COMPUTATIONAL ANALYSIS OF INTERNATIONAL INVESTMENT AGREEMENTS

Wolfgang Alschner / Dmitriy Skougarevskiy

Post-doctoral Researcher in International Law, World Trade Institute, Hallerstrasse 6, Bern, CH, and
Graduate Institute of International and Development Studies (IHEID), Maison de la Paix, Chemin Eugène-Rigot 2, Geneva, CH
wolfgang.alschner@graduateinstitute.ch

PhD Candidate in International Economics, European University, St. Petersburg, RU, and
Graduate Institute of International and Development Studies (IHEID), Maison de la Paix, Chemin Eugène-Rigot 2, Geneva, CH
dmitriy.skugarevskiy@graduateinstitute.ch

Keywords: ***International investment treaties, international arbitration, text-as-data, q-character-grams, Jaccard distance***

Abstract: *More than 3000 international investment agreements (IIAs) have been concluded by 2015 and virtually every country is a signatory. What makes these treaties special is their enforcement mechanism: private investors can sue states directly before international arbitration potentially winning multi-million dollar awards. Given its size and atomized nature, however, practitioners struggle to effectively navigate the IIA universe. To reduce investment law's complexity, this paper introduces a range of computational approaches relying on state-of-the-art technology. Implemented as a web-based tool, these approaches allow researchers, policy makers and litigators to assess similarities and differences between agreements quickly and intuitively helping them to navigate the investment treaty universe.*

1. Introduction

Close to 3000 international investment treaties (IIAs) protect foreign investors against the risks of expropriation, discrimination and unfair treatment.[1] Today, virtually every country is signatory to an IIA. What makes IIAs matter in practice is their strong enforcement mechanism. Private investors can bring claims directly to international arbitration in order to enforce an IIA's investment protection obligation against a host state. Over 600 of such investment treaty claims have been launched until today.[2] If successful, investors can win multi-million dollars worth of damages. As a result, IIAs have come to play a central role in international economic governance.

At the same time, due to its atomized and fragmented nature, states and investors alike struggle to effectively navigate the complex IIA universe: negotiators are sorting through hundreds of agreements to find common denominators in two countries' treaty practice; policy-makers strive to streamline a country's investment obligations scattered in scores of treaties and litigators are comparing hundreds of agreements to find useful distinctions or analogies to advance their case. Commercial platforms and empirical scholarship only provide limited assistance to researchers and practitioners. Legal information providers (e.g. www.investorstatelawguide.com) focus on awards rather than treaties and while political scientists and legal scholars have empirically coded treaty content, their datasets are either limited in scope to few treaty features or have not been made public.[3]

[1] UNCTAD, World Investment Report 2015: Reforming International Investment Governance, United Nations, Geneva (2015).

[2] *Ibid*.

[3] M.S. Manger, A Quantitative Perspective on Trends in IIA Rules, in: A. de Mestral & C. Lévesque (eds.), Improving International Investment Agreements, London, Routledge (2011). J. Chaisse and C. Bellak, Navigating the Expanding Universe of International

The *mappinginvestmenttreaties.com* project seeks to fill the ensuing gap by providing legal analytics to assist practitioners in navigating investment law's complexity. Using state-of-the-art text as data methods, the project reveals hitherto unknown patterns of similarities and differences in over 1600 international investment agreements and provides new web-based tools for academics and practitioners to engage interactively with the IIA universe.[4]

2. Data

This project sets out to build the most extensive dataset of English language IIAs to date. In this endeavor, we combined three sources of IIA full texts: Kluwer Arbitration (http://kluwerarbitration.com), Investment Claims (http://oxia.ouplaw.com), and the UNCTAD website (http://investmentpolicyhub.unctad.org), relying on UNCTAD's information on signatories and the date of IIA conclusion for all treaties. We then edited the texts both manually and automatically, removing annexes or side-letters and correcting typos, optical character recognition errors and other mistakes in the underlying data sources. To ensure replicability of our data cleaning procedure, we set up a version control system that tracked all the changes we introduced to the initial texts. We also unified treaty spelling, converting all British English words into their American English counterparts (e.g. «favour» to «favor») with the aid of spelling variant pairs from VarCon (http://wordlist.aspell.net/varcon/). In total, we gather over 1600 treaty texts spanning from 1959 (when the first Germany-Pakistan BIT was signed) to 2015. To the best of our knowledge, this is the largest structured data set of English international investment agreement texts in the literature.

We next engaged in a number of pre-processing steps of meta-data and text to facilitate our ensuing analysis. For bilateral investment treaties (BITs) we re-ordered parties in each treaty based on GDP per capita at the date of treaty signature grouping the richer treaty party first. Plurilateral agreements such as the North American Free Trade Agreement kept their name and are excluded from analyses that require dyadic data. We then split each treaty into articles by relying on pre-existing HTML mark-up from Kluwer and Investment Claims.com data and manually introduced mark-up for UNCTAD data. The procedure yielded almost 25 thousand article texts in total.

3. Methodology

We use our dataset of raw treaty texts to investigate similarities and differences across agreements and articles. We proceeded in three stages. First, we developed a similarity measure to calculate distances between agreements and articles. Second, we constructed heat maps to identify differences between treaties. Third, we employed diffs to color-code word-level variations among comparable articles thereby facilitating the manual detection of textual differences.

3.1. q-character gram representations of treaty texts and their Jaccard distance

To identify differences and similarities between agreements and articles, we follow an approach much akin to that employed in plagiarism detection software. First, we break down each treaty into its 5-character-long substrings and count the number of times each substring occurs in the document.[5] To illustrate, the imaginary document «shall not be permitted» will contain the following 5-character substrings: «shall», «hall_», «all_n», «ll_no», «l_not», «_not_», «not_b», «ot_be», «t_be_», «_be_p», «be_pe», «e_per», «_perm», «permi», «er-

Treaties on Foreign Investment: Creation and Use of a Critical Index, Journal of International Economic Law, 18 (2015), 79–115.

[4] An earlier version of this paper was presented at the 2015 JURIX Conference.

[5] A. Spirling, U.S. Treaty Making with American Indians: Institutional Change and Relative Power, 1784–1911, American Journal of Political Science 56 (2012), 84–97.

mit», «rmitt», «mitte», «itted» («_» signifies space). Second, we compute the *Jaccard* distance between two treaties based on the substrings that overlap between the pair. To continue with the above example, the document «shall not be permitted» and a second document «shall be permitted» will have similar substrings, except for «all_n», «ll_no», «l_not», «_not_», «not_b», «ot_be», «t_be_». This divergence is caused by the presence of «not» in the first document and can be quantified by counting the number of unique 5-character substrings appearing in both documents and dividing it by the total number of unique 5-character substrings in the two documents (and subtracting this figure from 1). Applying this method to our set of two documents would yield a *Jaccard* distance of 0.48 – a measure of dissimilarity between two documents with 1 involving two very different documents and 0 involving identical documents. Transposed to international investment treaties, this *Jaccard* distance allows us to determine what treaties are similar to each other and what treaties are farther apart revealing new clusters and patterns in our treaty data.

3.2. Heat map representation of Jaccard scores

To make differences in the *Jaccard* distance between treaties immediately visible, we display them in a heat map. Each tile in our heat map is a comparison between two treaties. The heat map color-codes short distances (very similar treaties) as red while large distances (very different treaties) are represented by yellow. Users can then visually investigate similarity through color patterns.

3.3. Diffs and article-level comparisons of treaty texts

To go beyond mere similarity scores, we introduced a diff-based comparator. First, we devised an automated matching algorithm that identified corresponding articles for each treaty pair to be compared. The algorithm first compared article headers to identify articles on the same subject in both treaties (e.g. the «Expropriation» article in treaty A would be matched to the «Expropriation» article in treaty B). However, even where article headers diverge the article may in fact concern the same subject matter (e.g. the «Expropriation» article in treaty A may correspond to the «Nationalization» article in treaty B). To catch such correspondence, we also calculate *Jaccard* distances between article texts matching them if their distance is below a set threshold. Articles that cannot be matched are excluded from the subsequent comparison. Second, we produced a diff on matched articles. To enhance readability and to focus on legally significant differences, we employed a stop word list of terms. The diff algorithm disregards the terms listed and thereby exclusively displays legally relevant variation between texts. What results is a comparison between two treaties on the article-level where words that are unique to each document are color-coded.

4. Applications

Our representation of investment treaty texts and their similarities yields a number of applications of interest to both researchers and practitioners.

4.1. Systemic and country-level comparisons

Our similarity representations allow users to compare treaties both at the global and at the country level. The global level is suitable for identifying systemic trends. In a separate study, for instance, we use the *Jaccard* distance measure to trace consistency and innovation in the BIT universe demonstrating, amongst others, that developed countries tend to possess internally-coherent treaty networks suggesting that they are the system's rule-makers whereas developing countries displaying low coherence scores are the rule-takers.[6] In the same

[6] W. Alschner and D. Skougarevskiy, Consistency and Legal Innovation in the BIT Universe, Stanford Public Law Working Paper, No. 2595288 (2015), available at: http://papers.ssrn.com/abstract=2595288.

paper, we also highlighted the usefulness of tracing developments at the country level. Without requiring external input, *Jaccard* distances can reveal when countries shift from one treaty template to another allowing researchers to identify changes of a country's evolving investment policy.

4.2. Treaty- and article-level comparisons

On the treaty-level, *Jaccard* distances can trace policy diffusion processes and allow for the placing of a treaty in the wider universe of agreements. An analysis of the Trans-Pacific Partnership's (TPP) Investment Chapter concluded in 2015, for instance, revealed that 82% of the Chapter's main text has been copied and pasted from the United States-Colombia Free Trade Agreement (FTA) signed in 2006.[7] In the same paper, we also highlighted that article-level *Jaccard* distances can be used to identify the articles responsible for differences between two agreements. In the case of the TPP, for instance, the National Treatment clause is virtually identical (almost 95% of similarity) to the corresponding clauses in the United States-Colombia FTA, while the Minimum Standard of Treatment clauses only share a 70% similarity suggesting that the latter but not the former is a driver of dissimilarity.

4.3. Applications in negotiations and litigation

By allowing for a fine-grained comparison of treaties, our approach has applications beyond academia. Indeed, knowledge about similarity and differences among agreements can prove vital for practitioners. Negotiators can use such information to identify convergence and divergence in their respective country's policy preferences. Moreover, in a multilateral setting, textual similarity can help structuring negotiations around a common denominator text. Textual similarity can also help litigators to distinguish or analogize treaty language. It thus becomes easier for lawyers to find the treaty most helpful to their client's case.

5. Web-based Tool and Future Developments

In order to allow users to engage directly with our *Jaccard* distance representation of BITs, we have developed an interactive web-based tool that can be accessed via www.mappinginvestmenttreaties.com. The tool allows users to quickly identify differences and similarities between treaties and to interpret them substantively using diffs. Steps to expand the analysis are in the pipeline. In particular, we aim at moving from textual to semantic similarity to more accurately depict legal differences.

6. References

Alschner,W. and D. Skougarevskiy, Consistency and Legal Innovation in the BIT Universe, Stanford Public Law Working Paper, No. 2595288 (2015), available at: http://papers.ssrn.com/abstract=2595288.

Alschner,W. and D. Skougarevskiy, The New Gold Standard? Empirically Situating the TPP in the Investment Treaty Universe, CTEI Working Paper, No. 2015-08 (2015).

Chaisse, J., and C. Bellak, Navigating the Expanding Universe of International Treaties on Foreign Investment: Creation and Use of a Critical Index, Journal of International Economic Law, 18 (2015), 79–115.

Manger, M.S., A Quantitative Perspective on Trends in IIA Rules, in: A. de Mestral & C. Lévesque (eds.), Improving International Investment Agreements, London, Routledge, 2011.

Spirling, A. U.S. Treaty Making with American Indians: Institutional Change and Relative Power, 1784–1911, American Journal of Political Science 56 (2012), 84–97.

UNCTAD, World Investment Report 2015: Reforming International Investment Governance, United Nations, Geneva, (2015).

[7] W. Alschner and D. Skougarevskiy, The New Gold Standard? Empirically Situating the TPP in the Investment Treaty Universe, CTEI Working Paper, No. 2015-08 (2015).