Humans versus computers: Manual and computational estimates of the design of international institutions

Andreas Dür [1] Lisa Lechner [2]

[1]University of Salzburg, andreas.duer@sbg.ac.at
[2]University of Salzburg, lisa.lechner@sbg.ac.at

Abstract

International relations scholars invest a lot of effort in manually coding treaties, with the aim of arriving at reliable and valid measures of the scope and depth of international institutions. Since manual coding is labor intensive, difficult and potentially unreliable, we assess whether computational text analysis can substitute or at least complement human coders. We do so by applying various types of computational text analysis, from simple word counts to topic models, to the measurement of the institutional design of preferential trade agreements. A comparison of the resulting measures with manual coding indicates that computational text analysis can indeed be helpful in measuring the design of international institutions. Some approaches, however, work better than others, largely independent of treaty characteristics.

International institutions strongly vary in their design (Koremenos et al., 2001). Recognition of this variation has inspired empirical research that tries to measure different dimensions of institutional design, including agreements' depth, flexibility, scope, and degree of centralization (Allee and Peinhardt, 2014; Baccini et al., 2015; Hooghe et al., 2016; Koremenos, 2016; Kucik, 2012; Mitchell, 2003; Young and Zürn, 2006). Much of this work relies on the manual coding of legal texts. This approach, however, comes with several drawbacks. First, the manual coding of international agreements is labor-intensive. Whether scholars study the design of international organizations, bilateral investment treaties, environmental treaties or preferential trade agreements (PTAs), manually coding hundreds of agreements is a daring task. Second, manually coding legal texts is difficult. Even for experienced researchers it can be tricky to assess whether, for example, a services chapter in a PTA follows a positive or a negative list approach. Finally, manual coding begs the question of reliability: would a replication exercise that repeats the manual coding lead to the same results?

These drawbacks of manual coding make it desirable to look for alternative or at least complementary approaches. Computational text analysis may offer a solution to many of the problems identified above. This approach has already been employed to derive the positions of political parties (Proksch and Slapin, 2009a, 2010; Slapin and Proksch, 2008), to measure interest groups' positions on specific issues (Klüver, 2009), and to identify judicial opinions (Dyevre, 2016; Evans et al., 2007). So far, however, computational text analysis has hardly been used to analyze international treaties, with the exception of a few attempts at studying treaty similarity (Allee and Elsig, 2016; Alschner and Skougarevskiy, 2016; Kim, 2015). As far as we know, there are no attempts at applying this method to the measurement of specific dimensions of the design of international agreements. If automated or semi-automated content analysis could be used to assess the design of international institutions, this would make future research in this field much easier.

Three reasons likely explain why research on the design of international institutions has yet to embrace computational text analysis. First, international agreements are legal texts that are written in a technical jargon making it hard to identify underlying, latent dimensions. Texts capturing positions of actors provide obvious dimensions: a left-right positioning for political parties, a pro-anti scale for interest groups, and a guilty-not-guilty scale for judges. Defining dimensions of international agreements is potentially more challenging. Second, international agreements can vary widely in terms of text structure, which complicates computational analysis. Third, international agreements also vary with respect to other aspects. For example, some include annexes whereas others do not, and some are bilateral whereas others are multilateral. There is also a time dimension to most sets of international agreements, with older agreements possibly following a different template than newer agreements. These differences can make comparisons across agreements difficult.

We explore whether despite these challenges, automated or semi-automated text analysis can be a useful tool to measure the design of international institutions. Concretely, we compare different approaches to measuring the scope and depth of international agreements. By scope, we understand the number of legal areas covered by an international agreement. A PTA, for example, can cover areas such as services, foreign direct investments, competition policy, and environmental protection. By depth we understand the extent to which a specific issue is regulated within an international agreement. Depth is likely to be even more difficult to capture than scope using computational approaches, as it is less manifest. Focusing on these two concepts thus not only has the advantage that they are prominent in the literature on the design of international institutions, but also that they allow us to have tests of varying difficulty for the computational methods. We compare various (semi-)automated approaches to manually

coded measures of scope and depth.[1] In selecting the approaches for our comparison, we were as comprehensive as possible and included all that appeared promising. As a result, we cover both fully automated and semi-automated approaches. Specifically, we consider topic models, section tagging, simple word counts, weighted word counts, Wordfish, and Wordscores.

The text corpus that we use consists of 390 PTAs and derives from the DESTA project (Dür et al., 2014).[2] PTAs are particularly useful for our purpose of examining the utility of computational text analysis because they are a hard case for the application of computational approaches to the measurement of the design of international institutions, since they are quite technical and heterogeneous. There is a good chance, then, that what works for PTAs also works for other international institutions, such as bilateral investment treaties or environmental agreements. With only few exceptions, our analysis covers the universe of PTAs that were signed between 1990 and 2016. We limit ourselves to this period because most PTAs signed before 1990 are narrow and shallow, meaning that assessing their scope and depth is relatively trivial. Moreover, the Wordfish and Wordscores approaches that we introduce below assume that word meanings remain constant over time. This assumption is more likely to hold for our shorter time period.

1 Measuring the scope of international agreements

Scope refers to the number of issues covered by an international institution. Measuring the scope of international agreements is important for the field of International Relations as issue linkage is a key strategy in international negotiations. A key problem in that respect is that there is no agreed upon definition of what an "issue" is. Is "environmental protection" an issue, or do we need to distinguish between provisions on climate change and provisions on endangered species? Since deductive and inductive approaches may provide divergent answers to this question, different approaches may identify a greater or smaller number of issues in an agreement. What we are interested in here is not to arrive at the exact same number of issues per agreement using different approaches, but to see whether the classification of agreements as narrow or broad correlates across approaches.

In the following, we discuss four approaches to measuring the scope of international agreements. For one, we present a theory-guided classification of topics solely based on the manual coding of treaty texts. Second, we propose a method where human coders tag sections in international agreements before the computer counts the number of sections in each agreement. Third, we look at the number of unique words in a treaty

[1] While we are not the first ones comparing manual coding and computation content analysis (e.g. Morris, 1994; Young and Soroka, 2012), our approach is novel in that we focus on legal texts and include a large set of different computational approaches in our comparison.

[2] A list of the PTAs covered can be found in Appendix A1.

text, as this may indicate the scope of an agreement. Fourth, we rely on topic models, a purely inductive and unsupervised method for detecting the latent semantic structure of a text body (Roberts et al., 2016, 2014). The four methods for measuring the scope of international agreements thus range from a purely manual coding via a human-computer collaboration to an unsupervised computational approach. The PTA texts that we use to compare these approaches with each other are all in English. Although we focus on the case of PTAs, the methods that we discuss should also be applicable to other text-types.[3]

1.1 Four approaches to measuring scope

1.1.1 Manual coding

To arrive at a measure of the scope of international agreements via manual coding, it is first necessary to draw up a codebook that covers all issues that can be contained in an agreement (and that for theoretical reasons are considered relevant). Human coders then look for provisions that relate to these issues. As this process is error-prone, it is essential that the agreement texts are double coded. Depending on the number of agreements under consideration, and the number of issues to be identified, this can be a labor-intensive process.

For the set of PTAs under analysis here, we rely on the data collection by Dür et al. (2014) and Lechner (2016). Based on these two datasets, we can count how many of the following 15 topics are included in a PTA: trade in services, foreign direct investments, intellectual property rights, public procurement, sanitary and phytosanitary (SPS) measures, technical barriers to trade, competition policy, anti-dumping, safeguards, subsidy provisions, dispute settlement procedures, civil and political rights, economic and social rights, environmental protection, and security. For all of these issues, the two datasets contain many sub-items (such as whether a PTA contains an MFN clause for services trade), but here we only use dichotomous indicators of whether an issue is mentioned in an agreement.[4] The coding was done manually, with two coders coding all provisions independently of each other, and a third coder resolving inconsistencies. Inter-coder reliability was generally high, with most variables scoring a value of 0.75 or higher on Cohen's kappa. In order to measure scope with these data, we count the number of issues covered by a PTA (*Scope manual*).

1.1.2 Section tagging

Another approach at measuring the scope of an international agreement is to rely on text structures. Most international agreements provide section (or chapter) titles that

[3] We expect similar results for bilateral investment treaties and bilateral tax treaties, and possibly also for human rights treaties and environmental protection treaties.

[4] The detailed codebook is reported in appendix A2.1.

allow a section-wise splitting of the texts. Of course, such splitting can only capture issues that are dealt with in separate sections. This approach cannot identify an issue that is dealt with in different sections, but that is not the dominant issue in any of them.

For our body of PTA texts, we split sections by having two coders insert a text-tag at the beginning of each section in the PDF files containing the PTA full texts, with the tag reflecting the contents of the section (e.g. "chapterchapterSERVICESchapterchapter" for a section on trade in services). Relying on manual coders was necessary because the formatting varies strongly across PTA texts, making a purely computer-based splitting impossible. For other text forms that are well structured, as for instance bilateral tax treaties or bilateral investment treaties, automated section tagging seems possible.[5] We used PDF files instead of text files in this step because human coders are faster and make fewer errors when they code formatted rather than plain text. We identified a total of 48 separate sections that can be contained in a PTA.[6] This step required some discretion by the coders, as not all PTAs use the same titles for the sections, but a section should be identified as, for example, "trade in services" independent of whether its title is "cross-border trade in services" or "supply of services". The resulting scope measure is the number of different sections contained in a PTA, which we counted using the software R (*Scope section*).

1.1.3 Count of unique words

The number of unique words in an international agreement may also inform about the number of issues covered. Adding an issue to an agreement generally means that negotiators need to use terms that otherwise would not have been used. Illustratively, when adding provisions on environmental protection to a trade agreement, words such as "climate", "hazardous" or "pollution" are likely to be added to the text of the agreement, which otherwise would not appear. Broader agreements thus should exhibit a larger number of unique words. A drawback of this approach is that it neither results in an actual number of issues contained in an agreement nor identifies the issues included. If it works, it can only locate agreements somewhere on the dimension between narrow and deep. Moreover, this approach requires a cleaning of the texts, to avoid contaminating the analysis with differences in writing style or formatting of the documents.

For the PTAs that we use here, we relied on the R-package *quanteda* for the cleaning of the texts (Benoit, 2017). This is a fast and flexible tool, which provides features for managing, processing, and quantitatively analyzing textual data. Following

[5] Both bilateral tax treaties and bilateral investment treaties are much more homogeneous in terms of structure and headings than PTAs. This is not least due to the fact that these treaty types often rely on model texts.

[6] The detailed codebook can be found in Appendix A3.

standard practices (Hopkins and King, 2010), we first removed bullet points, hyphens, enumerations, punctuations, and stopwords ('the', 'and', 'but', etc.). We added terms specific to legal texts, such as 'article', 'title', 'chapter', and 'paragraph', to the list of English stopwords.[7] To enable comparisons across PTAs, we further replaced names of cities and countries, regions, and regional trading blocs with neutral terms, namely 'place', 'region', and 'RTA', respectively. Another preprocessing step taken was the lower-casing of all letters. Whether the computer reads 'Member State' or 'member state' should not matter for the substantive meaning of the text.

In the next step, we dropped words that are shorter than two letters or appear in fewer than 3 percent or more than 97 percent of all documents. These specific thresholds make sense given our aim of getting rid of noise while keeping all the substantial information in the data. Assuming that untreated text covers more noise than reduced text and that this relationship is non-linear (exponential), we expect that noise decreases more strongly when moving from the 1 to the 2 percent threshold than when moving from the 3 to the 4 percent threshold. Information, in contrast, is assumed to be lower in reduced than in untreated text, again with a non-linear (exponential) relationship. Thus, we expect to lose more information when moving from a 2 to a 3 percent threshold than when moving from a 1 to a 2 percent threshold. In search for the optimum cut, we needed to find the point where the trend in terms of difference in the drop of unique words reverses. The results indicate that the drop in unique words when moving from a 0 to a 1 percent threshold is nine times higher than the respective drop from 1 to 2 percent, and the drop from 1 to 2 percent is twice as large as the drop from 2 to 3 percent (see Figure A3 in the Appendix). This trend reverses with the 4 percent threshold. For this reason, the 3 percent cut, e.g. keeping words that appear in more than 3 percent and fewer than 97 percent of all documents, is the optimum that minimizes noise while maximizing information.

Finally, we only kept word stems. Stemming reduces the number of different words in and thus the complexity of the texts (Porter, 1980). For example, the words 'restricts', 'restriction', and 'restricting' share a common stem, namely 'restrict'. For all of this, we solely used the main texts after eliminating attached annexes. The annexes are mainly composed of tariff lists for specific products and schedules for liberalization. These parts of the PTAs are document specific and hinder comparisons across agreements.

Although we base the choice of preprocessing steps on theoretical grounds, we additionally employ an inductive method developed by Denny and Spirling (2016) to check whether any of the above described preprocessing steps produces abnormal results. This method calculates pairwise distances between all document term matrices that result from any combination of the preprocessing steps. The cosine distances between the document term matrices, in our case 64, measure how unusual the document term

[7] For a list of all the stopwords, see the Appendix.

matrix resulting from a specific combination of preprocessing steps is relative to all other combinations. Based on these distances, Denny and Spirling (2016) derive *preText* scores that capture the influence of each of the preprocessing steps. The results of this test suggest that none of our preprocessing steps creates problems for the further analysis.[8] The resulting measure of scope is the number of unique words per document (*Scope count*).

1.1.4 Topic models

A final approach to capture the scope of international institutions is to rely on structural topic models (Blei et al., 2003; Roberts et al., 2016). Structural topic models are a purely inductive and unsupervised approach to text analysis that builds on classical methods in natural language processing. In contrast to the latter (e.g. unigram models, mixture of unigram models, Latent Semantic Analysis), topic models allow for document memberships in multiple topics instead of just one (Airoldi et al., 2008). The underlying idea of topic models is that a text consists of several topics, with each word having a certain probability of belonging to a topic. This probability, in turn, is derived from distances between words in the texts. A topic is said to be present in a specific text if many of the terms that have a high probability of belonging to this topic are found in the text.

For our application of structural topic models, we use the R-package *topicmodels* (Grün and Hornik, 2011). Here, the variational expectation-maximization algorithm is used to fit a latent Dirichlet allocation (LDA) model (Blei et al., 2003).[9] The assumption of the LDA model is that topics are uncorrelated. For three reasons, we chose the LDA model instead of the correlated topic model (CTM, Blei and Lafferty, 2007), which accounts for across-topic correlations. First, it is the most frequently applied technique (Grün and Hornik, 2011). Second, since the other methods for the measurement of scope discussed above (*Scope manual* and *Scope section*) ignore correlations between topics, using the LDA model facilitates comparisons across different approaches. Third, we found that the LDA model, in contrast to CTM, converges in all cases and implies significantly shorter computational times than the CTM.

The LDA model relies on a generative process which first determines the term distribution that corresponds to the Dirichlet distribution for each topic; then defines the proportions of topics in each document; and finally allocates each word to a topic and chooses "a word from a multinomial probability distribution conditioned on the topic" (Grün and Hornik, 2011, p. 3). Given this generative process, the variational

[8] Figure A4 in the Appendix reports the detailed results.

[9] An alternative R package for topic models is *lda* (Chang, 2015). This package relies on Gibbs sampling instead of the variational expectation-maximization algorithm and thus differs from the methods proposed by the original papers introducing topic models.

expectation-maximization algorithm, which is a procedure used if the quantities cannot be tractably computed, allows to derive the optimal β, which is the probability of a given word w to occur in latent topics.[10]

As an input, we used the PTA texts that were cleaned as described above for the word count measure of scope. In order to calculate topic models, the number of topics k contained in the texts must be defined a priori. To determine the number of topics that best describes the PTA texts, we proceeded inductively. First, we randomly split the corpus in a training (80 percent of the documents) and a test set (20 percent of the documents).[11] We then ran models on the random training set that assume that k topics exist, where $k = \{10, 15, 25, 35, 45, 60, 80\}$. The perplexity measure helped us evaluate how the models that resulted from the training sets perform on random test sets relative to each other. We conducted this procedure ten times for each k parameter and took the average of the resulting ten perplexity scores to arrive at the measure that we use for the performance evaluation.[12] The optimal number of topics that we arrived at is 15.

Using this optimal number of topics, we fit a single topic model using the command *LDA* of the R package *topicmodels* on the entire corpus.[13] The command returns an object, which contains inter alia information on the probabilities of each word belonging to a topic and thus the most important keywords for each topic. Using this information, we assigned labels to each topic based on the most frequent terms associated with them. For example, the stem "procur" stands for public procurement and "agri" indicates that a topic captures trade in agriculture products.

To measure scope on this basis, we coded a topic as being included in an agreement if more than half of the 20 terms most specific to that topic are contained in the agreement's text.[14] A word's specificity to a topic is calculated as the term frequency within a topic divided by the total term frequency over all topics (Chuang et al., 2012).[15] The scope measure that derives from this approach is the sum of topics covered in a PTA (*Scope topic*).

1.2 Comparing the measures of scope

In the following, we compare the automated or semi-automated measures to *Scope manual*, which for this purpose we consider as the "true" measure. As explained above,

[10] For details on this process see Wainwright and Jordan (2008).

[11] Denny and Spirling (2016) and Gilardi et al. (2017) use a similar approach. Grün and Hornik (2011) do not give precise percentages to split the corpus, but nevertheless recommend using a training and a test set to determine the optimal number of topics.

[12] Details on all perplexity measures can be found in Table A3 in the appendix.

[13] The input for this is a document term matrix, where the rows correspond to the documents and the columns to the terms.

[14] In Appendix A5.2, we report alternative specifications for scope derived from topic models.

[15] The correlation between the probability of a word belonging to a topic and the specificity of a word in a topic is high with r=0.59.

Scope manual is the result of two coders coding the agreements independently of each other, and then a third coder resolving inconsistencies. Moreover, for the manual coding we are quite certain about the validity of the measure. So overall this is a high standard to test the various automated approaches against. If we can closely approximate this measure, computational text analysis can be considered superior to a single human coding agreements, as a single human is likely to produce a coding that is less reliable than *Scope manual*.

In Figure 1, we show the correlations between *Scope manual* on the one hand and the other measures of scope on the other hand. To facilitate comparisons, we rescaled all indices to measures ranging from 0 to 1. The figure shows that all automated or semi-automated measures are positively correlated with *Scope manual*. In fact, the correlation coefficients are quite high, ranging from 0.62 [95 percent confidence intervals: 0.56, 0.68] for *Scope count* to 0.74 [0.69, 0.78] for *Scope topic* and even 0.76 [0.72, 0.80] for *Scope section*. For all three measures, however, the relationship to *Scope manual* is not linear. From about 2000 words (or the rescaled index of 0.54) onwards, an increase in the number of unique words in a text is no longer reflected in an increase in the *Scope manual* measure. Reflecting this non-linearity, the correlation between *Scope count* and *Scope manual* increases to 0.68 [0.62, 0.73] when taking the log of *Scope count*. The pattern is similar for *Scope section*, although less pronounced. By contrast, the *Scope topic* measure does not work well for agreements that score low on *Scope manual*. Only few PTAs are badly classified by the three measures. The Preferential Trading Arrangement of the South Asian Association for Regional Cooperation (SAPTA, 1993) is an outlier in three and the Canada Jordan PTA in two comparisons. Either the computational measures systematically overestimate the scope of these agreements or our manual coding scheme fails to capture certain aspects of scope that the automated approach is able to detect. None of the PTAs is systematically underestimated.

2 Measuring the depth of international agreements

Depth refers to the extent to which an international institution constrains the activities of a member state in a policy area. Whereas scope captures the number of issues regulated in an international agreement, depth measures the extent of cooperation in an issue area. Illustratively, a national treatment clause in a PTA services chapter restricts the ability of states to choose among different policies more than just a vague statement that member states should strive towards the liberalization of services trade.

Since the depth of cooperation may vary strongly across issue areas, and in line with the recommendation by Slapin and Proksch (2008, 711-712) to divide texts into specific dimensions, we assess depth by issue. Our analysis of scope, however, showed that different approaches arrive at different numbers of issues. Therefore, for each

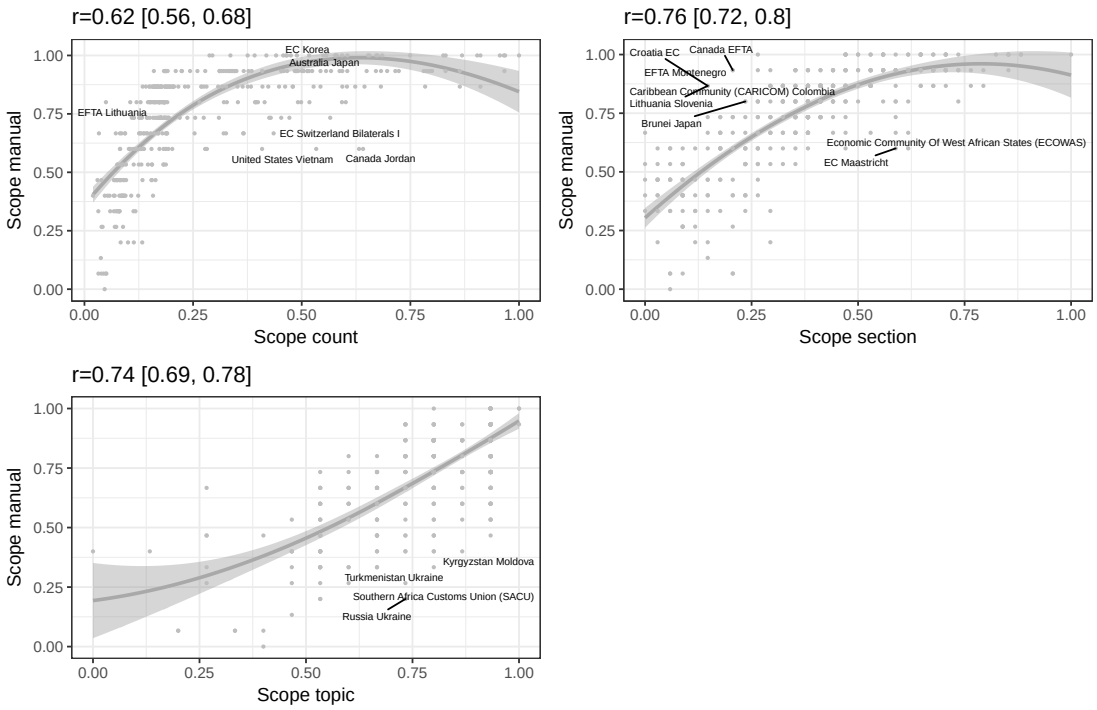

Figure 1: Comparison of scope measures

computational approach that we introduce, we calculate two sets of depth measures: for the sections that we split using section tags and for the keywords that we identified via the topic model.

2.1 Five approaches to measuring depth

2.1.1 Manual coding

To arrive at measures of depth using manual coding, it is necessary to code provisions related to each issue. A simple measure of depth for an issue then can be the number of provisions related to an issue present in an agreement. Alternatively, the various items can be assigned a weight, as most likely some provisions are more restrictive than others. These weights can be arrived at either based on prior theoretical considerations or inductively.

The concrete measures of depth based on manual coding that we rely on in this paper use latent trait analysis to establish the relevance of each provision inductively. Latent trait analysis is a type of factor analysis for binary data (Bartholomew et al., 2011). Concretely, we apply the Rasch model that assumes that all items capture one underlying latent dimension, but with different discriminatory power (Rasch, 1980). We implement this procedure by dint of the commands *rasch* and *factor.scores* of the R package *ltm* (Rizopoulos, 2006). The number of manually coded items that we can include in the analysis varies across areas. For dispute settlement mechanisms, for

example, 30 variables are available in the DESTA dataset; for public procurement only seven. Via this approach, for each agreement we arrive at measures of depth for the 15 areas distinguished in *Scope manual*, for example for trade in services, intellectual property rights, and so on (*Depth manual*). We report the detailed coding scheme for this approach in Appendix A2.2.

2.1.2 Count of unique words

The simplest approach to arrive at depth measures by topic via computational means is to count unique words in an international agreement related to a specific issue. The assumption here is that a larger number of different words dedicated to a topic indicates deeper commitments with respect to that topic. For most cases, this is plausible: deeper cooperation should require more detailed provisions, which will mean more unique words per issue. In some cases, however, a negative list approach (e.g. all measures of type X are prohibited) may be deeper than a positive list approach (e.g. measures X1, X2, and X3 are prohibited), and at the same time require fewer unique words. To which extent this poses problems for this approach likely varies across issue areas. A possible objection to the count of unique words as a measure of depth can also be that this number actually captures scope (see our discussion above). We submit, however, that when applied to segments of texts that are specific to quite narrow topics, as we do here, it is plausible that this is a measure of depth.

For the case of PTAs dealt with in this paper, we counted the words included in the sections identified via the human tagging and the number of times the 100 keywords identified by the structural topic model as belonging to a topic show up in a text. For each agreement, we end up with up to 48 or up to 15 depth measures, depending on how many different sections and topics are contained in a PTA (*Depth count section* and *Depth count topics*).

2.1.3 Latent trait analysis on unique words

Another option is to not simply count the number of words per section or text part, but to discriminate between words according to their relevance for capturing the depth of commitments on a specific topic. Given the large number of words, assigning these weights based on theoretical considerations is not feasible. Instead, this can only be done using an inductive approach. Inductive approaches generally assume that words that are relatively rare are more relevant to discriminate among texts than words that are relatively frequent.

For the PTA texts, we relied on latent trait analysis to achieve this aim. This approach assumes that the rarer a certain word, the greater the difficulty of being selected and thus the greater the weight it should have in the final depth measure. As

for the *Depth manual* measure, we used the Rasch model for the analysis (Bartholomew et al., 2011; Rasch, 1980). The input for this model is the unique-word document matrix for each topic. Rows i are the documents and columns j the unique words included in sections or the keywords identified by the topic model. M_{ij} is either zero, if the word is not included in the (section of a) document, or one, if the word is covered in a respective (section of an) agreement. The number of depth measures that we have for each agreement again depends on its scope as captured by sections and topics covered (*Depth LTA section* and *Depth LTA topics*).

2.1.4 Wordfish

A recent innovation in quantitative content analysis is *Wordfish* (Slapin and Proksch, 2008). Wordfish is a statistical model that allows latent traits of uni-dimensional texts to be estimated simply by drawing on word frequencies in texts. The model is based on the assumption that words are distributed according to a Poisson distribution. The model is as follows:

$$y_{ij} \sim \ Poisson(\lambda_{ij})$$
$$\lambda_{ij} = exp(\alpha_i + \psi_j + \beta_j * \omega_i),$$

where y_{ij} is the count of word j in text i, λ_{ij} is the mean and the variance of the distribution, α is a set of document fixed effects, ψ is a set of word fixed effects, β is an estimate of a word specific weight capturing the importance of word j in discriminating between positions, and ω is the estimate of document i's location on a uni-dimensional scale.

For the case of PTAs, we assume that ω measures the depth of a specific issue area in PTAs, whereas β identifies the words that differentiate between the depth of different PTAs. We use the *textmodel_wordfish* command from the *quanteda* R package (Benoit, 2017; Proksch and Slapin, 2009b) that builds on the "Poisson scaling model of one-dimensional document positions using conditional maximum likelihood" by Slapin and Proksch (2008). The input is the document-term matrix of sections or keywords. A key problem for our application of this method is that shallow PTAs are mainly characterized by the absence of certain "deep" terms rather than the presence of specific "shallow" terms. Although this may create problems for the Wordfish approach, we still deem it useful to see to which extent this method can capture depth. As the underlying dimension that Wordfish identifies in some cases ranges from deep to shallow rather than from shallow to deep (which becomes clear when looking at which words have negative and which words have positive values for β_j), we invert these scores to make them

comparable with the others.[16] We again potentially have as many measures of depth as sections or topics contained in a PTA (*Depth wordfish section* and *Depth wordfish topic*), but for some sections/sets of keywords the Wordfish algorithm did not converge. These are mostly shallow agreements. Thus, the two measures mainly capture variation across deep PTAs instead of variation across all PTAs.

2.1.5 Wordscores

Wordscores is a method that assigns scores to documents on the basis of reference documents (Benoit and Laver, 2003; Lowe, 2008). For these reference texts, the score on a dimension needs to be known. Ideally, these reference texts are located at or close to the ends of the dimension. The method estimates scores for each word in the reference text, and then uses these scores to arrive at scores for new texts (which is the average score across all words). The advantage of this approach is that no functional or distributional assumptions are required. Lowe (2008), however, stresses several problematic aspects of this method, such as that it assumes that all words are equally informative about a text. Moreover, finding adequate reference documents is a practical problem that limits the usefulness of this approach.

For the case of PTAs, we selected three reference texts if four or more documents were available for one topic. In cases for which we only have three or fewer texts, we were not able to use Wordscores. We chose the texts automatically on the basis of the length of sections (for the section-split texts) or of the number of keywords covered (for the topics from the topic model). Concretely, we picked the shortest, the mean-length, and the longest texts (and equivalently for the keywords). The shortest documents contain only some words, the documents of mean length cover a paragraph, and the longest documents contain several paragraphs on a topic. We then assigned the shortest text a score of -1 (assuming that it is very shallow), the text of mean length a score of 0, and the very long text a score of 1 (assuming that it is very deep). Just as was the case for our application of Wordfish, the problem that emerges is that the absence of "deep" terms is a more important feature of shallow PTAs than the presence of "shallow" terms. Although this may create problems for Wordscores, rather than exclude the approach a priori we assess whether it still manages to distinguish between deep and shallow PTAs.

Again the document-term matrix of sections or keywords served as input on which we applied the R command *textmodel_wordscores* of the *quanteda* package (Benoit, 2017). This command captures the above described Wordscores method by Benoit and Laver (2003). For some texts the Wordscores algorithm did not converge. Similar

[16] Negative scores indicated depth in the following cases: competition policy, dispute settlement, economic and social rights, environmental protection, public procurement, and technical barriers to trade for the section-split texts; investments, public procurement, and trade remedies for the keyword-split texts.

to Wordfish, this introduces a bias towards deep PTAs, because mostly shallow agreements fall out of the sample. Still, we end up with a large number of values both for the sections and the topics (*Depth wordscores section* and *Depth wordscores topic*).

2.2 Comparing the measures of depth

We begin our comparison of the various depth measures by aggregating them up to the level of PTAs. Concretely, we average the values across issue areas for each agreement, after again rescaling all variables so that they range from 0 to 1. This gives us a first overview of how well the various computational measures fare when compared to *Depth manual*. For this comparison, we only consider the eleven areas that we identified in the manual coding and at least one of the section or keyword approaches, namely competition policy, dispute settlement, economic and social rights, environmental protection, investments, SPS measures, security, technical barriers to trade, trade in services, and trade remedies.

Figure 2 shows that the *Depth lta* approach comes closest to the manual coding when considering both the sections and the keywords (with correlation coefficients of 0.79 [0.75, 0.83] and 0.90 [0.88, 0.92], respectively). The relationship between *Depth manual* and *Depth lta topic* also is close to linear. The biggest outlier is the agreement between the European Free Trade Association and Lithuania from 1995, which has a very low value on *Depth lta topic*, but scores highly on *Depth manual*. The high value on the manual depth measure is driven by provisions on the environment, civil and political rights etc. included in the agreement's preamble, which are not picked up by our computational measure. The two *Depth count* measures perform well, too. However, the *Depth count section* measure is skewed to the left and the *Depth count topic* measure is skewed to the right. *Depth count section* thus tends to underestimate depth, whereas *Depth count topic* tends to overestimate depth. By dint of weighting, the latent trait analysis can compensate for this shortcoming.

The *Wordfish* and *Wordscore* measures produce weaker results than the simple word count and latent trait analysis measures. The *Wordfish* scores underestimate deep agreements and overestimate shallow ones. This is the case for both text types, sections and keywords extracted by dint of topic models. Both *Depth wordscores section* and *Depth wordscores topic*, moreover, assign shallow issue areas (according to the manual coding) too high scores. This trend is stronger for the section-split texts than for the keywords. In the former case, nearly all agreements scatter around a score of 0.8. The *Depth wordscores topic* measure is less skewed and also exhibits a higher correlation with the manual measure.[17] This suggests that Wordscores is better able to discriminate across texts on the basis of a limited amount of keywords than on the basis of a wide

[17] The agreement with a *Depth wordscores topic* score of 0 is the Armenia Georgia agreement that covers hardly any keyword.

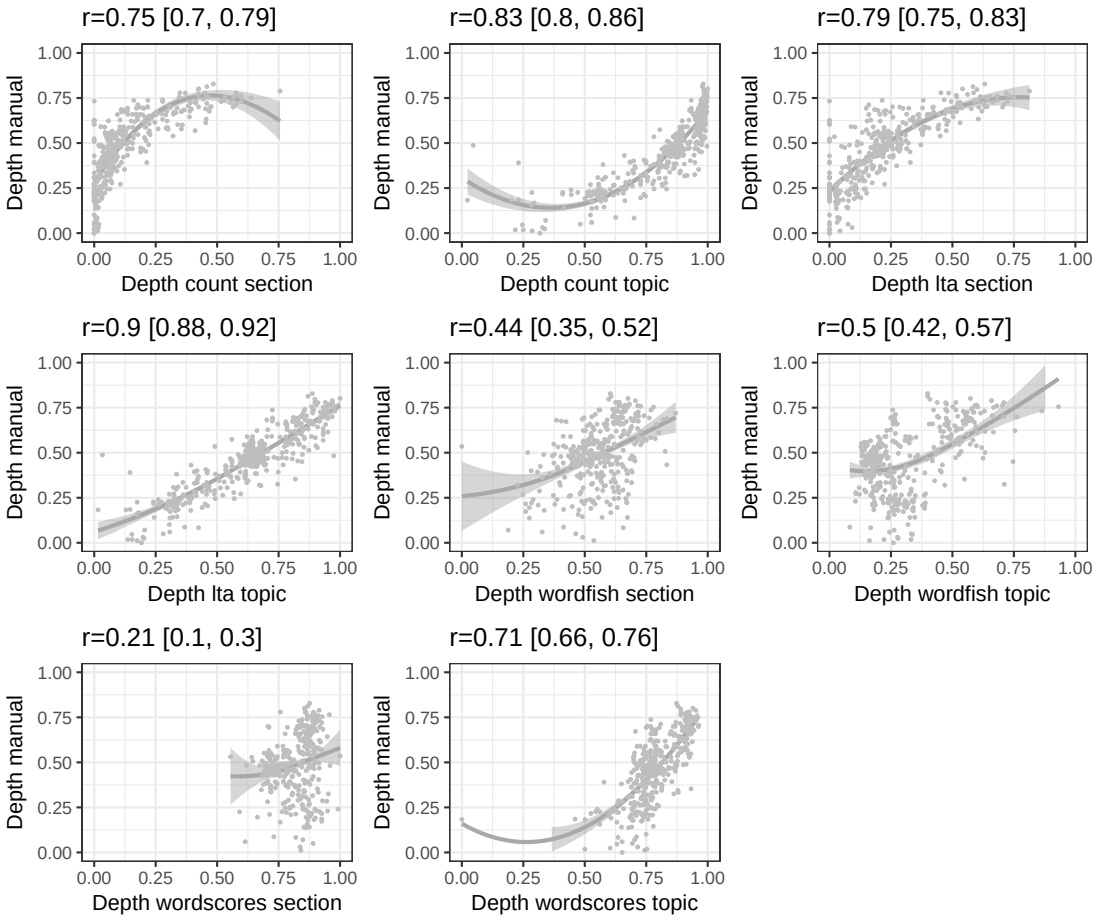

Figure 2: Comparison of depth measures (aggregated to the level of PTAs)

variety of words that appear in sections.

In the following, we provide further detail on the correlations between *Depth manual* and all other depth measures by breaking down the results by topic. In Table 1 we do so for the depth measures relying on the section-split texts. As can be seen, the correlations between *Depth manual* on the one hand, and *Depth count section* and *Depth lta section* on the other hand are at least decent for most of the eleven topics for which we can make comparisons. The count approach works particularly well for dispute settlement, trade in services, investments, and public procurement. The correlations are lowest for economic and social rights, security, and environmental protection. These are all so-called non-trade issues that are often referred to in the treaties' preambles, which do not enter the computational measures. The latent trait analysis approach produces particularly good results for trade in services and public procurement. The Wordfish approach works much better for some topics than for others. The correlation coefficients are ≥ 0.5 for five of the eleven topics, but negative for three topics. When comparing the Wordscores measures to the manual measures, we find correlations between 0.3 and 0.6 for six of eleven topic areas.

Table 1: Manual versus computational depth measures relying on sections

Manual versus...	Depth count section	Depth lta section	Depth wordfish section	Depth wordscores section
Competition	0.64 [0.60, 0.68]	0.69 [0.65, 0.72]	-0.19 [-0.28, -0.10]	0.39 [0.30, 0.46]
Dispute settlement	0.77 [0.73, 0.81]	0.74 [0.69, 0.78]	0.50 [0.39, 0.60]	0.17 [0.02, 0.30]
Economic & social rights	0.32 [0.23, 0.41]	0.35 [0.26, 0.43]	0.44 [0.13, 0.67]	0.27 [-0.06, 0.55]
Environmental protect.	0.43 [0.35, 0.51]	0.48 [0.41, 0.56]	0.64 [0.48, 0.76]	0.26 [0.04, 0.46]
Investments	0.71 [0.66, 0.76]	0.73 [0.68, 0.77]	0.79 [0.71, 0.85]	0.27 [0.10, 0.44]
Public procurement	0.67 [0.61, 0.72]	0.81 [0.77,0.84]	0.77 [0.70, 0.82]	0.58 [0.48, 0.67]
Security	0.48 [0.40, 0.55]	0.39 [0.30, 0.47]	-0.02 [-0.40, 0.37]	0.47 [0.10, 0.73]
SPS measures	0.53 [0.45, 0.60]	0.61 [0.55, 0.67]	0.24 [0.10, 0.38]	0.38 [0.24, 0.49]
TBT	0.50 [0.42, 0.57]	0.52 [0.44, 0.59]	0.40 [0.29, 0.50]	0.35 [0.23, 0.45]
Trade in services	0.76 [0.71, 0.80]	0.85 [0.82, 0.88]	0.60 [0.50, 0.69]	0.33 [0.19, 0.45]
Trade remedies	0.47 [0.39, 0.55]	0.51 [0.43, 0.58]	-0.06 [-0.17, 0.06]	0.13 [0.01, 0.24]

Table 2 shows the results for the comparison between the manual depth measure and the various computational measures that rely on the keywords from the topic model. For this comparison, we can only rely on four issues, as these are the only ones that match across the approaches. Again, the correlations are high or at least decent across all four areas for *Depth count topic* and *Depth lta topic*. For both approaches, the correlations are lowest for public procurement. Both Wordfish and Wordscores work better for the keywords than for the section-split texts. Wordfish produces the best results for investments and the weakest for trade remedies. *Depth wordscores topic* performs well for all four issue areas and scores highest on investment.

Table 2: Manual versus computational depth measures relying on keywords

Manual versus...	Depth count topic	Depth lta topic	Depth wordfish topic	Depth wordscores topic
Investments	0.61 [0.55, 0.67]	0.79 [0.75, 0.83]	0.70 [0.65, 0.75]	0.71 [0.66, 0.76]
Public procurement	0.56 [0.49, 0.63]	0.60 [0.54, 0.66]	0.32 [0.22, 0.40]	0.52 [0.44,0.59]
Trade in services	0.66 [0.61, 0.72]	0.76 [0.71, 0.80]	0.47 [0.39, 0.54]	0.60 [0.53,0.66]
Trade remedies	0.73 [0.68, 0.78]	0.75 [0.70, 0.79]	0.24 [0.14, 0.33]	0.47 [0.39, 0.54]

3 Treaty characteristics and the performance of the various approaches

So far, we have shown that some computational measures fare better than others in approximating manual measures of the scope and depth of international institutions. However, which computational measures work best may depend on treaty characteristics. To investigate whether any specific factors favor one or another approach, we thus ran multivariate regressions explaining deviations from *Scope manual* and *Depth manual*, respectively. The unit of analysis is the agreement-method (N=1,161) when assessing the scope measures and the agreement-topic-method (N=18,148) when assessing the depth measures.

To measure deviations from the two manual measures, we regressed them on the measures generated by the computational approaches. Then, we took the residuals from these regressions as dependent variables in two further models. In these models, we included dummies accounting for the various approaches on the right side of the equation.

In the case of depth, the model also contains dummies for the various topics for which we assessed depth (services etc.), and a measure of the number of topics covered within an agreement (*Scope manual*). Most importantly, we interact the dummies for the computational approaches with three treaty characteristics: *Signature year*, *North-South PTA*, and *Number of member states*. *Signature year* accounts for the increasing complexity of PTAs over time. Some approaches may work better for complex agreements than others. The *North-South PTA* variable captures power-asymmetries among member states. Highly asymmetric agreements may closely follow templates imposed by the stronger states, which may favor some approaches over others. The *Number of member states* variable proxies for the diversity of interests during negotiation processes, which might result in more or less complex text structures. The interactions between the approaches and these variables allow us to answer the question whether treaty characteristics condition the performance of the various computational approaches to measuring scope and depth.

Table 3 reports the coefficients for the interaction terms included in the model explaining the performance of the computational scope measures (where *Scope count* serves as reference category). The results indicate that the performance of the computational approaches to measuring scope is largely independent of treaty characteristics. We only find that *Scope section* works less well for newer agreements. None of the other coefficients for the interaction terms is statistically significant. When looking at the main effects (not shown in Table 3), we find a statistically negative and sizeable coefficient for *Scope section*, indicating that this approach comes closer to *Scope manual* than *Scope count*. Moreover, the main effect for *Signature year* is negative and statistically significant, meaning that *Scope count* works better for more recent agreements. Overall, this evidence indicates that researchers should prefer section tagging over word counts and topic models when trying to measure the scope of agreements, largely independent of treaty characteristics.

In Table 4 we report the coefficients for the interaction terms in the model assessing the performance of the various computational depth measures. Just as in the case of scope, we find that the performance of the various approaches is largely independent of treaty characteristics. Both *Depth count topic* and *Depth lta topic*, however, perform worse for newer agreements and for North-South PTAs. Moreover, *Depth wordscores topic* works better for agreements signed by a larger number of member states. Moving on to the main effects (not shown in Table 4), we find that *Depth lta topic* and *Depth count topic* more closely approximate *Depth manual* than *Depth count section*. With respect to topics, the various computational approaches work better for security provisions, competition policy, and trade remedies than for other topics. Especially for economic and social rights, and environmental protection it is difficult to substitute the manual coding with computational measures. Finally, and surprisingly, the computa-

Table 3: Conditional performance of scope measures

	× Signature year	× North-South PTA	× # of member states
Scope section...	0.002**	−0.001	0.001
	(0.001)	(0.016)	(0.001)
Scope topic...	0.001	−0.023	0.001
	(0.001)	(0.016)	(0.001)

Note: The model also contained main effects that are not shown here for space reasons. Negative values indicate a better fit. *Scope count* serves as reference category. N=1,161. Adjusted $R^2 = 0.047$. *p<0.1; **p<0.05; ***p<0.01.

tional measures perform better for agreements that are quite broad (as indicated by a negative and statistically significant coefficient for *Scope manual*). This means that relatively complex treaty texts should not keep researchers from relying on computational approaches.

4 Conclusion

To what extent can computational text analysis help in measuring the design of international institutions? Can the computer keep up with human coders in that task? With the aim of responding to these questions, we have used 390 full texts of PTAs to compare manual coding with various computational measures, including a fully inductive approach that can be completely automatized.

The results are encouraging. All approaches that we introduced with the aim of capturing the scope of PTAs – namely section tagging, simple word count, and topic model – offered similar results. In contrast to scope, not all approaches that we used to measure the depth of PTAs resulted in equally plausible results. Still, several approaches are promising. The latent trait analysis on unique words shows the highest correlation with the manual coding. Two more measures reached correlations with the manual coding of around 0.8. Also the Wordfish and Wordscores method on top of keywords derived from topic models showed a decent correlation with the manual coding. Only the Wordscores method applied on section-split text produced results that clearly diverged from the other approaches. We also showed that the performance of the various computational measures of both scope and depth is largely independent of treaty characteristics, making it plausible that our findings for PTAs can be generalized to other bodies of international agreements.

These findings make a key contribution to the literature on international institu-

Table 4: Conditional performance of depth measures

	× Signature year	× North-South PTA	× # of member states
Depth count topic...	0.003^{***}	0.021^{**}	−0.001
	(0.001)	(0.009)	(0.0005)
Depth lta section...	−0.0002	0.002	−0.00000
	(0.0004)	(0.006)	(0.0004)
Depth lta topic...	0.004^{***}	0.023^{**}	−0.001
	(0.001)	(0.009)	(0.0005)
Depth wordfish section...	0.00001	−0.0004	−0.0004
	(0.001)	(0.008)	(0.0004)
Depth wordfish topic...	0.0004	0.010	-0.001^{*}
	(0.001)	(0.009)	(0.0005)
Depth wordscore section...	−0.001	−0.006	-0.001^{*}
	(0.001)	(0.008)	(0.0004)
Depth wordscore topic...	0.0003	0.016^{*}	-0.001^{**}
	(0.001)	(0.009)	(0.0005)

Note: The model also contained main effects that are not shown here for space reasons. Negative values indicate a better fit. *Depth count section* serves as reference category. N=18,148. Adjusted $R^2 = 0.155$. *p<0.1; **p<0.05; ***p<0.01.

tions. With a large group of International Relations scholars investing time and money on manual coding to capture formal state cooperation, this paper has discussed possible alternatives to expensive human coders. We have introduced researchers of international institutions to a wide range of tools in textual analysis that are all relatively easy to implement. Importantly, we have demonstrated that fairly simple computational approaches perform well in substituting the manual data-collection process. In as such, the paper can serve as a text-as-data methods guide for scholars of international institutions that has yet been missing.

Our results also speak to a growing literature on the use of computational text analysis in political science (Grimmer and Stewart, 2013; Lucas et al., 2015; Monroe and Schrodt, 2008). In particular, we show that there is an even broader range of applications for these approaches than has been recognized so far. Even latent traits in legal documents can be captured fairly well using fully automated approaches. A key message to take away from our paper thus is the promise of computational text analysis to become a standard tool for all subfields of political science.

References

Airoldi, E. M., Blei, D. M., Fienberg, S. E., and Xing, E. P. (2008). Mixed Membership Stochastic Blockmodels. *Journal of Machine Learning Research*, 9:1981–2014. 7

Allee, T. and Elsig, M. (2016). Are the contents of international treaties copied-and-pasted? evidence from preferential trade agreements. August:1–43. 2

Allee, T. and Peinhardt, C. (2014). Evaluating three explanations for the design of bilateral investment treaties. *World Politics*, 66(01):47–87. 1

Alschner, W. and Skougarevskiy, D. (2016). Mapping the universe of international investment agreements. *Journal of International Economic Law*, 19(3):561–588. 2

Baccini, L., Dür, A., and Elsig, M. (2015). The politics of trade agreement design: Revisiting the depth-flexibility nexus. *International Studies Quarterly*, 59(4):765–775. 1

Bartholomew, D. J., Knott, M., and Moustaki, I. (2011). *Latent variable models and factor analysis: A unified approach*. Wiley series in probability and statistics. Wiley, Hoboken, N.J., 3rd ed. edition. 10, 12

Benoit, K. (2017). *quanteda: Quantitative Analysis of Textual Data*. R package version 0.99.22. 5, 12, 13

Benoit, K. and Laver, M. (2003). Estimating irish party policy positions using computer wordscoring: The 2002 election – a research note. *Irish Political Studies*, 18(1):97–107. 13

Blei, D. M. and Lafferty, J. D. (2007). A Correlated Topic Model of Science. *The Annals of Applied Statistics*, 1(1):17–35. 7

Blei, D. M., Ng, A. Y., and Jordan, M. (2003). Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3:993–1022. 7

Chang, J. (2015). Package lda: Collapsed Gibbs Sampling Methods for Topic Models. *CRAN*. 7

Chuang, J., Manning, C. D., and Heer, J. (2012). Termite: Visualization techniques for assessing textual topic models. *Advanced Visual Interfaces*. 8

Denny, M. J. and Spirling, A. (2016). Text Preprocessing for Unsupervised Learning: Why It Matters, When It Misleads, and What to Do about It. 6, 7, 8

Dür, A., Baccini, L., and Elsig, M. (2014). The design of international trade agreements: Introducing a new dataset. *The Review of International Organizations*, 9(3):353–375. 3, 4

Dyevre, A. (29.02.2016). The promise and pitfalls of automated text-scaling techniques for the analysis of judicial opinions. 2

Evans, M. C., McIntosh, W. V., Lin, J., and Cates, C. L. (2007). Recounting the courts? applying automated content analysis to enhance empirical legal research. *Journal of Empirical Legal Studies*, 4(4):1007–1039. 2

Gilardi, F., Shipan, C. R., and Wueest (2017). Policy Diffusion: The Issue-Definition Stage. 8

Grimmer, J. and Stewart, B. M. (2013). Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political Analysis*, 21(3):1–31. 21

Grün, B. and Hornik, K. (2011). topicmodels: An R Package for Fitting Topic Models. *Journal of Statistical Software*, 40(13):1–30. 7, 8

Hooghe, L., Marks, G., Lenz, T., Bezuijen, J., Ceka, , Besir, and Derderyan, S. (2016). *Measuring regional authority: A postfunctionalist theory of governance, Volume I*. Transformations in governance. Oxford University Press, Oxford. 1

Hopkins, D. J. and King, G. (2010). A method of automated nonparametric content analysis for social science. *American Journal of Political Science*, 54(1):229–247. 6

Kim, S. Y., editor (2015). *The Language of Institutional Design: Text Similarity in Preferential Trade Agreements*. 2

Klüver, H. (2009). Measuring interest group influence using quantitative text analysis. *European Union Politics*, 10(4):535–549. 2

Koremenos, B. (2016). *The continent of international law: Explaining agreement design*. Cambridge University Press, Cambridge, United Kingdom. 1

Koremenos, B., Lipson, C., and Snidal, D. (2001). The rational design of international institutions. *International Organization*, 55(4):761–799. 1

Kucik, J. (2012). The domestic politics of institutional design: Producer preferences over trade agreement rules. *Economics & Politics*, 24(2):95–118. 1

Lechner, L. (2016). The domestic battle over the design of non-trade issues in prefer-

ential trade agreements. *Review of International Political Economy*, 23(5):840–871. 4

Lowe, W. (2008). Understanding wordscores. *Political Analysis*, 16(4):356–371. 13

Lucas, C., Nielsen, R. A., Roberts, M. E., and Steward, B. M. (2015). Computer-Assisted Text Analysis for Comparative Politics. *Political Analysis*, 23(2):254–277. 21

Mitchell, R. B. (2003). International Environmental Agreements: A Survey of Their Features, Formation, and Effects. *Annual Review of Environment and Resources*, 28(1):429–461. 1

Monroe, B. L. and Schrodt, P. A. (2008). Introduction to the Special Issue: The Statistical Analysis of Political Text. *Political Analysis*, 16(04):351–355. 21

Morris, R. (1994). Computerized Content Analysis in Management Research: A Demonstration of Advantages and Limitations. *Journal of Management*, 20(4):903–931. 3

Porter, M. F. (1980). An algorithm for suffix stripping. *Program*, 14(3):130–137. 6

Proksch, S.-O. and Slapin, J. B. (2009a). How to avoid pitfalls in statistical analysis of political texts: The case of germany. *German Politics*, 18(3):323–344. 2

Proksch, S.-O. and Slapin, J. B. (2009b). Wordfish: Scaling software for estimating political positions from texts. Version 1.3. 15 January 2009. 12

Proksch, S.-O. and Slapin, J. B. (2010). Position taking in european parliament speeches. *British Journal of Political Science*, 40(03):587–611. 2

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests.* University of Chicago Press, Chicago. 10, 12

Rizopoulos, D. (2006). ltm: An r package for latent variable modelling and item response theory analyses. *Journal of Statistical Software*, 17(5):1–25. 10

Roberts, M. E., Stewart, B. M., and Airoldi, E. M. (2016). A model of text for experimentation in the social sciences. *Journal of the American Statistical Association*, 111(515):988–1003. 4, 7

Roberts, M. E., Stewart, B. M., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S. K., Albertson, B., and Rand, D. G. (2014). Structural topic models for open-ended survey responses. *American Journal of Political Science*, 58(4):1064–1082. 4

Slapin, J. B. and Proksch, S.-O. (2008). A scaling model for estimating time-series party positions from texts. *American Journal of Political Science*, 52(3):705–722. 2, 9, 12

Wainwright, M. J. and Jordan, M. I. (2008). *Graphical models, exponential families, and variational inference.* Now Publishers. 8

Young, L. and Soroka, S. (2012). Affective News: The Automated Coding of Sentiment in Political Texts. *Political Communication*, 29(2):205–231. 3

Young, O. R. and Zürn, M. (2006). The International Regimes Database: Designing and Using a Sophisticated Tool for Institutional Analysis. *Global Environmental Politics*, 6(3):121–143. 1