaries and supervised methods were only half as common in non-English text research. Inversely, reliance on manual classification increased from 69% to 78% (communication) and from 44% to 56% (political science) for non-English language materials.

Discussion

As our content analysis of recent research published in top ranking social science journals documents, CTAM are used both widely and selectively in cutting-edge social scientific textual research. On the one hand, especially more recent, high-powered CTAM - notably, advanced machine learning tools - remain rare in the studied research applications. On the other hand, social scientists have embraced a wide variety of CTAM and applied them to advance their research in various ways. While there is also some evidence of methodological choices that appear better explained by lacking computational literacy and confidence,

each gap contributes to delineating specific conditions under which social scientists tend to systematically opt out of using CTAM.

With regard to the first gap, we have argued that social science researchers might frequently prefer CTAM that permit far-reaching researcher control over operationalization and validation processes. Consistently with this expectation, we have documented social scientists' intense and pervasive efforts at validating CTAM wherever they are used. On the one hand, researchers exhibit a strong preference for fully researcher-controlled CTAM - notably, dictionaries. Tools that involve more demanding algorithms, less transparent modeling assumptions, and diminished opportunities for human oversight – notably, supervised CTAM – are comparatively less popular. On the other hand, social scientists' efforts to ascertain the validity of used CTAM markedly intensify as researcher control diminishes. An exception concerns social scientists' ready adoption of topic modeling, which accounts for almost one in five recorded uses of CTAM. Yet, despite the difficulty of validating inductively constructed patterns, such efforts were still undertaken in a majority of studies.

That said, validation concerns did not dominate universally. Especially for very simple applications of CTAM – for extracting links and formal contents, recognizing keywords or applying readily-available software packages – validation efforts were markedly reduced. Rather, such efforts appear to be triggered specifically when either complex measurement tasks or advanced CTAM bring potential validity issues into focus. While computational illiteracy or skepticism still likely contributes its share to the scarcity of more sophisticated CTAM, new tools seem to primarily redouble researchers' determiniation to validate.

The present study has documented a considerable, but very uneven presence of CTAM in social scientific textual research. Clearly, social scientists are aware of the promises and potentials of CTAM – especially considering that the alternative is manual classification. Yet, leading authors in the field have embraced topic modeling as a useful

addition, but identified no more than a handful of applications for supervised machine learning. For both unsupervised and supervised CTAM, there exist various easy-to-use software packages (Boumans & Trilling, 2016) – but, social scientists continue to prefer rule-based tools. Given the immense efforts needed to construct and validate dictionaries (Lind et al., 2019), social scientists clearly do not tend to go for easy solutions. In this light, social scientists' selective use of CTAM – and specifically, their limited adoption of supervised tools – appears to primarily reflect researchers' skepticism that the available methods meet present operational demands.

With regard to the second gap, we have argued that social science textual research routinely focuses on the measurement of textual contents that exceed the capacity of any one computational tool. In line with this proposition, our findings document a prevalent use of computational tools for measuring broad document-level properties such as topicality and sentiment, as well as relatively simple textual meanings - most notably, formal contents (e.g., links) and named entities. By contrast, social scientists' use of CTAM substantially decreases whenever multiple different textual contents, or internally complex constructs require measurement. Doubtlessly, social scientists still frequently compromise on operational validity to exploit the advantages of CTAM. Especially the widespread use of sentiment tools – whose measurement validity has recently come under intense scrutiny in social scientific methodological research (Boukes et al., 2020) – illustrates the point. Likewise, communication researchers' enthusiasm for measuring frames using CTAM, most of which remain unable to ascertain frames' constitutive internal structuring, often comes at the cost of rendering frames operationally indistinct from themes (Baden, 2018; Günther, 2020). For many variables of interest, available tools regularly incur major losses in measurement validity, while suitable computational tools simply don't exist. Even for English language text, the automated extraction of pragmatic information, propositional arguments or larger narratives remains an active construction site in computational development (Stalpouskaya, 2020; Tenney et al., 2019; Walter &

Ophir, 2019). As a consequence, social science researchers overwhelmingly continue to rely on manual classification whenever the variability and complexity of measured constructs increases. Contrary to claims that computational methods help researchers to reduce manual effort, researchers systematically opt *against* computational methods for more complicated, effortful classifications.

With regard to the third gap, finally, we have argued that the very uneven allocation of development efforts and resources in favor of English-language computational text analysis severely limits the utility of CTAM for the study of most other languages, as well as the study of multilingual and inter-lingually comparative text corpora. Consistently with this expectation, our findings document a substantial drop in the use of CTAM for all languages other than English. Except for the one exception of (Mandarin) Chinese, where a growing community of researchers has begun to construct dedicated computational tools, CTAM use steadily decreases with languages' increasing distance from English: While still reasonably common in the study of other Germanic languages, any step away from the English "default" case results both in a reduction of the overall use of computational tools, and a retreat toward increasingly basic (e.g., keyword counts) and linguistically agnostic methods.[8] The heavy dominance of English even beyond the share of authors based in anglophone countries and institutions documents researchers' pronounced tendency to prioritize studying English discourse to benefit from available computational tools. Of course, one possible factor in the disproportionate emphasis on English language text may also have to do with the legacy of many leading social scientific journals that have developed out of domestic journals in the U.S. and U.K. and still double as preferred outlets for research on domestic phenomena. Computational textual research on other languages may face additional hurdles breaking into the English- and Western-dominated

[8] However, as Bender (2011) convincingly argues, many ostensibly language-independent tools are still modeled upon the linguistic template of English discourse, and thus suffer performance losses when applied to very different languages.

flagship journals, and may be more commonly published in domestically oriented journals elsewhere (Wilson & Knutsen, 2020). Yet, computational studies on non-English text are underrepresented far beyond the overall under-representation of non-anglophone textual research in these privileged publication venues. Especially for non-anglophone (and even more so for non-Western) researchers, CTAMs' near-universal orientation toward English as default language presents a major hurdle for the application of computational tools.

Our analysis underscores that these issues are of concern primarily for scholars in communication and political science, who rely most heavily on quantitative text analyses. CTAM appear to have penetrated somewhat farther into political science text research than into communication. The price paid for this – or, inversely, the enabling condition – appears to be a much heavier focus on English-only content. In communication science, computational methods are less commonly used in leading publications, but they are applied to a notably wider variety of measured variables and languages. In both disciplines, alas, CTAM are applied selectively, enlisting the support of computational tools where these offer adequate validity and added value, while for more complex and demanding measurement tasks, researchers' need for operational control continues to lead them primarily toward manual classification.

Implications for Social Science Text Research. The presented three gaps raise important implications for the advancement of social science research, and the development of the field as a whole. To the extent that cutting-edge computational tools are becoming normalized in social scientific text research, their 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, confronted with the choice to either apply cutting-edge computational tools to study foreign events, in a non-native language, or to

make do with inferior tools or manual analyses in their native language, the imbalance disincentivizes researchers from diversifying textual research beyond the present dominance of WEIRD (Western, Educated, Industrialized, Rich & Developed) nations (Henrich, Heine, & Norenzayan, 2010).

Similarly, the disproportionate efforts and unavailable methodological knowledge required to computationally operationalize more demanding 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, issues, and other conceptually distinct phenomena, the progressing application of CTAM threatens to blur important theoretical distinctions (Baden, 2018; Overbeck et al., 2021). Moreover, due to the prevalent mode of CTAM development outside the social sciences, present difficulties in tailoring and concatenating computational tools to operationalize social scientific constructs privilege the study of phenomena for simply being measurable, compromising social science text researchers' control of the research agenda.

Especially 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. 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 many decades of social scientific, linguistic and other text-based research, computational tool developers are likely to miss, misconstrue or inadequately model important textual properties, and bound to laboriously reinvent the wheel through trial and error.

Research Agenda

That said, there are also several encouraging insights that can 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. Increasingly, social science methodologists have begun to systematically debate the operational implications of various computational procedures in light of existing methodological knowledge in textual research (e.g., Boukes et al., 2020; Maier, Niekler, Wiedemann, & Stoltenberg, 2020). Research teams in both social science and AI are beginning to concatenate different tools and methods in order to boost both 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); Maier, Baden, et al. (2020).

In the following, we wish to sketch several key desiderata for a future research agenda, which is suitable to narrow the presented gaps and address their problematic implications for social scientific research.

1. 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 Atteveldt & Peng, 2018). However, to communicate effectively, some shifts in perspective may be useful. One much needed shift concerns the perception, which seems widespread in the computational sciences, that views social scientists primarily as *users* (or, in worse cases, computationally illiterate *non-users*) of computational tools. While both variants of course exist, this perception overlooks the rich knowledge about text and discourse, language use and, not to mention, methodological challenges in text analysis, that exist in social science (and, importantly, linguistic) research. Social scientists can - and should - teach their computational colleagues many a

thing about a variety of relevant issues, ranging from the topical organization of different genres of discourse (Günther, 2020), to the cultural embedding of textual meaning (Maier, Baden, et al., 2020), to the non-random 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 tools that are much more nuanced, valid and useful for applied research - they can also skip quite a few detours that social scientists have amply explored in the past. Inversely, of course, social scientists would do well to abandon their perspective as CTAM users themselves and participate, if not necessarily in CTAM development, then at least in the strategic improvement of available tools (Chan et al., 2020). One key step here could be to shift emphasis in validation from demonstrating *that* a computational method perform certain tasks, to discussing exactly *how* it operationalizes relevant conceptual properties into algorithmic form (Liu & Zhang, 2012; Stalpouskaya, 2020), and how well such an implementation conforms to known operational demands (algorithmic auditing). Beyond validating tools to find them either wanting or "good enough", social scientists can generate valuable knowledge to inform the improvement of available tools – e.g., by analyzing misclassification patterns and identifying the sources of observed classification biases. In the same vein, social science methodologists are well-placed to scrutinize the differential behavior of available tools and algorithms, accumulating knowledge about the text-analytic implications of computational preprocessing and modeling choices (Günther, 2020; Maier, Baden, et al., 2020; Schoonvelde et al., 2019).

2. 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 few existing efforts at gathering diverse capabilities in unified text analysis platforms and packages (Benoit 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 better 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 additional attention (Baden et al., 2020). Inversely, also offering explicit conceptual definitions and operational models of textual meanings, and especially an explicit understanding of relatedness in social science text research can contribute valuably to this research agenda. In place of relying on human coders' intuitive grasp of grammatical, syntactical, and other forms of relatedness, especially linguistic and discourse research offer a rich vocabulary for adding precision to operational choices, enabling their better translation into algorithmic procedures (Baden, 2018).

3. 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., Tsarfaty et al., 2013), one key step concerns raising the visibility of non-English computational text research and the challenges that arise from such undertakings. At the Association for Computational Linguistics (ACL), for instance, dedicated forums and other organizational resources support and incentivize CTAM development in resource-poor languages. Following an important intervention by Emily Bender (2011), raising awareness especially among English-language tool developers that different languages work differently from English 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) that align or deviate from tools' (typically implicit) modeling assumptions – which of course requires that such assumptions are explicated better and exposed to methodological debate. To counter the inherent publication bias in favor of English language textual research, which can benefit from disproportionately rich computational resources and tools, editors, reviewers and also the authors themselves need to acknowledge the specific challenges and value 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; Maier, Baden, et al., 2020), 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.

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. In *The 27th international conference on computational linguistics (coling 2018).*

Baden, C. (2010). *Communication, contextualization, & cognition: Patterns & processes of frames' influence on people's interpretations of the EU constitution.* Delft, The Netherlands: 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* (p. 3-26). New York: Routledge.

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. doi: 10.1080/19312458.2020.1803247

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

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. Retrieved from `https://quanteda.io` doi: 10.21105/joss.00774

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with python: analyzing text with the natural language toolkit.* " O'Reilly Media, Inc.".

Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, *55*(4), 77–84.

Blei, D. M., & Lafferty, J. D. (2006). Dynamic topic models. In *Proceedings of the 23rd*

international conference on machine learning (pp. 113–120). doi: 10.1145/1143844.1143859

Boukes, M., van de Velde, B., Araujo, T., & Vliegenthart, R. (2020). What's the tone? easy doesn't do it: Analyzing performance and agreement between off-the-shelf sentiment analysis tools. *Communication Methods and Measures*, *14*(2), 83–104. doi: 10.1080/19312458.2019.1671966

Boumans, J. W., & Trilling, D. (2016). Taking stock of the toolkit. *Digital Journalism*, *4*, 8-23. doi: 10.1080/21670811.2015.1096598

Brady, H. E. (2019). The challenge of big data and data science. *Annual Review of Political Science*, *22*, 297–323.

Budanitsky, A., & Hirst, G. (2006). Evaluating wordnet-based measures of lexical semantic relatedness. *Computational Linguistics*, *32*(1), 13-47.

Burscher, B., Odijk, D., Vliegenthart, R., & de Rijke, C. H., M.and de Vreese. (2014). Teaching the computer to code frames in the news: Comparing two supervised machine learning approaches to frame analysis. *Communication Methods & Measures*, *8*, 190-206. doi: 10.1080/19312458.2014.937527

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 & Measures*. doi: 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.

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. doi: 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.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacl.*

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 (19328036)*, *13*.

Entman, R. M. (1993). Framing: Toward clarification of a fractured paradigm. *Journal of Communication*, *43*(4), 51-58. doi: 10.1111/j.1460-2466.1993.tb01304.x

Geiss, S., & Monzer, C. (2020). How effect size, sample size and coding accuracy jointly affect hypothesis testing in content analysis: A monte carlo simulation approach. In *70th ica annual conference.*

Goldberg, Y., & Elhadad, M. (2013). Word segmentation, unknown-word resolution, and morphological agreement in a hebrew parsing system. *Computational Linguistics*, *39*(1), 121-160. doi: 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*, 267-297. doi: 10.1093/pan/mps028

Günther, E. (2020). *Topic modeling: Theoretische einordnung algorithmischer themenkonzepte in gegenstand und methodik der kommunikationswissenschaft* (Unpublished doctoral dissertation). WWU Münster.

Günther, E., & Quandt, T. (2016). Word counts and topic models. *Digital Journalism*, *4*(1), 75-88. doi: 10.1080/21670811.2015.1093270

Hanitzsch, T., Hanusch, F., Ramaprasad, J., & de Beer, A. S. (2019). *Worlds of journalism: Journalistic cultures around the globe.* New York: Columbia University Press.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). Most people are not weird. *Nature*, *466*(7302), 29–29.

Hilbert, M., Barnett, G., Blumenstock, J., Contractor, N., Diesner, J., Frey, S., ... others

(2019). Computational communication science: A methodological catalyzer for a maturing discipline. *International Journal of Communication*, *13*. doi: 1932–8036/20190005

Hilbert, M., & Darmon, D. (2020). Large-scale communication is more complex and unpredictable with automated bots. *Journal of Communication*, *70*(5), 670-692.

Kantner, C., & Overbeck, M. (2020). Exploring soft concepts with hard corpus-analytic methods. *Reflektierte algorithmische Textanalyse. Interdisziplinäre (s) Arbeiten in der CRETA-Werkstatt*, 169–189. doi: 10.1515/9783110693973-008

Krippendorff, K. (2004). *Content analysis: An introduction to its methodology* [Book]. Thousand Oaks, CA: Sage.

Lazer, D., & Radford, J. (2017). Data ex machina: introduction to big data. *Annual Review of Sociology*, *43*, 19–39. doi: 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*, 4000-4020.

Liu, B., & Zhang, L. (2012). A survey of opinion mining and sentiment analysis. In *Mining text data* (pp. 415–463). Springer. doi: 10.1007/978-1-4614-3223-4_13

Maier, D., Baden, C., Stoltenberg, D., De Vries, M., & Waldherr, A. (2020). Assessing strategies for topic modeling of multilingual text collections in communication research. In *70st ica annual conference.*

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. doi: 10.5117/CCR2020.2.001.MAIE

Maier, D., Waldherr, A., Miltner, P., Wiedemann, G., Niekler, A., Keinert, A., ... Adam, S. (2018). Applying lda topic modeling in communication research: Toward a valid and reliable methodology. *Communication Methods & Measures*, *12*(2-3), 93-118. doi: 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*, 258-279. doi: 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. In *36th international acm sigir conference on research and development in information retrieval* (p. 889-892). doi: 10.1145/2484028.2484166

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. doi: 10.1080/10584609.2018.1517843

Munro, R. (2020). *Human-in-the-loop machine learning: Active learning, annotation, and human-computer interaction.* New York: Manning.

Nelson, L. K. (2017). Computational grounded theory: A methodological framework. *Sociological Methods & Research.* doi: 10.1177/00491241177297703

Nicholls, T., & Culpepper, P. D. (2020). Computational identification of media frames: strengths, weaknesses, and opportunities. *Political Communication*, 1–23. doi: 10.1080/10584609.2020.1812777

Ophir, Y., Walter, D., & Marchant, E. R. (2020). A collaborative way of knowing: Bridging computational communication research and grounded theory ethnography. *Journal of Communication*, *70*(3), 447–472. doi: 10.1093/joc/jqaa013

Overbeck, M., Baden, C., Aharoni, T., & Tenenboim-Weinblatt, K. (2021). Beyond sentiment: An algorithmic strategy for identifying evaluations within large text corpora. In *71st ica annual conference.*

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. doi: 10.1177/1940161207312676

Pipal, C., Schoonvelde, M., & Schumacher, G. (2019). Jst and rjst in political speech. OSF Preprints.

Possler, D., Bruns, S., & Niemann-Lenz, J. (2019). Data is the new oil–but how do we drill it? pathways to access and acquire large data sets in communication science. *International Journal of Communication (19328036)*, *13*.

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. doi: 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. In *Proceedings of the 2009 conference on empirical methods in natural language processing* (p. 248-256).

Rauh, C., & Schwalbach, J. (2020). The parlspeech v2 data set: Full-text corpora of 6.3 million parliamentary speeches in the key legislative chambers of nine representative democracies.

Reber, U. (2019). Overcoming language barriers: Assessing the potetial of machine translation and topic modeling for the comparative analysis of multilingual text corpora. *Communication Methods & Measures*, *13*(2), 102-125. doi: 10.1080/19312458.2018.155798

Roberts, M., Stewart, B., Tingley, D., Benoit, K., Stewart, M. B., Rcpp, L., ... KernSmooth, N. (2020). Package 'stm'. *Imports matrixStats, R. & KernSmooth*(2017).

Roberts, M. E. (2016). Introduction to the virtual issue: Recent innovations in text analysis for social science. *Political Analysis*, *24*(V10), 1-6.

Rudkowsky, E., Haselmayer, M., Wastian, M., Jenny, M., Emrich, S., & Sedlmair, M. (2018). More than bags of words: Sentiment analysis with word embeddings. *Communication Methods and Measures*, *12*(2-3), 140-157.

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 and Political Psychology*, *7*(1), 124–143.

Schumacher, G., Berk, N., Pipal, C., Kantorowicz, J., Schoonvelde, M., Traber, D., & de Vries, E. (2020). *EUSpeech V2.* Retrieved from `https://osf.io/preprints/socarxiv/a4uyw/` doi: 10.31235/OSF.IO/A4UYW

Segev, E. (2014). Visible and invisible countries: News flow theory revised. *Journalism*, *16*, 412-428. doi: 10.1177/1464884914521579

Stalpouskaya, K. (2020). *Automatic extraction of agendas for action from news coverage of violent conflict*. Munich: Ludwig Maximilian University.

Tenenboim-Weinblatt, K., & Baden, C. (2018). Gendered communication styles in the news: An algorithmic comparative study of conflict coverage. *Communication Research*. doi: 10.1177/0093650218815383

Tenney, I., Das, D., & Pavlick, E. (2019). Bert rediscovers the classical nlp pipeline. In *Proceedings of the 57th annual meeting of the association for computational linguistics* (pp. 4593–4601).

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.

Theocharis, Y., & Jungherr, A. (2020). Computational social science and the study of political communication. *Political Communication*, 1–22. doi: 10.1080/10584609.2020.1833121

Thor, A., Bornmann, L., Haunschild, R., & Leydesdorff, L. (2020). Which are the influential publications in the web of science subject categories over a long period of time? crexplorer software used for big-data analyses in bibliometrics. *Journal of Information Science*.

Trilling, D., Van De Velde, B., Kroon, A. C., Löcherbach, F., Araujo, T., Strycharz, J., ...

Jonkman, J. G. (2018). Inca: Infrastructure for content analysis. In *2018 ieee 14th international conference on e-science (e-science)* (pp. 329–330). doi: 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*, *13*, 15-22. doi: 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. doi: 10.1080/19312458.2018.1458084

van Atteveldt, W., Ruigrok, N., Takens, J., & Jacobi, C. (2014). Inhoudsanalyse met amcat. *Geraadpleegd via http://vanatteveldt. com/wp-content/uploads/amcatbook. pdf*.

van Atteveldt, W., Sheafer, 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*, 207-222. doi: 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*, 1–20. doi: 10.1080/19312458.2020.1869198

van Atteveldt, W., Welbers, K., & van der Velden, M. (2019). Studying political decision making with automatic text analysis. In *Oxford research encyclopedia of politics.*

Walter, D., & Ophir, Y. (2019). News frame analysis: an inductive mixed-method computational approach. *Communication Methods and Measures*, *13*(4), 248–266. doi: 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.

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. In *70th ica annual conference.*

Wilson, M. C., & Knutsen, C. H. (2020). Geographical coverage in political science research. *Perspectives on Politics*. doi: 10.1017/S1537592720002509

Windsor, L. C. (2020). Advancing interdisciplinary work in computational communication science. *Political Communication*. doi: 10.1080/10584609.2020.1765915

Yarchi, M., Baden, C., & Kligler-Vilenchik, N. (2020). Political polarization on the digital sphere: A cross-platform, over-time analysis of interactional, positional, and affective polarization on social media. *Political Communication*, 1–42. doi: 10.1080/10584609.2020.1785067

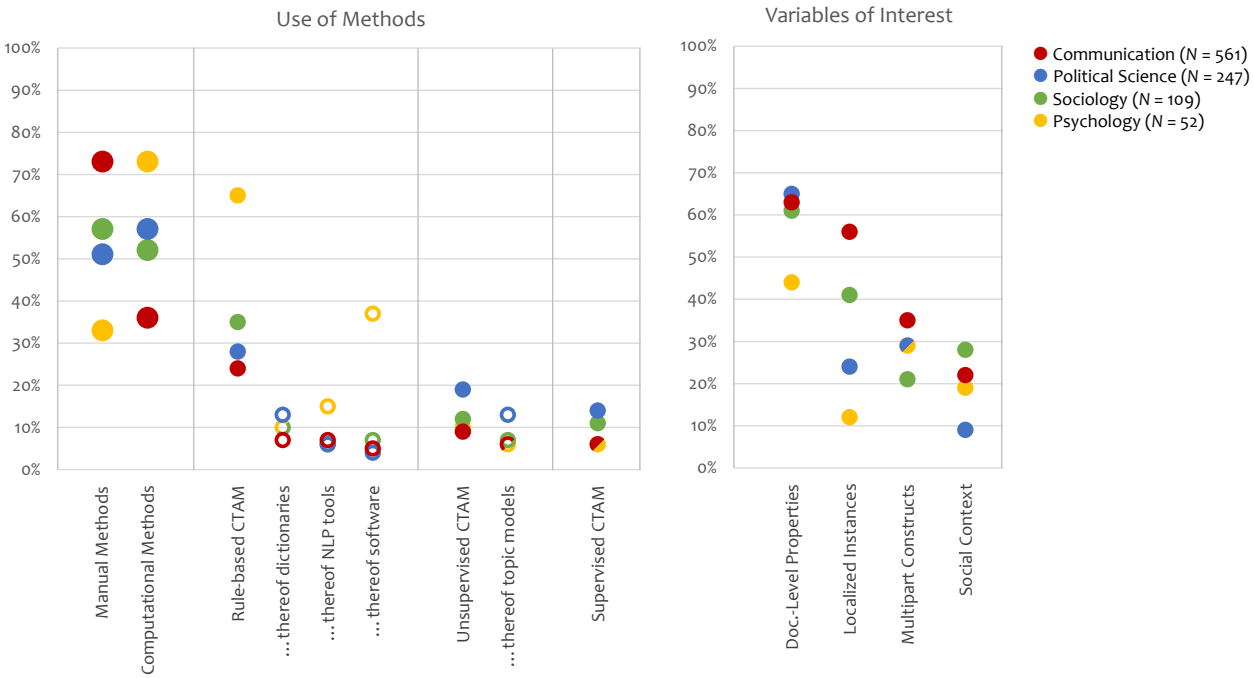
Table 1

Sample Composition

Field	Published Articles	Relevant Articles	Share (%)
Communication Science	6,262	561	9.0
Political Science	5,564	247	4.4
Sociology / Social Science Mathematical Models	10,995	109	1.0
Psychology (Multidisciplinary / Mathematical)	24,459	52	0.2
Total	45,437	854	1.9

Note: Values do not add up to Total because some journals are listed in multiple fields.

Figure 1. Use of Methods & Variables of Interest



Note: Shares relative to all articles using quantitative text analysis within each discipline (see Table 1). Open dots refer to specific subgroups of CTAM that are included within the figures representing the broader types of CTAM displayed to their immediate left (full dots). Values do not add up to 100% because the categories are not mutually exclusive.

Figure 2. Validation Efforts

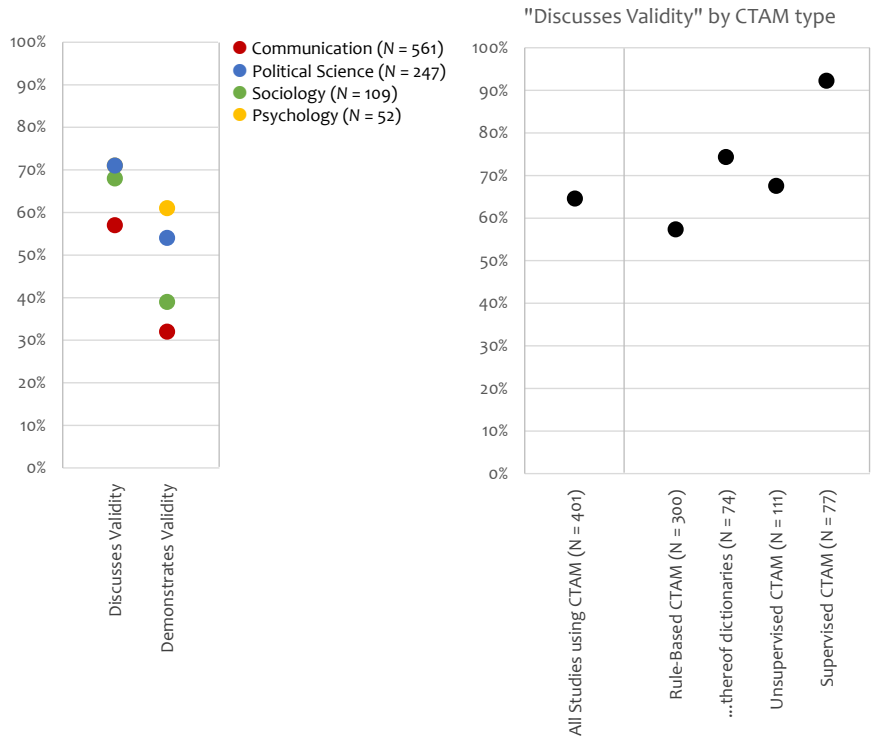


Figure 3. Languages in Studied Materials

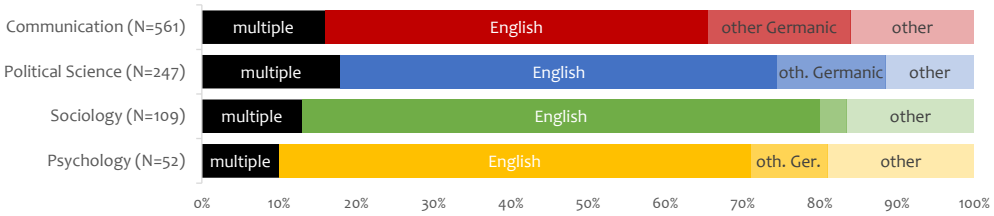
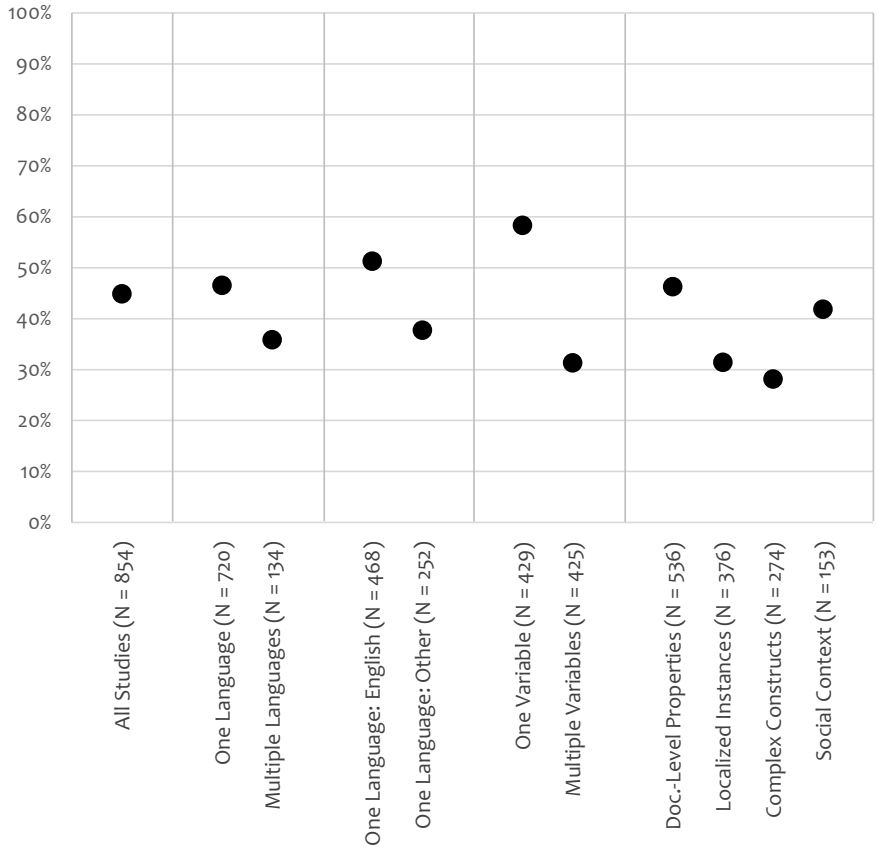


Figure 4. Use of CTAM by Language and Variable of Interest



Appendix A

Overview Included Articles

Table A1

Communication Science

Journal	n
Journal of Advertising	6
Political Communication ∗	39
Journal of Computer-Mediated Communication	5
Communication Methods & Measures	16
Journal of Communication	18
New Media & Society	48
Information, Communication & Society ∗	76
Digital Journalism	45
Communication Monographs	6
Communication Research	15
International Journal of Advertising	10
Human Communication Research	1
Comunicar *(excluded as it publishes in Spanish)*	
Journalism	96
Social Media & Society	63
Policy & Internet	20
International Journal of Press/Politics	46
Mobile Media & Communication	2
European Journal of Communication	44
Public Opinion Quarterly	5
TOTAL	561

∗ journal listed in multiple included categories

Table A2

Political Science

Journal	n
Political Communication *	39
International Organization	3
Environmental Politics	23
American Journal of Political Science	13
Political Analysis	19
American Political Science Review	19
Journal of European Public Policy	11
Annual Review of Political Science	1
British Journal of Political Science	16
Policy Studies Journal	21
Socio-Economic Review	3
Regulation & Governance	10
Journal of Public Administration Research & Theory	7
Political Psychology	19
Review of International Organizations	4
Comparative Political Studies	16
Political Behavior	8
New Political Economy	4
Global Environmental Politics	4
Policy & Politics	7
TOTAL	247

Table A3

Sociology

Journal	n
Annual Review of Sociology	0
American Sociological Review	2
Annals of Tourism Research	0
Information Communication & Society *	76
Sociological Methods & Research *	3
Socio-Economic Review	0
Sociology of Education	1
American Journal of Sociology	0
Work Employment & Society	0
Sociology	1
Scandinavian Journal of Hospitality & Tourism	1
Population & Development Review	0
British Journal of Sociology	2
Work & Occupations	0
Cornell Hospitality Quarterly	0
Gender & Society	2
Sociological Theory	0
Sociology of Sport Journal	0
Sociology of Religion	0
Annual Review of Law & Social Science	0
TOTAL	88

Table A4

Social Science Mathematical Models

Journal	n
Sociological Methods & Research ∗	3
Review of Economics & Statistics	0
Econometrica	1
Structural Equation Modeling	0
Risk Analysis	8
Financial Innovation	0
Journal of Business & Economic Statistics	0
EPJ Data Science	10
Multivariate Behavioral Research	0
Journal of Mathematical Psychology ∗	0
Mathematical Finance	0
Journal of The Royal Statistical Society Series A	1
Econometrics Journal	0
Finance & Stochastics	0
Journal of Educational & Behavioral Statistics ∗	0
Stata Journal	0
Psychometrika ∗	0
Journal of Applied Econometrics	1
System Dynamics Review	0
Journal of Causal Inference	0
TOTAL	24

Table A5

Psychology Mathematical

Journal	n
Behavior Research Methods	30
Psychonomic Bulletin & Review	0
Journal of Mathematical Psychology ∗	0
British Journal of Mathematical & Statistical Psychology	0
Journal of Educational & Behavioral Statistics ∗	0
Psychometrika ∗	0
Educational & Psychological Measurement	0
Applied Psychological Measurement	0
Methodology	0
Journal of Classification	0
Journal of Educational Measurement	0
Applied Measurement in Education	0
Nonlinear Dynamics Psychology & Life Sciences	0
TOTAL	30

Table A6

Psychology Multidisciplinary

Journal	n
Psychological Bulletin	0
Psychological Science in the Public Interest	2
Annual Review of Psychology	0
Psychological Inquiry	0
Psychological Methods	0
Perspectives on Psychological Science	0
Psychological Review	0
American Psychologist	0
Psychological Science	2
Environment & Behavior	0
Current Directions in Psychological Science	0
Computers in Human Behavior	18
European Journal of Psychology Applied to Legal Context	0
Annals of Behavioral Medicine	0
Emotion Review	0
Current Opinion in Psychology	0
Psychosocial Intervention	0
Suicide & Life-Threatening Behavior	2
Journal of Positive Psychology	0
Psychosomatic Medicine	0
TOTAL	22

Appendix B

Codebook

Quantitative Text Analysis Methods

Manual Approaches

- MAN – Manual content analysis, including crowdcoding: Every procedure where in the classification of textual contents is done by a person, for all coded documents[i.e., if only a training corpus is hand-coded and then extended by machine learning to a larger corpus, code as → SML.

Rule-based Approaches

- LNK – Extraction of links, hashtags, @-mentions and other contents that can be recognized by formal shape, but only if the extracted entities are then analyzed insome way [i.e., do not code if they served merely to identify/sample relevant documents/users].

- KEY – Keyword counts, that is, the use of researcher-defined [i.e., not taken from some existing tools] keywords [can be multigrams] whose occurrence is recognized in text; code only if the analysis takes place at the level of these keywords [i.e., how often these exact keywords are used; if multiple keywords refer to the same measured variables, or if there are additional criteria and disambiguation rules considered, code as → DICT].

- DCT – Dictionaries, that is, the use of of researcher-defined [i.e., not taken from some existing tools] dictionaries that operationalize one or more conceptual variables as the occurrence of multiple keywords or keyword combinations[including possible additional criteria and disambiguation rules].

- NLP – Natural Language Processing Tools, that is, any forms of analysis that apply formal algorithms to text in order to extract specific kinds of contents, but without pre-defining the content [e.g., word frequencies, concordances/KWICs, NER, extraction of specific POS, Word Embeddings . . .]

- SOF – Software, that is, any use of rule sets, algorithms and dictionaries that are included in available software or packages and serve to extract specific kinds of contents [most commonly, sentiment tools, LIWC and similar tools; code if them applying between conceptual meaning and textual indicator is part of the software].

Supervised Approaches

- SNA – Semantic Network Analysis, that is, any use of textual co-occurrence patterns used to represent ties between contents [can be hashtags; do not code if the networked entities aren't texts but users or other non-textual entities]

- TOM – Topic Modeling, including related clustering algorithms; code everything that uses textual co-occurrence patterns to obtain some kind of clustering solution of textual contents [i.e., do not code if clustering is applied to entire documents, whichmay be coded under → DSC or → Plag].

- DSC – Document Scaling, including related scoring algorithms; code everything that uses textual contents to rank or arrange documents on some continuous dimension defined by conceptual variables [do not code when the scoring focuses on document similarity regardless of conceptual variables, which is coded under → Plag].

- PLG – Plagiarism Tools, including related algorithms that use textual contents to score documents based on their similarity with one another, but without conceptual anchoring cases/variables (which are coded under → DSC), either using non-conceptual anchors (e.g., plagiarized originals, authored pieces in authorship recognition tools) or dyadic similarity metrics.

Unsupervised Approaches

- SML – Supervised Machine Learning, including active learning and other variants;code all uses wherein human coders make a limited number of classification decisions, which is then interpreted and applied to larger corpora by an algorithm of some sort.
- HCA – Hybrid Content Analysis; code if human classification takes place on the basis of textual patterns derived by an unsupervised algorithm, instead of raw documents.

Other Methods

- BOT – Bot Detection; code if existing software is used to classify entire document collections as bots, based on their textual contents
- MAT – Machine Translation; code if textual contents are subjected to automated translation

Validation Efforts

Discussion of Validity

- PRESENT: Code if the text contains any discussion of efforts made to establish or ascertain that used CTAM offer a valid measurement of textual meanings. Relevant efforts include any procedures aimed to improve the criteria in rule-based CTAM; algorithmic modeling choices justified by operational/validity concerns in unsupervised/supervised CTAM; efforts to establish the agreement of coded outputs with gold standard data, external data (convergent validity) or knowledge (face validity); and any form of algorithmic auditing. To count as validation effort, discussed considerations or procedures must target the textual measurement (not subsequent data analyses); apply measurement validity (not robustness or reliability)

as benchmark; and address validity as a question or challenge to be met, rather than merely assuming validity and interpreting obtained measurements.

- ABSENT: Code otherwise.

Demonstration of Validity

- PRESENT: Code if the text (including any attached or online appendices) contains any data intended to document the extent to which CTAM offer a valid measurement of textual meanings. Relevant data include parametric demonstrations (e.g., precision/recall and F1 scores, correlations for convergent validity); error analyses (e.g., confusion matrices, presentations of common errors); visual displays (e.g., data plots that juxtapose measurement and exteral data or otherwise enable a validity evaluation of measurement); exemplifications (e.g., of criteria designed to improve validity; of measurement outputs that otherwise enable a validity evaluation). To count as validity demonstration, such data must be presented with the express purpose of establishing the validity of applied CTAM.

- ABSENT: Code otherwise.

Languages

Germanic Languages

- DE – German
- DK – Danish
- EN – English
- IS – Icelandic
- NL – Dutch/Flemish
- NO – Norwegian
- SE – Swedish

Romance Languages

- CT – Catalan
- ES – Spanish
- FR – French
- IT – Italian
- PT – Portuguese/Brazilian
- RO – Romanian

Slavic Languages

- BH – Bosnian
- BG – Bulgarian
- CR – Croatian
- CZ – Czech
- PL – Polish
- RU – Russian
- SR – Serbian
- SK – Slovakian
- SL – Slovenian
- UA – Ukrainian

Other Indo-European Languages

- AL – Albanian
- FS – Farsi
- GR – Greek
- LT – Lithuanian
- LV – Latvian
- SN – Sinhalese

Semitic Languages

- AR – Arabic
- HE – Hebrew

Other Eurasian Language Families

- EE – Estonian
- HU – Hungarian
- SF – Suomi/Finnish
- TR – Turkish

East-Asian Language Families

- CN – Chinese
- JP – Japanese
- KR – Korean
- MY – Malay
- PH – Filipino
- TH – Thai
- TM – Tamil
- VN – Vietnamese

Other

- CC – Computer languages
- XX – Agnostic/Whatever
- UN – Impossible to determine

Variables of interest

This code does not (necessarily) capture the overall research focus of the study, but the kind of information that the study aimed to operationally extract from the analyzed

textual content – possibly to be further processed (e.g. to derive diversity scores or assemble larger patterns) or combined with other variables (e.g., from other methods used).

Document-Level Variables

- THEME – Themes, Issues, Topics and Events, any kind of information that focuses on the overall aboutness of a text, and not to specific, localized contents of these.Other than frames, themes do not presume a specific logical internal structure, but simply try to identify a kind of concern in the world that the text is about.
- SENTI – Sentiment or Tone, any kind of information regarding the overall tone or sentiment of a text (or extracted passages) as a whole (if only object-specific evaluations are of interest, code → ATTR). Code only if the distinction is simply on a scale from positive to negative; for more specific overall evaluations, code → QUAL.
- QUALI – Textual qualities, any kind of information regarding a text's over all tendency to express a certain kind of meaning; Qualities range from political ideologies over rhetorical styles to their conformity to theoretically modeled criteria(e.g., the text's quality can refer to its pertinence to certain journalistic styles, an article's possession of certain news values, or the expression of misogynic attitudes;code only if the quality of the text as a whole is in focus, if the focus is on specific instances of misogynic expressions or references to news values, this is probably an → ATTR).
- TYPES – Textual types, any kind of information regarding the classification of the entire text into discrete categories based on their contents; do not code if the categories are overlapping (e.g., a text may have more than one → QUAL); do not code if the text type is assigned based on external knowledge (e.g., that a paper is a tabloid), only if the type is derived from content (e.g., journalistic text sorts:commentary, report, ...)

Localized Variables

- FORMA – Formal properties, any information that can be obtained from the text by formal rules (e.g., counting properties, formatting-based recognition, reading data field) and recorded without the need for interpretation, and requires only interpretation to be used as operationalization of some theoretical construct; this can be anything from properties (e.g., length, position) over embedded data (e.g.,likes, comment counts), to embedded elements (e.g., use of visuals, hashtags).Note: what is of interest here is whether a study wanted to measure the use of hashtags as indicator of something; if formal properties are extracted in order to be further examined with regard to their encoded meaning or content, they do not appear as formal properties, but as whatever information they deliver (below).

- THING – Things, any information that pertains to specific instances wherein the text refers to named entities that have to be recognized/identified (e.g., by having lists of relevant entities, using named entity recognition algorithms) but then do not require interpretation (e.g., references to people, corporations, countries, etc.). This includes the case when the recognized entities are then further classified (e.g., by type, gender, ...) based on external knowledge (e.g., we know that Trump is a politician). If the classification of entities into broader conceptual groups is derived from the text itself (e.g., trying to decide if an actor is presented in a private or professional role), this is coded under → CNST.

- CONST – Constructs, any information that pertains to the specific instances where in the text refers to constructs that require some interpretation to decide whether a specific reference in the text qualifies. Constructs range from abstractions/groupings (e.g., freedom, politicians) to complex theoretically founded variables (e.g., warrants, emotions). Constructs are non-propositional, they can bethought of as complex units, not as claims/statements (which may be coded as → ATTR, PROP).

Multi-Component Variables

- ATTRI – Attributions, any information that attributes certain qualities to a thing or construct. Attributions include object-specific evaluations or issue positions, the presentation of an object in specified ways (e.g., refugees-as-deserving, refugees-as-threat); they consist of an object and something said about that object, but they are not propositional (→ PROP).

- PROPS – Propositions, any information that makes specific claims about a thing or construct. Propositions formulate a qualified link between a pair of objects (e.g.,whether a journalist works for a certain outlet, whether climate change causes wildfires) and cannot be reduced to merely qualifying an object.

- FRAME – Frames and friends, any information that exceeds specific propositions but specifies a more complex set of claims – e.g., by linking causes, states, evaluations;forming narratives, logics, etc. For a variable to be considered a frame, there needs to be some sense of an internal logical structure of the extracted information,wherein different textual contents contribute in different ways to the required information. Frames can refer to specific instances within a text, or to the framing of the text as a whole.

Other Variables

- PRAGM – Pragmatics and Behavior, any information that focuses on the communicative intention expressed in a text or the behavior reflected therein. Code if the focus is on extracting the kind of speech acts (e.g., directives, commissives, assertives), appeals (e.g., to fear, to support) or communicative behavior (e.g.,rebutting, disagreeing, supporting, asking questions).

- RELAS – Relations between texts, any kind of information that does not focus on determining the specific content of one text, but to determine how it relates to other texts (e.g., as a response to, coverage of, plagiarism of, etc., another text).

- OTHER – Other, any other kind of information derived from the text; noted subcategories include cases where the text is merely an instrument to derive information about the world (e.g., learning about users' social relations), and cases when the focus is on identifying in what context(s) specific constructs come up.The above categories are ordered from information derived by looking at attributes of the text, to information derived from looking into the text, to information derived by appraising the text as a whole, to information derived by looking beyond the text(at its relation to the speaker, audience, or other texts). Constructs are in many times nested: For instance, frames consist of propositions and attributions, which in turn involve constructs and things.

Code all that apply, but code at the highest applicable level (e.g., if the study is interested in frames, do not also code attributions that might be part of that frame, unless there are attributions that are of interest besides – and thus independently of – the extracted frames).