Transparency Across Analytic Approaches

Plain Text? Transparency in Computer-Assisted Text Analysis

https://doi.org/10.5281/zenodo.893085

David Romney
Harvard University

Brandon M. Stewart
Princeton University

Dustin Tingley
Harvard University

In political science, research using computer-assisted text analysis techniques has exploded in the last fifteen years. This scholarship spans work studying political ideology,[1] congressional speech,[2] representational style,[3] American foreign policy,[4] climate change attitudes,[5] media,[6] Islamic clerics,[7] and treaty making,[8] to name but a few. As these examples illustrate, computer-assisted text analysis—a prime example of mixed-methods research—allows gaining new insights from long-familiar political texts, like parliamentary debates, and altogether enables the analysis to new forms of political communication, such as those happening on social media.

While the new methods greatly facilitate the analysis of many aspects of texts and hence allow for content analysis on an unprecedented scale, they also challenge traditional approaches to research transparency and replication.[9] Specific challenges range from new forms of data pre-processing and cleaning, to terms of service for websites, which may explicitly prohibit the redistribution of their content. The Statement on Data Access and Research Transparency[10] provides only very general guidance regarding the kind of transparency positivist empirical researchers should provide. In this paper, we consider the application of these general guidelines to the specific context of computer-assisted text analysis to suggest what transparency demands of scholars using such methods.

We explore the implications of computer-assisted text analysis for data transparency by tracking the three main stages of a research project involving text as data: (1) acquisition, where the researcher decides what her corpus of texts will consist of; (2) analysis, to obtain inferences about the research question of interest using the texts; and (3) *ex post* access, where the researcher provides the data and/or other information to allow the verification of her results. To be transparent, we must document and account for decisions made at each stage in the research project. Transparency not only plays an essential role in replication[11] but it also helps to communicate the essential procedures of new methods to the broader research community. Thus transparency also plays a didactic role and makes results more interpretable.

Many transparency issues are not unique to text analysis. There are aspects of acquisition (e.g., random selection), analysis (e.g., outlining model assumptions), and access (e.g., providing replication code) that are important regardless of what is being studied and the method used to study it. These general issues, as well as a discussion of issues specific to traditional qualitative textual analysis, are outside of our purview. Instead, we focus here on those issues that are uniquely important for transparency in the context of computer-assisted text analysis.[12]

1. Where It All Begins: Acquisition

Our first step is to obtain the texts to be analyzed, and even this simple task already poses myriad potential transparency issues. Traditionally, quantitative political science has been dominated by a relatively small number of stable and publicly available datasets, such as the American National Election Survey (ANES) or the World Bank Development Indicators (WDI). However, a key attraction of new text methods is that they open up the possibility of exploring diverse new types of data, not all of which are as stable and publicly available as the ANES or WDI. Websites are taken down and their content can change daily; social media websites suspend users regularly. In rare cases, websites prohibit any scraping at all. Researchers should strive to record the weblinks of the pages they scraped (and when they scraped them), so that the data can be verified via the Wayback Machine[13] if necessary and available. This reflects a common theme throughout this piece: Full transparency is difficult to impossible in many situations, but researchers should strive to achieve as much transparency as possible.

David Romney is a PhD Candidate in the Department of Government at Harvard University. He is online at dromney@fas.harvard.edu and http://scholar.harvard.edu/dromney. Brandon Stewart is Assistant Professor of Sociology at Princeton University. He is online at bms4@princeton.edu and www.brandonstewart.org. Dustin Tingley is Professor of Government in the Department of Government at Harvard University. He is online at dtingley@gov.harvard.edu and http://scholar.harvard.edu/dtingley. The authors are grateful to Richard Nielsen, Margaret Roberts and the editors for insightful comments.

[1] Laver, Benoit, and Garry 2003.
[2] Quinn et al. 2010.
[3] Grimmer 2010.
[4] Milner and Tingley 2015.
[5] Tvinnereim and Fløttum 2015.
[6] Gentzkow and Shapiro 2010.
[7] Nielsen 2013; Lucas etal. 2015.
[8] Spirling 2011.
[9] King 2011.
[10] DA-RT 2014.
[11] King 2011.
[12] See Grimmer and Stewart (2013) for a review of different text analysis methods.
[13] http://archive.org/web/.

Sometimes researchers are lucky enough to obtain their data from someone who has put it in a tractable format, usually through the platform of a text analytics company. However, this comes with its own problems—particularly the fact that these data often come with severe restrictions on access to the actual texts themselves and to the models used for analysis, making validation difficult. For this reason, some recommend against using black-box commercial tools that do not provide access to the texts and/or the methods used to analyze them.[14] Nevertheless, commercial platforms can provide a valuable source of data that would be too costly for an individual researcher to collect. For instance, Jamal et al.[15] use Crimson Hexagon, a social media analytics company, to look at historical data from Arabic Twitter that would be difficult and expensive to obtain in other ways.[16] In situations where researchers do obtain their data from such a source, they should clearly outline the restrictions placed on them by their partner as well as document how the text could be obtained by another person. For example, in the supplementary materials to Jamal et al.,[17] the authors provide extensive detail on the keywords and date ranges used to create their sample.

A final set of concerns at the acquisition stage is rooted in the fact that, even if texts are taken from a "universe" of cases, that universe is often delimited by a set of keywords that the researcher determined, for example when using Twitter's Streaming API. Determining an appropriate set of keywords is a significant task. First, there are issues with making sure you are capturing all instances of a given word. In English this is hard enough, but in other languages it can be amazingly complicated—something researchers not fluent in the languages they are collecting should keep in mind. For instance, in languages like Arabic and Hebrew you have to take into account gender; plurality; attached articles, prepositions, conjunctions, and other attached words; and alternative spellings for the same word. More than 250 iterations of the Arabic word for "America" were used by Jamal et al.[18] And this number does not take into account the synonyms (e.g., "The United States"), metonyms (e.g., "Washington"), and other associated words that ought to be selected on as well. Computer-assisted selection of keywords offers one potential solution to this problem. For instance, King, Lam, and Roberts[19] have developed an algorithm to help select keywords and demonstrated that it works on English and Chinese data. Such an approach would help supplement researchers' efforts to provide a "universe" of texts related to a particular topic, to protect them from making *ad hoc* keyword selections.

[14] Grimmer and Stewart 2013, 5.

[15] Jamal et al. 2015.

[16] Jamal et al. 2015. In this case, access to the Crimson Hexagon platform was made possible through the Social Impact Program, which works with academics and non-profits to provide access to the platform. Other work in political science has used this data source, notably King, Pan, and Roberts 2013.

[17] Jamal et al. 2015.

[18] Jamal et al. 2015.

[19] King, Lam, and Roberts 2014.

Thus a commitment to transparency at the acquisition stage requires that the researcher either provide the texts they are analyzing, or provide the capacity for the text corpus to be reconstructed. Even when the texts themselves can be made available it is incumbent on the researcher to describe how the collection of texts was defined so that readers understand the universe of texts being considered. When such information is missing it can be difficult for readers and reviewers alike to assess the inferences we can draw from the findings presented.

2. The Rubber Hits the Road: Training and Analysis

Researchers also should be transparent about the decisions made once the data have been collected. We discuss three areas where important decisions are made: text processing, selection of analysis method, and addressing uncertainty.

Processing

All research involves data cleaning, but research with unstructured data involves more than most. The typical text analysis workflow involves several steps at which the texts are filtered and cleaned to prepare them for computer-assisted analysis. This involves a number of seemingly innocuous decisions that can be important.[20]Among the most important questions to ask are: Did you stem (i.e., map words referring to the same concept to a single root)? Which stemmer did you use? If you scraped your data from a website, did you remove all html tags? Did you remove punctuation and common words (i.e. "stop" words), and did you prune off words that only appear in a few texts? Did you include *bigrams* (word pairs) in your analysis? Although this is a long list of items to consider, each is important. For example, removing common stop words can obscure politically interesting content, such as the role of gendered pronouns like 'her' in debates on abortion.[21] Since inferences can be susceptible to these processing decisions, their documentation, in the paper itself or in an appendix, is essential to replication.

Analysis

Once the texts are acquired and processed, the researcher must choose a method of analysis that matches her research objective. A common goal is to assign documents to a set of categories. There are broadly speaking three approaches available: keyword methods, where categories are based on counts of words in a pre-defined dictionary; supervised methods, where humans classify a set of documents by hand (called the training set) to teach the algorithm how to classify the rest; and unsupervised methods, where the model simultaneously estimates a set of categories and assign texts to them.[22]An important part of transparency is justifying the use of a particular approach and clarifying how the method provides leverage to answer the research question. Each method then in turn entails particular transparency considerations.

[20] See Lucas et al. (2015) for additional discussion.

[21] Monroe, Colaresi and Quinn 2008, 378.

[22] Grimmer and Stewart 2013, Figure 1.

In supervised learning and dictionary approaches, the best way to promote transparency is to maintain, and subsequently make available, a clear codebook that documents the procedures for how humans classified the set of documents or words.[23] Some useful questions to consider are:

> How did the researcher determine the words used to define categories in a dictionary approach?
>
> How are the categories defined and what efforts were taken to ensure inter-coder reliability?
>
> Was your training set selected randomly?

We offer two specific recommendations for supervised and dictionary methods. First, make publicly available a codebook which documents category definitions and the methods used to ensure intercoder reliability.[24] Second, the researcher should report the method used to select the training set of documents—ideally random selection unless there is a compelling reason to select training texts based on certain criteria.[25]

When using unsupervised methods, researchers need to provide justifications for a different set of decisions. For example, such models typically require that the analyst specify the number of topics to be estimated. It is important to note that there is not necessarily a "right" answer here. Having more topics enables a more granular view of the data, but this might not always be appropriate for what the analyst is interested in. The appendix provides one such example using data from political blogs. Furthermore, as discussed by Roberts, Stewart, and Tingley,[26] topic models may have multiple solutions even for a fixed number of topics.[27] These authors discuss a range of stability checks and numerical initialization options that can be employed, which enable greater transparency.

Uncertainty

The final area of concern is transparency in the appropriate incorporation of uncertainty into our analysis. Text analysis is often used as a measurement instrument, with the estimated categorizations used in a separate regression model to, for example, estimate a causal effect. While these two stage models (measure, then regress) are attractive in simplicity, they often come with no straightforward way to incorporate measurement uncertainty. The Structural Topic Model (STM),[28] building on previous work by Treier and Jackman,[29] provides one approach to explicitly building-in estimation uncertainty. But regardless of how estimation uncertainty is modeled, in the spirit of transparency, it is important to acknowledge the concern and specify how it has been incorporated in the analysis.

Another form of uncertainty derives from the research process itself. The decisions discussed in this section—decisions about processing the text, determining dictionary words, or determining categories/the number of categories—are often not the product of a single *ex ante* decision. In reality, the process is iterative. New processing procedures are incorporated, new dictionary words or categories are discovered, and more (or fewer) unsupervised topics are chosen based on the results of previous iterations of the model. While many research designs involve iterative analysis, using text as data often involves more iterations than other research designs. Iteration is a necessary step in the development of any text analysis model, and we are not advocating that researchers unyieldingly devote themselves to their initial course of action. However, we argue that researchers should clearly state when the process was iterative and which aspects of it were iterative. This documentation of how choices were made, in combination with a codebook clearly documenting those choices, helps to minimize remaining uncertainty.

3. On the Other Side: Presentation and Data Access

It is at the access stage of a project that we run into text analyses's most common violation of transparency: not providing replication data. A common reason for not providing these data is that doing so would violate the intellectual property rights of the content provider (website, news agency, etc.), although sometimes the legal concerns are more complicated, such as in the case of research on the Wikileaks cables[30] or Jihadist texts.[31] Sometimes other reasons prevent the researcher from providing the data, such as the sensitivity of the texts or ethical concerns for the safety of those who produced them. Researchers who seek to maximize transparency have found avariety of ways around these concerns. Some provide the document term matrix as opposed to the corpus of texts, allowing for a partial replication of the analysis. For example, the replication material for the recent article by Lucas et al.[32] provides the document term matrix for fatwas used in their analy-

[23] Usually those following the instructions in such a codebook are either the researchers themselves or skilled RAs; however, we note with interest a new approach proposed by Benoit et al. (2015), where this portion of the research process is completed via crowdsourcing. In situations where this approach is applicable, it can potentially make replication easier.

[24] See Section 5 of Hopkins et al. (2010) for some compact guidance on developing a codebook. A strong applied example of a codebook is found in the appendix of Stewart and Zhukov (2009).

[25] Random selection of training cases is necessary to ensure that the training set is representative of the joint distribution of features and outcomes (Hand 2006, 7–9; Hopkins and King 2010, 234). Occasionally alternate strategies can be necessary in order to maintain efficiency when categories are rare (Taddy 2013; Koehler-Derrick, Nielsen, and Romney 2015); however, these strategies should be explicitly recognized and defended.

[26] Roberts, Stewart, and Tingley 2015.

[27] Topic models, like Latent Dirichlet Allocation (Blei, Ng, and Jordan 2003) and the Structural Topic Model (Roberts, Stewart, and Tingley 2015), are non-convex optimization problems and thus can have local modes.

[28] See Roberts et al. 2014.

[29] Treier and Jackman 2008.

[30] Gill and Spirling 2015.

[31] Nielsen 2013.

[32] Lucas et al. 2015.

sis. One reason is that some of these fatwas are potentially under copyright and reproducing them would infringe on the rights of their owners, whereas using them for text analysis falls under standard definitions of "fair use" and is not an infringement on copyright. Another possible problem is that disseminating the complete text of the jihadist fatwas in the Lucas et al. data set may raise ethical concerns or legal concerns under US anti-terrorism law—releasing document-term matrices avoids this issue as well. Others provide code allowing scholars who seek to replicate the analysis to obtain the same dataset (either through licensing or web scraping).[33] These are perhaps the best available strategies when providing the raw texts is impossible, but they are each at least partially unsatisfying because they raise the cost of engaging with and replicating the work, which in turn decreases transparency. While intellectual property issues can complicate the creation of replication archives, we should still strive to always provide enough information to replicate a study.

There are also other post-analysis transparency concerns, which are comparatively rarely discussed. We cover two of them here. One of them concerns the presentation of the analysis and results. Greater steps could be taken to allow other researchers, including those less familiar with text analysis or without the computing power to fully replicate the analysis, the opportunity to "replicate" the interpretive aspects of the analysis. As an example, the researcher could set up a browser to let people explore the model and read the documents within the corpus, recreating the classification exercise for those who want to assess the results.[34] With a bit of work, this kind of "transparency through visualization" could form a useful transparency tool.

The second presentation issue relates to the unit of analysis at which research conclusions are drawn. Text analysis is, naturally, done at the text level. However, that is not necessarily the level of interest for the project. Those attempting to use Twitter to measure public opinion, for instance, are not interested in the opinions of the Tweets themselves. But when we present category proportions of texts, that is what we are measuring. As a consequence, those who write the most have the loudest "voice" in our results. To address this issue, we recommend that researchers be more transparent about this concern and either come up with a strategy for aggregating up to the level of interest or justify using texts as the level of analysis.

[33] For example, O'Connor, Stewart, and Smith (2013) use a corpus available from the Linguistic Data Consortium (LDC) which licenses large collections of texts. Although the texts cannot themselves be made publicly available, O'Connor, Stewart, and Smith (2013) provide scripts which perform all the necessary operations on the data as provided by the LDC, meaning that any researcher with access to an institutional membership can replicate the analysis.

[34] One example is the stmBrowser package (Freeman et al., 2015); see also the visualization at the following link: http://pages.ucsd.edu/~meroberts/stm-online-example/index.html.

4. Conclusion

Computer-assisted text analysis is quickly becoming an important part of social scientists' toolkit. Thus, we need to think carefully about the implications of these methods for research transparency. In many ways the concerns we raise here reflect general concerns about transparent and replicable research.[35] However, as we have highlighted, text analysis produces a number of idiosyncratic challenges for replication—from legal restrictions on the dissemination of data to the numerous, seemingly minor, text processing decisions that must be made along the way.

There is no substitute for providing all the necessary data and code to fully replicate a study. However, when full replication is just not possible, we should strive to provide as much information as is feasible given the constraints of the data source. Regardless, we strongly encourage the use of detailed and extensive supplemental appendices, which document the procedures used.

Finally, we do want to emphasize the silver lining for transparency. Text analysis of any type requires an interpretive exercise where meaning is ascribed by the analyst to a text. Through validations of document categories, and new developments in visualization, we are hopeful that the interpretive work can be more fully shared with the reader as well. Providing our readers the ability to not only evaluate our data and models, but also to make their own judgments and interpretations, is the fullest realization of research transparency.

Appendix: An Illustration of Topic Granularity

This table seeks to illustrate the need to be transparent about how researchers choose the number of topics in an unsupervised model, discussed in the paper. It shows how estimating a topic model with different numbers of topics unpacks content at different levels of granularity. The table displays the results of a structural topic model run on the poliblog5k data, a 5,000 document sample of political blogs collected during the 2008 U.S. presidential election season. The dataset is contained in the stm R package (http://cran.r-project.org/web/packages/stm/). The model was specified separately with 5 (left column), 20 (middle column), and 100 (right column) topics.[36] Topics across the different models are aligned according to the correlation in the document-topic loadings—for example, document-topic loadings for "energy" and "financial crisis" in the 20-topic model were most closely matched with that of "economics" in the 5-topic model. With each topic in the left and middle columns, we include words that are highly correlated with those topics. While expanding the number of topics would not necessarily change the substantive conclusions of the researcher, the focus does shift in a way that may or may not be appropriate for a given research question.[37]

[35] King 2011.

[36] We used the following specification: stm(poliblog5k.docs, poliblog5k.voc, K=5, prevalence=~rating, data=poliblog5k.meta, init.type="Spectral") with package version 1.0.8

[37] More specifically topics are aligned by correlating the topic-

Economics (tax, legisl, billion, compani, econom)	Energy (oil, energi, tax, economi, price)	**Jobs**
		Off-shore Drilling, Gas
		Taxes
		Recession/Unemployment
		Fuel Pricing
	Financial Crisis (financi, bailout, mortgag, loan, earmark)	**Earmarks**
		Inflations/Budget
		Bailouts
		Mortgage Crisis
		Unions
Elections (hillari, poll, campaign, obama, voter)	**Parties** (republican, parti, democrat, conserv, Pelosi)	**Parties**
		Congressional Leadership
		Elections
		Corruption/Pork
	Congressional Races (franken, rep, coleman, smith, Minnesota)	**Minnesota Race**
		Oregon Race
		Martin Luther King
	Biden/Lieberman (biden, joe, debat, lieberman, senat)	**Lieberman Campaign**
		Debate Night
		Obama Transition Team
		Calls/Meetings
		Senate Votes
		Biden as Running Mate
	Primaries (poll, pennsylvania, virginia, percent, margin)	**Polls**
		Battleground States
	Republican General (palin, mccain, sarah, john, Alaska)	**Attack Adds**
		Joe the Plumber
		Gibson Interview
		McCain Campaign
		Palin
		Giuliani
	Candidates (hillari, clinton, deleg, primari, Edward)	**Clinton**
		Republican Primary Field
		DNC/RNC
		Democratic Primary Field

document loadings (theta) and choosing the topic pairings that maximize the correlation. Topics are then annotated using the 5 most probable words under the given topic-word distribution. We assigned the labels (in bold) based on manual inspection of the most probable words. The most probable words were omitted from the 100 topic model due to space constraints.

References

Benoit, Kenneth, Drew Conway, Benjamin E. Lauderdale, Michael Laver, and Slava Michaylov. 2015. "Crowd-Sourced Text Analysis: Reproducible and Agile Production of Political Data." *American Political Science Review*, forthcoming (available at http://eprints.lse.ac.uk/62242/, last accessed 7/5/2015).

Blei, David M., Andrew Y. Ng, and Michael I. Jordan. 2003. "Latent Dirichlet Allocation." *The Journal of Machine Learning Research* vol.3 no.4/5: 993–1022.

"Data Access and Research Transparency (DA-RT): A Joint Statement by Political Science Journal Editors." 2014. At http://media.wix.com/ugd/fa8393_da017d3fed824cf587932534c860ea25.pdf (last accessed 7/10/2015).

Freeman, Michael, Jason Chuang, Margaret Roberts, Brandon Stewart, and Dustin Tingley. 2015. *stmBrowser: An R Package for the Structural Topic Model Browser*. https://github.com/mroberts/stmBrowser (last accessed 6/25/2015).

Gentzkow, Matthew, and Jesse M. Shapiro. 2010. "What Drives Media Slant? Evidence from US Daily Newspapers." *Econometrica* vol. 78, no. 1: 35–71.

Gill, Michael, and Arthur Spirling. 2015. "Estimating the Severity of the Wikileaks U.S. Diplomatic Cables Disclosure." *Political Analysis* vol. 23, no. 2: 299–305.

Grimmer, Justin. 2010. "A Bayesian Hierarchical Topic Model for Political Texts: Measuring Expressed Agendas in Senate Press Releases." *Political Analysis* vol. 18, no. 1: 1–35.

Grimmer, Justin, and Brandon M. Stewart. 2013. "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts." *Political Analysis* vol. 21, no. 2: 1–31.

Hand, David J. 2006. "Classifier Technology and the Illusion of Progress." *Statistical Science* vol. 21, no. 1: 1–15.

Hopkins, Daniel and Gary King. 2010. "A Method of Automated Nonparametric Content Analysis for Social Science." *American Journal of Political Science* vol. 54, no. 1: 229–247.

Hopkins, Daniel, Gary King, Matthew Knowles, and Steven Melendez. 2010. "ReadMe: Software for Automated Content Analysis." Online at http://j.mp/1KbkQ9j (last accessed 7/5/2015).

Jamal, Amaney A., Robert O. Keohane, David Romney, and Dustin Tingley. 2015. "Anti-Americanism and Anti-Interventionism in Arabic Twitter Discourses." *Perspectives on Politics* vol. 13, no. 1: 55–73.

King, Gary. 2011. "Ensuring the Data Rich Future of the Social Sciences." *Science* vol.331 no.6018: 719–721.

King, Gary, Jennifer Pan, and Margaret E. Roberts. 2013. "How Censorship in China Allows Government Criticism but Silences Collective Expression." *American Political Science Review* vol.107 no.2: 326–343.

Contentious Issues (guy, wright, church, media, man)	Online (linktocommentspostcount, postcounttb, thing, guy, think)	**Apologies**
		Liberal/Conservative Think Tanks
		Media/Press
		Books Writing
		Religious Words
		Internet Sites
	Social Issues (abort, school, children, gay, women)	**Stem Cell Research**
		Gay Rights
		Education
		Abortion (Religious)
		Abortion (Women's Rights)
		Health Care
		Family
	Hollywood (doesn, film, didn, isn, eastern)	**Radio Show**
		Talk Shows
		Emotion Words
		Blogging
		Memes ("Messiah", "Maverick")
		Films and Hollywood
		Eating/Drinking
	Obama Controversies (wright, ayer, barack, obama, black)	**Obama Fundraising**
		Ayer Issues
		Speeches
		Jewish Community Organization
		Wright
	Climate Change/News (warm, publish, global, newspap, stori)	**Climate Change**
		Newspapers
		Pentagon Stories
		Violence in News
		Environmental Issues
Legal/Torture (court, investig, tortur, justic, attorney)	**Torture** (legisl, tortur, court, constitut, law)	**Bipartisan Legislation**
		CIA Torture
		Rule of Law
		FISA Surveillance
		Supreme Court
		Guantanomo
	Presidency (rove, bush, fox, cheney, white)	**Fox News/ Rove**
		Cheney Vice Presidency
		Websites
		Bush Legacy
		White House
	Voting Issues (immigr, acorn, illeg, union, fraud)	**Voter Fraud**
		California Gun Laws
		Illegal Immigration
	Blagojevich and Scandals (investig, blagojevich, attorney, depart, staff)	**Blagojevich**
		Steven
		Jackson
		Lobbying
		Attorney Scandal
		Johnson
Foreign Military (israel, iran, iraqi, troop, Russia)	**Middle East** (israel, iran, hama, isra, iranian)	**Israel**
		Iran Nuclear Weapons
		Saddam Bin Laden Link
		Terrorism in Middle East
	Iraq/Afghanistan Wars (iraqi, troop, iraq, afghanistan, pentagon)	**Iraqi Factions**
		Pakistan/Afghanistan
		Withdrawal from Iraq
		Surge in Iraq
		Veterans
	Foreign Affairs (russia, world, russian, georgia, democracy)	**Russia and Georgia**
		Nuclear North Korea
		Rice and Foreign Policy
		Opposition Governments
		American Vision

King, Gary, Patrick Lam, and Margaret E. Roberts. 2014. "Computer-assisted Keyword and Document Set Discovery from Unstructured Text." Unpublished manuscript, Harvard University: http://gking.harvard.edu/publications/computer-Assisted-Keyword-And-Document-Set-Discovery-Fromunstructured-Text (last accessed 7/5/2015).

Koehler-Derrick, Gabriel, Richard Nielsen, and David A. Romney. 2015. "The Lies of Others: Conspiracy Theories and Selective Censorship in State Media in the Middle East." Unpublished manuscript, Harvard University; available from the authors on request.

Laver, Michael, Kenneth Benoit, and John Garry. 2003. "Extracting Policy Positions from Political Texts Using Words as Data." *American Political Science Review* vol 97, no. 2: 311–331.

Lucas, Christopher, Richard Nielsen, Margaret Roberts, Brandon Stewart, Alex Storer, and Dustin Tingley. 2015. "Computer Assisted Text Analysis for Comparative Politics." *Political Analysis* vol.23 no.2: 254–277.

Milner, Helen V. and Dustin H. Tingley. 2015. *Sailing the Water's Edge: Domestic Politics and American Foreign Policy*. Princeton University Press.

Monroe, Burt L., Michael P. Colaresi, and Kevin M. Quinn. 2008. "Fightin' Words: Lexical Feature Selection and Evaluation for Identifying the Content of Political Conflict." *Political Analysis* vol. 16, no. 4: 372–403.

Nielsen, Richard. 2013. "The Lonely Jihadist: Weak Networks and the Radicalization of Muslim Clerics." Ph.D. dissertation, Harvard University, Department of Government. (Ann Arbor: ProQuest/UMI Publication No. 3567018).

O'Connor, Brendan, Brandon M. Stewart and Noah A. Smith. 2013. "Learning to Extract International Relations from Political Context." In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, Volume 1: Long Papers* (Sofia, Bulgaria): 1094-1104. Online at http://aclanthology.info/events/acl-2013 and http://aclweb.org/anthology/P/P13/P13-1108.pdf (last accessed 7/10/2015).

Quinn, Kevin M., Burt L. Monroe, Michael Colaresi, Michael H. Crespin, and Dragomir R. Radev. 2010. "How to Analyze Political Attention with Minimal Assumptions and Costs." *American Journal of Political Science* vol. 54, no. 1: 209–228.

Roberts, Margaret, Brandon Stewart, and Dustin Tingley. 2015. "Navigating the Local Modes of Big Data: The Case of Topic Models." In *Computational Social Science: Discovery and Prediction*, edited by R. Michael Alvarez. Cambridge: Cambridge University Press, forthcoming.

Roberts, Margaret E., Brandon M. Stewart, Dustin Tingley, Christopher Lucas, Jetson Leder-Luis, Shana Kushner Gadarian, Bethany Albertson, and David G. Rand. 2014. "Structural Topic Models for Open-Ended Survey Responses." *American Journal of Political Science* vol. 58 no. 4: 1064–1082.

Spirling, Arthur. 2011. "US Treaty Making with American Indians: Institutional Change and Relative Power, 1784–1911." *American Journal of Political Science* vol. 56, no. 1: 84–97.

Stewart, Brandon M., and Yuri M Zhukov. 2009. "Use of Force and Civil-Military Relations in Russia: An Automated Content Analysis." *Small Wars & Insurgencies* vol. 20, no. 2: 319–343.

Taddy, Matt. 2013. "Measuring Political Sentiment on Twitter: Factor Optimal Design for Multinomial Inverse Regression." *Technometrics* vol. 55, no. 4: 415–425.

Treier, Shawn, and Simon Jackman. 2008. "Democracy as a Latent Variable." *American Journal of Political Science* vol. 52, no. 1: 201–217.

Tvinnereim, Endre and Kjersti Fløttum. 2015. "Explaining Topic Prevalence in Answers to Open-Ended Survey Questions about Climate Change." Nature Climate Change, forthcoming (available at http://dx.doi.org/ 10.1038/nclimate2663, last accessed 7/5/2015).

Transparency Standards in Qualitative Comparative Analysis

Claudius Wagemann
Goethe University, Frankfurt

Carsten Q. Schneider
Central European University, Budapest

When judging the usefulness of methods, it is not only their technical principles that matter, but also how these principles are then translated into applied practice. No matter how well developed our techniques and methods are, if their usage runs against their spirit, they cannot be what the originally ancient Greek word "method" literally means: a "way towards a goal." Standards of best practice are therefore important components of methodological advancement, if such standards are recognized for what they ought to be: transitory condensations of a shared understanding that are valid until improved.

The more popular a specific method becomes, the greater the need for shared understandings among users. This was our motivation for proposing a "code of good standards" for Qualitative Comparative Analysis (QCA).[1] Due to the transitory nature of any such list, we subsequently provided an update.[2] Transparency is one of the major underlying themes of this list.

QCA is the most formalized and widespread method making use of set-analytic thinking as a fundamental logical basis for qualitative case analysis.[3] The goal consists of the identification of sufficient and necessary conditions for outcomes, and their derivates, namely INUS and SUIN conditions.[4] Almost by default, QCA reveals conjunctural causation (i.e., conditions that do not work on their own, but have to be combined with one another); equifinality (where more than one conjunction produces the outcome in different cases); and asymmetry (where the complement of the phenomenon is explained in

Claudius Wagemann is Professor for Political and Social Sciences and Dean of Studies at Goethe University, Frankfurt. He is online at wagemann@soz.uni-frankfurt.de and http://www.fb03.uni-frankfurt.de/politikwissenschaft/wagemann. Carsten Q. Schneider is Full Professor and Head of the Political Science Department at the Central European University, Budapest. He is online at schneider@ceu.edu and http://people.ceu.edu/carsten-q_schneider.

[1] Schneider and Wagemann 2010.

[2] In Schneider and Wagemann 2012, 275ff.

[3] Goertz and Mahoney 2012, 11.

[4] INUS stands for "Insufficient but Necessary part of a condition which is itself Unnecessary but Sufficient for the result" (Mackie 1965, 246). SUIN stands for "Sufficient but Unnecessary part of a factor that is Insufficient but Necessary for the result" (Mahoney, Kimball, and Koivu 2009, 126).