Legal N-Grams? A Simple Approach to Track the 'Evolution' of Legal Language[*]

Daniel Martin Katz[a], Michael J. Bommarito II[b], Julie Seaman[c],
Adam Candeub[d] & Eugene Agichtein[e]

[a, d] *Michigan State University - College of Law*
[b] *University of Michigan - Center for the Study of Complex Systems*
[c] *Emory University - School of Law*
[e] *Emory University - Department of Mathematics and Computer Science*

Abstract. In this paper, we highlight the potential of *n-grams* as a vehicle to explore the 'evolution' of the law and legal language. Using the full text corpus of decisions of the United States Supreme Court (1791-2005), we explore the n-gram space, offer some initial results based upon our calculations and highlight the beta version of our n-gram search interface (www.legallanguageexplorer.com).

Keywords. N-Grams, Computational Linguistics, Law, Courts, Judges, Evolution of Law, Legal Memes, Time Series

1. Introduction

Modern legal systems feature bodies of highly technical vocabulary that lawyers, judges and legal academics use to construct legal ideas and legal doctrines [1]. Phrases such as *substantive due process*, *clear and present danger*, *custodial interrogation*, etc. characterize and map onto distinct legal concepts. Common law judges use these concepts to both explain the reasoning in the given case and to offer guidance to those involved in future disputes.

Of course, not all cases or judges are created equal. For most judges, their proposed ideas, doctrines and proffered language combinations are scarcely followed by their brethren and are quickly forgotten [2][3][4]. However, some judges and the legal ideas they champion persist. The law (similar to many other intellectual pursuits) is characterized by a relatively small number of highly influential cases and jurists whose conceptualizations of law come to dominate [5][6].

Judges who create dominant common law rules often use words in new and novel ways. Distinctive language and verbal formulae are the raw materials by which judges help establish their reputation as intellectual thought leaders. In the entrepreneurial business of judging, innovative use of language and metaphor is in part how Holmes became "Holmes" and how Posner became "Posner."

Although judges introduce individual terms in discrete cases and those terms enter the common law lexicon through a process of information diffusion that likely mirrors the spread of other memes, as both an empirical and theoretical matter, the model(s) by which these key phrases become dominant is still drastically underspecified. Indeed,

[*] A previous version of this paper was published in the conference proceedings of *Jurix: The 24th International Conference on Legal Knowledge and Information Systems (Vienna 2011).*

while there is basic knowledge at the level of rough intuition, a more rigorous account would allow for deeper understanding of a number of open questions including:

- What are the patterns of adoption for legal language?
- In different time periods, which authors and which phrases dominate?
- Can we specify the process or processes by which this occurs?
- Can we develop a taxonomy using distinct signatures of these dynamic processes?
- Is there any evidence that language authored in prior periods operates to constrain the actions of jurists in future periods?
- Can accounts of evolution in language map onto the evolution of legal rules?
- Can accounts of evolution in language create an account of judicial influence?

The decision corpus is our archeological record, and that record can be usefully explored using the tools of computational linguistics.

2. Toward the 'Evolution' of Legal Language?

A series of recent projects demonstrate that the study of *memes* offers a scientifically tractable approach to explore cultural adoption and cultural evolution [7][8]. At the same time, there exist a number of computational linguistic techniques that have been fruitfully applied to help enrich positive legal theory [9][10][11][12]. Taken together, these contributions highlight the potential of a computational linguistics-enabled memetics style study of law's evolution. While this is not the first project to engage the concept of legal memetics [13][14][15][16], legal evolution [17][18][19][20] or the transfer to particular legal ideas [21][22], this paper is the first empirical application of the ideas at broad scale, highlighting the potential of *legal n-grams* as a simple, yet powerful, empirical method to broadly explore the evolution of legal language.

3. Legal N-Grams: Data Acquisition and Processing

While well known to linguistics scholars for some time, the concept of *n-grams* has recently been popularized by the release of the Google N-gram explorer [23]. A byproduct of the Google Books Library Project, the *n-gram explorer* allows end users to use language as a prism to explore institutional and cultural transformation processes. Building upon the approach applied in [7][8][23], this paper represents an initial effort to explore the *legal n-gram* space.

Although in principle we could map the *n-grams* contained within any digitized corpus of legal documents, we demonstrate the calculation and implementation of *legal n-grams* in one important corpus of judicial decisions - the entire full text corpus of decisions of the United States Supreme Court (1791-2005). This range begins with the first decision of the Jay Court and terminates with the final decision of the Rehnquist Court.

Data for this project was obtained from *http://bulk.resource.org/* which offers free, high quality access to a wide variety of governmental information including information from the U.S. Securities and Exchange Commission, the U.S. Patent and Trademark Office, public safety and administrative codes as well as judicial opinions from various courts [24].

Leveraging *bulk resource.org*, we extracted the opinion text from each case and store it as a sequence of characters. Humans naturally segment such sequences of characters into at least three types of groupings: words, sentences and paragraphs. In order to analyze the *n-gram space*, we converted this sequence of characters into a sequence of words through tokenization. Word tokenization is a common procedure in computational linguistics and natural language processing, and we apply the popular Treebank tokenization algorithm developed by scholars at the University of Pennsylvania [25].

Once we have obtained a sequence of words, generating the n-gram mapping is trivial. For an *n-gram* of size N, we iterate over words in the length M sequence from index 1 to index M + 1 – N and subset the sequence. This procedure can be understood from the following pseudocode:

```
W : length M sequence, N : size of desired n-grams
    for i = 1:M+1-N
    g := W_{i:i+N-1}
    update_frequency_map(g)
    end
```

Alternatively, consider the *n-gram assignment* process as visualized in Figure 1 below:

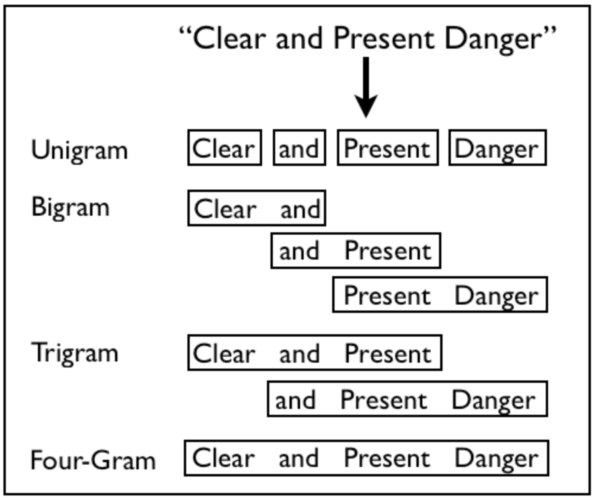


Figure 1. The N-Gram Assignment Process for "*Clear and Present Danger*"

Figure 1 highlights the *n-gram* calculation for Justice Holmes famous phrase from *Schenck v. United States* [26]. As Figure 1 above reveals, this phrase can be subdivided into 4 unigrams, 3 bigrams, 2 trigrams and a single four-gram. Applying this process across the full corpus of United States Supreme Court dispositions (1791-2005), yields a database populated with millions *n-grams* (although a much smaller subset are of high frequency use).

4. The Rise (and Fall) of Particular Legal Phrases

With a complete mapping of the *n-gram* space in hand, a variety of useful applications become possible. For example, legal researchers are often interested in the origins of a phrase and its subsequent usage. While context is of course critical to classifying various forms of usage, a simple starting point for a more detailed inquiry is the construction of a frequency plot. While the presence (or absence) of a phrase is a somewhat noisy signal, tracking temporal variation in rates and patterns of usage can provide quick aggregate insights regarding the relative importance of particular legal questions during different time periods. Consider Figures 2 through 4, which plot the temporal frequency of three meaningful legal phrases: (a) *unconstitutional,* (b) *privacy* and (c) *interstate commerce*.

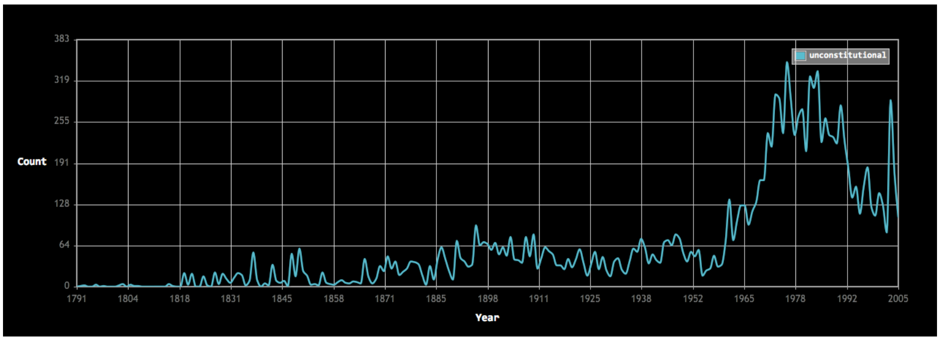


Figure 2. The Time Series for "Unconstitutional" (1791 - 2005)

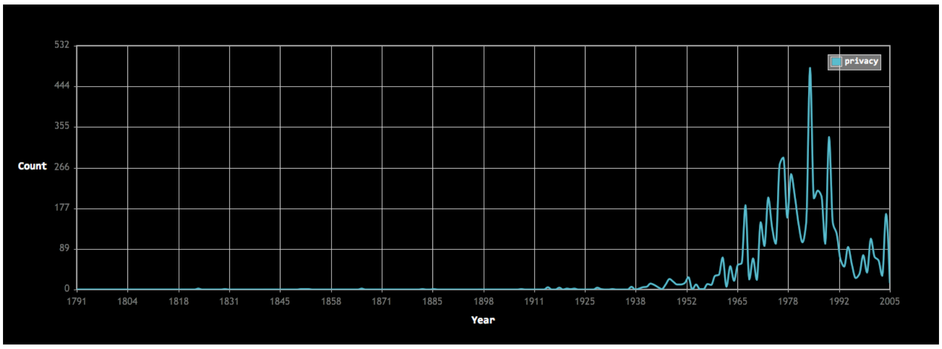


Figure 3. The Time Series for "Privacy" (1791 - 2005)

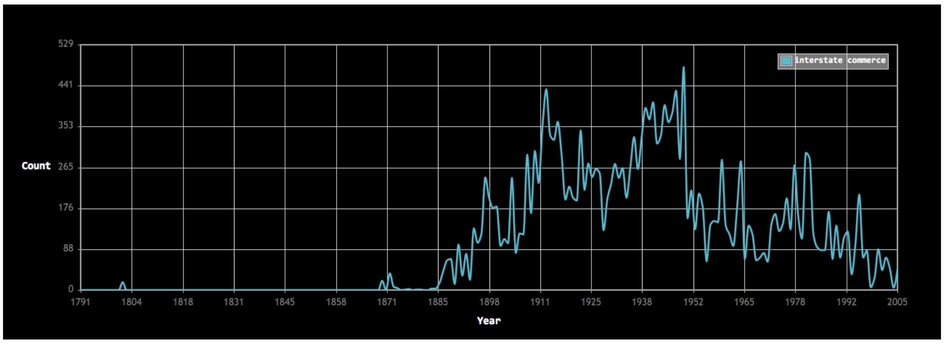


Figure 4. The Time Series for "Interstate Commerce " (1791 - 2005)

Figures 2, 3 & 4 above display the number of instances per year where the phrases, *unconstitutional, privacy* and *interstate commerce*, were used. However, these plots merely report the unnormalized raw counts and do not control (normalize) for important factors such as the volume of cases on the docket in a given year. Figure 5 offers an alternative presentation of the time series for *interstate commerce* where the results are normalized to the yearly rate for the respective term.

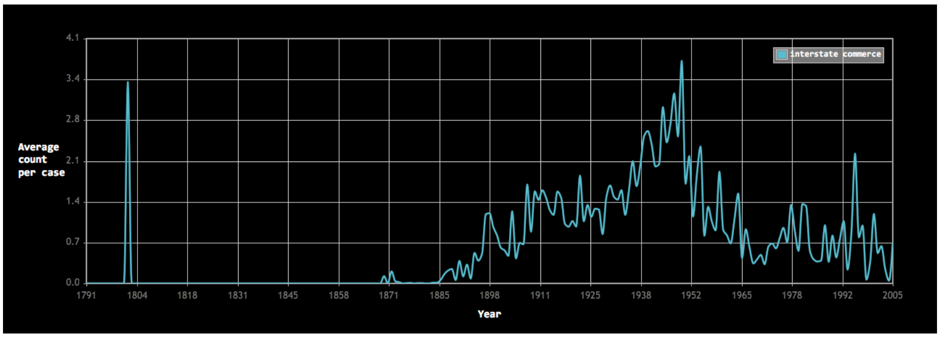
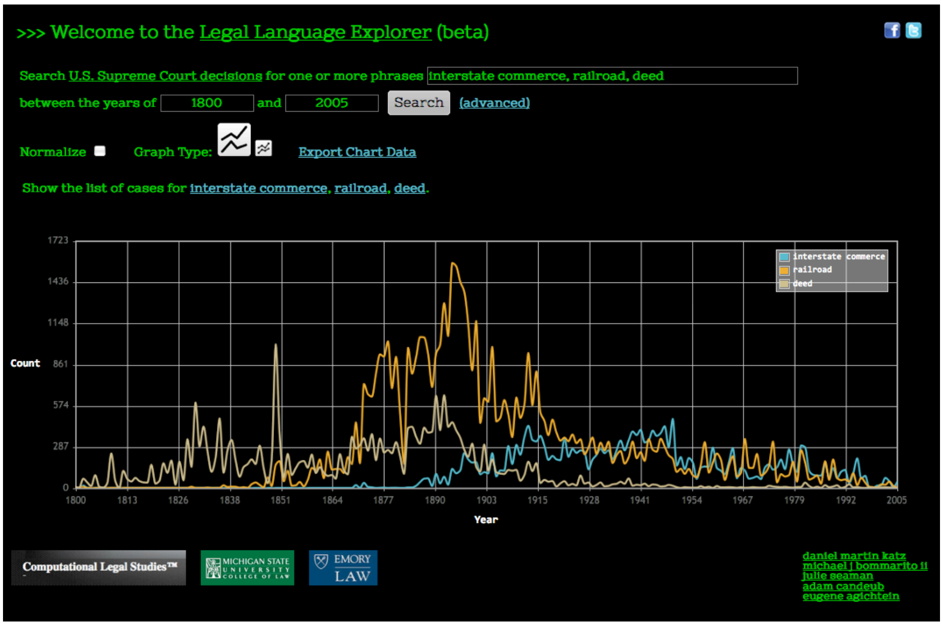


Figure 5. The Normalized Time Series for "Interstate Commerce " (1791 - 2005)

5. Legal Language Explorer.com (Beta Version)

Through this project we have developed a *beta version* of a web-interface that - among other things - will allow researchers (or any interested person) to trace the origin and temporal frequency of *any* arbitrary legal phrase (http://legallanguageexplorer.com/) within the decisions of United States Supreme Court. The site allows basic plotting as well as a number of advanced features. In addition, our site allows the end user to access a case list with proper citations and embedded URLs that link to the full text version of the case at *http://bulk.resource.org/*.

Figure 6 illustrates the beta version of the interface with the search terms *interstate commerce*, *railroad* and *deed*.



>>> Welcome to the Legal Language Explorer (beta)

Search U.S. Supreme Court decisions for one or more phrases interstate commerce, railroad, deed

between the years of 1800 and 2005

Normalize □

Show the list of cases for interstate commerce, railroad, deed.

Figure 6. The Beta Version of Legal Language Explorer.com Interface

6. Conclusion and Future Directions

In this paper, we briefly highlight the potential of *n-grams* as a vehicle to better understand the development of legal language. In addition, we introduce a free web interface designed to allow end users to explore the history of various phrases. Indeed, we believe that the calculation and analysis of *judicial n-grams* will provide a powerful tool not only for legal scholars, but for social and political scientists more generally because a significant number of the phrases contained within the wider universe of *n-grams* are associated with particular legal ideas, doctrines and paradigms. This is especially true in common law systems where the strength and importance of a legal rule often correlates with its pervasiveness in opinions and its persistence through time.

While our approach could be used to explore other judicial decisions or legal documents, the initial results offered here point to a much wider scope of potential future research that might be undertaken. Published cases include a large amount of metadata about a decision, including parties, lawyers, docket history, syllabi, and other supplementary portions. For example, resources such as the *Supreme Court Database (SCDB)* [27] feature useful metadata for each document within our corpora authored after the start of the Vinson Court. These include the author of each opinion, the votes of each justice in response to this language, the *Spaeth topic code* for each case, etc.

Thus while in this version of the project we focus solely on the language of the opinions we expect to add a number of exciting additional search features in the near future.

[1] E. Mertz, The Language of Law School, Oxford University Press, Oxford, UK (2007).

[2] T. Smith, The Web of the Law, San Diego Law Review, **44** (2007) 309-354.

[3] D. Post & M. Eisen, How Long is the Coastline of the Law? Thoughts on the Fractal Nature of Legal Systems, *Journal of Legal Studies*, **29** (2000) 545-584.

[4] D. M. Katz & D. Stafford, Hustle and Flow: A Social Network Analysis of the American Federal Judiciary, *Ohio State Law Journal*, **71** (2010) 457-510.

[5] D. M. Katz, Power Laws, Preferential Attachment and Positive Legal Theory: Part 2, *available at* http://computationallegalstudies.com/2009/08/12/power-laws-preferential-attachment-and-positive-legal-theory-part-2/

[6] D. M. Katz, How Long is the Coastline of the Law? Additional Thoughts on the Fractal Nature of Legal Systems, *available at* http://computationallegalstudies.com/2011/05/13/how-long-is-the-coastline-of-the-law-additional-thoughts-on-the-fractal-nature-of-legal-systems-repost/

[7] J. Leskovec, L. Backstrom, J. Kleinberg, Meme-tracking and the Dynamics of the News Cycle, *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2009.

[8] Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K. Gray, William Brockman, The Google Books Team, Joseph P. Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, Steven Pinker, Martin A. Nowak, & Erez Lieberman Aiden. Quantitative Analysis of Culture Using Millions of Digitized Books, *Science* **331** (2011), 176-182.

[9] Michael Evans, Wayne McIntosh, Jimmy Lin & Cynthia Cates, Recounting the Courts? Applying Automated Content Analysis to Enhance Empirical Legal Research, *Journal of Empirical Legal Studies* **4** (2007), 1007–1039.

[10] M. J. Bommarito II & D. M. Katz, A Mathematical Approach to the Study of the United States Code, **389** *Physica A* (2010) 4195-4200.

[11] K.D. Ashley & S. Bruninghaus, Automatically Classifying Case Texts and Predicting Outcomes **17** *Artif. Intell. & Law* (2009) 125–165.

[12] P. C. Corley, The Supreme Court and Opinion Content: The Influence of Parties' Briefs, **61** *Political Research Quarterly* (2008) 468-478.

[13] Michael Fried, The Evolution of Legal Concepts: The Memetic Perspective, *Jurimetrics*, **39** (1999) 291- 316.

[14] Simon Deakin, Evolution for Our Time: A Theory of Legal Memetics, *Current Legal Problems*, **55** (2002) 1-42.

[15] O. R. Goodenough, Cultural Replication Theory in Law: Proximate Mechanisms Make a Difference, Vermont Law Review, **30** (2006) 989-1005.

[16] J. E. Stake, Are We Buyers or Hosts? A Memetic Approach to the First Amendment, *Alabama Law Review,* **52** (2001) 1213-1268.

[17] E. D. Elliott, *The Evolutionary Tradition in Jurisprudence*, Columbia Law Review, **85** (1985) 38-94.

[18] D. M. Katz, D. Stafford & E. Provins, Social Architecture, Judicial Peer Effects and the Evolution of the Law: Toward a Positive Theory of Judicial Social Structure, Georgia State Law Review, **24** (2008) 977-1002.

[19] A. C. Hutchinson, Evolution and the Common Law, Cambridge University Press, Cambridge, UK (2005).

[20] H. Hovenkamp, *Evolutionary Models in Jurisprudence*, Texas. Law Review, **64** (1985) 645-686.

[21] D. M. Katz, J. Gubler, J. Zelner, M.J. Bommarito II, E. Provins & E. Ingall, Reproduction of Hierarchy? A Social Network Analysis of the American Law Professoriate, Journal of Legal Education, **61** (2011) 76-103.

[22] S.W. Waller, The Law and Economics Virus, *Cardozo Law Review*, **31** (2009) 367-403.

[23] http://ngrams.googlelabs.com/info

[24] http://ftp.resource.org/courts.gov/c/

[25] http://www.cis.upenn.edu/~treebank/

[26] *Schenck v. United States,* 249 U.S. 47 (1919).

[27] http://scdb.wustl.edu/