Meaning Variation and Data Quality in the Corpus of Founding Era American English

Dallas Card
University of Michigan School of Information*
dalc@umich.edu

Abstract

Legal scholars are increasingly using corpus based methods for assessing historical meaning. Among work focused on the so-called founding era (mid to late 18th century), the majority of such studies use the Corpus of Founding Era American English (COFEA) and rely on methods such as word counting and manual coding. Here, we demonstrate what can be inferred about meaning change and variation using more advanced NLP methods, focusing on terms in the U.S. Constitution. We also carry out a data quality assessment of COFEA, pointing out issues with OCR quality and metadata, compare diachronic change to synchronic variation, and discuss limitations when using NLP methods for studying historical meaning.

1 Introduction

Alongside dictionaries and other reference material, legal scholars are increasingly turning towards historical text corpora in order to make arguments about the historical meanings of terms, in so far as they are relevant to modern legal questions. While much of this work involves manually inspecting usages of individual terms, here we demonstrate the use of more recent NLP methods for assessing meaning change and variation, relying on masked language models (MLMs) and focusing on terms in the U.S. Constitution.

The main corpus used in investigating historical legal meanings from the so-called founding era (roughly the second half of the 18th century), is the Corpus of Founding Era American English (COFEA; Hashimoto, 2023), which brings together a series of text collections, including broadsides, legal statutes, debates, and letters. Because COFEA has become so central to arguments about legal meaning, it is essential to investigate both the limitations of the corpus, and what we can learn from it with using state of the art NLP techniques. Using a full-text copy of the corpus, we first investigate and describe its contents, noting issues with unreliable metadata and the quality of the optical character recognition (OCR). In contrast to the in-depth study of individual words, as is more commonly done, we then present a broader automated investigation of meaning, along with an extensive discussion of the limitations of this approach.

The structure of this paper is as follows: we first describe and investigate COFEA (§3), noting several issues and limitations, especially with respect to the OCR and metadata (§4). Using MLMs, we then measure change and variation in the meaning of terms in the U.S. Constitution (§5), both across time and across collections, finding the former to be more extensive than the latter. Finally, we consider the inferred meanings of individual terms in the specific context of the Constitution, finding suggestive evidence of an overall bias towards more formal, as opposed to popular, meanings (§5.3), and end with a discussion of limitations. Data processing and analysis code, as well as interactive versions of all figures, are available online.[1]

2 Background

The U.S. Constitution was drafted in the summer of 1787, and quickly embraced as the fundamental law of the United States. As described in Gienapp (2018), disagreements began almost immediately as to how it should be interpreted, and continue to this day. In recent decades, much of this debate has been with respect to the theory of *originalism*, which is broadly the idea that the original meaning of the Constitution should remain in force.[2]

Independent of fundamental legal theories and questions, most scholars recognize the importance and difficulty of knowing and characterizing the

*Part of this work done while at Stanford University.

[1] https://dallascard.github.io/cofea

[2] There is an enormous literature on originalism which we do not attempt to survey here, but note there are a range of positions which all go under this name; see Solum (2019).

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 841–856

meaning of language. Among other tools used by legal scholars for this purpose is that of *corpus linguistics*—using a corpus of texts to attempt to understand the meaning of particular terms in context (Mouritsen, 2010). Since COFEA was introduced, in part to help answer questions about "ordinary meaning" in the founding era (Hashimoto, 2023), it has quickly become the main reference corpus for this type of investigation.[3]

Indeed, a popular genre of legal scholarship focuses on the meanings of specific terms or phrases, such as "citizens" (Stout et al., 2020), "carry" (Mouritsen, 2010), or "bear arms" (Goldfarb, 2019). Most of these papers tend to use relatively straightforward methods, such as colocations, keyword-in-context, or manual annotations (e.g,. Barclay et al., 2019; Stout et al., 2020), but increasingly we are seeing legal scholars experiment with more advanced NLP techniques (e.g., Nyarko and Sanga, 2022; Livermore et al., 2024).

In this paper, we draw particular inspiration from Lee and Phillips (2019), who focus on handful of terms, like *commerce* and *domestic violence*, using collocations and manual coding, to examine both how the meanings of these terms have changed since the founding era, and the extent of variation within COFEA. Specifically, we leverage the capability of MLMs to assess meaning, considering all terms in the Constitution. While remaining agnostic with respect to debates about originalism, our purpose is both to illustrate the power of such methods, but also to demonstrate and discuss the limitations that these data and methods entail.

3 Data and Preprocessing

We first acquire a full text copy of COFEA from its creators. Although the corpus is made up of six collections, the vast majority of words come from three of these: books, pamphlets, and broadsides (EVANS), letters of the U.S. "Founding Fathers" (FOUNDERS), and a collection of mostly legal documents provided by HeinOnline (HEIN). Smaller collections include laws and resolutions enacted by Congress (STATUTES), records of the 1787 Constitutional Convention (FARRANDS), and a collection of the corresponding debates in state conventions (ELLIOTS). To narrow our focus, we limit our analysis to the period from 1760–1800, which is the period of greatest overlap across sources,

[3]COFEA is publicly searchable via a web interface at https://lawcorpus.byu.edu/cofea.

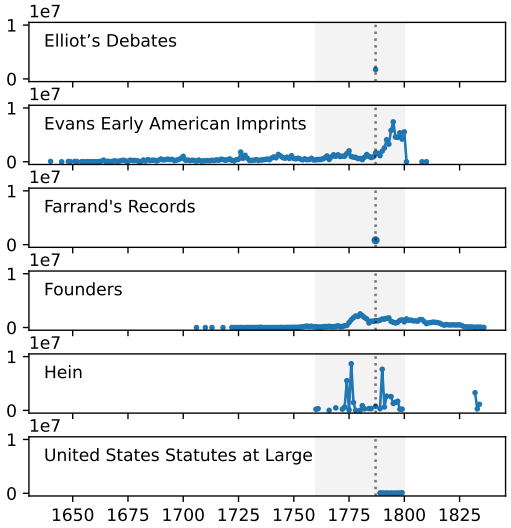

Figure 1: Number of tokens per year in the six collections that comprise COFEA. Grey bands indicate the period studied in this paper, and the dotted line shows the year in which the U.S. Constitution was written. Note that years associated with HEIN documents may in some cases by unreliable (see §4).

as shown in Figure 1. For additional information about COFEA, please refer to Appendix A.

To augment COFEA, we also include all newspaper articles from those years published in *The Pennsylvania Gazette* (TPG), which has previously been used for investigating historical legal language (Strang, 2018). For a modern reference corpus, we use the Corpus of Contemporary American English (COCA; Davies, 2008). For the U.S. Constitution, we use the text provided by the National Archives, including the Bill of Rights. After some initial filtering and preprocessing of COFEA, we tokenize all documents using BERT (Devlin et al., 2019), do spelling normalization, and curate a list of bigrams based on normalized pointwise mutual information, which we augment with the bigrams from Lee and Phillips (2019). Additional details on sources and preprocessing are given in Appendices A and B.

4 Data Quality Assessment

Before investigating meaning change, we first assess the quality of the source documents in COFEA. During initial filtering, we note the presence of a number of documents that contain modern text, such as editorial notes, which we exclude (see Appendix B.1). In addition, the years associated with some documents seem to be incorrect or imprecise. This is especially the case for documents in HEIN, some of which collect together

texts written over multiple years. For our purposes, this makes little difference, as most texts in HEIN appear to date from roughly 1760–1800. However, this could be a more serious issue for investigations which consider precise timing, such as before and after the Constitution was written, or documents from a particular year (see Appendix B.2).

To measure the quality of the OCR for each corpus, we use a simple but commonly used technique of checking for coverage using a dictionary (Springmann et al., 2014; Neudecker et al., 2021).[4] After lemmatizing all documents, we measure the proportion of lemmas that appear in the Webster's 1913 dictionary (Porter, 1913), enhanced with additional terms such as titles, names, places, and abbreviations (see Appendix C.1 for details).[5]

Figure 2 shows a plot of the OCR quality of each corpus based on this measure, where each dot represents one document, sorted by score, and evenly distributed across the x-axis, on a log scale. As can be seen, some of the worst documents are in EVANS and FOUNDERS, but overall HEIN has the worst quality over most of the range. Inspecting the data, the most common tokens in HEIN that are not in the dictionary appear to be poor OCR renderings of terms like "shall" (e.g., "thall") and "justice" (e.g., "juflice"), as well as subwords from improper splitting, like "tion". This means that many occurrences of some terms in the HEIN collection will effectively be missed by a simple keyword search, which is the primary way that scholars access COFEA. For alternative methods of assessing OCR quality, with broadly similar results, please refer to Appendix C.

5 Change and Variation in Meaning

5.1 Measuring changes in meaning

To investigate both change and variation in word meaning, we make use of masked language models, specifically `bert-large-uncased`. In particular, we borrow the technique from Card (2023), which uses changes in the most probable substitutes for terms as informative of word meaning, as this approach allows for both aggregate comparisons,

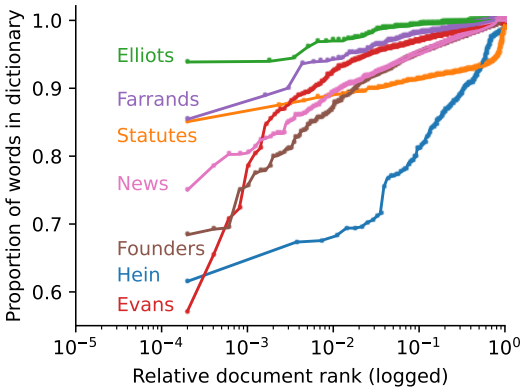

Figure 2: OCR quality across corpora as measured by coverage in the augmented Webster's 1913 dictionary. Each point represents one document, showing the proportion of words in that document found in the dictionary, ranked and distributed across the x-axis.

and inspecting individual term mentions.[6]

For all terms that occur in the Constitution, we sample up to 4000 instances from each of COFEA and COCA, along with samples of additional random words, and pass each word, masked and in context, through the model. Following Eyal et al. (2022), we save the top-k most probable substitutes (with $k = 10$), excluding stopwords (details in Appendix D). For measuring changes in word meaning, we take the total counts of the top-k substitutes for all instances of the word in a corpus, to get a distribution over the vocabulary, and then compute the Jensen-Shannon divergence (JSD) between the distributions for each corpus.

5.2 Diachronic Changes in Meaning

Using the method described above, we first compare COFEA documents from 1760–1800 to modern documents COCA. In doing so, we validate that there have indeed been considerable changes in meaning. As illustrative examples, the terms with the largest measured change in meaning over time are listed in Table 1, along with common probable substitutes from the model. We note that some of those identified involve different grammatical senses (e.g., *captures* as a noun rather than a verb), whereas others represent different meanings (e.g., *quartered* as stationed rather than divided).

Among constitutional terms that have changed the most, many of these had historical military meanings (e.g., *captures*, *quartered*, *training*, *ar-*

[4] According to Hashimoto (2023), documents in COFEA are being evaluated for OCR accuracy, but these scores are not yet available, to the best of our knowledge.

[5] Although this dictionary is from more than a century after most of the COFEA documents, dictionaries are known to lag behind linguistic change, and preserve some historical meanings, making this a reasonable choice.

[6] For comparison to a simpler method of measuring change in meaning, we also include a parallel analysis using the approach of Hamilton et al. (2016) in Appendix D.

Term	Founding era	Modern era	JSD
captures	captures prizes capture seizures prize	captures reflects shows represents describes	0.91
domestic violence	invasion insurrection violence invasions	violence abuse rape crime assault	0.91
marque	mar truce protection war commission	porte salle junta grange crescent	0.90
capitation	direct poll land general state	medicare payment insurance compensation	0.90
quartered	stationed posted kept placed lodged	sliced chopped peeled seeded trimmed	0.86
affirmation	declaration oath certificate deposition	expression recognition acceptance assertion	0.86
training	training raising bringing exercising trained	training education instruction practice	0.85
piracies	crimes offenses piracy murders treason	crimes crime abuses offenses acts	0.85

Table 1: Constitutional terms with the largest meaning change from the founding to modern era, with most common substitutes, and corresponding JSD values. For an expanded list, see Table 7 in the Appendix.

senals). In addition, compared to a random set of background terms, we find that the constitutional terms have changed slightly but significantly more than other terms since the founding era, even after correcting for frequency (see analysis in Appendix D). As a sanity check, we also find that the terms showing the *least* change in meaning tend to be numbers and names of months or days, as was previously exploited by past work (Nyarko and Sanga, 2022). In this case, all but one of the top 20 lowest JSD scores correspond to numbers or temporal terms (e.g., *days*, *years*).

5.3 Specialized vs. Popular Meanings

An enduring debate about the Constitution is the extent to which it was written using a specialized legal vocabulary, as opposed to an accessible popular vernacular (McGinnis and Rappaport, 2018; Gienapp, 2018). Looking first at word counts, we find that, on average, constitutional terms appear more frequently in legal corpora than others. Figure 3 shows the relative frequencies of all terms in the Constitution in Popular, Legal, and Founders documents, projected onto the simplex.[7] As can be seen, most of the terms in the Constitution are relatively evenly represented across these three types of sources. However, on average, there are more terms that appear more commonly in legal sources, rather than popular sources or Founders' papers.

Frequency alone, however, does necessarily tell us about meaning. To that end, we carry out a brief comparison to provide suggestive evidence on this question. To do so, we use the same technique as above for studying change in meaning, but here apply it to variation across parts of the COFEA corpus representing legal vs. popular documents.

Overall, we find the synchronic variation between legal and popular sources is much less than

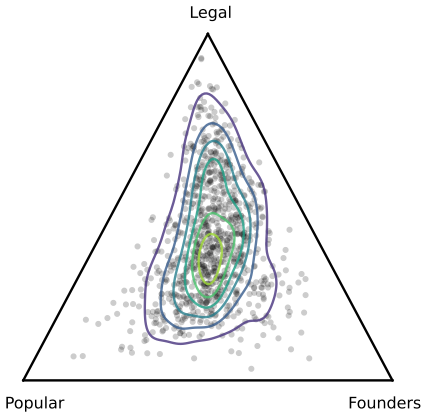

Figure 3: Overall, constitutional terms are more common in legal than other sources. Each point represents one term, with relative frequencies across the three types of sources projected on to the simplex, with contours from kernel density estimation overlaid on top.

the diachronic change in meaning from the founding to the modern era: the mean JSD across all constitutional terms is 0.45 for the former and 0.62 for the latter. Nevertheless, there are many terms which show clear differences in terms of primary meaning across sources. The terms with the largest differences are shown in Table 2, excluding those due to obvious OCR errors.

The above analysis tells us about the most common meanings, but in principle, MLMs also allow us to investigate the specific usage of terms in the Constitution itself. To get at this, we repeat the process described above to embed each specific token in the Constitution, and compare the top ten replacements suggested for the constitutional context to the most common replacements from each of the legal and popular subsets. For each mention, we count the token as leaning towards a specialized (legal) vs. popular usage if the set overlap is at least two greater for one than the other; otherwise, we count the mention as indeterminate. For words that

[7] For legal documents, we use all of STATUTES, FARRANDS, and ELLIOTS, and the parts of HEIN marked as "Legal". For popular sources, we use EVANS and TPG.

Term	Legal	Popular	JSD
tender	tender payment demand currency money	tender kind soft generous great	0.80
dock	dock ship navy naval docks	dock market water street front	0.75
bankruptcies	commerce trade religion slaves debts	losses debts commerce trade failures	0.71
affecting	affecting respecting touching concerning	affecting interesting awful melancholy	0.70
repassed	rejected amended approved repealed	passed crossed entered left ascended	0.70
resignation	resignation removal refusal appointment	resignation submission patience obedience	0.70
sign	sign receive make deliver take	sign head tavern foot signs	0.69
searches	search searches seizures arrests attacks	searches search knows sees inquiries	0.69

Table 2: Constitutional terms with the largest difference in meaning between legal and popular sources in COFEA, (excluding those due to obvious OCR errors), shown with most common substitutes, and corresponding JSD values. All terms here lean towards typical legal meanings when used in the Constitution, according to the analysis described in section §5.3. See Table 10 in the Appendix for an expanded list.

occur more than once in the Constitution, we take the majority outcome, defaulting to indeterminate in the case of ties (see Appendix D.3 for details).

Inspecting the top 40 words with greatest difference between sources, we find only two that lean towards a popular meaning, and in both cases (*aid* and *fix*), these can be attributed to OCR errors. Of the remainder, 26 lean towards a specialized legal meaning, and the remaining 12 are indeterminate. Thus, for the terms in the Constitution that show the greatest variation in meaning across sources, most of these align more with how the terms were used in legal documents, and almost none with usage in more popular sources.

6 Discussion

Despite the popularity of corpus linguistics for assessing historical legal meanings, numerous criticisms have been directed at this approach in general, and COFEA in particular. With respect to COFEA itself, a major concern has to do with it being a non-representative sample of text. In particular, a vastly disproportionate amount of text in COFEA comes from a handful of people (the so-called "Founding Fathers"). As such, COFEA as a whole over-represents elite voices, and under-represents more popular sources.

With respect to the use of corpus linguistics in law, additional critiques have focused on biases in the sources included (Drakeman, 2020; Jennejohn et al., 2021), subjectivity in interpretation (Gries, 2020), placing too much weight on frequent usages (Herenstein, 2017), the historical or linguistic competence required in interpretation (Slocum and Gries, 2020), and lack of clear and reproducible methodology (Henderson et al., 2024). While our more computational approach has the advantage of being fully reproducible, other criticisms still apply, especially with respect to source bias.

Beyond the reasons for caution noted by others, we additionally demonstrate the presence of some modern text in COFEA, as well as some OCR errors and potentially unreliable dates, especially in HEIN. In addition, it is important to note that the methods we have used here primarily establish a single, relatively coarse sense for each time period or collection, and may miss more subtle nuances of meaning. While other methods could potentially exhibit greater sensitivity to such nuances, there is inherently only so much information in a single occurrence of a word. We can thus have much more confidence in aggregate comparisons, even though they lack some sensitivity to context.

Nevertheless, without implying any particular legal relevance, we can say that the extent of differences between specialized vs. popular meanings in the founding era is small compared to meaning changes over time, and that the constitutional terms that differ in meaning across source types seem to broadly lean towards legal senses. Confident judgements about any specific terms would require more input from legal and historical scholars, but overall our results highlight the importance of combining careful manual inspection with reproducible computational analyses on known data.

7 Conclusion

This paper provides the first external assessment of data quality in COFEA, and the first computational analysis of meaning change and variation for all terms in the U.S. Constitution. While noting limitations with respect to both methods and data, especially issues involving dates and OCR quality for the HEIN collection, we find that variation in meaning across sources is much less than change in meaning over time, with weak evidence in favor of the broadly specialized (legal) character of the language in the U.S. Constitution.

Limitations

Although the use of more sophisticated techniques than word counting or manual coding can help to shed some light on historical meanings, we should nevertheless note considerable limitations with the analyses presented here.

First, building on the limitations noted in the Discussion, COFEA itself is far from representative of all text produced during the founding era. Although EVANS claims to be close to comprehensive of certain media (broadsides, books, and pamphlets), even the language used in these is still no doubt far from how people of the time typically spoke in conversation. Moreover, COFEA as a whole is missing newspaper articles, which were an important venue for public communication. Here, we have augmented COFEA with *The Pennsylvania Gazette* because it is easily accessibility, but other sources would be useful to consider.

In terms of the methods used here, our approach provides one dominant sense per time period (or per source), when in fact many terms may exhibit multiple meanings within any collection of documents. As such, our analyses are primarily getting at the most common meanings, and may miss less common senses. It is entirely possible, for example, that the more common sense of a term in the legal sources was also used in popular sources, but with less frequency than another sense. Although one could try to inspect individual usages in each corpus, inferences based on single usages will be much noisier than collection averages.

In addition, the meanings inferred via our approach are relatively coarse. For example, a notable legal case hinged on whether "carrying" a weapon should include transporting it in a vehicle (Mouritsen, 2010). Looking only at common replacements for *carry* in COFEA can help us to get a general sense of its meaning (e.g., bring, take, put), but this may tell us relatively little about the subtleties of implied meaning. As such, this approach is unlikely to be an adequate substitute for historical analysis; at best, computational methods can complement other approaches.

To compare with past work, we opted to use some common phrases, but how to best identify bigrams or other multiword expressions is open to debate. Although the method we use gives reasonable results, we note that the bigrams we included from Lee and Phillips (2019) are quite far from meeting our criteria. This raises the question of whether such phrases (e.g., *public use*) are best understood as distinct concepts or not, and illustrates the importance of having some criteria for deciding what to include. Because meaning is partly but not entirely compositional, studying the meaning of a particular word as opposed to a phrase across all mentions could lead to different results.

Regarding what can be inferred, it is also worth noting that the evidence for certain terms is extremely limited. For example, there are fewer than 100 mentions of the phrase *domestic violence* in COFEA. Of these, only a handful appear in documents that were published before the U.S. Constitution appeared. Thus, COFEA provides limited evidence on the meaning of the phrase as it existed prior to its usage in the Constitution.

While identifying broad changes in meaning is relatively easy, inferring the meaning of an individual mention of a word is inherently difficult, and sometimes underdetermined; our judgements as to the tendency of terms in the Constitution to reflect specialized (legal) as opposed to popular meanings should thus be interpreted cautiously. Ultimately, it is inherently difficult to know precisely what an author meant when they chose to use a word in a particular context. Although the combination of frequency differences and differences in inferred meanings suggest the presence of numerous terms in the Constitution that drew heavily on a specialized legal vocabulary, that does not necessarily apply to the Constitution as a whole.

Finally, and most importantly, we do not address broader questions about originalism, including how much it matters what a particular author meant, how a word would have been understood at a particular time, or whether such questions are even relevant to applying the law in a modern context. Independent of legal implications, we hope our work will add to the discussion of limitations associated with corpus linguistics generally, and the use of COFEA for this purpose in particular.

Acknowledgements

Many thanks to Ben Stone, Nicole Coleman, and Dave Armond for facilitating access to COFEA, and to Accessible Archives for providing access to *The Pennsylvania Gazette*. Additional thanks to Candice Laine Penelton for help with initial data exploration, and to Peter Henderson, Dan Jurafsky, and anonymous reviewers for thoughtful feedback on earlier drafts.

References

Stephanie H. Barclay, Brady Earley, and Annika Boone. 2019. Original meaning and the establishment clause: A corpus linguistics analysis. *Ariz. L. Rev.*, 61:505–560.

Dallas Card. 2023. Substitution-based semantic change detection using contextual embeddings. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*.

Mark Davies. 2008. The Corpus of Contemporary American English (COCA). Available online at https://www.english-corpora.org/coca/.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Donald L. Drakeman. 2020. Is corpus linguistics better than flipping a coin? *The Georgetown Law Journal*, 109:81–101.

Matan Eyal, Shoval Sadde, Hillel Taub-Tabib, and Yoav Goldberg. 2022. Large scale substitution-based word sense induction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4738–4752, Dublin, Ireland. Association for Computational Linguistics.

Jonathan Gienapp. 2018. *The Second Creation: Fixing the American Constitution in the Founding Era*. Harvard University Press.

Neal Goldfarb. 2019. A (mostly corpus-based) linguistic reexamination of D.C. v. Heller and the second amendment. *SSRN Electronic Journal*.

Stefan Th. Gries. 2020. Corpora and legal interpretation: Corpus approaches to ordinary meaning in legal interpretation. In Malcolm Coulthard, Alison May, and Rui Sousa-Silva, editors, *The Routledge Handbook of Forensic Linguistics*, 2 edition, chapter 38, pages 628–643. Routledge.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1489–1501, Berlin, Germany. Association for Computational Linguistics.

Brett Hashimoto. 2023. Corpus of Founding Era American English: designing a corpus for interpreting the United States Constitution. *Corpora*, 18(1):1–14.

Kenneth Heafield. 2011. KenLM: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197, Edinburgh, Scotland. Association for Computational Linguistics.

Peter Henderson, Daniel E. Ho, Andrea Vallebueno, and Cassandra Handan-Nader. 2024. Corpus enigmas and contradictory linguistics: Tensions between empirical semantic meaning and judicial interpretation. *Minnesota Journal of Law, Science & Technology*, 25(2).

Ethan J. Herenstein. 2017. The faulty frequency hypothesis: Difficulties in operationalizing ordinary meaning through corpus linguistics. *Stan. L. Rev.*, 70.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Matthew Jennejohn, Samuel Nelson, and D. Carolina Núñez. 2021. Hidden bias in empirical textualism. *Georgetown Law Journal*, 109:767–811.

Thomas R. Lee and James C. Phillips. 2019. Data-driven originalism. *University of Pennsylvania Law Review*, 167(2):261–335.

Michael A. Livermore, Felix Herron, and Daniel N. Rockmore. 2024. Language model interpretability and empirical legal studies. *Journal of Institutional and Theoretical Economics (JITE)*, 180(2):244–276.

Gregory E. Maggs. 2012. A concise guide to the records of the Federal Constitutional Convention of 1787 as a source of the original meaning of the Convention of 1787 as a source of the original meaning of the U.S. Constitution. *Geo. Wash. L. Rev*, 1707:1–43.

John O. McGinnis and Michael B. Rappaport. 2018. The constitution and the language of the law. *William and Mary Law Review*, 59(4):1321–1412.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.

Burt L. Monroe, Michael P. Colaresi, and Kevin M. Quinn. 2017. Fightin' words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis*, 16(4):372–403.

Stephen C. Mouritsen. 2010. The dictionary is not a fortress: Definitional fallacies and a corpus-based approach to plain meaning. *Brigham Young University Law Review*, 2010(5):1915–1979.

Clemens Neudecker, Konstantin Baierer, Mike Gerber, Christian Clausner, Apostolos Antonacopoulos, and Stefan Pletschacher. 2021. A survey of OCR evaluation tools and metrics. In *Proceedings of the 6th International Workshop on Historical Document Imaging*

and Processing, HIP '21, page 13–18, New York, NY, USA. Association for Computing Machinery.

Julian Nyarko and Sarath Sanga. 2022. A statistical test for legal interpretation: Theory and applications. *Journal of Law, Economics, and Organization*, 38(2):539–569.

Noah Porter, editor. 1913. *Webster's Revised Unabridged Dictionary*. C. & G. Merriam Co.

Brian G. Slocum and Stefan Th. Gries. 2020. Judging corpus linguistics. *Southern California Law Review Postscript*, 94:13–31.

Lawrence B. Solum. 2019. Originalism versus living constitutionalism: The conceptual structure of the great debate. *Nw. U. L. Rev.*, 113(6):1243–1296.

Uwe Springmann, Dietmar Najock, Hermann Morgenroth, Helmut Schmid, Annette Gotscharek, and Florian Fink. 2014. OCR of historical printings of latin texts: Problems, prospects, progress. In *Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage*, DATeCH '14, page 71–75, New York, NY, USA. Association for Computing Machinery.

Abigail Stout, Diana Coetzee, and Ute Römer. 2020. "We the citizens?": A corpus linguistic inquiry into the use of "people" and "citizens" in the founding era. *Georgia State University Law Review Georgia State University Law Review*, 36(5):665–697.

Lee J. Strang. 2018. The original meaning of "religion"' in the first amendment: A test case of originalism's utilization of corpus linguistics. *Brigham Young University Law Review*, 2017(6):1683–1750.

Appendix

A Details on Corpora and Sources

A.1 COFEA

COFEA comprises six subcorpora, each of which we briefly describe here. Token counts for each full corpus and the time period of interest are given in Table 3.

EVANS: The Evans Early Imprint Series, created by the Text Creation Partnership, attempts to collect a nearly comprehensive set of books, pamphlets, and broadsides printed in America from the late 17th to early 19th century. As such, it best represents language of more general usage, compared to the other corpora in COFEA.

FOUNDERS: The National Archives Founders Papers Online project collects together the correspondence, personal papers, and private writings from John Adams, Benjamin Franklin, Alexander Hamilton, Thomas Jefferson, James Madison, and George Washington. Although these documents represent much more informal and personal language, compared to the legal documents in COFEA, they are nevertheless unusual for being written by some of the most powerful and privileged people in the United States at that time.

HEIN: COFEA contains a series of collections from HeinOnline, including: the U.S. Treaties and Agreements Library, U.S. Congressional Documents, American Indian Law Collection, and Session Laws Library. As indicated by the titles, these are primarily statutes and legislative records.

ELLIOTS: Compiled and printed by Jonathan Elliot, Elliot's Debates contains the text of debates held at state conventions in relation to adopting the federal Constitution.

STATUTES: Like HEIN, the US Statutes at Large corpus, contains the text of laws enacted by Congress from the end of the 18th century.

FARRANDS: Farrand's Records of the Federal Constitution of 1787 collects together documentary material related to the Constitutional Convention.

A.2 The Pennsylvania Gazette

The Pennsylvania Gazette (TPG) was a prominent U.S. newspaper throughout the 18th century (Strang, 2018). Although it is smaller than the largest COFEA corpora, the digitized version was hand-keyed, and thus avoids many of the problems that arise with OCR, which we leverage in Appendix C.2. We obtained a copy of The Pennsylvania Gazette from Accessible Archives.[8] Token counts for TPG are also given in Table 3.

Subset	Years	Tokens	Tokens (1760-1800)
EVANS	1640–1810	117926408	78327884
FOUNDERS	1706–1836	67070491	41131670
HEIN	1760–1834	43957047	39230854
ELLIOTS	1787	1758037	1758037
STATUTES	1789–1799	531556	531556
FARRANDS	1787	835837	835837
TPG	1725–1815	28671182	19148702

Table 3: Token counts in each subcorpus

A.3 The U.S. Constitution

There are multiple versions of the text of the U.S. Constitution available online; here, we use the electronic copy provided by the National Archives.[9,10] Specifically, we include the preamble and the seven articles, (but not the signatories), along with the Bill of Rights (Amendments I through X). We do not include Article 1 of the 1789 Joint Resolution of Congress (never ratified) or Article 2 (ratified in 1992). Each numbered section of each Article is processed as a separate document (or the entire Article in the case of Articles without numbered sections).

A.4 COCA

The Corpus of Contemporary American English (Davies, 2008) is a large and balanced dataset containing over 1 billion words of American English from multiple genres. Here, we use the text from the academic, fiction, magazine, and newspaper genres, for the years 1990–2017.

B Data Preprocessing

B.1 Initial filtering

Inspecting the documents in COFEA, it is immediately apparent that there are a number of errors, both in terms of metadata, and the text retrieved. For example, several documents in the FOUNDERS corpus consist only of the phrase "> We were unable to find any matches for your search.", which

[8] https://www.accessible.com/accessible/

[9] https://www.archives.gov/founding-docs/constitution-transcript

[10] Interestingly, there are subtle variations in spelling in different versions of the Constitution provided by the websites for different branches of the U.S. government.

is presumably the result of a scraping error. We exclude these documents from our analysis, along with those with the title "Editorial Note", and those with a body less than nine characters long.

To restrict our analysis to those documents that are primarily in English, we use `fasttext` to assess the primary language of each document, and discarded those classified as anything other than English.[11]

B.2 Metadata Assessment and Correction

All documents in COFEA have an associated year. In the case of EVANS, these are typically the year in which the document (such as a broadside) was published. For FARRANDS, by contrast—which provides a record of the Federal Convention of 1787—the documents are associated with the year in which the debates took place, even though this full collection was not published until 1911 (Maggs, 2012). Similarly, the letters in FOUNDERS are typically dated according to the date of the letter.

For the most part, these dates seem adequate to historical investigation. One collection, however, deserves additional scrutiny, namely HEIN. Although HEIN is one of the three large collections in COFEA by number of tokens, it consists of only 285 documents, in the version of COFEA we gained access to. Obviously many of these documents are quite long, including entire books. In this case, using the date of publication presents more of an issue, if the goal is to capture the language of a particular era. In particular, some documents in HEIN are historical texts, but include a modern preamble. In other cases, the documents collect together laws that were written over many years, such as "Laws of the State of New York".

We briefly reviewed all of the documents in HEIN, and try to associate each document with the most recent text within it, to avoid contaminating historical text with modern usages. Fortunately, the given dates seem to be mostly reliable with respect to whether documents were written during the time period 1760–1800 or not. Thus, the issue with dates has very little effect on our selection of documents to study. However, a somewhat larger number of documents appear to have potentially unreliable dates relative to the year in which the Constitution was written (i.e., they may have been written before 1787, but dated later, or vice versa). As such, one should be especially cautious when trying to look for usages which pre-date the writing of the Constitution itself.[12]

B.3 Spelling normalization

Because some terms appear with varied or archaic spellings (e.g., "choose" vs "chuse"), we begin by normalizing the spelling of all occurrences of terms that appear in the Constitution for which we can find a relevant alternate form. In more detail, we first use word2vec (Mikolov et al., 2013) to train static word vectors on documents from COFEA (from all years), after converting all text to lower case. For all terms in the Constitution, we find the top 100 most similar word vectors, in terms of cosine similarity, and compute the edit distance between the corresponding pair of words. We filter out those that have cosine similarity less than 0.5 or an edit distance of more than two. We then manually inspect all candidates for each term, and keep those that can be identified as legitimate alternate spellings. We then replace all version of the alternate spellings everywhere in our corpora with a standardized form. Ultimately, at most one alternate spelling was found for each term, mostly reflecting differences between British and American spellings. The full list of alternate spellings identified is given in Table 4.

B.4 Identifying bigrams

Because past work has focused on certain bigrams in the Constitution (Lee and Phillips, 2019), we curate a small list of multi-word expressions, to which we add past bigrams that have been considered. To construct our list, we compute the normalized pointwise mutual information (NPMI) for all bigrams that appear in COFEA (not just in the Constitution). We then keep noun-noun and adjective-noun bigrams with NPMI ≥ 0.6 and at least 2000 mentions in COFEA. We also exclude bigrams with incorrectly tagged parts of speech, as well as names of people, quantities, streets, cities, counties, or bodies of water, and greetings, such as "obedient servant". Finally, we add in the three bigrams from Lee and Phillips (2019) for comparison with past work, even though they would not meet our selection criteria for inclusion. The resulting bigrams are given in Table 5, along with the ones added from past work.

[11] https://fasttext.cc/docs/en/language-identification.html

[12] Note that the token counts per year showing in Figure 1 uses the uncorrected dates.

Word	Alternate spelling
ambassadors	embassadors
among	amongst
authorized	authorised
behavior	behaviour
cannot	canot
choose	chuse
choosing	chusing
compel	compell
control	controul
controversy	controversie
days	dayes
defense	defence
domestic	domestick
increase	encrease
entered	entred
expel	expell
favor	favour
guarantee	guaranty
habeas	habeus
honor	honour
inferior	inferiour
judgment	judgement
labor	labour
limited	limitted
massachusetts	massachusets
misdemeanors	misdemeanours
needful	needfull
net	nett
offense	offence
offenses	offences
organizing	organising
payment	paiment
pennsylvania	pensylvania
piracies	pyracies
privilege	priviledge
privileged	priviledged
privileges	priviledges
public	publick
receive	recieve
repel	repell
rhode	rhoad
secrecy	secresy
soldier	souldier
supreme	supream
swear	sware
tranquility	tranquillity
trial	tryal
tried	tryed
until	untill
useful	usefull
vessels	vessells
welfare	wellfare
writs	writts

Table 4: Alternate spellings identified for normalization

Bigram	Tags	NPMI	Count
nova scotia	NN	0.99	3010
united states	NN	0.93	145086
rhode island	NN	0.88	14042
new york	NN	0.82	66155
south carolina	NN	0.82	15966
west indies	NN	0.82	8809
head quarters	NN	0.78	12090
reasonable charges	JN	0.78	8750
u. s.	NN	0.76	3466
north carolina	NN	0.76	12787
yellow fever	JN	0.75	2291
small pox	JN	0.75	4602
minister plenipotentiary	NN	0.75	2930
fellow citizens	JN	0.73	8278
great britain	NN	0.72	35462
common pleas	NN	0.71	4388
west india	NN	0.70	3965
new jersey	NN	0.70	16867
new hampshire	NN	0.68	10289
continental congress	NN	0.67	12059
east india	NN	0.67	2446
lieutenant colonel	NN	0.66	2819
court martial	NN	0.66	5150
grand jury	NN	0.65	2101
military stores	JN	0.64	2656
indian corn	NN	0.64	2149
dwelling house	NN	0.63	6707
human race	JN	0.63	2006
fellow creatures	NN	0.62	2088
pounds sterling	NN	0.62	2236
quarter master	NN	0.61	4532
holy scriptures	NN	0.61	2669
foreign affairs	NN	0.60	4620
french republic	NN	0.60	3219
vice president	NN	0.60	3998
domestic violence	JN	0.36	92
natural born	JN	0.35	368
public use	JN	0.20	738

Table 5: Bigrams included in our analysis

C OCR Quality Assessment

C.1 Dictionary-based OCR assessment

To assess the quality of documents according to coverage in an appropriate dictionary, we make use of the 1913 edition of Webster's Revised Unabridged Dictionary (Porter, 1913), which is available in tabular form online.[13] Because dictionaries typically do not include proper names, we augment the list of terms in this dictionary with the names of all countries, as well as common names of U.S. cities, states, and counties.[14] We additionally add in titles (e.g., "mr", "mrs", "esq") and common person names, including names from the Old Testament,[15] and common baby names based on data from the U.S. Social Security Administration.[16]

We then preprocess each document in COFEA and TPG by using spaCy (Honnibal and Montani,

[13] https://github.com/ahacop/websters-dict-1913-stardict

[14] https://github.com/grammakov/USA-cities-and-states/

[15] https://github.com/hadley/data-baby-names/blob/master/old-testament.txt

[16] https://github.com/hadley/data-baby-names/blob/master/baby-names.csv

2017) to convert all words to lemmas. From this, we drop punctuation, and convert tokens to lowercase. Based on inspecting the remaining tokens, we additionally add to the dictionary common abbreviations for months (e.g., "feb"), roman numerals (e.g., "xvi"), ordinals (e.g., "1st"), common abbreviations (e.g, "servt"), and obvious misspellings (e.g., "recieve"). Given these exclusions and adjustments, we finally compute the proportion of remaining words in each document that appear in the dictionary. In doing so, we further allow for certain suffixes that may have been missed by the lemmatizer (e.g., "ly"), as well as accommodating common British spellings (e.g., substituting "our" for "or" in a word). For full details, please refer to online replication code.

C.2 Language model OCR assessment

As an alternative way of assessing OCR quality, we also try using a character language model to estimate the perplexity of each document. Using `kenlm` (Heafield, 2011), we train a trigram character language model on the union of TPG and the similarly hand-keyed (Text Creation Partnership) portion of the Eighteenth Century Collections Online project (ECCO-TCP), both of which can be assumed to be of high quality.[17] We then use this model to assess the perplexity of each document in each subcorpus in COFEA. Because `kenlm` is designed for word-level models, we convert all documents to sequences of characters, separated by spaces, with spaces converted to a `<space>` token. We also use the "discount fallback" option, because the distribution of characters differs dramatically from that of words.

The results of this assessment are show in Figure 4. As with the dictionary-based assessment, we find that the HEIN subcorpus again appears to be among the worst, in so far as the best documents (farthest to the right) in HEIN are worse than the best document in other collections. Surprisingly, FOUNDERS and FARRANDS appear to be worse than HEIN in the worst cases. In the case of FARRANDS, this appears to be due to the fact that some documents are very short, and contain a high number of capitals letters and punctuation (which are improbable given our training documents), such as "SATURDAY, JUNE 9, 1787". More generally, the presence of abbreviations, symbols, and roman

[17] https://textcreationpartnership.org/tcp-texts/

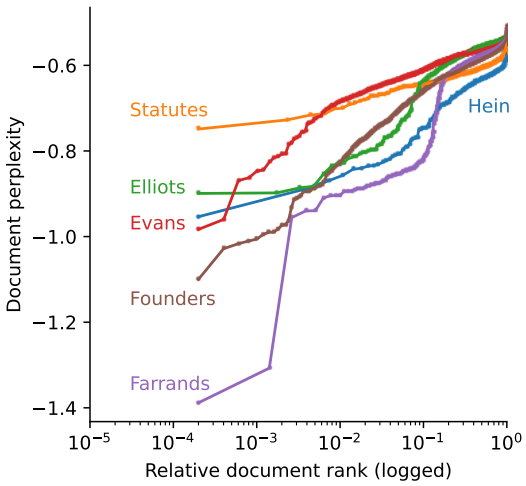

Figure 4: OCR quality assessment made using a trigram character language model.

numerals referring to other sections also hurts perplexity.

The low quality documents in FOUNDERS can similarly be attributed to very short and terse documents (e.g., "July 31. 13. 25.lb brown sugar Th:J."), as well as text that does not resemble normal paragraphs of text (e.g., lists of names or citations). The key takeaway here is that these documents do not seem to have issues with the transcription (such that keyword searches should not be affected), but that certain terms may occur in unusual contexts in these documents.

C.3 Word frequencies across subcorpora

As a final way of identifying potential OCR problems, we examine raw frequency differences between collections in COFEA. Using the log-odd technique from Monroe et al. (2017), we identify the tokens that are most over-represented in each of the three largest collections (EVANS, FOUNDERS, and HEIN), relative to the others.

From this analysis, it is clear that there are differences between the COFEA collections in terms of formality, orthography, languages used, and OCR quality. For example, the most frequent unique terms in EVANS are distinctive punctuation (e.g., •, !, >, <, ?, "), religious terms (e.g., *God*, *Christ*, *church*), and archaic forms (e.g., *thy*, *thou*, *ye*). Distinctive punctuation also appears in FOUNDERS (e.g., & -, ", ", [,]), as well as terms of address (e.g., *sir*, *excellency*), and various short forms (e.g., *obedt*, *servt*, *genl*, *yr*).

For HEIN, many of the most distinctive terms are a mix of governmental terms (e.g, *court*, *commit-*

Term	Count
shall	137392
thall	29932
fliall	12263
ihall	11966
flall	7633
flhall	5527
lhall	3111

Table 6: Frequent misspellings of the term "shall" in HEIN, illustrating the prevalence of OCR errors.

tee, *report*), and apparent OCR mistakes, revealing differences in how these different datasets were digitized. For example, multiple of the most frequent distinctive terms appear to be misspellings of the word "shall", such as "fhall" (see Table 6). This likely results in part from the use of the "long S", which many OCR engines may misinterpret as an "f" or "t". In other words, we should be aware that the documents from each subset may be partially identifiable based purely on orthographic features, independent of content.

D Measuring Change in Meaning

As described in the main paper, we measure meaning change by getting probable replacements for masked terms, and looking at how the distribution of these replacements differs between periods (or collections). As per Eyal et al. (2022), we first adapt a pretrained language model to the data, by doing continued masked language model pretraining, starting with `bert-large-uncased`. In particular, we continue training for five epochs, either on the union of COFEA and COCA, or just COFEA, for measuring change and variation, respectively. We then index the relevant corpora, to collect all occurrences of each term, and sample up to 4000 occurrences of each. We include all the terms in the Constitution (treating bigrams in Table 5 as single words), along with 10,000 random background terms, selected using stratified sampling, to over-represent more common tokens. In addition, for background terms, we limit ourselves to those with at least 50 occurrences, minimum length of 2, and exclude punctuation.

To get replacements, we mask the term of interest (replacing with it with a single mask token, even if it is composed of multiple tokens), and provide the mask, along with up to 50 tokens of context to either side of the fine-tuned model. We then collect the top 10 most likely replacement tokens, only counting whole word replacements with length greater than one, and excluding stopwords from the Snowball sampler. To measure meaning change between two corpora, we collect all replacements for each word from each corpora (e.g., COFEA vs. COCA), and measure the Jensen-Shannon divergence of the two distributions. All processing was run on one A6000 GPU.

Because there is no definitively best method for assessing changing in meaning, we also run an evaluation using the word vector-based approach from Hamilton et al. (2016). In particular, we use `gensim` to fit 100-dimensional word2vec vectors, with a window size of 5, for each corpus or collection to be compared. We then align word vectors using the Procrustes alignment, and use the cosine distance between vectors for a given word as an estimate of the difference in meaning.

D.1 Change in Meaning Over Time

As described in the main paper, we first measure change over time, by comparing COFEA to COCA using the method described above. A more complete list of the terms with the greatest change in meaning, is given in Table 7, along with corresponding JSD values.

To assess the amount of change in meaning experienced by terms in the Constitution, we compare these to the sampled set of random background terms mentioned above. Because of the inverse correlation between meaning change and term frequency noted in past work (Hamilton et al., 2016; Card, 2023), we attempt to measure the association with being in the Constitution while accounting for frequency. In particular, we model change in meaning using linear regression as a function of logged term frequency, and whether the term is in the Constitution. More precisely, we fit,

$$\text{JSD}_t = \beta_0 + \beta_1 \cdot \log(\text{count}_t) + \beta_2 \cdot \mathbb{I}[t \in C] + \epsilon_t,$$

using all terms in the Constitution, along with a large set of background terms. JSD_t is the JSD for term t, count_t is the count of term t in both COFEA and COCA combined, ϵ_t is the error associated with term t, and C is the set of terms in the Constitution. The output of this regression is given in Table 8, showing that the coefficient associated with constitutional terms is indeed significant, although the effect is small.

Figure 5 shows the measured change for constitutional terms (in orange) compared against this random sample of other terms (in blue), with select

Term	Founding era	Modern era	JSD
captures	captures prizes capture seizures prize	captures reflects shows represents describes	0.91
domestic violence	invasion insurrection violence invasions	violence abuse rape crime assault	0.91
marque	mar truce protection war commission	porte salle junta grange crescent	0.90
capitation	direct poll land general state	medicare payment insurance compensation	0.90
quartered	stationed posted kept placed lodged	sliced chopped peeled seeded trimmed	0.86
affirmation	declaration oath certificate deposition	expression recognition acceptance assertion	0.86
training	training raising bringing exercising trained	training education instruction practice	0.85
piracies	crimes offenses piracy murders treason	crimes crime abuses offenses acts	0.85
natural born	natural native free good born	serial cop born psycho professional	0.85
arsenals	magazines fortifications stores barracks	weapons forces capabilities arsenal arms	0.83
emolument	advantage profit benefit interest happiness	finance money profit advantage brown	0.83
counterfeiting	altering making printing destroying signing	fraud theft crime smuggling terrorism	0.83
presentment	complaint trial information indictment report	consolidation verdict processing hearing	0.82
test	test trial proof late bankrupt	test tests testing exam assessment	0.82
reprisal	commissions commission protection passports	retaliation retribution violence punishment	0.82

Table 7: Constitutional terms with the largest meaning change from the founding to the modern era, shown with most common substitutes, and corresponding JSD values.

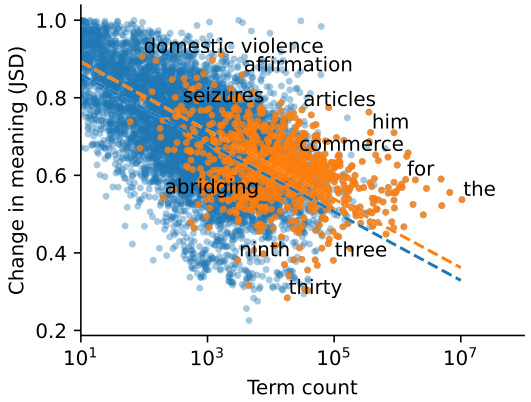

Figure 5: Change in meaning between founding and modern eras vs. term counts in both corpora combined, with constitutional terms shown in orange, random background terms in blue, and select terms labeled.

Variable	Coefficient	Std. err.	p-value
Intercept	0.9480	0.003	< 0.001
In Constitution	0.0328	0.004	< 0.001
log(Count)	-0.0384	< 0.001	< 0.001

Table 8: Regression results modeling change in meaning (JSD) as a function of term frequency and whether or not the term is in the Constitution. As shown, constitutional terms have changed slightly but significantly more than random background terms since the Founding era.

terms labeled. It also shows the fitted regression model, giving an intuitive sense of the difference between constitutional and background terms with respect to change in meaning.

For comparison, Table 9 shows the 15 terms with the greatest change over time, according to the method of Hamilton et al. (2016). This list has many broad similarities with the results of our method, but also some differences. Four terms

Term	Cosine Distance
captures	1.19
respecting	0.90
whereof	0.87
likewise	0.84
vest	0.84
compact	0.83
natural born	0.82
writ	0.82
erection	0.79
emolument	0.79
magazines	0.79
capitation	0.77
compulsory	0.77
corpus	0.77
cannot	0.76

Table 9: Constitutional terms with the largest meaning change from the founding to the modern era, according to the method of Hamilton et al. (2016).

overlap with the top 15 from Table 7. Other terms here are similarly ranked by both methods (e.g., *magazines* is ranked 20th by the method in our paper). Surprising here is the inclusion of *likewise* (ranked in the bottom half of terms by the method in our paper), and the absence of *domestic violence* (which has clearly changed dramatically over time). Overall, the results from the two metrics have a correlation of 0.64.

D.2 Specialized vs. Popular Meanings

As described in the main paper, we assess differences in meaning between legal vs. popular sources by using the method described above to compare legal sources in COFEA (FARRANDS, ELLIOTS, STATUTES, and the documents in HEIN marked as "Legal") vs. popular sources (EVANS, and The Pennsylvania Gazette; TPG). The top 15 terms with the largest difference in meaning across sources are

Term	Legal	Popular	JSD	Lean
aid*	aid said laid id hid	aid assistance support help protection	0.80	P
tender	tender payment demand currency money	tender kind soft generous great	0.80	L
fix*	fix three four two five	fix establish set determine settle	0.78	P
dock	dock ship navy naval docks	dock market water street front	0.75	L
fled*	fled ed led ted fed	fled retired escaped returned went	0.72	I
bankruptcies	commerce trade religion slaves debts	losses debts commerce trade failures	0.71	L
affecting	affecting respecting touching concerning	affecting interesting awful melancholy	0.70	L
repassed	rejected amended approved repealed	passed crossed entered left ascended	0.70	L
resignation	resignation removal refusal appointment	resignation submission patience obedience	0.70	L
sign	sign receive make deliver take	sign head tavern foot signs	0.69	L
searches	search searches seizures arrests attacks	searches search knows sees inquiries	0.69	L
domestic	domestic national public foreign internal	domestic private public social family	0.69	I
escaping	escaping returning flying taken escape	escaping escape avoiding returning escaped	0.68	I
vest	vest invest vested place leave	vest coat jacket shirt hat	0.67	L
training	training muster organizing regulating exercising	training raising bringing trained building	0.67	L

Table 10: Constitutional terms with the largest difference in meaning between legal and popular sources in COFEA, shown with most common substitutes, and corresponding JSD values. Terms marked with an asterisk (*) appear to be primarily due to common OCR errors in Hein. The Lean column shows whether the specific mentions of the term in the Constitution appear to lean more towards the typical legal meaning (L), popular meaning (P), or are indeterminate (I).

shown in Table 10, with those which seem to be due to OCR error in HEIN marked with an asterisk. As can be seen, while some differences are related to different syntactic forms (e.g., *sign* as a verb as opposed to a noun), others show a subtle variation in meaning between legal and popular sources (e.g., *resignation*, *searches*, *domestic*, *escaping*).

For comparison, Table 11 shows the 15 terms with the greatest meaning variation across sources, according to the method of Hamilton et al. (2016). These results, compared against Table 10 are again broadly similar. This time, there are six terms in common among the top 15, but the correlation between scores is only 0.53. As with the results in the main paper, the lower values here again indicate less variation across sources within COFEA, relative to change over time.

Term	Cosine Distance
affecting	0.87
tender	0.81
bankruptcies	0.79
sign	0.79
discoveries	0.76
training	0.76
confession	0.71
sundays	0.70
aid	0.70
high	0.70
likewise	0.69
privileged	0.68
reexamined	0.67
comfort	0.67
nobility	0.66

Table 11: Constitutional terms with the largest meaning change from the founding to the modern era, according to the method of Hamilton et al. (2016).

D.3 Assessing meaning in the Constitution

Although inferring the meanings of individual mentions of terms is inherently difficult, we attempt to characterize the broad meanings of terms as they occur in the Constitution. To do so, we rely on the overall most common meanings inferred from comparing the legal to the popular sources.

For each mention of a term in the Constitution, we gather the k most common replacements, (with $k = 10$), and then compare this set to the corresponding set of k overall most common replacements for each subset (legal vs. popular). That is, let R_c^w be the set of ten most common replacements for word w from corpus c, for $c \in \{\text{Legal } (L), \text{Popular } (P)\}$. For each mention of w in the Constitution, we use the same process to collect the ten most probable replacements for that specific mention, which we denote R_m^w. We then compute the overlap between the set for that mention with corresponding replacements set for legal and popular sources. That is, let $o_c^w = |R_m^w \cap R_c^w|$. If $o_L^w > o_P^w + 1$, then we count that mention of word w as a specialized legal usage. If $o_P^w > o_L^w + 1$, then we count it as a popular usage. If neither is satisfied, we count it as indeterminate.

This gives us an imperfect but still useful measure of meaning alignment. For example, consider the word *tender*. This term has numerous meanings, encompassing nouns, verbs, and adjectives. In the popular sources, the dominant meaning has to do with kindness (kind, soft, generous, great). In

legal sources, by contrast, the dominant meaning has to do with payment (payment, demand, currency, money, security). The word occurs once in the Constitution ("No State shall . . . make any Thing but gold and silver Coin a Tender in Payment of Debts"), from which we can infer that it is being used in the sense of payments. This too is captured by our metric; the top replacements suggested by the model (medium, fund, currency, standard) do not perfectly align with either usage, but are ultimately more similar to the financial sense, which is more common in legal documents, and thus $o_L^{\text{tender}} > o_P^{\text{tender}} + 1$. The same applies to terms like *faith* and *sign*, although many are less clear cut.

D.4 Effect of Continued MLM Training

Unsurprisingly, because BERT was trained primarily on contemporary text, it is better at predicting masked words from modern documents (i.e., COCA), rather than founding era documents. To mitigate this issue, we make use of continued masked language model training, to adapt the base model to our data, as described in the main paper. As a way of quantifying the impact of this, we compute the proportion of mentions included in our analyses in which the original (masked) word is included in the set of predicted terms, and then compare this to using the corresponding off-the-shelf model.

Using the vanilla `bert-large-uncased` model, (without continued training), we find that there is indeed an imbalance between the early and modern text. The proportion of mentions in which the masked term is included in the set of 10 most likely predictions is 60.0% in the modern documents and only 40.0% in the founding era documents. When using the model that has been adapted to these data (which is what we use for obtaining the results reported in this paper), the corresponding values are 64.6% and 60.0%, respectively, demonstrating that the continued MLM training has been effective in adapting the model to our data.