

JULY 2017

UNCTAD Research Paper No. 5
UNCTAD/SER.RP/2017/5/Rev.1**Wolfgang
Alschner**

University of Ottawa

**Julia
Seiermann**Division on
International Trade in
Goods and Services,
and Commodities
UNCTAD**Dmitriy
Skougarev
skiy**European University
in St. Petersburg &
Graduate Institute of
International and
Development Studies

Text-as-data analysis of preferential trade agreements: Mapping the PTA landscape

Abstract

Preferential trade agreements (PTAs) form an intricate web that connects countries across the globe. In this article, we introduce a PTA text corpus and research tools for its fine-grained, automated analysis. Recent computational advances allow for efficient and effective content analysis by treating text as data. We digitize PTA texts and use textual similarity tools to assess PTA design patterns on the global, national, and chapter level. Our descriptive analysis reveals, inter alia, that PTAs are more heterogeneous as a group than, for instance, bilateral investment agreements, but that they converge in regional or inter-regional clusters of similarly worded agreements. Following our descriptive account, we provide three concrete, interdisciplinary examples of how text-as-data analysis can advance the study of trade economics, politics, and law. In trade economics, similarity measures can provide more detailed representations of PTA design differences. These allow researchers to capture more meaningful variation when studying the economic impact of PTAs. In trade politics, scholars can use treaty similarity to trace design diffusion more accurately and test competing explanations for treaty design choices. Finally, in trade law, similarity measures offer new insights into the processes of normative convergence between legal regimes such as trade and investment law.

Key words: Text-as-data, preferential trade agreements

UNITED NATIONS

The findings, interpretations, and conclusions expressed herein are those of the author(s) and do not necessarily reflect the views of the United Nations or its official Member States. The designations employed and the presentation of material on any map in this work do not imply the expression of any opinion whatsoever on the part of the United Nations concerning the legal status of any country, territory, city, or area or of its authorities, or concerning the delimitation of its frontiers and boundaries.

This paper represents the personal views of the author(s) only, not the views of the UNCTAD secretariat or member States. The author(s) accept sole responsibility for any errors. Any citation should refer to the author(s) and not the publisher. This paper has not been formally edited.

Contents

Acknowledgements.....	2
Introduction.....	3
1. Investigating Treaty Design: From Hand-Coding to Text-as-Data.....	4
2. Mapping the PTA Landscape.....	6
3. Trade Economics - Going Deeper than "Depth".....	17
4. Trade Politics - Detecting Design Diffusion.....	20
5. Trade Law - Tracking Legal Convergence.....	23
Conclusion.....	26
References.....	27

Acknowledgements

The authors are grateful to seminar participants in Villars, Geneva and Lausanne for useful comments and suggestions on previous versions of the paper.

We thank Veronika Zhirnova and Kseniia Tumasova for research assistance and gratefully acknowledge the funding support of the SNSF project “Convergence versus Divergence? Text-as-data and Network Analysis of International Economic Law Treaties and Tribunals” (Grant Number: 162379).

Introduction

Preferential trade agreements (PTAs) have proliferated globally since the 1990s and now form a complex network of agreements connecting virtually every country in the world (Pauwelyn and Alschner 2015). While their count can diverge depending on the definition that is used, the World Trade Organization's (WTO) repository of regional trade agreements (RTAs) notified to the WTO lists 291 PTAs in force, which can be considered a conservative lower bound.¹ Many of these PTAs are hundreds of pages long and contain an intricate web of detailed obligations. Due to its growing size and complexity, the PTA universe has become increasingly hard to navigate for scholars, policy-makers, negotiators, and litigators. At the same time, the need for understanding PTAs to better assess their impact, importance, and evolution has never been greater with continued deadlock in WTO Doha Round negotiations and the emergence of new mega-regional deals such as the Comprehensive Economic and Trade Agreement (CETA) between the European Union and Canada and the on-going negotiations of a Tripartite Free Trade Area in Africa or of the Regional Comprehensive Economic Partnership (RCEP) in South East Asia.

In this article, we introduce a new structured corpus of digitized PTA full texts drawn from the WTO RTA database, and apply text-as-data tools to map the design of PTAs. We argue that textual similarity measures are particularly suitable to capture fine-grained treaty design differences and find that the term PTA regroups a set of very heterogeneous agreements, which vary systematically in scope, content and language. Moreover, in contrast to bilateral investment treaties (BITs), which largely follow the design of country-specific model agreements of developed countries, PTAs cluster in regional or inter-regional groups of similarly worded agreements. These agreements furthermore display a greater degree of customization amidst standardization than BITs. Finally, even PTAs that are similar in overall design typically display important textual variation in specific chapters.

In addition to providing a descriptive account of the design variation of PTAs, this article illustrates how similarity measures can be harnessed for specific disciplinary research applications. In trade economics, textual similarity can complement existing measures of depth to capture more of the meaningful variation of treaty design when estimating the impact of PTAs on trade flows, as we highlight for the Trans-Pacific Partnership (TPP) Agreement after the withdrawal of the United States of America. In trade politics, textual similarity metrics can help trace the diffusion of treaty design, which we exemplify by tracking the diffusion of NAFTA throughout the Pacific Rim. Finally, in trade law, textual similarity can shed new light on normative convergence processes. Investment chapters in PTAs, for instance, are usually perceived to be contributing to the convergence of trade and investment law, but, as our analysis shows, potentially also fragment the two fields.

1. Investigating Treaty Design: From Hand-Coding to Text-as-Data

The empirical study of the design of trade agreements has made major advances over the past decade. From quantitative studies that treat PTAs as a 0 or 1 depending on whether they are absent or present in an inter-state dyad, we have moved towards an ever more detailed assessment of the design of ever more treaties. Projects relying on traditional hand-coding have spearheaded these efforts. Yet, semi- or fully-automated methods that treat PTA texts as data have now become available. They have the potential of adding new scale to the empirical investigation of treaty design by analysing vast amounts of information more efficiently and effectively.

¹ WTO RTA database, available at: <http://rtais.wto.org/UI/PublicMaintainRTAHome.aspx>. Even though the WTO has adopted the term "RTA", the database also contains inter-regional preferential trade agreements. In the WTO context, the abbreviation PTAs is used for unilateral preferential tariff arrangements accorded by developed to developing countries. In this article, we use the term "preferential trade agreement" for reciprocal trade agreements between two or more partners.

1.1 Text-as-data as the new frontier in PTA content analysis

The past decade has seen a push towards a more fine-grained understanding of design variation across preferential trade agreements. In 2010, Horn, Mavroidis and Sapir investigated the content of 28 PTAs concluded by the European Union and the United States of America across 52 subject areas distinguishing between “WTO-plus” (where PTAs provide more ambitious commitments than the WTO agreements) or “WTO-extra” (policy areas covered in a PTA but not in the WTO agreements) and their legal enforceability (Horn, Mavroidis, and Sapir 2010). Subsequent work by the WTO in 2011 (WTO 2011) and, more recently, by Hofmann, Osnago, and Ruta has extended this mapping to all PTAs notified to the WTO until 2015 (Ruta, Hofmann, and Osnago 2017). Parallel efforts have been undertaken by Baccini, Dür, and Elsig to study PTAs in their own right (rather than in reference to the WTO baseline). As part of their Design of Trade Agreements (DESTA) project, researchers have coded over 600 PTAs across more than 100 content variables (Dür, Baccini, and Elsig 2014). Together, these studies have provided unprecedented insights into the design of trade agreements and spurred further research into the causes of varying PTA design as well as its effects (e.g. Dür, Baccini, and Haftel 2015; Pauwelyn and Alschner 2015; Felbermayr, Aichele, and Heiland 2016; Baccini, Pinto, and Weymouth 2017).

Hand-coding of PTAs has both advantages and disadvantages. On the one hand, if done by subject matter experts, hand-coding can capture nuanced differences in treaty design that are difficult to spot for laymen or computers. Moreover, hand coders can, through their background knowledge, differentiate between meaningful variation in language and mere stylistic variation more easily than computers. On the other hand, hand-coding is labourious, expensive and time consuming and thus less scalable. Resource constraints limit the number of variables that can be coded and make it cumbersome to extend the coding to new treaties or new variables. Hand-coding is also a deductive enterprise as researchers identify the content variation they want to investigate *ex ante* or after a sample coding. As a result, content variables outside the contemplation of the original coders or the test sample risk to be systematically omitted.

To overcome these shortcomings and to add a new scale to investigations, empirical scholars in the social sciences increasingly turn to semi- or fully automated means of content analysis as part of the emerging field variously known as “text-as-data” research (Grimmer and Stewart 2013; Gentzkow, Kelly, and Taddy 2017; Alschner, Pauwelyn, and Puig 2017), “computational social science” (Lazer et al. 2009) or the “digital humanities” (Berry 2012). What these approaches have in common is that they rely on recent advances in computer science to process social or textual data. The standardized nature of the legal language of international treaties provides particularly fertile ground for the deployment of text-as-data approaches from rule-based key word searches (Manger and Peinhardt 2017) to similarity measures akin to those employed in plagiarism detection software (Alschner and Skougarevskiy 2015) or machine learning applications (Alschner and Skougarevskiy 2016a).

Text-as-data approaches have three advantages over traditional hand-coding. First, provided that properly formatted and machine-readable text exists, text-as-data tools can analyse thousands of pages in a matter of minutes or even seconds. They are thus scalable and highly efficient when dealing with design variation in large text corpora. Second, text-as-data measures are easily adjustable and extendable making them versatile, effective, and attractive for use across projects that investigate similar data, but from different research or disciplinary angles. Third, text-as-data measures are not tied to any particular conceptual framework. They can thus be employed in concert with a multitude of theoretical lenses as well as inductively to reveal latent patterns in texts that were previously unknown. Given these advantages and potential, text-as-data can be seen as a promising new frontier in the empirical study of treaty design.

1.2 Textual similarity analysis of PTAs

Text-as-data analysis comes with an entire toolkit of approaches that can be applied to the study of PTA design (Grimmer and Stewart 2013). For instance, dictionary methods used to study tone and sentiment of texts by compiling lists of positively or negatively connoted words can be repurposed to detect characteristics in legal texts such as their legal bindingness. Supervised machine learning can be harnessed to automatically code treaty features after having been trained on a sample of hand-coded data. Unsupervised machine learning

tools, in turn, can be employed to automatically cluster treaties in groups of similar agreements. Which approach (or combination thereof) is chosen depends on the research question pursued and the level of sophistication required.

In the interest of parsimony, we restrict ourselves to one simple but powerful text-as-data method in this article: textual similarity. According to Alschner and Skougarevskiy textual similarity metrics can “reveal latent treaty design differences and trends that we did not know to exist nor actively looked for beforehand” (Alschner and Skougarevskiy 2016b: 564).² To illustrate why detecting latent design differences may be important, just ask yourself how many variables it takes to account for meaningful design variation in PTAs. DESTA uses over 100 variables to code the content of entire PTAs (Dür, Baccini, and Elsig 2014). The mapping project of the United Nations Conference on Trade and Development (UNCTAD), in contrast, already relies on over 120 variables just to code design variation of individual investment chapters within larger PTAs.³ So how many variables does it take to comprehensively map the design of an entire PTA – 500, 1000 or more? Not only is this question difficult to answer in the abstract, but responses are likely to vary between disciplines as researchers differ on the design variation they consider to be of theoretical or practical importance. Textual similarity allows us to side-step this question by using textual variation as a proxy for fine-grained design distinctions that can then be adapted to specific disciplinary research contexts.

The intuition underlying textual similarity assessments of treaty design is simple. Where language overlaps, PTAs follow similar design. Where language diverges, so does PTA design. Prior research suggests that textual similarity is a useful proxy for meaningful treaty design differences. With respect to bilateral investment treaties, Alschner and Skougarevskiy demonstrate that such metrics can expose power asymmetries in rule-making, track consistency and innovation in national treaty programs, trace diffusion of treaty elements and assess a treaty’s novelty (Alschner and Skougarevskiy 2016b). In the PTA context, similar techniques have been employed to show that the United States of America wrote the rules of the TPP (Alschner and Skougarevskiy 2016c; Allee and Lugg 2016), to highlight innovation in the Comprehensive Trade and Investment Partnership (CETA) (Allee, Elsig, and Lugg 2017a) or to account for the presence of WTO references in PTAs (Allee, Elsig, and Lugg 2017b).

There are a number of reasons why textual similarity metrics are a particularly effective proxy for design variation in the legal domain. First, legal language tends to be more uniform than natural language. Contracts, statutes and treaties are highly formulaic, structured and standardized documents full of terms of art that carry their own legal meaning. While this makes such documents harder to read for humans, they become more amenable to computational analysis as they tend not to vary for stylistic reasons alone.

Second, lawyers place a premium on predictability given that legal language has legal consequences. As a result, they employ tried and tested (“boilerplate”) language (Baird 2006). Even when they innovate, legal drafters tend to try to build on existing formulations rather than inventing new terms from scratch (see e.g. Vandeveld 2009, 108). Moreover, specifically in relation to PTAs, multilateral treaties, such as the WTO agreements, provide important reference language that is often replicated in PTA texts, adding to their homogeneity and path dependency (Allee, Elsig, and Lugg 2017b).

Third, the texts of PTAs emerge from a domestic political and institutional context that privileges textual uniformity and continuity across consecutive agreements. Legislative directives, such as the Trade Promotion Authority bill passed by the United States Congress in 2002, create clear expectations on what trade agreements should and should not contain, limiting the flexibility of the Executive.⁴ Furthermore, bureaucracies tend to rely on the language of existing treaties when negotiating new ones (Puig 2013). All these reasons make uniformity among PTA texts common – and deviations meaningful. Textual similarity is thus a particularly promising instrument from the toolkit of text-as-data methods to study treaty design variation across PTAs.

² See also www.mappinginvestmenttreaties.com.

³ UNCTAD’s IIA University Mapping Project, available at: <http://investmentpolicyhub.unctad.org/Upload/Documents/UNCTAD%20IIA%20MAPPING%20PROJECT%202013-2014.pdf>.

⁴ On the role of parliaments in creating templates for negotiations see Hornbeck and Cooper (2011) (for the US); Fecák (2011) (for Czechia).

At the same time, we acknowledge that textual similarity measures also have obvious limitations. In contrast to theoretically derived concepts describing institutional design such as a treaty's degree of commitment, flexibility or delegation (Koremenos, Lipson, and Snidal 2001; van Aaken 2009), there are no prior expectations as to how similar PTAs should be or what dissimilarity means in legal or policy terms. That is because similarity is a relational concept, which requires comparison and context to become meaningful. In consequence, textual similarity assessments are particularly useful and intuitive when studying relational processes with clear benchmarks such as negotiation outcomes based on initial positions, diffusion of texts or the novelty of language (see e.g. Berge and Stiansen 2016). Another shortcoming of textual similarity is that textual variation does not always imply a variation of meaning. Some linguistic choices have large repercussions ("shall" vs "should") while others do not. Textual similarity measures thus benefit from being combined with other text-as-data tools, hand-coding or contextual legal analysis to guide the interpretation of textual differences.

In short, textual similarity is an imperfect proxy for treaty design variation. At the same time, it is also a useful and easily scalable one. It can reveal meaningful latent design variations difficult to detect through traditional methods and can conveniently complement legal analysis and detailed feature coding. By mapping the PTA universe and showcasing a set of interdisciplinary research applications, we hope to illustrate the potential textual similarity assessment has to offer.

2. Mapping the PTA Landscape

Turning now to our empirical mapping of PTAs, we will first introduce our text corpus. We then characterize our data through more traditional proxies of treaty design such as PTA type, treaty length and feature content before investigating the textual similarity between the agreements of our corpus. We find that PTAs are heterogeneous as a group – much more so than bilateral investment treaties. This variation in terms of agreement type, scope, content and language provides important opportunities for empirical studies on PTA design in trade economics, politics, and law as we show in the next section.

2.1 Dataset

To map the PTA landscape, we have created a text corpus based on the WTO RTA database, the primary official repository of preferential trade agreements.⁵ That dataset comprises 447 PTAs that WTO members have notified to the organization. These agreements have been signed between 1948 and 2015 and involve 202 parties. 60% of them are currently in force; the remaining treaties have been signed and await ratification or have been replaced or suspended. To that list, we added the Transpacific Partnership Agreement (TPP) to account for important recent trends in PTA-making.⁶ Of these treaties, 414 are in English full texts, 33 in Spanish and one in French.

The WTO RTA database comprises four types of PTAs⁷ (with multiple assignments for the same PTA being possible in some combinations): (1) "Customs Unions" (CUs) pursuant to Article XXIV:8(a) of the GATT or Article 2(c) of the Enabling Clause (between developing countries); (2) "Goods Free Trade Agreements" (FTAs) pursuant to Article XXIV:8(b) of the GATT or Article 2(c) of the Enabling Clause (between developing countries); (3) "Services FTAs" pursuant to Article V of the GATS (listed in the RTA database as "Economic Integration Agreement"); and (4) "Partial Scope Arrangements" (PSAs) between developing countries liberalizing trade on a more limited number of products based on Article 2(c) of the Enabling Clause.

For each of these treaties, the WTO RTA database contains meta-data and links to their full texts, which we exploit to build our corpus. After collecting all full texts from the stipulated sources and manual search in cases

⁵ Regional Trade Agreements Information System, available at <http://rtais.wto.org/UI/PublicMaintainRTAHome.aspx>.

⁶ We intend to add further agreements outside the WTO RTA dataset in the future.

⁷ See the WTO RTA Database User Guide: http://rtais.wto.org/UserGuide/RTAIS_USER_GUIDE_EN.html.

of broken links, we deleted annexes and schedules to focus on the main body of each agreement. Through optical character recognition, text structure information and manual work, we then transformed treaties from diverging formats into a unified and machine-readable marked-up text format (XML). This type of digitized textual data allows us to capture the structure of agreements, distinguishing between chapters, articles and paragraphs as well as headers and full text for our subsequent analysis, and permits the deployment of text-as-data tools.

The WTO RTA database only lists those agreements that have been notified by WTO members. The Secretariat itself acknowledges that over 70 known PTAs that WTO members should ordinarily have notified are missing from the list.⁸ Yet, patterns existing in the WTO RTA dataset are likely to be indicative of structures present in the wider universe of non-notified PTAs. Pending future efforts to expand our corpus, the sample of the WTO RTA dataset, thus provides a useful starting point to study the PTA landscape.

2.2 The scope, content and evolution of PTAs

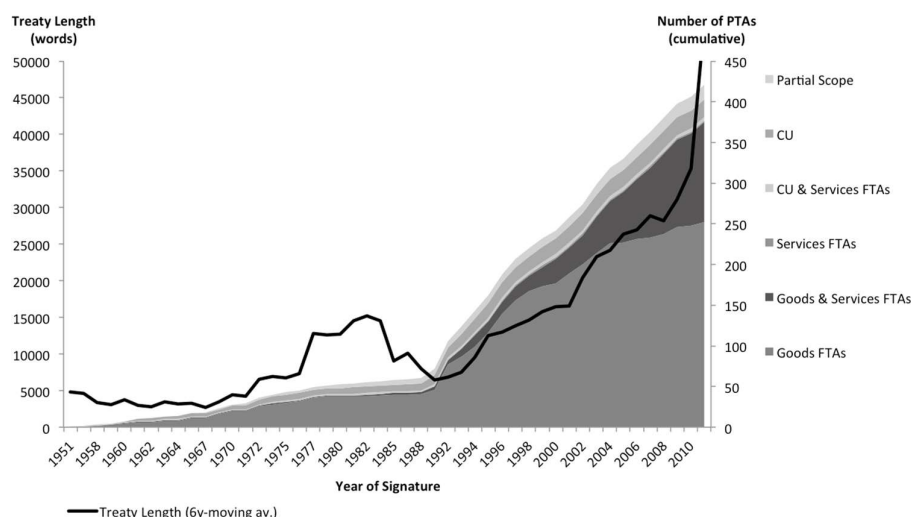
In this section, we offer a preliminary characterization of the scope, content and design evolution of the PTAs in our database. We begin by looking at treaty length. The number of words in an agreement is a first indicator of a treaty's coverage and complexity. Treaty length varies systematically across PTA types. As can be seen in Table 1, Partial Scope Agreements are shortest, with an average of 2545 words per treaty, followed by Customs Unions and Goods FTAs. Agreements that contain services components are wordier and Goods & Services FTAs are by far the longest agreements in our dataset.

Table 1. PTA type counts, length and share of texts with chapter structure

Type of PTA	# of PTAs (en, sp, fr)	mean PTA length (words)	% of PTAs with Chapters
Partial Scope Arrangements	18 (15,3,0)	2545	33%
Customs Unions	22 (19,2,1)	5672	83%
Customs Unions & Services FTAs	7 (7,0,0)	10095	71%
Goods FTAs	256 (255,1,0)	5063	59%
Services FTAs	2 (2,0,0)	15095	100%
Goods & Services FTAs	142 (116,16,0)	36360	96%

Average treaty length has increased tenfold over the past 60 years, from a mean of about 5000 words in the 1950s to over 50 000 words in the 2010s. As illustrated by Figure 1, this increase in length is driven in large parts by the conclusion of wordier Goods & Services FTAs starting in the early 2000s, which continued to proliferate while the number of existing Customs Unions stagnated and Goods FTA grew only modestly.

⁸ Secretariat Note to the Committee on Regional Trade Agreements, List of RTAs which have appeared in factual presentations and have not yet been notified to the WTO, 21 June 2016, WTO Doc Ref. WT/REG/W/104.

Figure 1. Evolution of PTAs per type and increase in length

Treaty length, however, tells us little about treaty content. To investigate content, we make use of the fact that contracting states structure most PTAs along subject-specific chapters, e.g. on services, intellectual property rights or investment, to better manage their complexity. This structure is present in 319 of our 447 English-language PTA texts (71%).⁹ In total, these PTAs contain 2688 chapters, which we classify by subject matter into 73 substantive categories, such as “Investment”, “Financial Services”, or “Competition”.¹⁰ Each chapter is assigned only one such category. Aggregated by treaty this chapter-level coding provides a snapshot of the content of a PTA.

To compare the scope of content features across PTA types, we count the number of unique categories in each treaty. Plotted in Figure 2, the feature coding for PTAs with chapters corroborates the scope differences between PTA types, but adds further nuances. Goods & Services FTAs cover considerably more issues than all other types of PTAs with a median of 17 features per treaty. In comparison, the median of Goods FTAs is lowest with 4 features, while Partial Scope Arrangements, Customs Unions, and Services FTAs have slightly higher, but still single-digit median scores. While generally word counts and feature counts of PTAs are strongly and positively correlated,¹¹ this relationship holds less well with respect to Partial Scope Agreements. This is due, in part, to the fact that word counts understate the coverage of agreements, which, like the Mercosur–India PTA (2004) or the El Salvador–Cuba PTA (2011), reaffirm WTO commitments (e.g. on trade remedies, sanitary measures or technical barriers) or incorporate them by reference rather than spelling them out.

⁹ For the purpose of this paper we consider a “chapter” to be the highest level in the hierarchy of an agreement. Most

预览已结束，完整报告链接和二维码如下：

https://www.yunbaogao.cn/report/index/report?reportId=5_9286

