David S. Law*

The Global Language of Human Rights: A Computational Linguistic Analysis

https://doi.org/10.1515/lehr-2018-0001

Abstract: Human rights discourse has been likened to a global lingua franca, and in more ways than one, the analogy seems apt. Human rights discourse is a language that is used by all yet belongs uniquely to no particular place. It crosses not only the borders between nation-states, but also the divide between national law and international law: it appears in national constitutions and international treaties alike. But is it possible to conceive of human rights as a global language or lingua franca not just in a figurative or metaphorical sense, but in a literal or linguistic sense as a legal dialect defined by distinctive patterns of word choice and usage? Does there exist a global language of human rights that transcends not only national borders, but also the divide between domestic and international law?

Empirical analysis suggests that the answer is yes, but this global language comes in at least two variants or dialects. New techniques for performing automated content analysis enable us to analyze the bulk of all national constitutions over the last two centuries, together with the world's leading regional and international human rights instruments, for patterns of linguistic similarity and to evaluate how much language, if any, they share in common. Specifically, we employ a technique known as topic modeling that disassembles texts into recurring verbal patterns.

The results highlight the existence of two species or dialects of rights talk—the universalist dialect and the positive-rights dialect—both of which are global in reach and rising in popularity. The universalist dialect is generic in content and draws heavily on the type of language found in international and regional human rights instruments. It appears in particularly large doses in the constitutions of transitional states, developing states, and states that have been heavily exposed to the influence of the international community.

The positive-rights dialect, by contrast, is characterized by its substantive emphasis on positive rights of a social or economic variety, and by its prevalence in lengthier constitutions and constitutions from outside the common

***Corresponding author: David S. Law,** Professor, Charles Nagel Chair of Constitutional Law and Political Science, Washington University, St Louis, Missouri, USA; Sir Y.K. Pao Chair in Public Law, The University of Hong Kong, Hong Kong, Hong Kong, E-mail: davidlaw@wustl.edu

law world, especially those of the Spanish-speaking world. Both dialects of rights talk are truly transnational, in the sense that they appear simultaneously in national, regional, and international legal instruments and transcend the distinction between domestic and international law. Their existence attests to the blurring of the boundary between constitutional law and international law.

Keywords: human rights discourse, universalist dialect, positive-rights dialect, dialects of rights talk, global lingua franca

Introduction: A Linguistic View of A Linguistic Metaphor

Human rights discourse has been likened to a global lingua franca,[1] and in more ways than one, the analogy seems apt. A lingua franca is a bridge language used by speakers of other languages to communicate.[2] Its defining characteristic is that it is used in many places but is not native to any particular place.[3] The original lingua franca—literally, the "Frankish language"—was used in the Mediterranean Basin as the language of commerce and diplomacy.[4] Human rights discourse, a species of rights talk with explicitly international aspirations,[5] is a form of legal and ideological expression widely used around the world. It is a language employed by public lawyers everywhere. Like the

1 *See, e.g.*, Michael Ignatieff, Human Rights as Politics and Idolatry 53 (2001) ("Human rights has become the major article of faith of a secular culture that fears it believes in nothing else. It has become the lingua franca of global moral thought, as English has become the lingua franca of the global economy."); Kenneth Cmiel, *The Recent History of Human Rights, in* The Human Rights Revolution: An International History 27, 32 (Akira Iriye et al. eds., 2012) ("[H]uman rights talk communicates across cultures in ways similar to money, statistics, pidgin English, or a discussion of soccer. … [H]uman rights have become one of the linguae francae of a globalized world[.]").

2 *See* Jean-Benoît Nadeau & Julie Barlow, The Story of French 29 (2006) ("Today a lingua franca is any common language used in economics, diplomacy or science, in a context where it is not a mother tongue.").

3 *See id*. ("The Mediterranean *lingua franca* never evolved into anyone's mother tongue, which is why there are very few written traces of it.").

4 *See id*. ("In the Mediterranean region, fishermen, sailors and merchants used a rudimentary version of *langue d'oc* mixed with Italian that people called the *lingua franca* ("Frankish language"), and over time this spoken language soaked up influences from Italian, Spanish, and Turkish.").

5 *See* Mary Ann Glendon, Rights Talk 13 (1991); Samuel Moyn, The Last Utopia 210–11 (2010).

original Frankish language, it is widely used and shaped by many influences; it is used by all yet belongs uniquely to no one. The same language is found in national constitutions and international agreements alike. In other words, this language crosses not only the borders between nation-states, but also the divide between national law and international law. Its use has spread both horizontally (across national borders) and vertically (across national, regional, and international legal orders).

But is it possible to conceive of human rights as a global language or lingua franca in a literal sense? Is it accurate to speak of a global discourse or language of human rights not just in a figurative or metaphorical sense, but also in a linguistic or semantic sense? What might we mean by language, and what would the characteristics of this language be? To tackle these questions, we must determine (1) where to look; (2) what to look for; and (3) how to look for it. In other words, (1) what data should we examine? (2) What might fairly be characterized as a human rights "language"? And (3) what methodology might be appropriate for identifying its existence and characteristics?

With respect to the first question—how to identify appropriate data for analysis—a natural starting point would be the text of the world's various national constitutions and international human rights agreements. There are, of course, other documents that could also be examined in the domain of constitutional law and international law, such as judicial rulings, but constitutions and treaties are foundational and authoritative texts that lie at the heart of their respective legal orders and lend themselves to empirical analysis. Collectively, they offer a well-defined and tractable universe of texts that are analogous in function and global in range.

On the second question, there is obviously no full-blown language of human rights in the sense that English, Spanish, and Japanese are languages. It is possible, however, to evaluate from a linguistic perspective how constitutions and treaties discuss human rights. There are three possible patterns that we might observe: *linguistic uniformity*, *linguistic diversity*, and *linguistic dialects*. Linguistic uniformity describes a lack of linguistic variation: it means that different countries and different international organizations tend to use the same words with the same frequency and in the same combinations when addressing human rights. Such uniformity would make it fair to speak of the existence of a global language of human rights that transcends not only national borders, but also the divide between domestic and international law. Linguistic diversity is simply the opposite: it describes a world in which constitutions and treaties discuss human rights in highly heterogeneous ways and it is consequently difficult to speak of a global language of human rights from a linguistic perspective.

The remaining possibility—that of linguistic dialects—falls between the first two. In this scenario, different countries and international organizations do not all use similar language. Nor, however, do they exhibit much originality or vary randomly. Instead, constitutions and treaties are characterized by a few variants, or dialects, of rights talk, where each dialect is defined by distinctive and recurring patterns of word choice, word frequency, and word co-occurrence. We might find, for example, that some treaties and constitutions fall in one linguistic camp, while the rest fall in another. Alternatively, there could exist a divide between international law and constitutional law: constitutions might exhibit one set of linguistic patterns while treaties might exhibit a different set, even when discussing the same concepts.

This brings us to the third question: what methodology might we use to identify and study linguistic patterns of this type? To date, linguistic analysis remains rare in empirical legal scholarship. Quantitatively oriented or "large-N" empirical research on constitutional drafting has made great strides over the last decade,[6] but one of the weaknesses of this genre thus far has been its inattention to language. This may reflect a lack of well-suited tools for studying linguistic patterns. Conventional techniques for conducting quantitative empirical analysis of constitutions and other legal documents require scholars to convert text into numbers for statistical analysis via the process known as "coding." In the process of translating language into numbers, however, the language itself is discarded.

At least in principle, it is clear that empirical constitutional scholarship ought to be attentive to the language that drafters use. Lawyers know that language matters. The language found in legal texts such as constitutions and treaties conveys more than just legal concepts. It is also a medium of emphasis, tone, rhetoric, and style; it contains telltale signs of the inspirations and influences acting upon the drafters. But research of this sort is easier said than done. Subtle patterns of language can be difficult to identify, much less quantify, with the naked eye. Neither old-fashioned reading nor traditional forms of quantitative analysis that rely on coding are well suited to identifying such patterns across large numbers of documents in a systematic way.

But there is now another way. New methodologies developed by computational linguists and refined by social scientists for performing automated content analysis do not discard the text but instead treat the raw text itself as the data.

---

6 *See* Ran Hirschl, Comparative Matters: The Renaissance of Comparative Constitutional Law 267–81 (2014) (discussing the recent boom in "large-N" scholarship on constitutions); David S. Law, *Constitutions, in* The Oxford Handbook of Empirical Legal Research 376, 379 (Peter Cane & Herbert Kritzer eds., 2010).

These techniques excel at identifying and analyzing subtle, complex semantic patterns that are difficult, if not impossible, for manually coded data to capture.[7] They break down vast bodies of text into their component verbal patterns in a rapid, systematic, and unbiased way.

This Article takes advantage of a form of automated content analysis known as topic modeling to scan the bulk of all national constitutions over the last two centuries, together with the world's leading regional and international human rights instruments, in order to ascertain what type of language they share in common. Specifically, we employ a technique known as topic modeling that breaks constitutions down into verbal patterns or "topics."[8] The results highlight the existence of two species or dialects of rights talk, both of which are global in reach and rising in popularity.

The first, which we call the universalist dialect, is generic in content and draws heavily on the type of language found in international and regional human rights instruments. It appears in many constitutions but accounts for an especially sizable proportion of the constitutions of transitional states, developing states, and states that have been heavily exposed to the influence of the international community. The second, which might be called the positive-rights dialect, also features in international and regional human rights instruments, albeit not to the same extent as the universalist dialect and is distinguished in substantive terms by its emphasis on positive rights of a social or economic variety. It is more prevalent in lengthier constitutions and constitutions from outside the common law world, including in particular the Spanish-speaking world. These dialects of rights talk are truly transnational, in the sense that they appear simultaneously in national, regional, and international legal instruments and transcend the distinction between domestic and international law. Their prevalence attests to the blurring of the boundary between constitutional law and international law.

I Automated Content Analysis: What Is It, and Why Should We Use It?

The field of automated content analysis (ACA) is in its infancy but has the potential to revolutionize both the practice and the study of law. Computational linguistics and computer science have pioneered techniques for

7 *See* Kevin Quinn et al., *How to Analyze Political Attention with Minimal Assumptions and Costs*, 54 Am J. Pol. Sci. 209, 213–14 (2010) (discussing how topic modeling works and applying it to legislative data).

8 *Id*.

performing ACA that are already attracting enormous interest in the social sciences and humanities. Although legal scholarship has been slow to embrace ACA[9] compared to other fields such as political science[10] or even the humanities,[11] law is an especially obvious candidate for the introduction of ACA.[12] Legal scholars and practicing lawyers alike spend the bulk of their time digesting and analyzing large volumes of text in the form of cases, constitutions, statutes, regulations, treaties, contracts, corporate filings, briefs, and so forth.[13] ACA offers a means of performing these bread-and-butter tasks at unprecedented speed and in novel ways.

There are two fundamental differences between ACA and conventional quantitative techniques for analyzing documents, both of which have far-reaching consequences. First, whereas conventional techniques involve a lengthy and potentially error-prone process of converting or "coding" text into numeric data

9 As of 6 August 2015, a search of Westlaw's database of law reviews and journals for the terms "automated content analysis" and "text analysis" uncovers no examples of the use of such methods. A search for the term "topic model," the specific type of content analysis employed in this Article, yields two results, a co-authored corporate law article from 2014 and a student note that appeared in 2013 in the *Yale Law Journal*. *See* Jonathan Macey & Joshua Mitts, *Finding Order in the Morass: The Three Real Justifications for Piercing the Corporate Veil*, 100 CORNELL L. REV. 99 (2014); Daniel Taylor Young, Note, *How Do You Measure a Constitutional Moment? Using Algorithmic Topic Modeling to Evaluate Bruce Ackerman's Theory of Constitutional Change*, 122 YALE L.J. 1990 (2013). That note was authored in collaboration with Brandon Stewart, who is one of the creators of the software package used in this Article and whose assistance is very gratefully acknowledged.

10 *See, e.g.*, Justin Grimmer & Brandon M. Stewart, *Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts*, 21 POL. ANALYSIS 267 (2013); Christopher Lucas et al., *Computer-Assisted Text Analysis for Comparative Politics*, 23 POL. ANALYSIS 254 (2015); Margaret E. Roberts et al., *Structural Topic Models for Open-Ended Survey Responses*, 58 AM. J. POL. SCI. 1064 (2014).

11 The creation of campus-wide, interdisciplinary initiatives in the "Digital Humanities" (the academic buzzword for a broader universe of computer-assisted approaches to the study of humanities, including ACA) is all the rage among leading universities. *See, e.g.*, Stanford Humanities Center: Digital Humanities, http://shc.stanford.edu/digital-humanities; digital.humanities@oxford, https://digital.humanities.ox.ac.uk; UCL Centre for Digital Humanities, http://www.ucl.ac.uk/dh; University of Chicago Digital Humanities initiative, https://humanities.uchicago.edu/articles/2014/11/digital-humanities-uchicago.

12 Arthur Dyevre, *The Promise and Pitfalls of Automated Text-Scaling Techniques for the Analysis of Judicial Opinions* (10 April 2015), *available at* https://ssrn.com/abstract=2626370 or http://dx.doi.org/10.2139/ssrn.2626370 (last accessed 10 October 2016).

13 *See* George S. Geis, *Automating Contract Law*, 83 N.Y.U. L. REV. 450, 478 (2008).

that can then be analyzed using statistical techniques,[14] ACA treats the text itself as the data.[15] To be clear, ACA is not a substitute for human interpretation or judgment. It does, however, free us from having to search each and every document for phenomena of interest and enables us to focus our attention where it is most needed and valuable—namely, on interpreting the results of the automated analysis.

Second, rather than reading and understanding text in the way that a human coder would, current ACA techniques rely on the identification and quantification of semantic patterns (namely, the frequency with which certain words appear in conjunction with each other). These patterns are proxies for the phenomena of interest to the researcher (such as a particular topic of discussion, or the writing style of one author as opposed to another). The software need not—and cannot—understand the meaning or significance of the words that it encounters. It can tell us, for example, that the words "donald" and "trump" often appear in conjunction with each other, for example, or that these words are also positively correlated with the appearance of the words "republican," "campaign," "wiretap," and "russia." It cannot tell us that these words appear together because Donald Trump is a Republican president who has accused his predecessor of illegal wiretapping and is said to have benefited from Russian sabotage of his campaign opponent. It can, however, ascertain the probability with which these co-occur, and it can identify clusters of words that co-occur often enough to signify a distinctive topic or idiom.

To illustrate by way of analogy how word co-occurrence can be used to analyze text, suppose that we are hosting a giant potluck dinner at which guests have brought numerous dishes that need to be sorted into the salad bar, main dishes, accompaniments, and the dessert table. Just as a human would sort unknown documents by reading them, a human would sort unknown dishes by tasting them. What topic modeling does with a mass of documents might be likened to a food-sorting computer that disassembles each dish into its individual ingredients, calculates the exact quantity of each ingredient in each dish with inhuman precision, and lumps certain dishes together with others because they draw more heavily on one cluster of ingredients as opposed to another. The computer would not understand that the first cluster tastes sweet to a human while the second cluster tastes salty, and it would not know what to call each of the tables of food that it generates. But it would probably do a pretty good job of

14 *See* Joshua B. Fischman & David S. Law, *What Is Judicial Ideology, and How Should We Measure It?*, 29 WASH. U. J. L. & POL'Y 133, 156–66 (2009) (discussing competing approaches to, and various challenges involved in, the coding of judicial decisions).

15 *See* Grimmer & Stewart, *supra* note 10, at 272 (developing the idea of text as data).

sorting the dishes based on the fact that chocolate tends to appear together more often with sugar than with pork or paprika, and so on.

Of course, not all of our hypothetical computer's sorting decisions will match those that a human would make. Based on its ingredient-clustering analysis, the computer might assign pancakes and cornbread to the dessert table, for example, or gazpacho to the salad bar. Even sorting errors of this variety, however, would rest on some objective basis and might, in fact, invite us to examine the peculiarities of our own conceptual categories.

The type of ACA used in this Article is called topic modeling. The underlying idea is that a corpus of text can be modeled as a collection of "topics," where a "topic" consists of a distribution of probabilities over a set of words.[16] To give a simplified example, one topic might consist of an 90% likelihood of the word "trump" appearing, together with an 85% likelihood of the word "donald" and a 60% likelihood of the word "campaign," whereas another topic might consist of an 80% likelihood of the word "disney," together with a 90% likelihood of the word "cartoon," a 50% likelihood of the word "donald," and a 50% likelihood of the word "duck." Each of these "topics" is, technically speaking, just a collection of words that tends to appear together, but these collections of words correspond to a distinctive topic in the conventional sense.

Topic modeling software estimates the probability that a given word will appear whenever a particular topic is being discussed, and to the extent that the actual distribution of words in a particular document matches the distribution of words that the software predicts for a particular topic, the software will identify that document as discussing that topic.[17] For a given text corpus, topic modeling software estimates both topic content (the words and probability distributions associated with each topic) and topic prevalence (the proportion of each document that is composed of each topic).[18] The fact that a certain word (in this example, "donald") appears in conjunction with multiple topics does not pose grave difficulties: because it takes into account the surrounding words, topic modeling has an inherent ability to assign the ambiguous word to the correct topic.[19]

Topic modeling is an unsupervised, as opposed to supervised, method of ACA.[20] A supervised method is one in which the researcher furnishes the

16 *Id*. at 283.

17 *See* David S. Law, *Constitutional Archetypes*, 95 Tex. L. Rev. 153, 190–91 (2016).

18 *See id*. at 191.

19 *See id*. at 204–05 (noting the ability of topic modeling to disambiguate ambiguous terms such as "charter" and "convention" based on the words that accompany them).

20 *See* Lucas et al., *supra* note 10, at 260–61 (noting that there are "essentially two approaches to automated text analysis: supervised and unsupervised methods, each of which *amplifies* human effort in a different way"); Roberts et al., *supra* note 10, at 1066 ("Topic models are often

software with examples of the topics or patterns that he or she wishes to locate within a larger corpus of text.[21] The software then classifies or rates documents within the larger text corpus based on their similarity to the examples. With an unsupervised method such as topic modeling, by contrast, the researcher does not tell the software what patterns to seek. Instead, the software breaks down the text corpus into underlying, naturally occurring semantic patterns, and it falls upon the researcher to ascertain what, if anything, these patterns signify.[22]

Precisely because topic modeling embodies no assumptions on the part of the researcher as to what topics exist in the corpus—in other words, the researcher does not *supervise* the model—it offers an escape from existing conceptual frameworks and are well suited to identifying subtle or complex patterns that might previously have been invisible or unknown to the researcher. It can also identify aspects of a text that are not substantive at all and much harder for hand-coding to capture in an objective and reliable way, such as style, tone, and rhetoric.

Therein, however, also lies one of the greatest challenges associated with this approach. The results of a topic model can be difficult to interpret or, equivalently, open to multiple interpretations. Precisely because the researcher does not tell the model what to look for, the researcher may not be able to recognize what the model finds. The same kinds of subtle semantic patterns that are indicative of rhetorical tone (which is potentially of great interest to researchers) may instead be indicative of the idiosyncrasies of a particular translator (which may be of intrinsic interest to literary scholars, but probably not to legal scholars) or may simply be the equivalent of meaningless verbal noise. Because the computer cannot understand or interpret what it has found, the computer cannot distinguish topics that are meaningful from topics that are mere noise. That responsibility falls upon the researcher.

---

referred to as 'unsupervised' methods because they *infer* rather than *assume* the content of the topics under study[.]" (emphasis in original)).

21 *See* Grimmer & Stewart, *supra* note 10, at 292–94 (contrasting "supervised" and "unsupervised" methods for placing documents on a political spectrum based on automated analysis of their content); Lucas et al., *supra* note 10, at 260–61 ("In supervised methods, we specify what is conceptually interesting about documents in advance, and then the model seeks to extend our insights to a larger population of unseen documents. ... In unsupervised methods, such as topic modeling, we do not specify the conceptual structure of the texts beforehand. Instead, we use the model to find a low-dimensional summary that best explains observed documents given some set of assumptions.").

22 *See* Lucas et al., *supra* note 10, at 261 (noting that, in the case of unsupervised methods such as topic modeling, "human effort shifts ...to interpretation of the model results").

Another challenge arises when some of the text corpus has been translated from another language,[23] as is true of many of the national constitutions under analysis. Reliance on translated texts creates the risk that the results of the model will reflect the effects of translation rather than the characteristics of the texts themselves. The extent of this risk depends in part on the type of distortion introduced by the translation. For instance, if different translators prefer different terms for similar underlying ideas (e. g., "privacy" as opposed to "autonomy," or vice versa), multiple terms will appear in the text corpus in lieu of a single term. Fortunately, topic modeling is inherently robust against this type of relatively subtle translation error: as long as the various translations of the same term appear in similar contexts and are used interchangeably, they will tend to be lumped together into the same topic, precisely as they should be.[24]

A different problem arises when entirely unrelated concepts are confused in translation (e. g., "president" as opposed to "precedent"). But this too is not fatal: gross translation error of this variety tends to call attention to itself and is therefore easy to diagnose. If a particular topic happens to be found only in documents that have been translated from a particular language, for example, that topic can and should be examined more closely for words that simply do not fit.[25] Conversely, translation-induced distortion can be especially troublesome if the vocabulary differences among topics are very subtle (which can happen as the number of topics increases), or if translation-related quirks are clustered together in ways that resemble distinct topics from a semantic perspective. Even in such cases, however, the use of translations from a mix of sources and publishers (as in the case of this study) can mitigate the impact of translation error. Topic modeling is less likely to associate words with the wrong topics if some translations are correct, or if different translations make different errors, than if the same word is consistently mistranslated in a particular way. To the extent that it has the effect of randomizing translation error, reliance on translations from different sources can actually be a virtue rather than a vice.

23 *See*, *e. g.*, *id*. at 270–71 (using STM to analyze a text corpus consisting of social media postings in both Chinese and Arabic, noting that "differences in word rate use [may] arise due to linguistic differences or errors in translation," and correcting for these differences by incorporating the document's original language into the model as a content covariate).

24 *See* Law, *supra* note 17, at 228–31.

25 *See id*. (explaining why certain translation errors tend to be either harmless or easy to diagnose in the context of topic modeling).

II An Unsupervised Topic Model of Constitutional Texts

A Data and Methodology

The raw material of our analysis is a corpus of 615 constitutional texts, drawn from a variety of sources and reflecting roughly two-thirds of all new or interim constitutions ever produced.[26] For purposes of facilitating comparison across documents that perform constitutional functions or serve as models for constitutional drafters, we also include in the corpus a number of prominent international and regional human rights agreements–namely, the United Nations Declaration of Human Rights (UDHR),[27] the International Covenant on Civil and Political Rights (ICCPR),[28] the International Covenant on Economic, Social, and Cultural Rights (ICESCR),[29] the European Convention on Human Rights (ECHR),[30] the African Charter of Human and Peoples' Rights,[31] the American Declaration of the Rights and Duties of Man ("American Declaration"),[32] and the Charter of Civil Society for the Caribbean Community ("Caribbean Charter")[33]– bringing the total to 622 texts. A majority of the texts are English translations of constitutions originally written in other languages.

We estimate a type of unsupervised topic model called the Structural Topic Model using the STM package for R.[34] The package performs several standard pre-processing steps to prepare the text corpus for analysis, such as stemming, or the reduction of words to their stems. In addition to simplifying the necessary computations, stemming sensibly ensures that conjugations of the same word, such as "vote" and "voting," or "right" and "rights," are not double-counted as

26 Comparative Constitutions Project, Chronology of Constitutional Events, v. 1.2, April 2014, http://comparativeconstitutionsproject.org/download-data/.

27 G.A. Res. 217 (III) A, Universal Declaration of Human Rights (10 December 1948).

28 G.A. Res. 2200A (XXI), International Covenant on Civil and Political Rights (16 December 1966).

29 G.A. Res. 2200A (XXI). the International Covenant on Economic, Social, and Cultural Rights (16 December 1966).

30 European Convention for the Protection of Human Rights and Fundamental Freedoms, as amended by Protocols Nos. 11 and 1, E.T.S. 005.

31 African Charter on Human Rights, 21 I.L.M. 58 (1982).

32 O.A.S. Res. XXX, reprinted in Basic Documents Pertaining to Human Rights in the Inter-American System, OAS/Ser.L/V/I.4 Rev. 9 (2003); 43 AJIL Supp. 133 (1949)

33 Adopted by the Conference of Heads of Government of the Caribbean Community at Inter-Sessional Meeting, 1997.

34 Roberts et al., *supra* note 10, at 1067.

unrelated concepts. The automated pre-processing also eliminates extremely common and substantively uninteresting words such as articles and prepositions, as well as words that appear only once in the entire corpus and therefore cannot be analyzed in a meaningful way.

More controversial is the fact that STM ignores the order in which words appear. For computational reasons, the current generation of ACA and topic modeling software treats word order as immaterial: it draws no distinction between, for example, "man bites dog" and "dog bites man." The assumption that words can be sorted into topics without regard to the order in which they appear is known as the "bag of words" assumption,[35] and it obviously circumscribes the appropriate range of applications for the current generation of ACA techniques. It is wholly unrealistic, for example, if the goal is to identify legal propositions or answer legal questions (e. g., "are term limits constitutional"). There are many applications, however, in which word order matters relatively little or not at all.[36] The bag of words assumption is much more viable if—as in the present case—our goal is simply to identify what topics are discussed (e. g., "are term limits discussed in the constitution") using what kind of language (e. g., "do different constitutions use the same kind of language when discussing term limits"). In the present case, most of the actual topics generated by the model correspond to recognizable concepts of the kind captured by expert coding schemes. Thus, the substantive implications of the bag of words assumption for the present analysis should not be overestimated.

Computerized estimation of the model requires specification by the researcher of the number of topics that the model will contain (or, in other words, the value of K, where K denotes the number of topics). In the context of a regression, certain combinations of variables will do a better job of explaining the variance in the dependent variable (or offer better goodness-of-fit) than other combinations of variables. Analogously, certain values of K will offer a better fit to the data (do a better job of explaining semantic variation within the text corpus) than other values. There are different ways of evaluating the goodness-of-fit of a topic model, and in the case of our text corpus, the diagnostics for evaluating different values of K suggest more than one possibility.[37]

35 Lucas et al., *supra* note 10, at 257.

36 For example, many constitutional preambles are so heavy with rhetorical language and lofty ideals that they already resemble a potpourri of ideological buzzwords. *See* Law, *supra* note 17, at 199.

37 The usual diagnostic favored by computer scientists has been held-out likelihood, a measure of raw goodness-of-fit that, all other things being equal, tends to favor models with more topics over models with fewer topics (much as, all other things being equal, a regression model that contains more variables tends to explain more of the variance in the dependent variable than a

Ultimately, however, there is no substitute for expert judgment as to what value of K yields topics that are substantively meaningful and lend themselves to interpretation. Because only the researcher can attach meaning to the topics identified, a model that defies interpretation by the researcher is not a useful model. The test of interpretability is all the more important where, as here, the quantitative diagnostics suggest more than one viable candidate. Here, we report the results from a thirty-topic model (K = 30) because this model not only scored highly on multiple diagnostics, but also produced topics that lent themselves to substantive interpretation and differentiation.

The labels attached to the topics are not generated by the software but instead reflect the researcher's understanding of the substantive meaning or significance of the topic. Labeling the topics involves a degree of subjectivity and judgment; mislabeling the topics is akin to misinterpreting the results of the model. To enhance reliability and incorporate the benefit of additional expert judgment, the labels were devised in conjunction with Tom Ginsburg, who also furnished the bulk of the raw text corpus used in the analysis. Each of us independently arrived at a tentative substantive interpretation of, and label for, each topic. These tentative labels coincided between 30 % and 50 % of the time. Through discussion, we reconciled our interpretations and arrived at a mutually agreed label for each topic.

model that contains fewer variables). In this case, the held-out likelihood measure favors a large number of topics (K ≥ 90), as does the lower-bound measure. However, many of the topics produced by a K = 100 model are difficult to interpret or to distinguish from one another, and diagnostics other than held-out likelihood and lower bound weigh in favor of a model with fewer topics. Among the other diagnostics that STM also computes is a measure of semantic coherence that captures the extent to which words that have been grouped under the same topic actually appear together within the same documents. Because it is a proxy for the internal substantive coherence of topics, semantic coherence ought to be of particular interest to legal scholars. *See* Law, *supra* note 17, at 197 (explaining and contrasting semantic coherence with other goodness-of-fit measures for ascertaining the appropriate number of topics). In the present case, semantic coherence heavily favors a lower number of topics. A seven-topic model performs the best on this metric, while models with more than ten or fifteen topics perform significantly worse. Semantic coherence is not the only measure that favors a lower number of topics. As with the held-out likelihood measure, the residuals measure tends to favor a higher number of topics, but in the present case, the residuals bottom out (meaning that goodness-of-fit is maximized) between thirty and fifty topics and begin to increase again (meaning that goodness-of-fit begins to deteriorate) once the number of topics exceeds fifty.

B Results of the Topic Model

The two-word clouds below in Figures 1 and 2 are examples of the topics generated by the topic model. Each word cloud depicts the vocabulary associated with a particular topic. Within each word cloud, the physical arrangement of the words is random, but the size of each word corresponds to the degree of probability that the word will appear when the topic is present. Consequently, many of the words that feature most prominently in the word clouds are words that appear quite frequently regardless of the topic, such as "may" or "shall," simply because these basic verbs are difficult for a constitution to avoid regardless of the topic under discussion.

Figure 1: High-probability words from topic 20 ("federalism").

Because generic words tend not to capture what is most distinctive about each topic, STM also calculates and reports frequency-exclusivity ("FREX") scores.[38] Frequency refers to the frequency with which a word appears in a particular

38 Edoardo M. Airoldi & Jonathan M. Bischof, *A Poisson Convolution Model for Characterizing Topical Content with Word Frequency and Exclusivity*, arXiv:1206. 4631 [CS], *available at* https://arxiv.org/pdf/1206.4631.pdf.

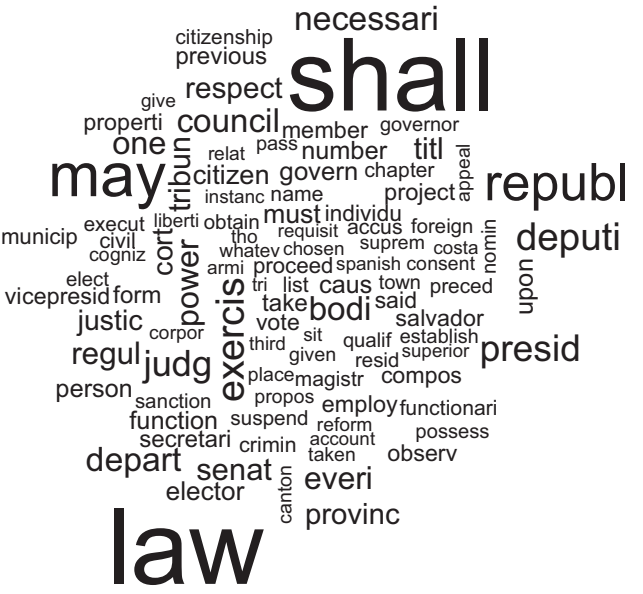

Figure 2: High-probability words from topic 13 ("legal system").

topic (which the word clouds capture), whereas exclusivity refers to the extent to which a word appears only in that topic and not in others (which the word clouds do not capture). Thus, words with high FREX scores are both relatively frequent within, and exclusive to, a particular topic.

The first word cloud depicts a topic that was both easy to label and easy to distinguish from other topics ("federalism"). Words that featured prominently in this topic but not others included "federal," "canton," "region," "confederation," and "jurisdiction," while the two countries that appear by name in the word cloud–Nigeria and Somalia–are both federal states. Another telltale sign is the presence of the word "sabah," the name of a state that enjoys heightened autonomy within Malaysia's federal system.[39] The second word cloud depicts a topic that resists easy labeling owing to its combination of fairly generic words and terms drawn from a range of possible topics. A plurality of high-probability words relating to the legal or judicial system (such as "judge," "justice," "preside," "magistrate," "supreme," "sanction," "proceeding," "accuse," "criminal," and of course "law") led us ultimately to label this topic as pertaining to the

39 *See* Andrew Harding, The Constitution of Malaysia 147–48 (2013).

"legal system." The FREX words for this topic, however, could support an alternative label pertaining to Latin America or the Spanish-speaking world. Both the high-probability and FREX words for all topics in both models are reported (in their stemmed form) in Table 1.

Table 1: Computer-identified keywords and user-defined topic labels for all topics.

Topic number and label (alternative label)	Highest-probability words	FREX words
1: parliament	parliament, law, elect, right, provis, declar, deputi, constitut, day, speaker, duti, parliamentari, first, condit, form, divis, head, set, administ, number, rhodesia, emerg, defenc, held, particular	parliament, rhodesia, parliamentari, roll, nauru, hungari, statutori, bangladesh, divis, administ, declar, truste, cardin, entrench, tribal, speaker, affirm, european, emerg, trust, head, subdivis, registrar, loyalti, interpret
2: legislative power	art, law, may, member, right, provis, govern, deputi, accord, must, matter, decis, elect, council, session, regul, one, general, case, vote, administr, riksdag, determin, budget, concern	riksdag, art, diet, realm, print, commune, haitian, swedish, sweden, chancellor, howev, putnam, offens, american, press, insofar, editor, grand, albanian, peasle, haiti, minut, czechoslovak, church, finnish
3: municipalities	law, may, paragraph, shall, municip, organ, council, servic, provid, within, offic, vote, educ, provis, judg, exercis, period, entiti, except, elector, section, suprem, tribun, purpos, person	ecuadoran, ecuador, contentious, administr, guatemalan, comptroller, plurin, ecuadorian, department, entiti, amparo, brazilian, pertin, guatemala, intendent, cuban, agrarian, bosnia, wage, herzegovina, labor, altern, cuba, branch, honduran, concess
4: civil service	clause, commiss, provis, servic, function, part, judg, minist, bill, council, year, schedul, respect, member, relat, made, fund, force, refer, date, cabinet, chief, justic, unless, proceed	clause, ghana, schedul, auditorgener, notwithstand, india, singapor, commenc, consolid, sri, lanka, cabinet, assent, anyth, gratuiti, samoa, repeal, discret, commission, payabl, satisfi, hand, proclam, write, immedi

(*continued*)

Table 1: (*continued*)

Topic number and label (alternative label)	Highest-probability words	FREX words
5: local government	member, kenya, govern, elect, parti, accord, constitut, sierra, counti, leon, majesty, right, committee, nepal, commiss, matter, condit, island, will, follow, seychell, act, pursuant, number, unit	nepal, seychelles, saeima, kenya, sierra, leone, panchayat, christoph, zanzibar, tanzania, counti, latvia, ile, raj, parliamentarian, coven, nairobi, sabha, island, par, marshal, pursuant, ilot, majesty, secretary-general
6: officeholder	offic, person, public, appoint, purpos, author, servic, hold, time, power, exercis, make, prescrib, govern, citizen, day, chapter, upon, mean, offenc, whether, appropri, take, specifi, applic	offic, person, hold, purpos, appoint, offenc, whether, prescrib, done, impos, make, qualifi, mean, time, convict, author, public, ceas, applic, appropri, appear, practic, judgment, chapter, secretari
7: Commonwealth courts	shall, subsect, function, person, appeal, relat, proceed, respect, appli, advic, public, part, judg, polic, reason, justic, question, fund, freedom, without, unless, properti, protect, held, oath	swaziland, lesotho, malawi, gambia, tuvalu, subsect, kiribati, zambia, bahama, botswana, contravent, advic, malta, pursuanc, saint, grenada, director, barbado, polic, appeal, puisn, constru, descript, justifi, jamaica
8: governor-general (royal representative)	shall, council, governor, hous, act, majesty, time, provinc, respect, order, law, may, member, new, general, said, senat, repres, canada, trinidad, tobago, provis, subject, provid, part	turkey, quebec, canada, tobago, trinidad, turkish, papua, guinea, ontario, queen, equatorial, nova, hebrides, ireland, scotia, lieutenant, commonwealth, zealand, irish, greek, coast, brunswick, majesty, liberia, governor
9: commissions and tribunals	may, court, constitut, presid, subject, order, provid, period, includ, high, speaker, vote, paragraph, suprem, interest, direct, determin, decis, made, entitl, question, chairman, otherwis, charg, thereof	subject, high, includ, period, entitl, court, speaker, chairman, may, vest, otherwis, thereof, order, confer, charg, behalf, authoris, provid, less, continu, expir, instrument, manner, question, interest

(*continued*)

Table 1: (*continued*)

Topic number and label (alternative label)	Highest-probability words	FREX words
10: social, economic, and cultural rights	public, social, regul, guarantee, function, econom, develop, respect, author, general, protect, activ, establish, interest, institut, privat, condit, cultur, well, resourc, plan, control, work, family, relat	social, sector, program, guarantee, norm, autonomy, resourc, family, plan, privat, scientif, coordin, evalu, quality, criteria, artist, econom, consum, ident, prioriti, integr, develop, cultur, train, characterist
11: legislative rules	shall, presid, elect, member, peopl, suprem, duti, term, upon, amend, provid, major, offici, pass, held, submit, event, declar, justic, majli, unless, accord, special, rule, three	shall, majli, event, maldiv, presid, discharg, suprem, assum, amend, impeach, major, bhutan, elect, conven, druk, term, quorum, held, gyalpo, pass, upon, pertain, offici, reconsider, ballot
12: Latin America	will, law, republ, presid, may, deputi, accord, exercis, must, function, offici, municip, organ, relat, necessari, concern, venezuelan, judg, venezuela, section, minist, respect, elect, compet, district	venezuela, dominican, venezuelan, panama, will, dictat, panamanian, caraca, affin, terna, consanguin, gaceta, celebr, cassat, trujillo, cogniz, miranda, aragua, feminin, salin, faculti, indic, attribut, masculin, esparta
13: legal system	law, shall, may, republ, exercis, deputi, presid, judg, necessari, depart, bodi, council, one, power, everi, regul, senat, tribun, titl, respect, provinc, cort, justic, citizen, must	republ, salvadorean, publick, cort, chili, salvador, agreeabl, costa, equatorian, spaniard, spain, granadin, directori, mexican, ecclesiast, peruvian, ternari, aud, consul, bull, chilian, apostol
14: socialism	peopl, organ, council, work, nation, state, social, assembl, citizen, right, committe, countri, develop, duti, socialist, may, popular, interest, labour, democrat, worker, communiti, econom, republ, execut	self-manag, revolut, yugoslavia, sociopolit, mozambiqu, struggl, revolutionari, mozambican, popular, peopl, socialist, realiz, liber, labour, maputo, front, socialpolit, worker, republican, task, pereira, guinea-bissau, fax, email, expand

(*continued*)

Table 1: (*continued*)

Topic number and label (alternative label)	Highest-probability words	FREX words
15: judiciary	state, constitut, court, case, establish, govern, justic, public, power, approv, judici, follow, matter, properti, one, day, without, arm, bill, account, form, tax, servic, receiv, jurisdict	state, establish, case, justic, judici, court, approv, arm, properti, tax, receiv, account, without, follow, constitut, jurisdict, least, vicepresid, foreign, matter, institut, employ, assign, render, convent
16: legislative chambers	state, hluttaw, law, imperi, union, constitut, russian, region, right, sec, emperor, empir, princ, kenesh, accord, jogorku, pyithu, repres, concern, minist, power, member, yuan, elect, session	hluttaw, emperor, kenesh, jogorku, pyithu, yuan, kyrgyz, myanmar, reichstag, duma, alth, sec, russian, empir, imperi, selfadminist, philippin, german, princ, hospodar, bavaria, lama, reichsrat, hsien, duke
17: public order	year, public, order, author, except, legisl, present, minist, decre, without, act, first, can, two, appoint, forc, time, requir, within, administr, execut, respons, place, made, suprem	decree, can, present, year, militari, except, place, manner, order, fix, without, crime, grant, legisl, two, first, war, penalti, age, sentenc, enter, respons, also, anoth, discuss
18: government powers	articl, law, member, right, accord, provis, determin, general, offic, one, presid, relat, citizen, territori, decis, necessari, provid, freedom, individu, budget, judg, well, properti, compet, request	articl, determin, accord, stipul, individu, presidenti, general, valid, law, ratif, budget, moral, extradit, request, member, territori, minor, circumst, depriv, current, inviol, offens, long, choose, temporarily
19: territory	list, presid, uganda, follow, commiss, thenc, function, sudan, govern, servic, member, southern, river, zimbabw, offic, must, accord, forc, district, refer, eireann, provis, land, minist, boundari	sudan, zimbabw, cameroon, thalweg, oireachta, seanad, dail, summit, angolan, upstream, eireann, confluenc, uganda, thenc, nile, southern, straight, hill, downstream, angola, list, junction, titleright, commission, titl, river
20: federalism	feder, state, hous, law, nigeria, govern, may, governor, provis, legisl, territori, respect, council, part, region, matter, appeal, relat, canton, confeder, bodi, power, court, within, repres	feder, laender, abuja, swiss, kadi, lago, landtag, nigeria, confeder, ruler, somalia, canton, sabah, sharia, sarawak, commod, low, sfri, vienna, switzerland, bundesrat, customari, export, electr, malaysia

(*continued*)

Table 1: (*continued*)

Topic number and label (alternative label)	Highest-probability words	FREX words
21: republic	republ, right, law, constitut, govern, region, court, bodi, act, elect, freedom, repres, legal, procedur, local, deputi, judg, within, vote, decis, accord, forc, communiti, propos, special	kosovo, region, everyon, hong, kong, prosecutor, republ, bodi, communiti, namibia, organis, status, bank, legal, freedom, immun, dismiss, ethnic, confid, procedur, basic, right, programm, local, recognis
22: provinces	provincial, provinc, legislatur, member, act, council, presid, union, govern, law, term, refer, governor, pakistan, matter, legisl, execut, function, subsect, respect, bill, seat, accord, within, area	provincial, pakistan, provinc, africa, emir, legislatur, premier, south, portfolio, seat, contempl, alloc, union, burma, mutandi, mutati, tribal, african, central, surplus, item, top, libyan, assign, quota
23: monarchy	king, hous, council, repres, law, minist, member, constitut, section, state, right, govern, case, senat, one, royal, session, kingdom, vote, person, accord, approv, duti, provid, bill	thailand, king, royal, loya, shura, afghanistan, thai, regent, kingdom, throne, amir, jirga, counter, afghan, changwat, regenc, privi, diplomaci, unoffici, egyptian, islam, alshura, ife, heir, banovina
24: generic rights	nation, person, state, polit, must, chapter, parti, may, term, promot, legisl, protect, secur, respons, independ, principl, administr, organ, human, particip, exercis, educ, system, ensur, polici	promot, human, access, particip, environ, polit, discrimin, tradit, independ, principl, fair, fundament, media, chapter, secur, digniti, polici, equit, transpar, effici, parti, system, gender, divers, level
25: elections	assembl, nation, elect, presid, member, year, declar, power, paragraph, parti, duti, judg, local, candid, case, senat, polit, number, committe, one, three, first, commiss, administr, vice-presid	assembl, nation, candid, elect, paragraph, parti, grand, declar, three, ballot, vice-presid, year, local, member, presid, five, polit, replac, physic, full, duti, councillor, code, number, judg
26: foreign affairs	chamber, congress, nation, may, power, senat, execut, right, elect, session, case, declar, territori, vote, repres, general, one, constitut, follow, determin, offic, foreign, day, title, duti	congress, chamber, nicaragua, nicaraguan, absolut, senat, agent, diplomat, alien, colombian, colombia, session, paraguay, recess, disapprov, hondura, paraguayan, disturb, paper, execut, foreign, sent, inhabit, honduranean, sieg

(*continued*)

Table 1: (*continued*)

Topic number and label (alternative label)	Highest-probability words	FREX words
27: parliamentarism	section, member, act, law, provis, commiss, minist, hous, offic, accord, appoint, refer, elect, senat, prime, forc, governor-general, remov, case, repres, requir, perform, bill, establish, reason	section, governor-general, prime, remov, act, fiji, vacat, commiss, holder, refer, regist, hous, becom, opposit, virtu, provis, guyana, leader, mind, attorney-general, senat, commenc, qualifi, perform, recommend
28: generic constitutional language	right, constitut, project, copyright, compar, reserv, shall, may, state, council, minist, ukrain, presid, court, serbia, citizen, montenegro, statut, govern, deputi, beliz, protect, duti, vote, repres	ukrain, beliz, compar, montenegro, serbia, sejm, copyright, project, khmer, poland, reserv, croatia, ethiopia, estonia, oblast, selfgovern, croatian, version, uruguay, iraq, cambodia, haiti, finland, minor, ethiopian
29: communism	state, republ, peopl, citizen, suprem, azerbaijan, deputi, right, council, soviet, organ, minist, elect, work, ssr, law, court, local, socialist, committe, activ, ussr, chairman, constitut, presidium	azerbaijan, soviet, ssr, ussr, presidium, seima, slovak, moldavian, turkmenistan, mejli, latvian, hural, romania, kazakhstan, nakhichevan, rumanian, fpri, mongol, lithuania, korea, milli, estonian, vietnam, dprk, china
30: Francophonie	republ, presid, law, may, council, minist, constitut, govern, organ, right, court, session, exercis, prime, vote, within, deputi, titl, condit, function, day, high, determin, territori, adopt	congoles, burundian, bureau, burundi, national, regulatori, domain, incompat, conseil, agenda, assur, mandat, magistratur, text, togoles, round, congo, gabones, armenia, transit, censur, modal, guarantor, envoy, uniti

In lay terms, the word "topic" carries strong connotations: It connotes a discussion of a substantive idea or concept. In the context of topic modeling, however, "topic" is a term of art that refers simply to a set of words that have a particular probability of appearing in conjunction with each other. Some, but not all, of the "topics" identified by the analysis resemble "topics" in the conventional sense. These include subjects that are recognizable from the broader literature on

constitutional design and diffusion, such as "federalism," "elections," "parliamentarism," and "monarchy."

Other "topics" identified by the topic model, however, do not resemble topics in the conventional sense. They are more reminiscent of dialects, in the sense of particular styles or genres of language that may be more reflective of style, idiom, mood, or ideological framework than institutional design features or legal concepts. For many researchers—such as those who employ topic modeling as an automated substitute for human coding—semantic topics that do not map onto substantive topics may be the equivalent of verbal noise or unwanted error. One researcher's garbage, however, may be another researcher's gold. Topics of this type may function as linguistic markers of historical, ideological, or stylistic influences that would be difficult for human readers to identify, much less quantify.

The results of the topic model reveal an abundance of these subtle yet revealing linguistic patterns. On the whole, topics that are associated in some way with various hegemonic influences such as colonialism and socialism—we might call them hegemonic dialects—are more numerous and prevalent than topics associated with the kinds of institutional design choices on which the constitutional design literature tends to focus.[40] More constitutional verbiage is spent on the rhetorical tics and tropes of British or French colonialism or Soviet domination than the finer points of electoral systems or federalism. Likewise, notwithstanding the amount of attention that courts and scholars tend to lavish on constitutional rights, rights talk makes up a relatively small proportion of the actual language of the average constitution: together, the two rights-specific topics account for only 6 % of the overall text corpus.[41]

The topic prevalence measures produced by the topic models offer a quantitative measure of the overall impact of various hegemonic influences on constitutional semantics. Table 2 reports the average (mean and median) prevalence of each topic in each model, where prevalence is measured as the

40 *See* David S. Law, *Constitutional Dialects: The Language of Transnational Legal Orders, in* CONSTITUTION-MAKING AS TRANSNATIONAL PRACTICE (Greg Shaffer et al. eds., forthcoming 2018) (using a seven-topic model to highlight the massive impact that colonialism and socialism have exerted on the verbal content of constitutions over the last two centuries).

41 The mean prevalence of the universalist dialect across all texts is 3.4 %, while the positive-rights dialect accounts has a mean prevalence of only 2.7 %. *See infra* Table 2. The median prevalence of these topics is lower still (1.2 % for the universalist dialect and 1.1 % for the positive-rights dialect). Removal of the seven human rights treaties from the corpus would further decrease both the mean and median prevalence of these topics. Given that only seven of the 622 documents in the corpus are human rights treaties, however, any effect they have on the average is minute.

Table 2: Average prevalence of all topics across all texts.

topic	1: parliament	2: legislative power	3: municipalities	4: civil service	5: local gov't	6: officeholder	7: Commonwealth courts	8: governor-general	9: comm'ns & tribunals	10: social, economic & cultural rights	11: legislative rules	12: Latin America	13: legal system	14: socialism	15: judiciary
mean prevalence	0.0097	0.04315	0.0253	0.0182	0.0102	0.0369	0.0194	0.0163	0.0358	0.0273	0.0560	0.0276	0.05392	0.03695	0.0544
median prevalence	0.0002	0.0015	0.0003	0.0009	5.5368 E-05	0.0186	9.201 E-05	0.0001	0.0126	0.0112	0.0514	0.0014	0.0001	0.0013	0.0543

16: legislative chambers	17: public order	18: gov't powers	19: territory	20: federalism	21: republic	22: provinces	23: monarchy	24: generic rights	25: elections	26: foreign affairs	27: parliamentarism	28: generic constitutional language	29: communism	30: Francophonie
0.0178	0.0822	0.0436	0.0072	0.0119	0.0304	0.0109	0.0412	0.0343	0.0299	0.0517	0.0335	0.0149	0.0413	0.0773
3.7546E-05	0.0675	0.0276	0.0001	0.0001	0.00666	0.0006	0.0004	0.0124	0.0124	0.0068	0.0024	0.0002	4.1837 E-05	0.0044

proportion of a text attributable to a particular topic and thus ranges from 0 to 1.0. The tables show that a hypothetical "average" constitution (average, in the sense of containing the mean proportion of each topic) would consist primarily of semantic content associated with a handful of hegemonic influences. For example, the model associates an average of 25% of the semantic content with British colonialism,[42] 8% with French colonialism,[43] and another 8% with socialism and communism.

By contrast, substantive topics of the type that preoccupy the constitutional design literature tend on average to make up a smaller proportion of the text. The mean prevalence of the elections topic and the foreign affairs topic is only 3%, for instance, while that of the federalism topic is only 1%. Needless to say, however, actual constitutions can and do depart significantly from the average. Sometimes, the reasons for this variation are quite obvious. For instance, the constitutions of federal countries contain significantly more language related to federalism: the federalism topic accounts for a full 32% of the Swiss constitution, 31% of the Nigerian constitution, and 25% of the German and Austrian constitutions.

C The Language of International Organizations: the Universalist Dialect

The impact of international organizations and the international legal order on constitutional semantics becomes apparent from comparative analysis of the language found in international and regional human rights instruments, on the one hand, and the language that constitutional drafters use to address rights-related topics, on the other.[44] This comparative semantic analysis enables us to identify and inventory what are, in a literal sense, different varieties or dialects of "rights talk."[45] Of the thirty topics listed in Table 1, there are two topics in particular that correspond

42 This percentage is the sum of the mean prevalence of the "parliament," "parliamentarism," "Commonwealth courts," and "governor-general" topics. *See infra* Table 2.

43 This figure is the mean prevalence of the "Francophonie" topic. *See id*.

44 *See generally* Colin J. Beck et al., *World Influences on Human Rights Language in Constitutions: A Cross-National Study*, 27 INT'L SOCIOLOGY 483, 496–97 (2012); Zachary Elkins et al., *Getting to Rights: Treaty Ratification, Constitutional Convergence, and Human Rights Practice*, 54 HARV. INT'L L. J. 61, 69–81 (2013) (describing the mutual influence of constitutions and human rights treaties).

45 *See* GLENDON, *supra* note 5, at 11–13; Kenneth Cmiel, *The Recent History of Human Rights, in* THE HUMAN RIGHTS REVOLUTION: AN INTERNATIONAL HISTORY 27, 32 (Akira Iriye et al. eds., 2012) (observing that "human rights talk communicates across cultures in ways similar to

to different dialects of rights talk. The first is the topic labeled "generic rights" (topic 24), and the second is labeled "social, economic, and cultural rights" (topic 10). Each of these dialects is associated with a different set of constitutions.

The topic labeled "generic rights" is virtually synonymous with the language found in the human rights instruments produced by various transnational intergovernmental organizations. This semantic topic appears to be, in effect, the language of international human rights law or the language that international organizations use to address questions of rights. We might call it a *universalist dialect* of rights talk: it is a manifestation of the ideological view, made explicit in many constitutional preambles, that national constitutions must embody and respect universal norms.[46] This universalist dialect penetrates national constitutions to varying degrees but, on the whole, is growing in popularity, as shown in Figure 3.

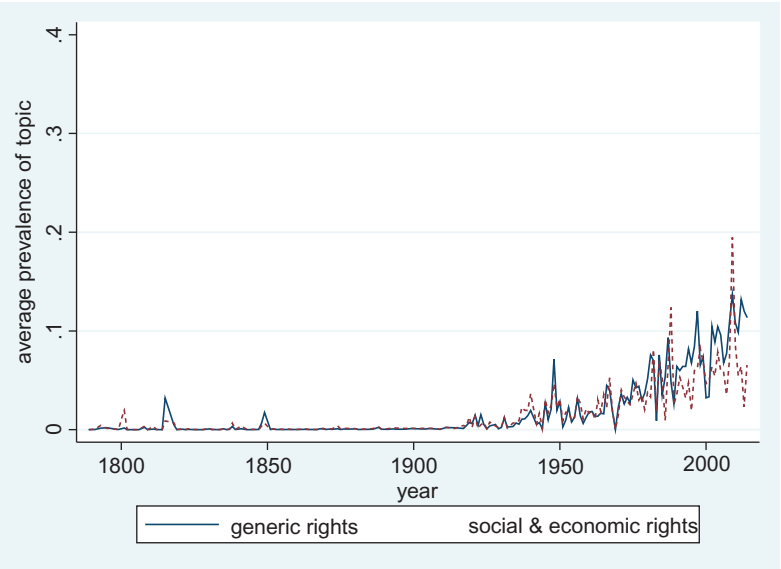

Figure 3: Average prevalence of "generic rights" and "social, economic, and cultural rights" topics in new constitutions over time.

money, statistics, pidgin English, or a discussion of soccer," and that "human rights have become one of the linguae francae of a globalized world").

46 *See* Law, *supra* note 17, at 164.

In terms of its substantive character, the keywords associated with this topic cover a range of rights both old and new. The vocabulary combines references to traditional negative rights ("discrimination," "dignity," "media," "police") with language that evokes rights of newer vintage ("environment," "education," "gender"). The unifying theme of this dialect of rights talk is that it centers on highly popular (and thus somewhat generic) rights that are found in a wide range of both international legal instruments and domestic constitutions.[47]

As Table 3 shows, all five of the texts with the highest proportion of this topic are international and regional human rights instruments. The next five highest-scoring constitutions on this topic are all national constitutions of recent vintage, with none older than 2002. These constitutions also share other characteristics that ought to have rendered them especially susceptible to the influence of the international community.[48] All but one belongs to African states (Angola, South Sudan, Somalia) or to states in which international organizations played a significant role in the constitutional drafting process (South Sudan, Timor).[49]

Table 3: Texts that contain the highest proportion of the "generic rights" topic.

Text	Year	Prevalence of topic
Caribbean Charter	1997	0.3521
ICESCR	1966	0.2486
UDHR	1948	0.2396
African Charter	1981	0.2196
American Declaration	1948	0.2052
Angola	2010	0.2036
East Timor	2002	0.1890
Ecuador	2008	0.1825
South Sudan	2011	0.1772
Somalia	2012	0.1684

47 *See* David S. Law & Mila Versteeg, *The Evolution and Ideology of Global Constitutionalism*, 99 CALIF. L. REV. 1163, 1200–02 tbl.2 (2011) (listing the rights that have appeared in the highest proportion of constitutions since World War II).

48 *See* Beck et al., *supra* note 44, at 492, 495 (finding that "emergent and peripheral countries are more susceptible to world influences" and therefore are more likely to incorporate global human rights discourse into their constitutions, as compared to "older regimes" and "[c]ore nations with strong national political traditions and identities formed in an earlier period").

49 *See* Markus Böckenforde & Daniel Sabsay, *Supranational Organizations and Their Impact on National Constitutions, in* ROUTLEDGE HANDBOOK OF CONSTITUTIONAL LAW 469, 471 (Mark Tushnet et al. eds., 2013); Kevin Cope, *The Intermestic Constitution: Lessons from the World's Newest Nation*, 53 VA. J. INT'L L. 667, 695 (2013).

Statistical analysis confirms the impressions left by Table 5. The STM software package includes support for covariate analysis. Like regression analysis, covariate analysis can be used to test whether a particular variable is correlated in a statistically significant way with the prevalence of a given topic, controlling for other variables. Four variables were tested: (1) the age of the document; (2) the length of the document (measured by word count); (3) the geographical region to which the document belongs[50]; and (4) the legal family to which the country that adopted the constitution belongs.[51] For purposes of evaluating the effect of region, the baseline for comparison is Central and Eastern Europe, while for the legal family variable, the baseline category is the American legal family.[52] The results of the covariate analysis are reported in Appendix A,[53] while Table 4 reports the results that are statistically significant with respect to the "generic rights" topic (or universalist dialect) specifically.

50 Each document was assigned to one of nine geographical categories: Central and Eastern Europe; East Asia; Latin America; the Middle East and North Africa; Oceania; South Asia; Sub-Saharan Africa; Western Europe, Canada, and the United States; and an "international" category (for any international or regional human rights instrument that was not assigned to any of the region-specific categories).

51 For purposes of this study, a "legal family" was defined as consisting of those countries that had been colonized by the same colonial power, together with the colonial power itself. Islamic law was treated as a separate legal family, as was international law (which consisted in this case of the seven international and regional human rights agreements included in the corpus.) In those cases in which a country's legal system had been influenced by multiple foreign systems, we attempted to ascertain which foreign system had been most influential. Any country that was never colonized or occupied by another country, did not model its own legal system on that of another country, and did not fall in the Islamic category was coded as making up its own legal family. The net result was a total of eighteen legal families or groupings: American; Belgian; British; Chinese; Danish; Dutch; French; German; Hungarian; Italian; Japanese; Portuguese; Russian; Scandinavian (other than Danish); Spanish; Turkish; Islamic; and the international category described above.

52 Because region and legal family are categorical variables, their effect can only be assessed relative to a baseline or reference category. In the case of legal family, the reference category is the American legal family (meaning the United States and other countries whose legal systems were most heavily influenced by the United States), while the reference category in the case of region is Central and Eastern Europe. Thus, for example, the table reports that membership in the French legal family is negatively correlated with the prevalence of the "generic rights" topic. This means that the "generic rights" topic is less prevalent among constitutions belonging to the French legal family than in constitutions belonging to the American legal family.

53 For reasons of space, Appendix I does not list the results for every legal family and geographical region but instead focuses on three legal families (British, French, and Spanish) and four geographical regions (Sub-Saharan Africa, the Middle East and North Africa, Latin America, and East Asia).

Table 4: Variables that are correlated with the prevalence of the "generic rights" topic.

Variable	Estimated effect on topic prevalence	Statistical significance	Positive or negative effect?
age	−4.26E-04	$p < 0.01$	negative
legal family: international	1.41E-01	$p < 0.01$	positive

Table 5: Texts that contain the highest proportion of the "social, economic, and cultural rights" topic.

Text	Year	Prevalence of topic
Bolivia	2009	0.2789
Ecuador	2008	0.2601
Ecuador	1998	0.2022
Spain	1978	0.1782
Caribbean Charter	1997	0.1646
Dominican Republic	2010	0.1642
Equatorial Guinea	1982	0.1547
Morocco	2011	0.1520
Venezuela	1999	0.1508
Ecuador	1997	0.1464

Table 6: Variables that are correlated with the prevalence of the "social, economic, and cultural rights" topic.[54]

Variable	Estimated effect on topic prevalence	Statistical significance	Positive or negative effect?
age	−4.26E-04	$p < 0.01$	negative
length	3.22E-07	$p < 0.10$	positive
region: East Asia	−0.0274	$p < 0.01$	negative
legal family: international	0.141	$p < 0.01$	positive
legal family: French	0.119	$p < 0.01$	positive
legal family: British	−0.0894	$p < 0.05$	negative
legal family: Chinese	−0.0279	$p < 0.10$	positive
legal family: Spanish	0.0340	$p < 0.05$	positive
legal family: Dutch	0.0303	$p < 0.10$	positive
legal family: Portugese	0.0336	$p < 0.05$	positive

54 As in the case of Table 4, the reference category for the legal family variable is the American legal family, while the reference category for the geographical region variable is Central and

Not surprisingly, authorship by an international organization is positively correlated with the prevalence of the universalist dialect. Consistent with its generic or universal character, it is not associated in a statistically significant way with any particular geographical region or legal family. It has also become more popular over time: newer constitutions tend to contain higher proportions of the universalist dialect than older constitutions. Although this species of rights talk has become so ubiquitous that it now seems generic, it was not as pervasive in the nineteenth century: as Table 7 indicates, the ten constitutional texts with the lowest proportion of this dialect all date back to the 1800s or earlier. Instead, as Figure 3 reveals, the extent to which constitutions incorporated this type of language surged in the immediate aftermath of World War II, at the same time as the establishment of the United Nations system. In other words, the universalist dialect has taken hold in tandem with the rise of the post-war international legal system.

Table 7: Texts that contain the lowest proportion of the two rights dialects.

Topic 24 ("generic rights")	Year	Topic prevalence	Topic 10 ("social rights")	Year	Topic prevalence
Canada	1791	1.7695E-06	Canada	1791	0.000001307
Paraguay	1813	1.57663E-05	Tanzania	1961	0.00001263
Mexico	1824	3.91781E-05	Lesotho	1966	0.00001771
Haiti	1806	4.66916E-05	New Zealand	1852	0.00002120
Paraguay	1844	4.70814E-05	Jamaica	1962	0.00002201
Chile	1833	4.96538E-05	Trinidad & Tobago	1962	0.00002261
Norway	1814	5.02947E-05	Malawi	1964	0.00002518
France	1804	5.14984E-05	Sierra Leone	1961	0.00002595
Peru	1828	5.60118E-05	Botswana	1966	0.00003292
Haiti	1811	5.68208E-05	Ghana	1954	0.00003333

These empirical findings support the view that international law has led the way in popularizing the universalist dialect, and that the influence of the international community on national constitution-making processes is growing over time, especially in transitional and periphery states.[55] They also suggest that the semantic topic that we call the universalist dialect is a proxy for—and thus

Eastern Europe. Thus, for example, the table reports that East Asia is negatively correlated with the prevalence of the positive rights topic, while no other regions are listed. This means that the topic is less prevalent in East Asian constitutions than in Central and Eastern European constitutions to a statistically significant extent, and that no other regions differ to a statistically significant extent from Central and Eastern Europe.

55 *See* Böckenforde & Sabsay, *supra* note 49, at 469, 473–75, 481–82 (describing growing influence over time).

serves as a quantitative measure of—the influence of international organizations and international human rights law on particular constitutions.

D Next-Generation Rights Talk: The Positive-Rights Dialect

The "social, economic, and cultural rights" topic, by contrast, might be characterized as a *positive-rights dialect* of rights talk that is concentrated within certain legal families as well as in newer constitutions. The substantive character of this dialect is evident from its inclusion of such high-frequency keywords as "social," "economic," "work," "family," and "culture." It appears in constitutions that contain second-generation rights that confer positive entitlements of a social or economic nature, such as education, health care, and subsistence, as well as third-generation or group rights that benefit certain groups or cultures.[56] The results of the covariate analysis, reported in Table 6, confirm that the positive-rights dialect appears in higher concentrations in newer and longer constitutions. Conversely, constitutions from the East Asian region or the British legal family tend to contain lower concentrations.[57]

Like the universalist dialect, the positive-rights dialect is international in character, in the dual sense that it appears in constitutions around the world as well as certain regional and international human rights instruments. However, it does not feature across as broad a range of international human rights instruments as the universalist dialect and tends instead to appear in a subset of instruments that have more of a positive-rights orientation. Comparison of Tables 3 with 5 reveals little overlap between the texts with a high proportion of the universalist dialect and those with a high proportion of the positive-rights dialect.

Of the ten texts that contain the highest proportion of the "social and economic rights" topic, eight are constitutions from the Spanish legal family, while one francophone African constitution (Morocco) makes the list. Only one of the ten—the Caribbean Charter—is the product of an international organization.[58] It is also the only document that ranks among the top ten texts

56 *See* Law & Versteeg, *supra* note 47, at 1191 (contrasting first-generation, second-generation, and third-generation rights).

57 The negative correlation between the positive-rights dialect and the British legal family is consistent with prior empirical findings. *See id*. at 1229–31 (finding that the constitutions of common law countries tend to emphasize negative over positive rights).

58 *See* CARICOM, CHARTER OF CIVIL SOCIETY (1992); Karel Vasak, *Human Rights: A Thirty-Year Struggle: The Sustained Efforts to Give Force of Law to the Universal Declaration of Human Rights*, 30 UNESCO COURIER 29, 29, 32 (1978) (distinguishing among the various generations of human rights, and between positive and negative rights).

for both the universalist dialect and the positive rights dialect. Examination of only the top ten, however, understates the extent to which the positive rights dialect appears in international human rights instruments. Both the American Declaration and the ICESCR make an appearance just outside the top ten (at fourteenth and fifteenth place, respectively), with the positive rights dialect accounting for just under 14% of both documents. Given the ICESCR's focus on positive rights, its high ranking is only to be expected and corroborates our assessment of the substantive character of the topic.[59]

It is equally unsurprising, and reassuring, that other international human rights instruments known for emphasizing negative rather than positive rights feature a much lower proportion of this topic. Although the ICCPR and ICESCR were both promulgated at the same time by the United Nations, the ICCPR ranks a distant 204th, which is consistent with its emphasis on first-generation negative rights as opposed to the second- and third-generation positive rights that are the focus of the ICESCR.[60] Likewise, the ECHR reflected a Cold War effort to advance a conception of human rights centered on civil and political rights as opposed to social and economic rights or decolonization[61]; at 330th place, it ranks even lower than the ICCPR in terms of the proportion of the positive rights dialect that it contains.

Like the generic rights topic or universalist dialect, the social and economic rights topic is associated with newer texts: eight of the ten texts were authored within the last twenty years. Conversely, as seen in Table 7, the ten texts that contain the least of this topic are all older constitutions from Commonwealth countries: every constitution on the list is at least fifty years old and belongs to a former British colony. On the whole, the prevalence of the positive-rights dialect has increased over time, but with much less consistency than that of the universalist dialect. These trends are captured by Figure 3, which depicts the average prevalence of the "generic rights" and "social and economic rights" topics in newly adopted constitutions over the last two centuries.

59 *See* PAUL GORDON LAUREN, THE EVOLUTION OF INTERNATIONAL HUMAN RIGHTS: VISIONS SEEN 228–33 (3d ed., 2011) (describing how the ideological fracturing of efforts to draft a comprehensive international human rights treaty led to the bifurcation of the ICCPR and ICESCR).

60 *See id*. (contrasting the substantive focus of the ICCPR with that of the ICESCR).

61 *See* Mikael Rask Madsen, *Human Rights and the Hegemony of Ideology: European Lawyers and the Cold War Battle Over International Human Rights, in* LAWYERS AND THE CONSTRUCTION OF TRANSNATIONAL JUSTICE 258, 268 (Yves Dezalay & Bryant Garth eds., 2012) (observing that the ECHR reflected a Cold War effort to advance a conception of human rights that emphasized civil and political rights rather than social and economic rights or decolonization).

Once again, the results of the covariate analysis confirm that these lists are indicative of trends that hold true across the text corpus as a whole. Prevalence of the positive-rights dialect within a given text is inversely correlated with the age of the text but positively correlated with the length of the text: the newer and longer that a constitution is, the more of the social and economic rights topic (or positive-rights dialect) that it is likely to contain. Consistent with Table 7, constitutions belonging to the American and British legal families tend to contain less of this topic than either international legal instruments or constitutions belonging to the Spanish, Portuguese, French, and Dutch legal families. There also exists a regional pattern: with the notable exception of China, East Asian constitutions tend to contain less of the positive-rights dialect than those in other regions.

Conclusion

The emergence of new automated techniques for analyzing raw text expands the arsenal of empirical legal scholarship by making it possible to identify and measure latent linguistic patterns with unprecedented speed and accuracy. It will take time for legal scholars to adopt these techniques and figure out all of the ways in which they can be used. But one especially fitting use of computational-linguistic techniques is to evaluate claims about legal discourse and legal language, such as the widespread notion of a global language of human rights. Analysis of the world's constitutions and human rights treaties suggests that there is indeed some linguistic basis to this notion. A more accurate description of the language used in constitutions and treaties, however, would acknowledge the existence of competing strains of human rights discourse or—to extend the linguistic metaphor still further—distinctive dialects of rights talk. In other words, the metaphor of a global language of human rights should be qualified by the recognition that this particular language is spoken in more ways than one.

It is fair to ask what, if anything, automated content analysis of constitutions or treaties can tell us about the actual world, apart from the accuracy of linguistic metaphors. The answer is that the kinds of verbal patterns identified by these techniques correspond to a variety of real-world phenomena. And this should not surprise us. The language found in constitutions and treaties is not merely a medium of communication, but also the crystallization of ways of thinking. Like the fingerprints left by a thief at the scene of a crime, verbal patterns are rich with telltale signs of authorship, influence, and outlook that

may escape even the author's own awareness. As the famed linguist Edward Sapir observed: "Language and our thought-grooves are inextricably interwoven, are, in a sense, one and the same."[62]

Today's automated content analysis techniques leave much room for improvement,[63] yet they are already capable of capturing subtle and revealing linguistic patterns beyond the detection of human readers. The fact that these patterns correspond to a number of known phenomena confirms the viability of automated content analysis as an empirical legal research method and opens an entire world of possibilities for scholars who have heretofore been able to do little with raw language and have instead been limited to extracting whatever their coding schemes will allow.

In this particular case, the results of the topic model are consistent with several developments that scholars have previously known or suspected. First, the finding that both dialects are growing in prevalence is consistent with the previously documented trend toward "rights creep," wherein constitutions incorporate a growing number of rights over time.[64] All other things being equal, an increase in the number of rights per constitution ought to manifest itself in a corresponding increase in the proportion of constitutional language spent discussing rights. This is precisely the trend that we observe. The consistency of the present findings with the past findings ought to boost our confidence in both the accuracy of the findings themselves and the ability of ACA to capture actual patterns in constitutional and treaty drafting.

Second, the fact that the same dialects run through both national constitutions and international treaties highlights a global trend that threatens to disrupt the traditional categories of legal scholarship, if not law itself. That trend is the mingling and interaction of constitutional law and international law. International law and constitutional law have become increasingly intertwined over the post-war period, to the point that scholars now speak of both the constitutionalization of international law, on the one hand, and the internationalization of constitutional law, on the other: international law has begun to focus on the same subjects and perform the same functions as constitutional law,[65] while constitutional law is increasingly international in

62 Edward Sapir, Language: An Introduction to the Substance of Speech 232 (1939).

63 *See supra* text accompanying note 35 (discussing the "bag of words" assumption).

64 *See* Law & Versteeg, *supra* note 47, at 1194–98.

65 *See, e.g.*, Bardo Fassbender, The United Nations Charter as the Constitution of the International Community 116–21 (2009); Antonio Cassese, *States: Rise and Decline of the Primary Subjects of the International Community, in* The Oxford Handbook of the History of International Law 49, 51, 62–69 (Bardo Fassbender et al. eds., 2012) (describing the "international order" that emerged from the 1648 peace of Westphalia as merely a "cluster of entities,

content.[66] The ability to identify and quantify linguistic similarities across large bodies of text offers a way of measuring empirically the pace and magnitude of these developments, if not legal convergence more generally.

Third, the existence of competing dialects of rights talk confirms and extends what international law scholars have long argued: human rights discourse is not monolithic but remains fractured along ideological lines that date back to the Cold War clash between the negative-rights conception of human rights favored by the United States and Western Europe and the positive-rights conception championed by the Soviet Bloc.[67] The linguistic evidence indicates not only that international law scholars are right to take this view with respect to human rights treaties, but also that the schism they identify at the international level is replicated at the national level in constitutional form.

At the same time, however, the results of the topic model also help to illustrate that there is more than one fault line running through the global language of human rights. Previous empirical work suggests that the mix of rights found in different constitutions is the product of variation along two dimensions.[68] One is the persistent and familiar Cold War divide between a liberal strain of constitutionalism that favors negative rights and individual rights, and a statist strain that leans more in the direction of positive rights and group rights. The other is the degree to which constitutions take a lagging or leading role in the emergence and growth of a generic, or universalist, strain

separate and unconnected" with each state exercising exclusive authority over its own territory, and tracing the subsequent movement toward a system in which states are increasingly accountable subject to legal restrictions on actions within their borders); Jeffrey L. Dunoff & Joel P. Trachtman, *A Functional Approach to Global Constitutionalism, in* RULING THE WORLD?: CONSTITUTIONALISM, INTERNATIONAL LAW, AND GLOBAL GOVERNANCE 3, 25–32 (Jeffrey L. Dunoff & Joel P. Trachtman eds., 2009); Erika de Wet, *The Constitutionalization of Public International Law, in* THE OXFORD HANDBOOK OF COMPARATIVE CONSTITUTIONAL LAW 1209, 1209–30 (Michel Rosenfeld & Andras Sajó eds., 2012).

66 *See*, *e.g.*, Wen-Chen Chang & Jiunn-Rong Yeh, *Internationalization of Constitutional Law, in* THE OXFORD HANDBOOK OF COMPARATIVE CONSTITUTIONAL LAW, *supra* note 64, at 1165; David S. Law, *Generic Constitutional Law*, 89 MINN. L. REV. 652 (2005); Anne Peters, *The Globalization of State Constitutions, in* NEW PERSPECTIVES ON THE DIVIDE BETWEEN NATIONAL AND INTERNATIONAL LAW 251 (André Nollkaemper & Janne E. Nijman ed., 2007); Mark Tushnet, *The Inevitable Globalization of Constitutional Law*, 49 VA. J. INT'L L. 985 (2009).

67 *See supra* notes 57–59 and accompanying text.

68 *See* Law & Versteeg, *supra* note 47, at 1221–26 (applying ideal-point estimation techniques to traditional, hand-coded data, and finding that a two-dimensional model explains 90% of the variation in the rights-related content of constitutions).

of rights talk.[69] Some constitutions limit themselves to the rights that are already generic, while others are more innovative and incorporate rights that are more novel or esoteric by today's standards but may over time become commonplace (thanks in no small part to the well-documented phenomenon of rights creep).[70]

The rights dialects identified by the topic model largely track these prior findings. The positive-rights dialect is the linguistic manifestation of the statist strain, while the universalist dialect encompasses both generic rights and esoteric rights. Constitutions that contain only generic rights speak a rudimentary version of the universalist dialect, while constitutions that introduce more novel and innovative rights into the global canon are expanding the universalist vocabulary and leading the evolution of the dialect.

It might be asked why the model does not also reveal a dialect of rights talk that corresponds to the liberal strain of constitutionalism and serves as the ideological foil to the positive-rights dialect. The simple answer is that there is, in fact, a topic that is clearly associated with liberal rights talk, although it is not explicitly labeled as such. The liberal strain is defined not just by its emphasis on negative liberties, but also by its heavy reliance on an independent judiciary as the guarantor of these liberties.[71] The topic labeled "Commonwealth courts" combines precisely these two elements: it consists of language associated with negative rights and judicial protection of the individual from the state, in the form of judicial imposition of reasonableness requirements on detention and other forms of government action.[72]

Whether this topic ought to be labeled differently is open to debate. Topic labeling can pose a difficult interpretive challenge: there may not be a single correct way of capturing the essence of a subtle or multifaceted topic in two or three words. But the problem is especially acute in the case of certain topics such as this one. As

69 *See id*. at 1243 (observing that "global constitutionalism has a strong and growing generic component" in the form of "a generic core of rights-related provisions that is gaining in both popularity and scope over time").

70 *See id*. at 1221–26.

71 *See id*. at 1224 (observing that "libertarian" constitutions—as opposed to "statist" constitutions—"are heavily oriented toward protecting an individual's interest in freedom from detention or punishment at the hands of the state, and they further enshrine the judicial process as the primary instrument for providing that protection); Law, *supra* note 17, at 166.

72 For example, as Table 1 shows, the words "reason" and "judge" are associated with this topic, as are various former British colonies in the tropics (such as Botswana, Grenada, Kiribati, and Swaziland). When these words appear in the constitutions of these countries, they often do so in connection with, inter alia, the conditions under which the government may deprive people of liberty, and the limits that courts must enforce upon the government in such cases (many of which involve requirements of reasonableness).

its label suggests, the "Commonwealth courts" topic appears mainly in constitutions that lean in a liberal direction or, more specifically, the constitutions of former British colonies that gained their independence after World War II.[73] Not surprisingly, these constitutions share much in common: they tend to cover similar subjects (such as parliamentarism, the role of the monarchy, and the judicial enforcement of negative rights) using similar language. The fact that multiple matters are consistently discussed in conjunction with each other poses a challenge for topic modeling: the software lacks the basis to determine whether it is faced with one giant topic or multiple topics that just happen to coincide. The result is that the constitutions of many former British colonies consist primarily of a small handful of topics that are rather broad and thus defy precise labeling, such as the "Commonwealth courts" topic.[74]

Identification of a linguistically distinct strain of liberal rights talk is also complicated by the fact that many liberal rights have over time become generic.[75] The more popular that liberal rights become, the harder that it becomes to distinguish liberal rights talk from universalist rights talk, and the more that the liberal dialect begins to resemble a subset of the universalist dialect. Various negative rights have become so popular that they are no longer closely identified with liberal ideology or, indeed, any other coherent conception of the role of the state. Insofar as liberal rights continue to graduate into the pantheon of universal rights, liberal rights talk blurs into universalist rights talk and becomes a victim of its own success.

Acknowledgements: This paper was initially prepared for presentation at the Boston College Conference on Global Constitutionalism and Human Rights.I am grateful to all participants at the conference. I am especially indebted to Tom Ginsburg, for sharing his corpus of full-text constitutions and collaborating on the analysis of the topic model described here; Ran Hirschl, for his insightful and trenchant feedback as my discussant; and J.P. Kuhn, Leah Robis, and Rosana Tse for their diligent research assistance, especially in preparing the corpus of constitutions for automated analysis.

73 *See* Law & Versteeg, *supra* note 47, at 1230–31.

74 Specifically, the constitutions with the highest proportions of topics 1 ("parliament"), 7 ("Commonwealth courts"), 8 ("governor-general"), 9 ("commissions and tribunals"), and 27 ("parliamentarism") belong disproportionately to former British colonies located in the tropics.

75 *See* Law & Versteeg, *supra* note 47, at 1200 & tbl.1 (listing the negative rights, such as freedom of expression and freedom of religion, that can now be found in virtually all constitutions).

Appendix A: Effect of selected covariates on topic prevalence

Covariate	Topic	Statistical significance	Effect of covariate on topic prevalence
age	3: municipality	$p < 0.01$	negative
age	6: officeholder	$p < 0.10$	positive
age	8: governor-general	$p < 0.10$	positive
age	10: social, economic & cultural rights	$p < 0.01$	negative
age	12: Latin America	$p < 0.01$	negative
age	13: legal system	$p < 0.01$	positive
age	14: socialism	$p < 0.05$	negative
age	15: judiciary	$p < 0.05$	negative
age	16: legislative chambers	$p < 0.01$	positive
age	17: public order	$p < 0.01$	positive
age	18: government powers	$p < 0.01$	negative
age	21: republic	$p < 0.01$	negative
age	23: monarchy	$p < 0.10$	positive
age	24: generic rights	$p < 0.01$	negative
age	26: foreign affairs	$p < 0.05$	positive
age	30: Francophonie	$p < 0.01$	negative
length	3: municipality	$p < 0.10$	positive
length	6: officeholder	$p < 0.01$	positive
length	7: Commonwealth courts	$p < 0.01$	positive
length	9: commissions & tribunals	$p < 0.01$	positive
length	10: social, economic & cultural rights	$p < 0.10$	positive
length	12: Latin America	$p < 0.10$	negative
length	14: socialism	$p < 0.10$	negative
length	17: public order	$p < 0.05$	negative
length	18: government powers	$p < 0.05$	negative
length	20: federalism	$p < 0.05$	positive
length	26: foreign affairs	$p < 0.10$	negative
length	27: parliamentarism	$p < 0.01$	positive
length	30: Francophonie	$p < 0.01$	negative
legal family: British	2: legislative power	$p < 0.10$	negative
legal family: British	6: officeholder	$p < 0.01$	positive
legal family: British	7: Commonwealth courts	$p < 0.01$	positive
legal family: British	9: commissions & tribunals	$p < 0.01$	positive

(*continued*)

(*continued*)

Covariate	Topic	Statistical significance	Effect of covariate on topic prevalence
legal family: British	20: federalism	$p < 0.10$	negative
legal family: British	21: republic	$p < 0.10$	negative
legal family: British	27: parliamentarism	$p < 0.01$	positive
legal family: Spanish	2: legislative power	$p < 0.10$	negative
legal family: Spanish	6: officeholder	$p < 0.10$	negative
legal family: Spanish	7: Commonwealth courts	$p < 0.10$	negative
legal family: Spanish	9: commissions & tribunals	$p < 0.05$	negative
legal family: Spanish	10: social, economic & cultural rights	$p < 0.05$	positive
legal family: Spanish	13: legal system	$p < 0.05$	positive
legal family: Spanish	18: government powers	$p < 0.05$	positive
legal family: Spanish	26: foreign affairs	$p < 0.10$	positive
legal family: Spanish	27: parliamentarism	$p < 0.05$	negative
legal family: French	10: social, economic & cultural rights	$p < 0.10$	positive
legal family: French	20: federalism	$p < 0.10$	negative
legal family: French	21: republic	$p < 0.10$	negative
region: Sub-Saharan Africa	4: civil service	$p < 0.10$	positive
region: Sub-Saharan Africa	6: officeholder	$p < 0.10$	positive
region: Sub-Saharan Africa	9: commissions & tribunals	$p < 0.05$	positive
region: Sub-Saharan Africa	20: federalism	$p < 0.10$	positive
region: Sub-Saharan Africa	21: republic	$p < 0.01$	negative
region: Sub-Saharan Africa	25: elections	$p < 0.10$	positive
region: Sub-Saharan Africa	29: communism	$p < 0.01$	negative

(*continued*)

(*continued*)

Covariate	Topic	Statistical significance	Effect of covariate on topic prevalence
region: Sub-Saharan Africa	30: Francophonie	$p < 0.01$	positive
region: Middle East/ N. Africa	6: officeholder	$p < 0.10$	negative
region: Middle East/ N. Africa	17: public order	$p < 0.05$	positive
region: Middle East/ N. Africa	21: republic	$p < 0.05$	negative
region: Middle East/ N. Africa	23: monarchy	$p < 0.05$	positive
region: Middle East/ N. Africa	29: communism	$p < 0.01$	negative
region: Latin America	3: municipality	$p < 0.05$	positive
region: Latin America	6: officeholder	$p < 0.05$	positive
region: Latin America	7: Commonwealth courts	$p < 0.05$	positive
region: Latin America	9: commissions & tribunals	$p < 0.10$	positive
region: Latin America	12: Latin America	$p < 0.01$	positive
region: Latin America	14: socialism	$p < 0.05$	negative
region: Latin America	15: judiciary	$p < 0.05$	positive
region: Latin America	17: public order	$p < 0.05$	positive
region: Latin America	18: government powers	$p < 0.10$	negative
region: Latin America	21: republic	$p < 0.01$	negative
region: Latin America	26: foreign affairs	$p < 0.01$	positive
region: Latin America	27: parliamentarism	$p < 0.05$	positive
region: Latin America	29: communism	$p < 0.01$	negative
region: East Asia	4: civil service	$p < 0.10$	positive

(*continued*)

(*continued*)

Covariate	Topic	Statistical significance	Effect of covariate on topic prevalence
region: East Asia	9: commissions & tribunals	$p < 0.05$	positive
region: East Asia	10: social, economic & cultural rights	$p < 0.01$	negative
region: East Asia	16: legislative chambers	$p < 0.10$	positive
region: East Asia	18: government powers	$p < 0.10$	negative
region: East Asia	21: republic	$p < 0.05$	negative
region: East Asia	23: monarchy	$p < 0.05$	positive
region: East Asia	29: communism	$p < 0.05$	negative
region: East Asia	30: Francophonie	$p < 0.05$	negative