1. Sense and similarity: automating legal text comparison*

Wolfgang Alschner

1. INTRODUCTION

As lawyers we often compare legal documents. We want to know which case is similar to the one we are currently litigating; we want to know where a contract deviates from an industry template; we want to know which part of a regulation was amended following a public comment procedure or how the commercial code in Canada differs from the commercial code in the United States. In short, a wide range of legal tasks involves a comparison between two or more legal texts. Improvements in technology now allow us to rethink and to scale up such legal document comparisons. Whereas human lawyers struggle to compare more than a few dozen contracts, statutes or cases in a day, computers can relate thousands of legal texts to each other in a few seconds. This creates a host of new opportunities for lawyers and legal researchers engaged in comparative legal analysis.

In this chapter, I will outline the technological underpinnings of automated legal text comparison and discuss the practical implications, applications and limitations arising from its increasing use. Automated legal text comparisons change comparative legal work in two ways. First, existing comparative tasks can be performed more efficiently through technology. Second, novel tools are added to the repertoire of legal scholars and lawyers to automatically identify patterns of similarities and differences in large legal corpora. Comparative text analysis of big legal data can thus reveal insights that were impossible to detect for human researchers using traditional methods. Automated text comparison is thus becoming a powerful and versatile tool in a range of legal research

* I gratefully acknowledge the research assistance of Samantha Lo and the financial support of the Ontario Early Researcher Award “Equipping Tomorrow’s Lawyers: Using Artificial Intelligence for Automatic Legal Document Production and Analysis”.

and technology applications. At the same time, these new tools primarily complement rather than substitute traditional comparative analysis, since they focus on comparing text tokens rather than their underlying meaning, interpretation and significance.

This chapter is structured as follows. First, I introduce the technology of modern automated text comparison and discuss its promises and limitations. In that context, I highlight the important distinction between similarity of text and similarity of meaning. I then explore how automated text comparison scales and improves the performance of existing legal work and introduces new tools to lawyers to carry out analyses on big legal data. Specifically, I discuss applications of automated legal text comparison for the study of (1) specific legal documents, (2) legal processes and (3) large legal corpora. The final section of the chapter concludes on how scholars and lawyers can acquire the skills to leverage automated legal text comparisons in their own work.

2. THE TECHNOLOGY OF AUTOMATING TEXT COMPARISON

To automate and scale legal document comparison, we have to re-conceptualize how we deal with legal texts. Lawyers and legal researchers compare legal texts through *close reading*. As humans, we carefully scan documents and can pick up a wide variety of commonalities between texts in the process, from similarity in content to related underlying legal concepts or parallels in writing style. Moreover, we can read between the lines, detect subtle irony and interpret words drawing from our legal background knowledge. What we cannot do, however, is to conduct legal document comparison at scale. For that we need what Franco Moretti termed *distant reading* (Moretti, 2013). In essence, we need to treat text as data to analyze more of it in less time. While we will lose some of the benefits associated with close reading in the process, distant reading not only offers efficiency gains but also an entirely new perspective on the law.

In this section, I explain how to transform text to data, how to compare the resulting text representations and how to visualize and interpret the results. Finally, I offer some reflection on the limitations and promises of automated text comparison.

In order to make the sometimes-technical discussion more tangible, I will lead with an example. Say you are a young associate in a tax law boutique in Toronto. One of the partners asks you to draft a memo on the evolution of the Canadian Federal Income Tax Regulation (C.R.C.,

c. 945) since the early 2000s.¹ The partner knows that the Regulation has been revised many times during that period but also knows that some revisions have been more important than others. For a report the firm is now working on, they specifically want to focus on major revisions to that Regulation, which you are tasked to identify. You quickly realize that the Regulation is extremely long, containing more than 360,000 words on 1217 pages in its current version. Moreover, the Regulation has been revised 72 times since 2004. In short, an impossible task for close reading. But, as we will see, not for distant reading.

1) Transforming Legal Text to Data

There are many methods to transform text to data, that is, to represent or model text quantitatively, to make it amenable for distant reading.² What these methods have in common is that, as Grimmer and Stewart put it, “all quantitative models of language are wrong” (Grimmer and Stewart, 2013, p.4). That is because a text is partially constructed through information outside of the text. We understand the semantics of a text – the sense of words, legal concepts, connotations, irony or puns – because we can tap into a wider pool of knowledge that computers (currently) do not have access to. As a result, any conversion from text to data will lead to some of that semantic information, which would be accessible to human readers, to be lost in translation. This is not necessarily a problem, though. As Grimmer and Stewart stress, quantitative models of language can be useful even if they are “wrong” (Grimmer and Stewart, 2013, p.4). The benchmark of any text-as-data representation is not how accurately it can represent meaning, but how useful it is in relation to the performance of a specific task.

We can think about available methods to convert text to data as a spectrum between model simplicity and complexity. From that spectrum, users can select the text-as-data model that corresponds most closely to their research needs. On one end of the spectrum are simple language models that entail a greater loss of semantic information, but that are easy and quick to implement. They include so-called bag-of-words approaches, which treat words as tokens whose frequency in a text is counted without regard to word order (Manning et al., 2008). Unigram language models,

¹ The Regulation is available at http://laws.justice.gc.ca/eng/regulations/C.R.C._c._945/index.html (retrieved 12 March 2019). The subsequent analysis uses point-in-time versions of that Regulation available at ftp://205.193.86.89/PITXML/regulations/C.R.C._c._945/ (retrieved 12 March 2019).

² For an introduction to natural language processing and text-as-data analysis, see Arnold and Tilton, 2015, ch. 9; Manning et al., 2008, ch. 12.

for instance, count the frequency of individual words, bigram models count the frequency of word pairs, trigrams consider three consecutive words, and so on. A step closer towards greater complexity are character-gram models, which split text into chunks of consecutive characters, for example of five, six, seven or more character length, and thereby incorporate word order (see e.g. Spirling, 2012). This can be particularly important in law where the phrase “guilty not to” shares a word frequency with “not guilty to”, but they differ in their character sequence. Character-gram models, however, take longer to compute as they segment text into more fine-grained components.

As we advance on the spectrum towards greater complexity, models incorporate syntactical elements through so-called part-of-speech tagging to tell verbs from nouns and identify word dependencies in text (Arnold and Tilton, 2015, pp. 134–43). The currently most sophisticated data-driven language models treat words as vectors. These *word2vec* representations of language deploy neural networks to learn the contexts in which words appear. Words that are textually unrelated, such as “cat” and “dog” or “father” and “son”, but that are semantically related given their shared word context are then placed closer together in a vector space (Mikolov et al., 2013). But even the most complex procedures currently remain approximations of the deeper semantic connections between words that a lawyer would be able to detect. Those engaged in text-as-data conversion thus face a trade-off: the closer one gets towards model complexity the more contextual information is incorporated reducing the information lost in the text-to-data translation, but at the same time parsimony declines and computational costs increase.

The above description is only a cursory overview. Moreover, each text-as-data method comes with a set of additional methodological choices (Denny and Spirling, 2018; Grimmer and Stewart, 2013). These choices include whether to disregard punctuation, whether to eliminate certain words (such as “and”, “is”, “or”) or to focus on certain words (e.g. only nouns), whether to reduce words to their stems to limit variation (e.g. “gone” and “going” become “go”) and so forth. As a result, there is no single or “true” representation of text as data.

The choice among different language representations will instead depend on the task at hand. Surprisingly to most lawyers who are trained to see significance in the placement of a comma (and sometimes rightfully so), many automated text-processing applications do not rely on sophisticated models of language to add value (Manning et al., 2008). In order to classify 1000 cases on whether they deal with a contract breach or a tort claim, for instance, we do not need to worry about commas and can rely on a simple unigram model. Similarly, if we want to detect which contracts

from a pool of 1000 contracts follow a specific template, a relatively simple character-gram model that incorporates word order will work just fine. In contrast, where we want to identify semantically related but textually distinct legal concepts such as good faith, the clean hands doctrine and abuse of rights, a more complex, semantically informed conversion model is necessary. In short, the task dictates how simple a language model can be to still be useful. But in most cases, simple models work surprisingly well.

Following this brief introduction into text-as-data approaches, we can now apply these techniques to the task of the young associate who is to investigate the evolution of the Canadian Federal Income Tax Regulation. Since the objective is to identify major revisions of that Regulation, there is no need, at this initial stage, to employ a sophisticated, semantically aware language model. As long as that model can represent changes to a text, it will suffice. Given the significant length of the Regulation and the ensuing computational costs, the young associate can settle on the simple model of a unigram representation of text that counts the frequency of each word for each document and stores results in what is called a document-term matrix where every row represents one version of the 72 iterations of the Income Tax Regulation since 2004 and every column represents a word and its frequency in each of these versions. This representation will then allow the young associate to compare the different iterations of the Income Tax Regulation texts.

2) Comparing Texts Computationally through Distance Measures

When lawyers compare texts, they assess qualitatively what two documents have in common. In order to automate that process, however, we need to express the commonality between texts quantitatively. In other words, we need to find a way to compute how different or similar two texts are. In mathematics the notion of similarity or difference can be expressed as the *distance* between two objects. The intuition is that in a virtual space two identical texts will occupy the same spot, whereas two very different texts will be located far apart.

Having represented text as data, we can calculate the distance between two texts relatively easily. A range of different metrics exists that require different data formats and use different techniques to perform this calculation. The most intuitive of them is the *Jaccard* distance. It simply counts the share of textual elements that a text pair has in common in relation to all text elements present in the pair. Subtracted from 1 this number yields the Jaccard distance, otherwise it indicates Jaccard similarity (Alschner and Skougarevskiy, 2016). A useful property of the Jaccard measure is that it is bounded between 0 and 1. Other distance metrics include

Euclidean, or more popular for text data, *Cosine* distances that measure the distance between documents, typically represented as word frequency vectors in a document-term-matrix, in a vector space.³

Given that textual similarity can vary starkly from one text corpus to another and that some of the distance measures (such as Euclidean) are not bounded, it is useful to interpret distance measures through intra-corpus benchmarking. How similar are the two most similar documents in the corpus? How different are the two most different documents? What is the average similarity and standard deviation in the corpus? On that basis, texts can be compared and their commonality assessed relative to other texts in the same corpus.

Going back to our example, the young associate can now compute the textual distance between the 72 versions of the Income Tax Regulation. To keep things simple, the young associate will calculate the Jaccard distance between versions of the Regulation. He thus essentially computes to what extent the same words are contained in a pair of regulations that is being compared. Given that we are dealing with the same Regulation over time, that distance should be comparatively low. After running the analysis, the young associate finds that the average distance between all 72 regulations is 0.12 with a standard deviation of 0.08. Recall that the Jaccard distance is bounded between 0 and 1, with 0 indicating perfect similarity. The minimum similarity in our corpus is 0, that is, at least two versions of the Regulation are essentially identical when it comes to words used and the maximum distance is 0.30, which means that 30 percent of unique words are different in two versions of the same Regulation. As we would expect, this latter, maximum distance is between the first (2004) and the last version (2018) of the Regulation in our dataset highlighting an incremental change of the text over time.

3) Visualizing Text Comparisons

Comparing raw similarities scores can be sufficient for some legal tasks. If the goal is to find the most similar case to the case at hand, for instance, raw scores are enough to identify the match. If the goal is to compare a large number of texts based on their similarity, to identify structural breaks where small changes give way to more drastic ones or if the analysis is to prepare a subsequent targeted close reading of texts, further visualizations

³ For an informative comparison of different distance measures, see the blog post Chris Emmerly, *Euclidean vs. Cosine Distance*, 25 March 2017, retrieved 12 March 2019 from <https://cmry.github.io/notes/euclidean-v-cosine>.

are helpful. Through past research, I have come to appreciate the use of so-called heat maps to visualize textual similarity across a large corpus of texts.

The output resulting from applying distance measures to text-as-data representations is a distance matrix. Each cell in that matrix is a pairwise comparison of the similarity of two texts. What a heat map does is to overlay the raw distance scores with a color-coding that makes it easy to differentiate low and high scores. Pairs of similar texts can thus be easily spotted. Moreover, as I will further explain below in the section on corpora analysis, heat maps can also facilitate the identifications of groupings of similarly worded documents when combined with clustering algorithms. Heat maps can therefore reveal patterns as well as candidates for further manual inspection (e.g. to identify remaining differences between the most similar pair of texts).⁴

Going back to the Income Tax Regulation example, the young associate can plot the distance matrix calculated as a heat map. The results are displayed in Figure 1.1. The heat map's axes are identical and list the versions of regulations chronologically by their enactment date. The map (and distance matrix) is symmetrical, meaning that in the diagonal line from top-left to bottom-right the same version of the Regulation is always compared to itself. The segments of the heat map above and below that diagonal are mirror images of each other. Where two versions of the Regulation are identical or very similar, the heat map will color-code them dark, where two versions are different, they will be color-coded bright.

The heat map visualizes the incremental evolution of the Regulation's text over time where small changes to each new version slowly change the Regulation, creating the greatest deviation between the first and the last version. The heat map, however, also allows us to spot a significant structural break between the February 2013 and June 2013 versions of the Regulation, where rather than small, incremental change, we see an abrupt and significant change corresponding to an 8 percent textual difference between the two versions.

The young associate has thus fulfilled part of the task. Canada's Federal Income Tax Regulation changed incrementally between 2004 and 2018. Over that period, 30 percent of the Regulation was gradually revised. Only one major revision occurred changing 8 percent of the Regulation between February and June 2013. The associate can subsequently focus

⁴ Such further inspection can also be facilitated by technology. Diff algorithms, for instance, are programs that display text side-by-side, color-coding differences between texts. There are free diff websites, which can be useful to quickly assess textual differences between two documents. See e.g. <https://www.diffchecker.com>.

on that specific revision and explore the legal changes brought about by it. Without going into too much detail, the 2013 major revision of the Canadian Federal Income Tax Regulation implemented changes resulting from the 2012 Technical Tax Amendments Act (S.C. 2013, c. 34). Among other things, the Regulation introduced a new regime relating to the taxation of Canadian multinational corporations with foreign affiliates (Section 5907) and changed taxation rules for taxpayers with interests in offshore investment fund property (Section 202(6.1)).⁵

4) Promises and Limitations of Automated Text Comparisons

The automated comparison of legal texts provides a scalable means to efficiently assess the similarity of large corpora of legal documents. But it also comes with limitations that stem from the fact that such analysis is rooted in the comparison of text tokens and not meaning. As a result, where similarity in meaning is not accompanied by similarity in text, the findings of automated text comparisons can be misleading. This is most obvious when comparing identical texts in different languages: an automated text comparison will then classify the same texts as very different. This limits the applicability of automated text comparisons across languages without intervening translation. Similarly, even within a language, different words can often express similar meaning. While more sophisticated word2vec approaches can pick up the underlying semantic similarity, the more common bag-of-words models cannot.

The above example of the Income Tax Regulation highlights some of these limitations. While we now know that a significant revision of the Canadian Federal Income Tax Regulation occurred between February and June 2013, we do not yet know what specific legal changes it entailed. Furthermore, we only looked for the magnitude of textual changes. Other amendments of the same Regulation may have brought about equally or even more significant *legal* changes, but they did not lead to major *textual* changes and were thus not uncovered using the approach adopted. In part, these limitations can be overcome through a second iteration of the analysis using a more sophisticated language model or by computationally identifying what sections and what words were changed.

More generally, however, these limitations point to the *complementarity*

⁵ See Technical Tax Amendments Act, 2012, S.C. 2013, c 34, Summary, and Department of Finance, *Explanatory Notes Relating to the Income Tax Act, the Excise Tax Act and Related Legislation*, October 2012, retrieved 12 March 2019 from <https://www.fin.gc.ca/drleg-apl/nwmm-amvm-1012n-05-eng.asp>.

of close reading and distant reading. The latter can provide insights that traditional analysis could not have delivered (such as the structural break in the Tax Regulation of 2013), while the former allows one to interpret the textual changes and evaluate their legal significance. Distant reading is therefore no substitute for close reading.

Another way of looking at this is that automated text comparison is most useful when it plays to its inherent strengths. Automated text comparison works especially well when the task involves comparing text rather than interpreting content. Diverse applications in law, which I will discuss below, actually look for common text rather than common meaning, making automated text comparison useful. Moreover, such comparisons work particularly well with transactional or statutory documents, where text and normative content is typically closely intertwined. That is because contracts and statutes are written in a stylized, standardized legal language that does not normally vary for stylistic reasons alone. In contrast, judicial decisions are written in natural language and, as a result, normative content and text are more loosely connected making semantically agnostic textual similarity analysis somewhat less insightful. Ultimately, however, it depends on the specific task and choice of technique that determines what contributions automatic text analysis can make and what limitations remain.

In the remainder of this chapter, I will therefore showcase some useful applications for automatic text comparisons in law differentiating between the analysis of (1) targeted documents, (2) legal processes and (3) large legal corpora.

3. APPLICATIONS I: SUPPORTING LEGAL DOCUMENT ANALYSIS

In this section, I begin with a review of applications for automated text comparison in (1) legal search, (2) authorship identification, (3) legislative implementation and influence detection and (4) template deviation assessment. This list is not exhaustive, but it highlights the contribution automated legal text comparisons can make to support targeted document analysis.

1) Similarity Analysis in Legal Search and Recommender Systems

Say you found a judicial decision and want to know whether there are other cases like it. In the past, legal search engines sought to address such queries by relying primarily on metadata. A human extracted key information from a judgment and this information was then linked to the

decision in a database as metadata. Similarity of cases is then expressed as the similarity of metadata. The efficacy of this approach is dependent on the quality of the metadata, but since cases can be similar across many dimensions, for example they share the same legal principle, same facts, same parties, same location, and so on, there is a risk that metadata entries do not capture all potentially useful information.

Automatically comparing the full text of legal documents has therefore become a useful tool for legal search applications and legal recommender systems (Winkels et al., 2014). One way to think about a search operation is to imagine that legal documents form a network that contains both formal links, such as cross-references and shared metadata, but also informal links, such as similar text content, which can be revealed through full text comparisons (Landthaler et al., 2016). Textual similarity metrics can then find close neighbors of the target case in the network by comparing its text to all other decisions in the database. Furthermore, similarity measures can be incorporated into legal recommender systems to suggest documents similar to the one returned through the initial search query. Thereby users can gain access to pertinent cases, which they would not have found through curated metadata connections.

Challenges for these types of applications include the above-discussed text versus meaning distinction. Users are typically interested in finding semantically or legally similar cases, whereas most computational approaches can only provide textually similar ones. Cases may thus be similar because the same judge wrote them, not because they cover identical issues. At the same time, this type of similarity may also be useful to litigants. A second challenge relates to the choice of a threshold. Similarity is a continuum between the extreme dissimilar and the extreme similar. On what point of that spectrum do two documents become similar enough for a legal search or recommender algorithm to consider them linked? Commercial legal search and recommender systems typically do not disclose their specific matching algorithm, but research suggests that a combination of textual similarity metrics with other information such as ontologies and manual annotation provides the most promising results to leverage both formal and informal connections between documents (Geist, 2008; Landthaler et al., 2016; Schweighofer and Geist, 2007).

2) Similarity and Authorship Detection

The detection of document authorship through text comparison is popular in the field of linguistic forensics, and has applications in law. In contrast to legal search applications discussed above, similarity here is not primarily used to detect content links but stylistic ties between two documents,

which is why the field is also known as stylometry. In the process, the text of a document of unknown authorship is automatically compared to texts of known authors to identify the writer among the group of potential candidates. Such stylometry is used in law to produce evidence in litigation as part of the larger field of legal forensics (Coulthard and Johnson, 2010). For instance, in the *Yukos v Russia* arbitration, which yielded the biggest investment arbitration award to date of USD 50 billion, Russia sought to set aside the arbitral award in Dutch courts on the grounds that a linguistic analysis had revealed that, illegitimately, the tribunal's secretary and not the three arbitrators had written large portions of the award.⁶

3) Legislative Drafting

Another application of automated text comparison lies in legislative drafting to assess whether one document influences the text of another. On the one hand, such influence may be explicit. For instance, where states implement an international treaty in domestic law, international law provides the normative content that is to be translated into municipal law. Such national implementation of international law is particularly common in Europe where directives of the supranational European Union (EU) are being implemented through national measures within the organization's member states. While the EU currently tasks specialized consulting companies to manually monitor the faithful implementation of such directives, researchers have shown that text-based similarity measures can identify national implementing statutes with reasonable accuracy across different EU member states (Nanda et al., 2018).

On the other hand, textual similarity may also reveal less explicit textual sources that leave traces in national legislation. Scholars at the University of Chicago, for instance, have used textual similarity to build a "Legislative Influence Detector" that compares how statutes in one US state are influenced by those of another state (Burgess et al., 2016). The project also investigated whether proposals submitted by lobby groups made their way into final legislation. Other scholars have tracked text reuse to trace where policy ideas that end up in US or EU legislation first originated (Cross and Hermansson, 2017; Wilkerson et al., 2015). When it comes to international law, political scientists have compared

⁶ Jarrod Hepburn, *Battling \$50 Billion Yukos Awards on Two Fronts, Russia Focuses on Claimants' Alleged Fraud and Linguistic Analysis of Tribunal Assistant's Alleged Role in Drafting Awards*, INV. ARB. Rep. (3 November 2015).

text of model bilateral investment with negotiated agreements to detect what party had greater say over the final outcome (Berge and Stiansen, 2016).

These approaches employ the same intuition as plagiarism detection software, which uses exact or near textual matches to detect academic fraud. The low probability of exact textual matches by chance suggests that two otherwise distinct documents with identical or highly similar language share a common author or, at least, are linked through a third factor. While such similarity-based analysis yields robust results when actual copy-and-pasting has taken place, it is unable, for the reasons discussed above, to detect legislative influence when it is manifested in the underlying policy ideas but not the actual texts.

4) Similarity, Boilerplate and Novelty

A fourth use of similarity compares targeted legal texts, typically contracts, statutes or treaties, to benchmark texts to assess to what extent they follow boilerplate language or deviate from established practice. The same technique can also be used to detect legal innovations or idiosyncrasies in texts that display passages not found in other texts of the same domain. This is particularly fruitful in relation to transactional and statutory documents, which tend to recycle legal language to benefit from established meaning of legal concepts, terms of trade and other industry terminology, making deviations from established wording noteworthy.

In the domain of international law, for instance, political scientists have compared the text of the Transpacific Partnership Agreement negotiated in 2016 to other trade agreements, to verify claims by the US administration that the agreement set new standards, finding instead that the text of the agreement had been taken in large parts from earlier agreements (Allee and Lugg, 2016). The same technique, however, also allows the researchers to identify those areas that were truly novel with no textual precedent in prior practice.

4. APPLICATIONS II: TRACKING LEGAL PROCESSES

A different type of legal application leverages automated text comparisons not to analyze individual documents but to track legal processes. The evolution of Canadian Federal Income Tax Regulations described above is an example of a process-oriented text comparison. These applications trace textual similarity patterns over time across large sets of comparators to

better understand, among others, negotiation dynamics as well as policy evolution and diffusion.

1) Dissecting Negotiation Processes

When distinct parties come to the bargaining table, their preferences and objectives initially diverge, but as negotiations unfold, the positions converge as each side gives and takes until they agree on a joint text. This convergence of views often leaves textual traces as parties table proposals, accept revisions and work towards the final document. Negotiations can thus be tracked through the lens of automated document comparison as a process, in which distinct texts are merged into a common document.

The negotiations of the North American Free Trade Agreement (NAFTA) in the early 1990s is a case in point. The 42 negotiation drafts of one of its chapters – Chapter 11 on investment protection – have been made public and allow researchers to track how the negotiations unfolded from the original proposal submitted by each party to the convergence in the final text. The comparison of these drafts suggests that no country dominated that negotiation as each had to compromise significantly based on their initial submission to reach a final outcome. Such text comparison also helps identify particularly contentious areas in negotiations. In the NAFTA talks consensus was reached quickly on many of the substantive protection provisions, but it took longer to find common ground on the role of escape clauses and dispute settlement (Alschner et al., 2018a).

Similarity analysis can thus cast new light on negotiating processes by comparing the language of drafts over time. But it also has shortcomings that again go back to the distinction between text and semantics. For the United States the inclusion of investor-state arbitration was the primary negotiation objective, while Canada and Mexico initially opposed it. In the end, the US got what it wanted, the inclusion of arbitration, even though it had to strongly compromise on the surrounding text. So by what metric should we evaluate whether the US lost or won that negotiation? Textual similarity can quantify the give-and-take in negotiations, but cannot discern what textual concession mattered more in legal or policy terms.

2) Policy Evolution and Diffusion

Another way automated document comparison can help understand legal processes is by tracking how policies evolve through the changing language in legal texts. Staying with the example of international treaties, NAFTA

Chapter 11 marked a clear departure from US bilateral investment treaty (BIT) practice (Alschner et al., 2017b). Prior and subsequent to NAFTA, the US concluded BITs that closely mirrored a common template. The NAFTA text remained an outlier and deviation from that template. Yet, after the US was sued by foreign investors based on NAFTA Chapter 11 and won a string of cases in the late 1990s, the country not only started to conclude other free trade agreements with investment chapters, but also changed its BIT practice to align it closer with the tried and tested NAFTA text (Alschner et al., 2017b).

While this US policy shift is anecdotally well known, automatic text comparison can also inductively reveal otherwise undocumented policy changes. An analysis of the investment treaty practice of Japan, for instance, reveals that the country undertook a major overhaul of its investment treaty practice in 2001 or that Spain and Finland introduced new template texts in 1996 and 1999 respectively. Hence, similarity analysis can reveal latent policy changes over time identifying periods of consistency and breaks of innovation. Moreover, the same type of analysis can help track policy diffusion patterns. By comparing treaty texts across countries, it is possible to identify when one country borrowed treaty language from another country and how new policies thereby spread from one state to the next (Alschner and Skougarevskiy, 2016).

A downside of focusing purely on similarity, however, is that it can tell us little about the direction of the policy change or the driver of diffusion. Strictly speaking, all we know is that texts at time 1 look different than those concluded at time 2 and that in the intervening period a change occurred. Automated text comparison tools therefore need to be combined with other social science research methods such as interview-based or archival research to understand the causal forces and direction behind patterns made visible through the text comparison.

5. APPLICATIONS III: PATTERN RECOGNITION IN LARGE TEXT CORPORA

Finally, automated text comparison can be used to reveal patterns in large legal corpora. On the one hand, such analysis can follow a deductive approach where we expect documents to vary systematically in similarity and difference as a function of some other variable, such as the underlying bargaining power of parties. On the other hand, the analysis can also be fully inductive letting the data speak for itself to reveal patterns that we did not know to exist (Alschner et al., 2017a). In each case, automated text comparison can confirm or reveal clusters and ordering in large legal corpora.

1) Deductive Pattern Recognition

A classic empirical research design can use automated text comparison deductively to test hypotheses on a given legal text corpus. For instance, we may hypothesize that bargaining outcomes in treaty negotiations are a function of power asymmetries. Where two parties each equipped with a treaty proposal enter negotiations, the hypothesis would suggest that the final text would be more similar to the more powerful party's proposal than to the weaker party's one. In previous research, Dmitry Skougarevskiy and I tested that hypothesis with respect to the negotiation of bilateral investment treaties. We found that developed countries tended to be the system's rule-makers, while developing countries were the system's rule-takers. Bargaining dynamics thus helped to explain the patterns of textual similarity and difference that we observed in the universe of investment agreements (Alschner and Skougarevskiy, 2016).

In the above example, the similarity or dissimilarity of texts in a legal corpus is the phenomenon to be explained. Similarity patterns can, however, also be used as explanatory variables on the right side of the equation to describe or explain a certain outcome. Ryan Whalen, for instance, has used the textual similarity of the corpus of patent applications to describe change in innovation styles (Whalen, 2018). By using textual distance to weigh citations between patent applications, he found that recent patents increasingly cite inventions from textually different domains indicating a shift in how creative recombination contributes to innovation (Whalen, 2018). Hence, automated text comparisons of large legal corpora can yield new, insightful weights and variables for deductive research design.

2) Inductive Pattern Recognition

Often we do not have a good intuition about whether texts in a legal domain are converging or diverging or what patterns characterize large legal text collections. After all, as lawyers we seldom engage with the big data environment of large legal corpora and instead traditionally prefer the small data environment of the close reading of selected texts. But by applying big data approaches to "let data speak for itself", we can detect patterns that enhance our understanding of these corpora (Mayer-Schönberger and Cukier, 2014, p. 14).

For instance, when colleagues and I investigated a corpus of 450 free trade agreements (FTAs), we first thought that it would behave like bilateral investment treaties clustering in groups along developed/

Figure 1.2 Deductive and inductive ordering of similarity groupings in large corpora

developing country boundaries. We thus visualized treaty similarity scores through a heat map ordered alphabetically by the richer party in each treaty (Figure 1.2a). While the figure suggests that there are groupings of similarly worded agreements, the map remained patchy. The FTA corpus unlike the investment treaty corpus could thus not be meaningfully described through a rule-maker/rule-taker pattern. Instead of testing another hypothesis, however, we decided to follow an inductive approach of letting the data speak for itself. We used a hierarchical clustering algorithm to reorder the same heat map. After interpreting the resulting structure depicted in Figure 1.2b we realized that regional clusters of similarly worded treaties characterized the FTA corpus: European states were concluding one type of agreement, while states in America and East Asia were following a different treaty template. This regionalization of FTA language was a structure we “discovered” inductively (Alschner et al., 2018b).

Using automated legal text comparison to understand how legal documents group together in large corpora then prepares the ground for new questions and theories. What makes investment treaties behave differently from FTAs? Why are some regions concluding more homogenous agreements than others? Inductive analysis can thus lead to questions that we did not conceive of before because we lacked a grasp of the underlying data structure. This, in turn, can give rise to new theories that formalize the explanations for the patterns we observe.

6. ACQUIRING THE SKILLS TO CONDUCT AUTOMATED TEXT COMPARISONS

The above overview has shown that automated text comparison enjoys a wide range of applications in the legal field. So how can lawyers and legal researchers not yet in possession of the necessary skills to conduct such analysis acquire them?

First, a growing number of online applications allow users to engage in simple corpus analytics and basic text comparisons. One of these tools is *Voyant*.⁷ It allows users to explore texts through an appealing interface and a wide range of analytics that draw on word frequency counts and collocations. As computational linguistics and natural language processing touches ever more aspects of the human, social and legal sciences, more such tools will become available that provide an entry point into corpus analytics without requiring any programming background.

Those thinking more seriously about applying automated text comparisons to legal questions should consider acquiring basic programming skills. Programming environments such as R allow users to webscrape, analyze and visualize legal texts. They allow a degree of customization that existing introductory websites do not offer and can thus be tailored to the specific task and research needs at hand. In addition, a growing body of resources including online materials, textbooks and a computational legal community with annual conferences allows all those interested in computational legal analysis to learn more about the field and to acquire the necessary skillset.⁸

As more legal texts become digitized, the corpus of machine-readable law is growing. Traditional legal tools of relying on close reading cannot be scaled and can thus not keep up with the growth of legal text information available. Particularly young lawyers and law students should therefore invest in acquiring distant reading skills to complement their close reading abilities in order to seize the many opportunities that increased use of automated text comparison creates in the legal domain.

⁷ <http://docs.voyant-tools.org/>.

⁸ See e.g. www.datascienceforlawyers.org, <https://www.computationallegalstudies.com>. International conferences include JURIX and ICAIL.

BIBLIOGRAPHY

- Allee, T., and Lugg, A. (2016). Who Wrote the Rules for the Trans-Pacific Partnership? *Research & Politics*, 3(3), 1–9, 2053168016658919. <https://doi.org/10.1177/2053168016658919>.
- Alschner, W., and Skougarevskiy, D. (2016). Mapping the Universe of International Investment Agreements. *Journal of International Economic Law*, 19(8), 561–88.
- Alschner, W., Panford-Walsh, R., and Skougarevskiy, D. (2018a). *What Can the Negotiations of NAFTA 1.0 Teach Us about the Fate of NAFTA 2.0?* Ottawa Faculty of Law Working Paper No. 2018-05.
- Alschner, W., Pauwelyn, J., and Puig, S. (2017a). The Data-Driven Future of International Economic Law. *Journal of International Economic Law*, 20(2), 217–31. <https://doi.org/10.1093/jiel/jgx020>.
- Alschner, W., Seiermann, J., and Skougarevskiy, D. (2017b). Text-as-Data Analysis of Preferential Trade Agreements: Mapping the PTA Landscape. UNCTAD Research Paper, UNCTAD/SER.RP/2017/5.
- Alschner, W., Seiermann, J., and Skougarevskiy, D. (2018b). Text of Trade Agreements (ToTA): A Structured Corpus for the Text-as-Data Analysis of Preferential Trade Agreements. *Journal of Empirical Legal Studies*, 15(3), 648–66. <https://doi.org/10.1111/jels.12189>.
- Arnold, T., and Tilton, L. (2015). *Humanities Data in R: Exploring Networks, Geospatial Data, Images, and Text* (1st ed.). Cham: Springer.
- Berge, T.L., and Stiansen, Ø. (2016). Negotiating BITs with Models: The Power of Expertise. Retrieved 12 March 2019 from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2851454.
- Burgess, M., Giraudy, E., Katz-Samuels, J., Walsh, J., Willis, D., Haynes, L., and Ghani, R. (2016). The Legislative Influence Detector: Finding Text Reuse in State Legislation. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 57–66). New York, NY: ACM Press, <https://doi.org/10.1145/2939672.2939697>.
- Coulthard, M., and Johnson, A. (eds). (2010). *The Routledge Handbook of Forensic Linguistics*. Milton Park, Abingdon, Oxon; New York, NY: Routledge.
- Cross, J.P., and Hermansson, H. (2017). Legislative Amendments and Informal Politics in the European Union: A Text Reuse Approach. *European Union Politics*, 18(4), 581–602. <https://doi.org/10.1177/1465116517717071>.
- Denny, M.J., and Spirling, A. (2018). Text Preprocessing for Unsupervised Learning: Why It Matters, When It Misleads, and What to Do about It. *Political Analysis*, 26(2), 168–89. <https://doi.org/10.1017/pan.2017.44>.
- Geist, A. (2008). The Open Revolution: Using Citation Analysis to Improve Legal Text Retrieval. *European Journal of Legal Studies*, 2(3), 137–45. Retrieved 12 March 2019 from <http://cadmus.eui.eu/handle/1814/15121>.
- 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), 267–97. <https://doi.org/10.1093/pan/mps028>.
- Landthaler, J., Walzl, B., and Matthes, F. (2016). Unveiling References in Legal Texts: Implicit versus Explicit Network Structures. *IRIS: Internationales Rechtsinformatik Symposium*, 8, 71–8.

- Manning, C.D., Raghavan, P., and Schütze, H. (2008). *Introduction to Information Retrieval* (1st ed.). New York: Cambridge University Press.
- Mayer-Schönberger, V., and Cukier, K. (2014). *Big Data: A Revolution That Will Transform How We Live, Work, and Think* (reprint edition). Boston: Eamon Dolan/Mariner Books.
- Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. *ArXiv Preprint ArXiv:1301.3781*. Retrieved 12 March 2019 from <http://arxiv.org/abs/1301.3781>.
- Moretti, F. (2013). *Distant Reading*. London New York: Verso.
- Nanda, R., Siragusa, G., Di Caro, L., Boella, G., Grossio, L., Gerbaudo, M., and Costamagna, F. (2018). Unsupervised and Supervised Text Similarity Systems for Automated Identification of National Implementing Measures of European Directives. *Artificial Intelligence and Law*, 27(1), 199–225 <https://doi.org/10.1007/s10506-018-9236-y>.
- Schweighofer, E., and Geist, A. (2007). Legal Query Expansion Using Ontologies and Relevance Feedback. In Casanovas, P., Biasiotti, M., Francesconi, E., and Sagri, M. (eds), *Proceedings of the 2nd Workshop on Legal Ontologies and Artificial Intelligence Techniques June 4th, 2007, Stanford University, Stanford, CA, USA* (pp. 149–60), available at: <http://ceur-ws.org/Vol-321/LOAIT07-Proceedings.pdf>.
- Spirling, A. (2012). U.S. Treaty Making with American Indians: Institutional Change and Relative Power, 1784–1911. *American Journal of Political Science*, 56(1), 84–97. <https://doi.org/10.1111/j.1540-5907.2011.00558.x>.
- Whalen, R. (2018). Boundary Spanning Innovation and the Patent System: Interdisciplinary Challenges for a Specialized Examination System. *Research Policy*, 47(7), 1334–43. <https://doi.org/10.1016/j.respol.2018.04.017>.
- Wilkerson, J., Smith, D., and Stramp, N. (2015). Tracing the Flow of Policy Ideas in Legislatures: A Text Reuse Approach. *American Journal of Political Science*, 59(4), 943–56. <https://doi.org/10.1111/ajps.12175>.
- Winkels, A.R., Boer, A., Vredebrecht, B., and Someren, A.V. (2014). Towards a Legal Recommender System. In Hoekstra R. (ed.), *JURIX 2014: International Conference on Legal Knowledge and Information Systems* (pp. 169–78). Amsterdam: IOS Press.