The Cognitive Underpinnings of Legal Complexity

by

Eric Martínez

B.A., Florida State University (2015)
J.D., Harvard Law School (2019)

Submitted to the Department of Brain and Cognitive Sciences
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY IN COGNITIVE SCIENCE

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2024

Authored by: Eric Martínez
Department of Brain and Cognitive Sciences
June, 2024

Certified by: Edward Gibson
Professor of Brain & Cognitive Sciences, Thesis Supervisor

Accepted by: Mark Harnett
Associate Professor of Brain & Cognitive Sciences
Head of the Graduate Program

The Cognitive Underpinnings of Legal Complexity

by

Eric Martínez

Submitted to the Department of Brain and Cognitive Sciences

on June, 2024 in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY IN COGNITIVE SCIENCE

ABSTRACT

Across modern civilization, societal norms and rules are codified and communicated largely in the form of written laws. Although principles of communicative efficiency and legal doctrine dictate that laws be comprehensible to the common world, legal documents have long been attested to be incomprehensible to those who are required to comply with them (i.e. everyone). Why? This thesis investigates this question using the tools of cognitive science.

Chapter II approaches the question from the comprehender side, documenting the cognitive and linguistic factors that make legal documents difficult to understand for non-lawyers. Corpus analyses reveal that legal contracts are laden with psycholinguistically complex structures at a strikingly higher rate than nine baseline genres of English. Experimental evidence further reveals that some of these structures, such as center-embedded syntax, inhibit recall and comprehension of legal content more than others, suggesting that difficulties in understanding legal content result largely from working-memory limitations imposed by long-distance syntactic dependencies as opposed to a mere lack of specialized legal knowledge.

Chapter III extends these results to other legal genres and investigates the cognitive and linguistic profile of law over time. Analyzing every law passed by congress between 1951 and 2022 with matched texts from four different genres, we find that laws have and continue to be disproportionately laden with psycholinguistically complex structures relative to baseline genres of English, suggesting that top-down efforts to simplify legal texts over this period have largely failed.

Chapters IV and V turn to the producer side, investigating why legal actors write in a complex manner in the first place. We find that lawyers likewise struggle to recall and comprehend legal content drafted in a complex register and prefer simplified legal documents to complex documents across virtually every dimension. We further find that people tasked with writing official laws write in a more convoluted manner than when tasked with writing unofficial legal texts of equivalent conceptual complexity, whereas people editing a legal document do not write in a more convoluted manner than when writing from scratch.

From a cognitive perspective, these results suggest law to be a rare exception to the general tendency in human language towards communicative efficiency. In particular, these results indicate law's complexity to be derived from its performativity, whereby low-frequency structures may be inserted to signal law's authoritative, world-state-altering nature, at the cost of increased processing demands on readers. From a legal perspective, these findings call into question the coherence and legitimacy of legal theories and principles whose validity rests on the notion of law being comprehensible to laypeople, such as ordinary meaning, fair notice, and modern variants of textualism. From a policy perspective, this work informs long-standing efforts to simplify legal documents for the public at-large, which, despite bipartisan support, have remained largely intractable. Finally, from a field-building perspective, this thesis lays the foundation for a broader interdisciplinary research program that uses insights from cognitive science to inform long-standing and cutting-edge questions of legal doctrine and policy.

Thesis supervisor: Edward Gibson

Title: Professor of Brain & Cognitive Sciences

Acknowledgments

First and foremost, thanks go to my advisor, Ted Gibson. Nothing in this dissertation would be possible were it not for Ted. I will forever be grateful to him for taking a chance on me as a grad student; for continuing to support and invest in me over the years as a mentor and collaborator; for having such a great sense of humor and for being an easy person to talk to about research and non-research-related things alike.

Next is Frank Mollica, who has been an indispensable collaborator on all of this work since its inception. I am particularly grateful to him for imparting his vast knowledge of computational methods in cognitive science; for his patience in the development of this research program; and for his ability to calmly and effectively brainstorm solutions to problems.

Thanks also go to my committee—Rebecca Saxe, Roger Levy, Nancy Kanwisher, Holger Spamann—for their flexibility, advice and encouragement to finish this thesis early and on short notice; and for their open-mindedness and enthusiasm in overseeing a research program as seemingly divergent from the core of the department as was the topic of this thesis.

Several other mentors have contributed greatly to my development in this field. In particular, I am grateful to Ev Fedorenko for giving me the opportunity to learn fMRI; and for establishing such a great lab environment between Evlab and Tedlab, which has been highly conducive to scientific collaboration and community.

Given that this thesis concerns the law, I would be remiss if I did not mention my academic law mentors. In particular, I am grateful to Michael Frakes, for inspiration to pursue empirical legal scholarship. To Thomas Lee, for helping me think more clearly about the role of language in law and legal interpretation. And to Kevin Tobia, for being an incredibly generous source of guidance and wisdom regarding legal scholarship and academia.

I am also indebted to my undergrad mentors—Brenda Cappuccio, Keith Howard, Carolina González, Lara Reglero, Delia Poey, Roberto Fernández, Antje Muntendam—who inspired and exhorted me to go to grad school to pursue language research (which ultimately did occur, even if later and of a different form than initially envisioned).

Since coming to MIT, I've been lucky to be surrounded by amazing friends and colleagues, who have helped make this institute the most rewarding intellectual environment I have been in, and my years in grad school some of the most meaningful of my life. Adequately acknowledging their individual contributions within the confines of this section is infeasible, though I appreciate every one of them and hope to adequately convey that through other means.

I am particularly thankful to my fellow Tedlab members, past and present, for their inspiring and infectious passion for language research; and to the broader TEvlab community, for creating such a warm and collaborative place to work.

I am grateful to Saxelab, for charitably permitting me to loiter around their lab space, attend lab gatherings, and take naps on their couch. And to Kanlab, for being such hospitable and prosocial lab neighbors, and for providing fun late-night discussions about scientific and non-scientific topics.

I am also grateful to my PhD cohort, for reminding me of the diverse array of research occurring in the department.

Penultimate thanks is owed to my family. I am grateful to my parents and

my grandparents, for always being supportive (and not overbearingly so) of my education. In particular, I am grateful to my dad, a practicing lawyer, for kindling my interest in law; and to my mother, not a practicing lawyer (an accountant), for encouraging me to go to grad school instead of practicing as a lawyer (or accountant). Thanks to my siblings—Alex and Ally—who likewise are (currently) not practicing lawyers or accountants, but rather are simply smart, interesting people whose existence I am grateful for.

Lastly, thanks to you, reader, for engaging with this thesis.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Since the dawn of modern civilization, humankind has codified and communicated societal norms and rules largely in the form of written laws. In order for people to understand and comply with societal norms and rules, it follows that legal content must be drafted in a way such that people can ultimately understand and comply with it.

Indeed, the principle that law should be understandable to the common world constitutes an implicit underlying assumption, if not an expressly core tenet, of modern legal doctrine. For example, under the *fair notice* principle of criminal and constitutional law, laws are mandated to provide proper warning of prohibited conduct "in language that the common world will understand," [1], [2] to "give the person of ordinary intelligence a reasonable opportunity to know what is prohibited, so that he may act accordingly." [3], [4]. Jurists have recently argued that such a fair notice principle may plausibly be satisfied only if ordinary people are able to "read and understand the law for themselves, without need to absorb distinctively legal training" [5], [6]. If so, one would expect in a (coherent) modern legal system that ordinary people would be able to understand the law for themselves, without need to absorb distinctively legal training.

Another example relates to the *ordinary meaning doctrine*, which has been referred to as "the most fundamental principle of legal interpretation" [7], [8], not only of United States law [2], [9]–[11] but of jurisdictions across the world [12]–[17]. The ordinary meaning doctrine has been argued to require that words in legal documents typically be interpreted according to how they are ordinarily understood by laypeople [2], [7], [10], [11], [15], [18], [19]. Given the existence of this doctrine, one would likewise expect that legal documents would ordinarily be understandable to laypeople.

The commitment to the accessibility of law to ordinary people is similarly reflected in the philosophy of textualism, which has become (a) the dominant interpretive approach of the United States Supreme Court [6]; (b) increasingly prevalent in the American judiciary writ-large [20]; and (c) widely endorsed by legal academics, as well [21]. According to Justice Amy Coney Barrett, one of the most prominent living jurists and practitioners of this theory, textualists "view themselves as agents of the people rather than of Congress" and "approach language from the perspective of an ordinary English speaker." [22] To the extent that an ordinary English speaker is unable to comprehend a legal document, it follows that this would likewise undermine the practice of legal interpretation as exercised by contemporary legal officials.

In addition to legal doctrine, principles of communicative efficiency likewise suggest that laws should be understandable to the common world. For example, a burgeoning psycholinguistics literature has uncovered various properties of human language that appear optimized for easing the communicative burden on speakers and listeners [23]–[34]. Two ways in which this efficiency manifests itself relate to word length and syntax. For example, words that are more frequent (such as "the") tend to be shorter than less frequent words (such as "accordion"), such that utterances tend not to be longer than necessary given one's communicative aims [35]. With regard to syntax, it has been observed

across languages that words that depend on each other tend to be close together in an utterance [36], so as to (by hypothesis) avoid overloading working memory capacity when interpreting an utterance [37]. Given that law is encoded in the form of natural language, one would therefore expect this encoding to similarly exhibit properties of communicative efficiency.

These principles notwithstanding, legal documents have long been observed to be notoriously difficult to understand and, relatedly, have been attested to be laden with features that are associated with psycholinguistic complexity [38]–[41]. Consider the following example from a Pennsylvania inkeeper act:

> *No innkeeper or hotelkeeper, which term, as used in this act, shall include apartment hotelkeepers, whether individual, partnership, or corporation, who constantly has in his inn or hotel, which term, as used in this act, shall include apartment hotels, a metal safe or suitable vault, in good order and fit for the custody of money, bank notes, jewelry, articles of gold and silver manufacture, precious stones, personal ornaments, railroad mileage books or tickets, negotiable or valuable papers, and bullion, and who keeps on the doors of the sleeping-rooms used by guests suitable locks or bolts, and on the transoms and windows of said rooms suitable fastening, and who keeps a copy of this section, printed in distinct type, constantly and conspicuously posted, in not less than ten conspicuous places in all, in said hotel or inn, shall be liable for the loss or injury suffered by any guest...* [17], [42]

The clausal material in red is embedded within the center of the main clause, separating important words from each other and leading to a structure that is notoriously difficult to process [43], [44].

In addition to center-embedded syntax, legal documents are also reportedly laden with other properties associated with increased psycholinguistic complex-

ity, including low-frequency jargon (*aforesaid*, *hereinafter*, and *to wit*: [45]), passive-voice constructions ("*The right to trial is waived by the parties*": [46]), and non-standard capitalization ("*ALL WARRANTIES ARE HEREBY DISCLAIMED*":[47]). However, there remains no systematic analysis of to what degree these features are in fact prevalent in legal documents relative to standard-language texts, and insofar as they are present, it remains unclear to what degree they contribute to the attested processing difficulties faced by readers tasked with reading legal documents.

A related question concerns the comprehensibility of legal documents over time. For example, in light of the attested mismatch between the ubiquity and impenetrability of legal documents, public officials have long acknowledged the need to simplify laws for the benefit of the public at-large. In the United States, top-down efforts to simplify government documents for the benefit of the public began at least as early as the 1970s, when Richard Nixon mandated that the Federal Registry be drafted in "layman's terms" and Jimmy Carter issued Executive Orders intended to make government regulations "easy-to-understand by those who were required to comply with them" [48], [49]. These and subsequent attempts to make government language more accessible have been collectively referred to as the "plain language movement" [50], [51]. One of the most recent call-to-arms, the Plain Writing Act of 2010, established formal guidelines regarding how to write government documents clearly for a lay audience [52].

The plain language movement spurred research exploring how best to write "plain-English" layperson summaries of official legal documents, such as jury instructions [53]–[56] and Miranda warnings [57], [58]. Many of the insights from this literature, as well as the general psycholinguistic literature, are now reflected in the Federal Plain Language Guidelines. While these studies have successfully demonstrated the feasibility and importance of using "plain-English" layperson summaries of legal documents to improve comprehension of legal content among

laypeople, these examples apply only to a small portion of the total corpus of legal language and appear less relevant to people's experience with the legal system than actual laws.[1]

However, there remains no systematic analysis of to what extent the plain-language movement impacted the accessibility of official legal documents, such as legislation or regulations. Moreover, on a more general level, there also remains no systematic evaluation of the accessibility of laws over time relative to baseline forms of English.

In addition to the question of *how* legal language is difficult to understand, it also remains an open question *why* legal language is hard to understand—that is, why do lawyers and lawmakers write in such a difficult to understand manner in the first place? Legal scholars, practitioners and commentators have long speculated as to why lawyers and lawmakers write in a complex manner. For example, some scholars have speculated, in line with what has been dubbed the "curse of knowledge" in other disciplines [61], [62], that the reason legal language is so difficult to understand is because lawyers do not realize that they write in an esoteric manner [63]. Other commentators have speculated that lawyers simply write in a complex register out of "habit, laziness" [64] or respect for "tradition" [39], that they "copy and paste" [65] from existing templates with old, complicated terms because that's the "quickest and cheapest way to produce a contract"[66]. A third set of commentators have hypothesized that lawyers write in legalese to be accepted by their peers, to sound more "lawyerly," to "mark themselves as members of the profession" [64]. A fourth set have maintained that lawyers write in legalese as a way of "preserving their monopoly" [67] on legal services and "justifying fees" [64], while others have asserted legal

[1]For example, although jury instructions can be an important part of cases that go to trial, a small and diminishing percentage of civil and criminal cases actually go to trial (as low as 3% for the former and 5% for the latter: [59], [60]). Moreover, while Miranda warnings provide crucial information to criminal suspects in police custody, the majority of individuals' contact with legal language takes place outside the context of criminal or civil suits.

language needs to be complex in order to satisfy certain communicative aims, such as conveying complex legal concepts in a way that "is far more precise than ordinary language" [39], to avoid ambiguity, and/or to ensure enforceability.

Despite the plethora of proposed hypotheses for the origins of the complexity of legal language, there remains no systematic empirical evaluations of these hypotheses.

The purpose of this thesis is to answer, to a first approximation, the answer to each of these questions, using methodological tools and insights from cognitive science.

Chapter II begins from the comprehender side, presenting the results of a corpus analysis and two pre-registered experiments to uncover the cognitive and linguistic factors that make legal documents difficult to understand for non-lawyers. In line with longstanding anecdotal observation, this work finds that legal contracts are laden with psycholinguistically complex structures at a strikingly higher rate than 9 baseline genres of English, and that contracts written with these features are more difficult to understand and recall than contracts of equivalent meaning without those features. This work further reveals that some features, such as center-embedded syntax, inhibited recall and comprehension of legal content more than other features, suggesting that such processing difficulties result largely from working-memory limitations imposed by long-distance syntactic dependencies as opposed to a mere lack of specialized legal knowledge.

Chapter III next replicates and extends the corpus results in laws as opposed to contracts, and analyzes the degree to which the cognitive and linguistic profile of legalese has changed over time, analyzing every law passed by congress between 1951 and 2022 with matched texts from four different genres. This work finds that laws have and continue to be laden with psycholinguistically complex structures relative to baseline genres of English, and that top-down efforts to simplify legal texts have largely failed.

The thesis then turns to the producer side, investigating why lawyers and lawmakers write in a complex manner in the first place. Chapter IV presents two pre-registered studies finding that lawyers (n=211) struggle to recall and comprehend legal content drafted in a complex register than content of equivalent meaning drafted in a simplified register, and prefer simplified legal documents to complex documents on virtually every dimension. In Chapter V, two more pre-registered studies find that people (n=280) tasked with writing official laws write in a more convoluted manner than when tasked with writing unofficial legal texts of equivalent conceptual complexity; and that people editing a legal document were not more likely to write in a convoluted manner than when writing the same document from scratch.

From a cognitive perspective, these results suggest law to be a rare exception to the general tendency in human language towards communicative efficiency. In particular, these findings indicate law's complexity to be derived from its performativity, whereby low-frequency structures may be inserted to signal law's authoritative, world-state-altering nature, at the cost of increased processing demands on readers. From a legal perspective, these results call into coherence and legitimacy of legal principles whose validity rests on the notion of laws being easily interpretable by laypeople. From a policy perspective, these results inform long-standing efforts to simplify legal documents for the public at-large, which, despite bipartisan support, have thus far remained largely intractable. Finally, from a field-building perspective, this work lays the foundation for a broader interdisciplinary research program that uses cognitive science to inform long-standing and cutting-edge questions of legal doctrine and policy.

Chapter 2

Poor Writing, Not Specialized Concepts, Drives Processing Difficulty in Legal Language

This chapter is adapted from the following publication (published under a CC BY 4.0 license):

Eric Martínez, Frank Mollica & Edward Gibson, *Poor Writing, Not Specialized Concepts, Drives Processing Difficulties in Legal Language*, 224 Cognition 1 (2022)

It is reproduced here with slight modifications.

Abstract

This chapter investigates the question of why, cognitively and linguistically, legal language is so difficult to understand for non-lawyers. To answer this question, we begin with a corpus analysis (n 10 million words), which revealed

that legal contracts contained startlingly high proportions of certain difficult-to-process features relative to nine baseline genres of English. Two experiments (N=184) further revealed that excerpts containing these features were recalled and comprehended at lower rates than excerpts of equivalent meaning without these features, even for experienced readers, and that center-embedded clauses inhibited recall more-so than other features. These findings (a) undermine the specialized concepts account of legal theory, according to which law is a system built upon expert knowledge of technical concepts; (b) suggest such processing difficulties result largely from working-memory limitations imposed by long-distance syntactic dependencies as opposed to a mere lack of specialized legal knowledge; and (c) suggest editing out problematic features would be tractable and beneficial for society.

2.1 Introduction

Contracts, such as online terms of service agreements, are at once ubiquitous and impenetrable, read by virtually everyone yet understood by seemingly no one, except lawyers. Dating as far back to the plain language movement in the 1970s, government officials have acknowledged the need to simplify public legal documents for the benefit of society at large. Since then, there has been a sizeable literature exploring how to best simplify public-facing legal language, such as jury instructions [53]–[56] and Miranda warnings [57], [58]. While these studies have successfully demonstrated the efficacy of identifying and replacing problematic features of legal text (such as archaic legal jargon and complex syntax) with "plain English" equivalents to increase comprehension rates among laypeople, they only apply to a small portion of the total corpus of legal language. For example, although jury instructions can be an important part of cases that go to trial, a small and diminishing percentage of civil and criminal

cases actually go to trial (as low as 3% for the former and 5% for the latter: [59], [60]). Moreover, while Miranda warnings provide crucial information to criminal suspects in police custody, the majority of individuals' contact with legal language takes place outside the context of criminal or civil suits and involves more than just public-facing documents, such as contracts and other private-facing documents.

In addition to their prevalence, contracts appear just as impenetrable, if not more, than other forms of legal language. Take the following example from a typical contract: *"In the event that any payment or benefit by the Company (all such payments and benefits, including the payments and benefits under Section 3(a) hereof, being hereinafter referred to as the 'Total Payments'), would be subject to excise tax, then the cash severance payments shall be reduced."*

The clausal material (*"all such payments and benefits, including the payments and benefits under Section 3(a) hereof, being hereinafter referred to as the 'Total Payments' "*) is embedded within the center of another clause, leading to a structure that is notoriously difficult to process [43], [44]. Un-embedding this clausal material into a separate sentence would be straightforward and intuitively easier to process, e.g. as follows: *"In the event that any payment or benefit by the Company would be subject to excise tax, then the cash severance payments shall be reduced. All payments and benefits by the Company shall hereinafter be referred to as the 'Total Payments.' This includes the payments and benefits under Section 3(a) hereof."*

In addition to center-embedded clauses, contracts are also reportedly laden with other properties associated with increased processing demands, including low-frequency jargon (*aforesaid*, *hereinafter*, and *to wit*: [45]), passive-voice constructions ("*The right to trial is waived by the parties*": [46]), and non-standard capitalization ("*ALL WARRANTIES ARE HEREBY DISCLAIMED*":[47]). However, there remains no systematic analysis of to what degree these features are

in fact prevalent in contracts relative to standard-language texts, and insofar as they are present, it remains unclear to what degree they collectively and individually result in language processing difficulties for the average layperson.

Additionally, contracts have become an increasingly prevalent part of the modern era, particularly with the rise of the internet and the constant exposure to online terms of service agreements [68]. While it seems intuitively obvious that very few understand (or even read) the content of these agreements [69], it remains an open question whether on a societal level the increased exposure to contracts might have mitigated the difficulty of reading legal texts, as well as whether on an individual level increased language exposure might mitigate this difficulty.

To address these questions, we conducted a comparative corpus analysis of contracts and a broad sample of standard-English texts and an experiment designed to test the effect of these features on comprehension and recall of legal content. Consistent with previous literature, we find that each of the complex psycholinguistic properties reportedly common in contracts–such as center embedding, low-frequency jargon, passive voice and non-standard capitalization–were strikingly more common in contracts relative to every genre of standard English that we compared it with, and that contracts containing these features were recalled and comprehended at a lower rate than contracts drafted without these features, independent of reading experience. We also found that center-embedding inhibited recall to a greater degree than other features.

2.2 Corpus Analyses

2.2.1 Corpus Materials

To determine the nature and source of processing difficulty in contracts, we first sought to systematically evaluate the degree to which contracts contain proper-

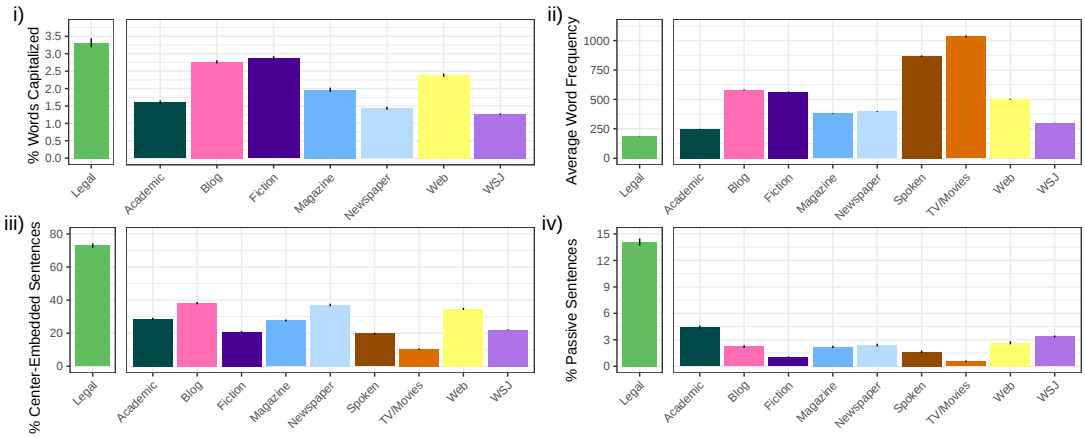

Figure 2.1: Comparison of indices of linguistic processing difficulty in contracts versus various genres of written and spoken English.

ties associated with processing difficulty relative to standard English texts. To do so, we expanded the corpus of contracts used by Gozdz-Roskowski ($n \approx 1$ million words; [70]) with an additional set of contracts ($n \approx 2.5$ million words) drawn from Westlaw's database of court documents published between 2018 and 2020. For our standard English corpus, we compiled: (a) a sample of Wall Street Journal articles ($n \approx 5$ million words) published in 1996, as part of the CSR Wall Street Journal corpus [71]; and (b) a broad sample ($n \approx 10$ million words) of TV/Movie scripts, spoken language, newspaper articles, blogs, magazine articles and web pages from the Corpus of Contemporary American English (COCA: [72]). For both corpora we extracted several linguistic structures at both the word level and sentence level. To determine which features to analyze, we performed a review of the legal language literature to investigate which features were purportedly most common among legal texts. We then reviewed both the legal and the general psycholinguistics literature to determine, of these features, which were attested to affect processing difficulty in legal and/or non-legal contexts.[1]

[1]Of course, there may be other features that we have missed, which may differ across the text types. That is, there may be more ways in which legal texts are more complex than control texts; and there may be ways that legal texts are simpler than control texts.

Here we further motivate and clarify individually each of the five features that our review led us to include in our analysis. Corpus processing and extraction details are provided in the SI.

Indices of Processing Difficulty

Capitalization Non-standard capitalization is ubiquitous in provisions such as warranty disclaimers and limitations of liability, which "must be conspicuous" in order to be legally upheld [73]. [47] found that most standard form agreements used by major companies contain at least one provision in all-caps. Although the use of all-caps provisions is ostensibly for the benefit of the reader, evidence suggests that they do not aid comprehension [47]. Here we sought to determine what percentage of words in contracts were in ALL CAPS relative to standard English.

Word frequency Words that are infrequently used in everyday speech cause processing difficulties for readers relative to higher-frequency synonyms [74]. Legal texts are reportedly laden with "archaic words" such as *aforesaid*, *herein*, and *to wit* [40], which have been shown to be frequently misunderstood by laypeople (e.g., [38]). To evaluate how frequently legal texts appeared in everyday speech relative to a baseline, we compared how frequently each of the content words in each of our corpora appeared in the SUBTLEX word frequency dictionary, a corpus of American film subtitles commonly used as a proxy for standard-English word frequency.

Word choice Insofar as legal terms are low frequency, some argue that this is a necessary consequence of the specialized concepts and corresponding terminology used to refer to those concepts by lawyers (cf. [75]). To evaluate this claim and determine to what extent legal jargon can be replaced by simpler terminology without a loss or distortion of information content, we calculated the proportion of content words in each corpus that had a higher-frequency

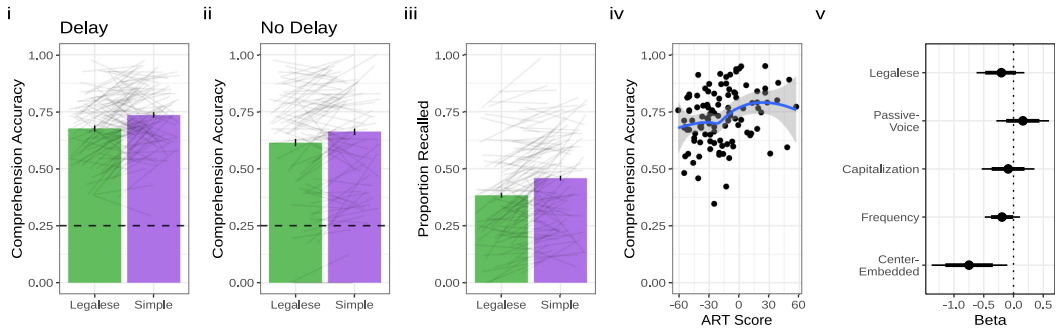

Figure 2.2: Effect of text register (legalese vs simple) on comprehension accuracy in the main experiment (i) and replication study (ii), and recall of legal content in the main study (iii). Effect of language experience (measured using Author Recognition Task) on comprehension accuracy (iii). Effects of linguistic structures on recall (iv). Outer line range reflects the 95% credible intervals over the interaction term, inner line range reflect the 80% credible intervals over the interaction term and points reflect medians.

synonym. In doing so we used two methods, a conservative analysis in which we assumed the authors intended the least common sense of each word used in a given corpus (the idea being that legal terms may resemble common words in form but have a more specialized meaning, such as the concept of "consideration" in contract law [73]), and an anti-conservative analysis in which we assumed the authors intended the most common sense of each word.

Center-embedding Center-embedded structures have long been observed to pose processing difficulties on a reader [43], [44]. The tendency for lawyers to "embed" legal jargon "in convoluted syntax" has been speculated not only to be prevalent in legal texts but as a potential badge of honor for those who wish to "talk like a lawyer" and be accepted by their profession [40]. Here we computed the prevalence of both center-embedding and right-branching embedding in contracts relative to standard English.

Active/Passive Voice Relative to their active-voice counterparts, passive-voice structures are acquired later by children than their active voice counterparts [76], and may continue to pose difficulties for adults [46]. [70] found passive

structures to be more prevalent in the contractual materials he used relative to other legal and non-legal genres (such as newspapers). Here we similarly sought to determine the prevalence of passive structures in contracts relative to standard language, both with regard to agentless passives ("*The right to a trial is waived*") and "by" passives (also known as reversible passives: "*The right to a trial is waived* ***by both parties***").

2.2.2 Results

Results are visualized in Figure 2.1. Descriptively, all of the metrics we looked at were prevalent to a greater degree in contracts than in the standard English corpus–both overall and within each standard-English subgenre. In most cases, this difference was striking—e.g., center embedding ($OR = 2.56; 95\%\text{CI}\ 2.48 - 2.63$), passive voice ($OR = 4.35; 95\%\text{CI}\ 4.11 - 4.61$), and non-standard capitalization ($OR = 1.78; 95\%\text{CI}\ 1.75 - 1.82$).

Capitalization In our analysis (Figure 2.1i), we find 2.780% of words in the contract corpus (95% CI: 2.752 to 2.806) were in all caps, as compared to 1.693% in the non-spoken portion of the standard-English corpus (95% CI: 1.684 to 1.703).

Word frequency As seen in Figure 2.1ii, content words in the contracts corpus occurred, on average, 187.531 times in the SUBTLEX corpus (95% CI: 185.455 to 189.847), compared to 451.399 occurrences for content words in the WSJ corpus (95% CI: 450.244 to 452.567).

Word choice Under the conservative method, the percentage of words with a higher-frequency synonym was 13.437% in the contract corpus (95% CI: 13.381 to 13.494) and 8.246% in the standard-English corpus (95% CI: 8.226 to 8.264). When considering only content words, the proportion was 23.556% in the contract corpus (95% CI: 22.467 to 23.650) and 16.128% in the WSJ corpus (95% CI: 16.090 to 16.165). We observe similar results in the anti-conservative anal-

ysis (see SI).

Center-embedding The prevalence of embedded clauses overall was 1.776290 per sentence in the contract corpus (95% CI: 1.754 to 1.797) and 0.841 clauses per sentence in the standard-English corpus (95% CI: 0.837 to 0.844). For center-embedding in particular (Figure 2.1iii), the mean value was 0.729 center-embedded clauses per sentence in the contract corpus (95% CI: 0.715 to 0.744) and 0.272 center-embedded clauses per sentence in the standard-English corpus (95% CI: 0.270 to 0.274).

Active/Passive Voice Passive voice structures occurred at a rate of 0.758 times per sentence in contracts (95% CI: 0.747 to 0.769) and 0.160 times per sentence in standard-English texts (95% CI: 0.159 to 0.162). When considering just "by" passives (Figure 2.1iv)–which are more likely to be replaceable with active voice structures without a loss or distortion of information content–the rate per sentence was 0.141 in contracts (95% CI: 0.137 to 0.145) and 0.0232 in the standard-English corpus (95% CI: 0.0227 to 0.0237).

2.3 Experimental Study

Having demonstrated the presence of complex psycholinguistic properties in contracts, we next conducted an experiment aimed at determining: (1) to what extent the presence of these features in contracts inhibited comprehension and recall of legal content; (2) whether any decline in performance is mitigated by increased language experience; and (3) to what extent certain individual linguistic structures inhibit recall more than others.

2.3.1 Methods

To do so, we developed a paradigm building on [41], with some deviations. We constructed 12 pairs of short contract excerpts. Each pair contained (a) one

excerpt drafted in a legalese register, containing several of the features analyzed in the corpus analyses; and (b) one excerpt drafted in a simple register, identical in content to the other excerpt but without the features analyzed in the corpus analyses. We also implemented the author recognition task (ART; [77], [78]) as a measure of individual differences in experience with language. Here we further elaborate the details of the materials, recruitment of participants, and experimental procedure.

Materials

Our primary materials consisted of 12 pairs of short contract excerpts of roughly 150 words each (see Supplemental Figure 1). First, twelve excerpts were constructed in a standard "legalese" register by the first author, a lawyer, who modeled the content and form of the materials after common naturalistic contracts. Each of the 12 texts for each condition corresponded to one of three types of common contract provisions (with four texts pertaining to each genre), including: (1) general contract provisions, specifying the basic terms of a contractual agreement; (2) liability and warranty provisions, specifying to what degree each party could be sued or held accountable for not adhering to the terms of the agreement; and (3) jurisdiction, venue and choice-of-law provisions, specifying how and where parties could sue or be held accountable for not fulfilling the terms of the agreement. We chose to include these genres so as to maximize generalizability, given that they are present in virtually every modern-day contract. We also sought to achieve a balance for the sake of generalizability with regard to the content within each provision; roughly half the provisions pertained to the lease or sale of goods, and the other half pertained to a services contract, as these types of agreements form the two general categories of contracts according to United States Contract Law [73], [79].

With regard to the language within each contract, each legalese text was

drafted to contain several instances of the features analyzed (and shown to be prevalent in naturalistic contracts) in the corpus analyses, including (a) low-frequency words, (b) center-embedded clauses, (c) passive-voice structures, and (d) non-standard capitalization. To ensure authenticity and minimize potential bias, the language in each legalese text was modeled after that in naturalistic contracts. For some provisions, the language used was virtually identical to that in naturalistic contracts. In other cases, further context was added to ensure that the content could be reasonably understood as an isolated excerpt in an experiment.

From the set of legalese materials, each passage was encoded in terms of legally relevant propositions. From these propositions, each passage was then translated into a "simple" version, which preserved the meaning of the original and differed only with respect to the four surface properties described above. Low-frequency words were replaced with high-frequency synonyms. Center-embedded clauses were "un-embedded" and re-drafted as separate sentences. Passive-voice structures were converted into active voice structures. Words in all caps were converted into standard capitalization. There were no other differences between the two texts. A subset of the texts–one pair from each genre–was reviewed by a licensed attorney (in addition to the first author, who is also a licensed attorney) who was not affiliated with the project, so as to further ensure that the two versions had the same meaning.

For each contract pair, 12-15 comprehension questions were drafted. The questions were multiple choice with four options. These questions both targeted comprehension of specific important legal propositions, as well as more general understanding of the legal content. To reduce a response bias for a given register, we controlled the overlap in form between contract excerpt and comprehension question. Both types of comprehension question were drafted in a "neutral" register. Passive/active structures were replaced by nominalizations.

For example, "*shipment of the goods on the part of merchant*" instead of "*the goods were shipped by merchant*" or "*merchant shipped the goods*"). High or low frequency synonyms were replaced with a third synonym (e.g. "*renter*" instead of "*lessee*" or "*tenant*").

In addition to our main experimental materials, we administered the Author Recognition Task (ART; [78]) as a proxy for language experience.

All experiment code, data and analysis scripts is available on OSF: https://osf.io/xcqd9/?view_only=325a9567b2f54dc99eff8e8d5683e1bf.

Participants and Procedure

Based on a pilot study (see SI), we found that 100 participants would provide us sufficient power ($> 80\%$) to detect our main effect of recall. Due to concerns about data validity with online collection, we actively assessed data quality during the experiment. Participants first completed three trials and were only allowed to complete the experiment if their comprehension was above chance performance. In total we recruited 186 participants for the first half, but we only retain 108 participants who completed the entire experiment for our analysis. All participants self-identified as native English users.

Retained participants were psuedorandomly assigned to six trials (3 legalese; 3 simple). Participants did not see the same contract in both a simple and legal register. Assignment of stimuli to participant was pseudorandom to ensure that across participants every trial was administered with approximately the same frequency. The order of trials was randomized for each participant.

A trial consisted of (a) reading an excerpt, (b) a subset of the ART, (c) recalling the excerpt, and (d) answering comprehension questions. For the reading component, participants were presented with exactly one excerpt, written in either legalese or plain English. They were asked to carefully read the text twice, and were given as much time as needed to do so. For the ART component,

participants were given the names of 50 individuals and were asked to select which names corresponded to real authors. We expanded the ART task to 300 trials in order to keep the timing of a trial consistent. The original items from the published ART were presented first. For the remaining trials, the participants were administered novel items that looked virtually the same as authentic materials (half of the names corresponding to real authors, the other half corresponding to high-school track stars). We do not use these novel items in our analysis as they have not been validated [80][2]. After being shown the ART materials, participants were asked to recall as much of the excerpt they had read as possible. They were told that they could use their own words, but that their version should stay true to the original. Finally, each trial ended with the comprehension questions corresponding to the excerpt.

Analysis Plan

Two trained research assistants coded whether a proposition was successfully recalled (see SI for details). Coders were unaware of whether a participant had seen or recalled the simple or legalese version of a text. Twenty percent of the retellings were coded by both coders so as to assess inter-rater reliability using Cohen's kappa coefficient [81], [82]. For our regression analyses, we perform both a conservative analysis and an anti-conservative analysis, with regard to ties. Our results do not qualitatively change, so we only report the conservative analysis in text (see SI for anti-conservative analysis).

[2]Our results do not qualitatively change if we use the full itemset

2.3.2 Experimental Results

Comprehension

Figure 2.2i illustrates the comprehension accuracy across registers and Figure 2.2ii depicts comprehension accuracy as a function of ART score. Descriptively, participants were more accurate in the simple register (73.5%) than in the legalese register (67.7%).

We first conducted a mixed effect logistic regression, with register (sum coded), standardized ART score and their interaction as fixed effects and comprehension question, excerpt, and participant as random effects, each with a random slope for register. Using likelihood ratio test to compare to a model without the interaction term, we found no significant interaction between standardized ART score and register. Therefore, we report the results of the model fit without the interaction term. Replicating [41], we find a significant decrease in comprehension accuracy for a legalese register compared to a simple register ($\beta = -0.179$, SE$= 0.052, p < 0.05$). Note that for 94.5% of question items, mean accuracy was above chance (25%) in both versions. When removing items for which participants' overall comprehension in either version was below chance, we still find a main effect of register ($\beta = .165$, SE$= 0.049, p < 0.05$), indicating that the effect was not driven by items where participants systematically interpreted a different meaning in the simple register versus the legalese register.

While we did not find an interaction between language experience and register, we do find that participants with less language experience (lower ART scores) have worse comprehension accuracy than participants with more language experience ($\beta = 0.229$, SE$= 0.080, p < 0.05$).

Recall of Legal Propositions

Our two coders agreed on approximately 85% of overlapping judgments. Cohen's Kappa (unweighted) was measured to be 0.719 ($z = 47.1$; $p < 0.05$), indicating substantial agreement.

Figure 2.2iii displays the proportion of propositions recalled across registers. Overall, the average recall among participants was 41.1%, which is slightly better than recall rates for previous studies using text materials but a longer delay [83]. Descriptively, propositions from excerpts in a simple register (42.4%) were recalled more than propositions presented in a legalese register (35.3%).

As for comprehension, we first conducted a mixed effect logistic regression with register (sum coded), standardized ART score and their interaction as fixed effects and excerpt and participant as random effects with register as a random slope for each. Using likelihood ratio tests, we fail to find a significant interaction and, thus, report here a simpler model without the interaction term. Again, fewer legal propositions were recalled when they were presented in a legalese register compared to a simple register ($\beta = -0.17914$, SE$= 0.050, p < 0.05$). Unlike our comprehension results, we do not find an effect of language experience on recall.

Exploring the effect of linguistic structures

While surface properties of a text seem to be forgotten relatively quickly (e.g., within an hour; [84]) compared to propositional content (lasting weeks), it seems intuitive that they might appreciably influence memory for more abstract representations of content. If a reader can't understand or mis-parses the text, it's unlikely that they make the intended inferences and have a full grasp of the situation. Therefore, we expect linguistic structures known to incur processing difficulties to reduce the proportion of legal propositions recalled. Here, we focused on four kinds of structures purportedly common in a legal register

and manipulated in our materials: center-embedded clauses, passive voice, frequency of lexical choice and capitalization. As we wanted to keep the materials close to natural, and to keep the task short and feasibly annotatable, we do not have sufficient power to assess the generalizability of each structure's influence on recall. Instead, we provide a descriptive estimation of each structure's effect using Bayesian mixed effect logistic regression.

For every proposition, we included a main effect of condition and coded the interactions between condition and center-embedding, voice, word frequency or capitalization. We conducted a mixed effect logistic regression predicting recall with condition and surface form properties as a fixed effects and random intercepts for excerpt and participant with complete random slopes. Our fixed effect was coded so that each coefficient reflects either an increase or decrease in recall rate for a legalese register relative to the average recall rate of a simple register. Figure 2.2iv represents the 95% and 80% credible intervals over the regression coefficient for each surface property. We find the strongest effect of register for propositions that differed in center-embedding, a smaller effect for frequency, and no effect of capitalization and voice. In the SI, we find convergent results in a similar exploratory analysis of the comprehension data.

2.3.3 Replication Study

In real-world scenarios involving legal documents, laypeople may not always be forced to remember the content of a legal document without having it directly in front of them. People may also plausibly be more motivated to understand the content of a legal document if there are real-world financial or legal stakes that depend on that understanding. To test if either of these factors might mitigate the observed results in our main experiment, we conducted a replication experiment of our comprehension results, with two key deviations. First, instead of answering comprehension questions about a text without having the

text in front of them, participants were given a legal text along with all of the comprehension questions associated with that text and given as much time as desired to read the text and answer the comprehension questions. Second, participants were given an additional monetary incentive if they correctly answered at least 90% of comprehension checks correctly across each of their trials.

As with our original experiment, we find a main effect of register on comprehension, with participants scoring significantly lower on texts written in legalese as compared with texts written in a simple register ($\beta = -.2223$, SE$= 0.0690, p < 0.05$). We also find that this main effect holds when removing items with below chance (25%) accuracy ($\beta = .4568$, SE$= 0.1362, p < 0.05$). Full results reported in the SI.

2.4 Discussion

Our study aimed to better understand the reason why legal texts can be difficult to understand for laypeople by assessing to what extent: (a) difficult-to-process features that are reportedly common in contracts are in fact present in contracts relative to normal texts, and (b) such features–insofar as they are present–cause processing difficulties for laypeople of different reading levels. Here we discuss in turn the extent to which our results successfully answer these questions, as well as the implications of our results from both a scientific and policy perspective.

With regard to (a), our corpus analysis revealed that features such as center embedding, low-frequency jargon, passive voice and non-standard capitalization–all associated with processing difficulty–were more prevalent in contracts relative to all other texts genres that we looked at. In most cases, this difference was striking. Prior to our study, there had been long-standing speculation and anecdotal accounts of the presence of these features in legal texts, and more recent studies had to some degree identified the prevalence of passive voice [70] and

non-standard capitalization [47] in legal contracts, either on a smaller scale or with regard to specific types of contracts. Our study provides the first large-scale systematic account of the presence of all of these features in legal texts, both overall and relative to a baseline.

With regard to (b), our experimental study revealed that contracts drafted with all of these features were more difficult to both comprehend and recall than contracts drafted without all of these features, while our analyses of individual linguistic structures revealed that some of the features–such as center-embedding and low-frequency words–present greater difficulties in the context of recall than others, such as passive voice. Although language experience–as measured by ART–predicted comprehension performance, there was no correlation between ART and recall performance, nor was there a significant interaction between register and performance on ART in predicting recall or comprehension. Taken together, these results suggest that these features collectively present processing difficulties for readers of all levels of experience.

From a cognitive science perspective, our results provide insight into the long-puzzling issue of why contracts and other legal texts appear so difficult to understand for laypeople. Some legal theorists have taken the position that "law is a system built upon expert knowledge of technical concepts," such as *habeas corpus*, *promissory estoppel*, and *voir dire* [85]. As a result, the processing difficulty of legal texts is simply a natural result of not knowing specialized legal concepts. Others have argued that "law is a system built upon ordinary concepts," such as *cause*, *consent*, and *best interest* [85], [86]. In which case, processing difficulty could be explained by psycholinguistic factors.

Our findings better align with an *ordinary concepts* account of legal language. Previous work in the general psycholinguistics literature has suggested that center-embedded clauses are difficult to process due to the working memory constraints they impose on readers. Correspondingly, the fact that center-

embedded clauses were more than twice as prevalent per sentence in the contract corpus than in the standard-English corpus, and inhibited recall to a greater degree than other features in our experimental study suggests that the cause of the processing difficulty of legal texts may be largely related to working memory costs as opposed to a mere lack of understanding of specialized legal concepts.

Furthermore, if certain concepts are not known by those without expert legal training, then one would not expect to find many words to describe those concepts aside from the low-frequency jargon used by legal experts (just as there are no higher-frequency synonyms for terms such as *quark* or *electron* in physics, for example). Consequently, the fact that our corpus analysis revealed that contracts contained even more cases of words with high-frequency synonyms than standard English texts undercuts the view that processing difficulty is driven merely by lack of specialized knowledge. Although it is conceivable that specialized concepts contribute to the perceived processing difficulty of legal texts, our results suggest that insofar as low-frequency legal terminology presents processing difficulty for laypeople, this often results not from unfamiliarity with the concept underlying that terminology but with the terminology itself (such as the phrases *ab initio* and *ex post facto*, which in many cases respectively can be simplified to "from the start" and "after the fact").

From a policy perspective, these findings also provide insight into the long-standing issue of how to ease the processing difficulty of legal texts for laypeople. Efforts to simplify legal language over the last 50 years have focused largely on public legal documents, despite the fact that contracts and other private legal documents are more commonly encountered by laypeople–and increasingly so with the rise of the internet and online terms of service agreements. The fact that contracts contain a stunningly high proportion of features that incur processing difficulty in laypeople that can be feasibly replaced with easier-to-process alternatives underscores the importance for efforts to simplify legal language to

not neglect private legal documents. Moreover, the fact that certain features that are common to legal texts–such as center embedding and low-frequency words–appear to inhibit recall to a greater degree than others, such as passive voice, suggests that lawyers interested in simplifying legal texts for the benefit of readers ought to prioritize unpacking clauses into separate sentences and opting for higher frequency synonyms when possible.

The main effect of language experience on comprehension performance suggests that those with less language experience have a harder time understanding legal texts. Given that those with less reading experience as a group tend to be of lower socioeconomic status [87], [88], and those of lower SES face greater disenfranchisement from the legal system [89], this suggests that simplifying contracts may have non-trivial access to justice implications, particularly as their prevalence increases. At the same time, the fact that those with higher reading experience also struggled to comprehend and understand contracts written in legalese suggests that redrafting texts into a simpler register would have beneficial effects for those of all reading levels.

To better understand how to integrate these findings, we should aim to understand why lawyers choose to write in such an esoteric manner in the first place. One possibility is that legal language must be written so as to maintain communicative precision. This possibility is undercut by our results and previous findings that show comprehension of legal content with a simplified register (e.g., [41]). While it seems entirely plausible that certain legal jargon is inevitable, our results suggest that in many instances such jargon can be replaced with simpler alternatives that increase recall and comprehension while preserving meaning.

Another possibility is that lawyers choose to write in a complex manner to convey their priorities. For example, if a lawyer prioritizes the user's responsibilities they may focus on making them clear at the expense of other content

(e.g., company's obligations). If the lawyer's priorities differ from the reader's priorities they may even do this implicitly as opposed to engaging in an outright "conspiracy of gobbledegook" [67]

Lastly, lawyers may not *choose* to write in an esoteric manner. Similar to the "curse of knowledge" [61], [62], they may not realize that their language is too complicated for the average reader to understand [90]. This hypothesis appears to be supported by previous findings that show an effect of features such as prior knowledge and reading skill on the processing of specialized texts [91]–[95]. Similarly, one might predict that lawyers would be equally likely to comprehend contracts if they were drafted in an esoteric style as they would if they were drafted in a simpler register, which may render them less able to appreciate the difficulty of these features for those without legal training. Further work into the plausibility of these hypotheses could yield insight into how best to persuade lawyers to integrate the findings of our and similar studies and help alleviate the growing mismatch between the ubiquity and impenetrability of legal texts in the modern era.

Chapter 3

So Much for Plain Language: An Analysis of the Accessibility of United States Laws Over Time

This chapter is adapted from the following publication:

Eric Martínez, Frank Mollica & Edward Gibson, *So Much for Plain Language: An Analysis of the Accessibility of US Federal Laws Over Time*, 153 J. Exp. Psychol. Gen. 1153 (2024)

Abstract

This chapter builds upon the previous chapter by investigating to what extent, if at all, law's complexity has changed over time. Over the last 50 years, there have been efforts on behalf of the US government to simplify legal documents

for society at-large. However, prior to this study, there had been no systematic evaluation of how effective these efforts have been. Here we conduct a large-scale longitudinal corpus analysis (n 225 million words), comparing every law passed by congress between 1951 and 2022 with a matched sample of English texts from four different baseline genres. We also compared the entirety of the United States Code with a large sample of recently published texts from six baseline genres of English. The paper found that laws remain laden with features associated with psycholinguistic complexity relative to each of the baseline genres of English, and that the prevalence of these features has not meaningfully declined since the initial onset of the plain-language movement. These findings suggest top-down efforts to simplify legal texts have thus far remained largely ineffectual, despite the apparent tractability of these changes, and call into question the coherence and legitimacy of legal doctrines whose validity rests on the notion of laws being easily interpretable by laypeople.

3.1 Introduction

Ignorantia juris non excusat is an ancient maxim of the law which holds that "ignorance of the law is no excuse" [96]. This ancient maxim remains at the heart of modern legal systems, which typically presume that the public understands the entirety of the legal doctrine and, consequently, do not typically allow ignorance or mistakes of the law as a defense to a crime [97], [98]. Of course, the presumption that a nation's citizenry is aware of the content of its laws does not appear to be well-grounded in fact. While part of the public's ignorance of the law may be attributed to a mere lack of exposure, it seems intuitively obvious that when the public does attempt to understand legal documents they have difficulty doing so. Indeed, the difficulty of reading legal texts has long been acknowledged not just by those tasked with reading these documents but

by those creating these documents as well. Sporadic attempts to draw up laws in "simple language, using words that everyone could understand" date back as far back as the eighteenth century in Europe [99], but have mostly been ignored [100].

In the United States, top-down efforts to simplify government documents for the benefit of the public began as early as the 1970s, when Richard Nixon mandated that the Federal Registry be drafted in "layman's terms" and Jimmy Carter issued Executive Orders intended to make government regulations "easy-to-understand by those who were required to comply with them" [48], [49]. These and subsequent attempts to make government language more accessible have been collectively referred to as the "plain language movement." The most recent call-to-arms, the Plain Writing Act of 2010, established formal guidelines regarding how to write government documents clearly for a lay audience [52].

The plain language movement spurred research exploring how best to write "plain-English" layperson summaries of official legal documents, such as jury instructions [53]–[56] and Miranda warnings [57], [58]. Many of the insights from this literature, as well as the general psycholinguistic literature, are now reflected in the Federal Plain Language Guidelines. While these studies have successfully demonstrated the feasibility and importance of using "plain-English" layperson summaries of legal documents to improve comprehension of legal content among laypeople, these examples apply only to a small portion of the total corpus of legal language and appear less relevant to people's experience with the legal system than actual laws.[1]

With regard to official legal documents, recent work has found that private contracts, such as online terms of service agreements, remain laden with complex

[1]For example, although jury instructions can be an important part of cases that go to trial, a small and diminishing percentage of civil and criminal cases actually go to trial (as low as 3% for the former and 5% for the latter: [59], [60]). Moreover, while Miranda warnings provide crucial information to criminal suspects in police custody, the majority of individuals' contact with legal language takes place outside the context of criminal or civil suits.

psycholinguistic features, including center-embedding and low-frequency jargon [101]. Recent experimental work has also found that people are less able to understand and recall legal documents drafted with these features relative to legal documents of equivalent meaning drafted without these features [101], [102].

With respect to public legal documents, however, there remains no systematic analysis of to what extent the plain-language movement impacted the accessibility of federal laws. Moreover, on a more general level, there also remains no systematic evaluation of the accessibility of federal laws over time relative to baseline forms of English. In addition to comparing legal texts to forms of "standard" or "plain" English, such as newspaper articles or popular press books, comparing the accessibility of laws relative to more conceptually complex forms of writing, such as academic texts, might reveal the extent to which the inaccessibility of legal texts can be attributed to inherently complex concepts as opposed to needlessly complex psycholinguistic structures. Given that academics are also tasked with establishing and communicating complex ideas that are relevant to the general public, such a comparison could also provide useful insight regarding how well the academic community is successfully achieving that aim relative to lawmakers.

As alluded to above, the potential inaccessibility of official legal documents poses problems not just for those tasked with reading legal documents but for the validity of the documents themselves, as well as the coherence and legitimacy of legal doctrines that either expressly assert or implicitly assume that legal documents are or should be easily interpretable by laypeople.

For example, in United States constitutional law, the *Fair Notice Doctrine* requires "that laws give the person of ordinary intelligence a reasonable opportunity to know what is prohibited, so that he may act accordingly" [1], [3], [103], [104]. Insofar as laws are incomprehensible to the typical layperson, this would

arguably imply that laws are not giving laypeople fair notice, which would in turn undermine both the constitutionality of those laws and the legitimacy of the fair notice doctrine.

Meanwhile, the *Ordinary Meaning Doctrine*, which has been referred to as "the most fundamental principle of legal interpretation," not only of United States law but of jurisdictions across the world, requires that words in legal documents typically be interpreted according to how they are ordinarily understood by laypeople [2], [7], [10], [11], [15]. However, insofar as legal documents are not ordinarily understood by laypeople, the coherence and legitimacy of this doctrine would also be undermined.

To address the above questions, we first conducted a corpus analysis of (a) every law passed by congress between January 1951 and May 2022 (as well as concurrent resolutions not signed into law and proclamations issued by the president), and (b) a large sample of magazine articles, newspaper articles, nonfiction books and fiction books published over roughly the same time span. We analyzed a variety of linguistic and stylistic features, whose use is (a) discouraged by the Federal Plain Language Guidelines [49], (b) associated with language processing difficulty in psycholinguistics research [41], [101], [102], and (c) purportedly common in legal documents [75]. We found that most of these features had not meaningfully decreased in prevalence since the start of the plain-language movement, although we find an increase in the variability of some features post-2010. Nonetheless, compared to time-matched baseline texts, each of the features remain strikingly more prevalent in public legal documents.

We additionally conducted a comparison between (a) the entirety of the United States Code (an official compilation of every federal law currently in force); and (b) a broad sample of five baseline texts from the Corpus of Contemporary English. We found that even compared to academic articles, laws contained higher rates of every complex psycholinguistic feature we looked at,

suggesting that the inaccessibility of laws may be the result of needlessly complex linguistic structures as opposed to inherently complex concepts.

3.2 Materials and Methods

3.2.1 Corpus Materials

For our primary analysis we constructed an exhaustive corpus of every public law, private law, concurrent resolution and proclamation issued by the American federal government between the 1951 and May 2022 using publicly available online resources from the United States library of congress [105]. As a baseline, we extracted a comparably-sized sample of English texts drawn from the Corpus of Historical American English [106], which consisted of a broad sample of fiction books, non-fiction books, magazine articles, and newspaper articles also published between 1951 and 2009. Because the Corpus of Historical American English only extends to 2009, we did not directly compare laws from 2010 to 2022 with our baseline corpus.

In addition to our primary materials, we also collected two additional corpora to compare the linguistic complexity of current laws relative to a baseline of contemporary texts, including those of comparable conceptual complexity. These two additional corpora consisted of (a) the 2021 edition of the United States Code, the official compendium of all federal laws that are in effect in the United States; and (b) a comparably sized sample of academic texts, fiction books, magazine articles, newspaper articles, and spoken English drawn from the Corpus of Contemporary American English.[2]

[2]According to COCA documentation, the academic texts were drawn from more than 200 different peer-reviewed journals and cover the full range of academic disciplines, with a good balance among education, social sciences, history, humanities, law, medicine, philosophy/religion, science/technology, and business. The texts were published between 1990 and 2012.

3.2.2 Pre-Processing Tools and Indices

To process and analyze our corpora, we used a number of natural language processing tools. One of the primary tools we used was the Stanford Stanza natural language package [107], a state-of-the-art NLP toolkit which we used to tokenize each document into sentences, lemmatize and tag each word by part of speech, and syntactically parse each tokenized sentence. Stanza has been shown to achieve over 90% accuracy on a variety of NLP tasks [107]. To verify its accuracy on our specific corpora and for our specific metrics, we spot-checked a random sample of 1000 sentences across our corpora by (a) hand-coding whether a given sentence had a passive-voice structure or a center-embedded clause, and (b) for each sentence comparing whether the parser's judgments aligned with the hand-coded judgments. Using this method, we found that the parser was 97.93% accurate at detecting by-passive structures (95% CI: 97.04 to 98.82) and 88.95% accurate at detecting center-embedding structures (95% CI: 86.98 to 90.73).

In addition to Stanza, we also used the SUBTLEX word frequency dictionary [108], which we used to get a word frequency estimate as a proxy for how common a given word in each corpus appears in everyday speech. The SUBTLEX frequency values themselves are derived from a large-scale corpus of American film subtitles have been show to correlate with reading-time behavior [108]. Finally, we also used WordNet [109], which, in tandem with SUBTLEX, was used to estimate whether a given word could have been replaced by a higher frequency word with the same meaning.

Pre-processing for all corpora were identical. Sentences were first tokenized and dependency-parsed using the Stanford Stanza NLP package. We then removed sentences without punctuation, as well as those with fewer than 10 words so as to remove headings, which are not really sentences but would otherwise be counted as such. We also removed sentences with 3+ consecutive punctuation

marks or related symbols (such as '@') so as to get rid of more non-sentences in both corpora. The total number of words after filtering was 150,393,499 (47,769,955 words for the legal corpus and 102,623,544 for the non-legal corpus). After filtering out non-sentences, we then dependency-parsed each corpus, lemmatized and tagged each word by part of speech and computed our indices of processing difficulty, which we further clarify and motivate below.

Word frequency. For each of our corpora we sought to determine, on average, how frequently the words in said corpora occur in everyday speech. Words that are infrequently used in everyday speech cause comprehension difficulties for readers relative to higher-frequency synonyms [74]. Legal language is reportedly laden with low-frequency jargon, such as *aforesaid*, *hereinafter*, and *to wit* [45], and recent work has shown the language in contracts to be lower-frequency than that of other genres of English [101]. According to the official plain-language guidelines, government writing should avoid the use of such low-frequency "dry legalisms" and "jargon" [49].

Frequency values were extracted from the SUBTLEX corpus of American film subtitles [108], commonly used as a proxy for standard-English word frequency and which has been shown to correlate with reading time behavior. Because the impact of word frequency on reading times is logarithmic as opposed to linear, we used the Zipf values (which are both logarithmic and standardized) as opposed to raw counts [110].[3]

To avoid including non-content words, we limited our analysis of frequency to the words in our corpora marked as a verb, noun, adjective or adverb according to Stanza. Proper nouns and other words that did not appear in the SUBTLEX corpus received a score of `NA`.

Word choice. Although many argue that the processing difficulty of unfamiliar language is a necessary consequence of the specialized concepts and

[3]The Zipf values we used can be found here: https://osf.io/djpqz/files/osfstorage

corresponding terminology used to refer to those concepts by lawyers (cf. Tobia, 2020), recent work suggests that private legal documents contain a high-proportion of overly complicated language that can be replaced with simpler terms that have the same meaning [101]. The official plain-language guidelines encourage the use of "familiar or commonly used'' words over such "unusual,'' "obscure'' or "unnecessarily complicated language" [49]. Here we sought to quantify the amount of unnecessarily complicated language in federal laws by calculating the percentage of words in each corpus that could have been replaced with a higher-frequency synonym.

We operationalize word choice difficulty as the proportion of content words in each corpus that had a higher-frequency synonym. We conducted three versions of this analysis using three separate assumptions. Under the first version, we make the conservative assumption that the authors intended the least common sense of each word used in a given corpus because while legal terms may resemble common words in form, they may have a more specialized meaning, such as the concept of "consideration" in contract law [73]. Under the second version, we make an anti-conservative assumption that the authors intended the most common sense of each word in a given corpus. Under the third version, we assume that the authors intended neither the most common nor least common sense of each word but rather a random sense of a word in a given corpus. Again, we limit our analysis to verbs, common nouns, adjectives and adverbs.

For all three methods, we determined the least common (conservative), most common (anti-conservative 1) or random (anti-conservative 2) meaning/sense of that word according to WordNet [109]. For all words sharing that meaning/sense (i.e., synonyms), we looked up the SUBTLEX frequency value and coded whether the SUBTLEX frequency value of any synonym was higher than that of the actual word used in the text (1=Yes; 0=No). Results of the conservative method are reported in the main text. The results of the two anti-

conservative versions are visualized in the SI, along with three additional versions that extend the analysis to all words as opposed to content words. All six versions yielded converging results.

Capitalization. In each corpus we computed the percentage of words that contained non-standard capitalization (specifically, those that were in ALL CAPS). Although the plain-language guidelines do not discourage the use of all-capitalization in government writing, evidence suggests that non-standard capitalization ("*ALL WARRANTIES ARE HEREBY DISCLAIMED*") is common in certain types of private legal documents [101] and has shown to inhibit comprehension in older readers [47], relative to standard capitalization.

We coded a word as being in "all-caps" by calculating the proportion of alphabetic word tokens that were marked by Stanza as being entirely in uppercase letters.

Sentence length. Plain-language guidelines encourage the use of shorter sentences so as to "break the information up into smaller, easier-to-process units'' [49]. Legal texts, especially laws and other public documents, are reportedly filled with long sentences [111], [112]. Although some evidence suggests sentence length is less of a predictor of processing difficulty than center-embedding and other types of syntactic complexity [113], words per sentence remains a consistent measure of processing difficulty in the reading literature [114], [115].

Here we computed sentence length by calculating the number of alphabetic words in each sentence as determined by Stanza.

Center-embedded clauses. Plain-language guidelines discourage the use of "convoluted" sentences, particularly those that are "loaded with dependent clauses'' and which separate the "essential parts'' of a sentence from each other (i.e. the subject, verb and object). The most notorious examples of such sentences contain center-embedded structures, in which a sentence or clause is em-

bedded within the center of another sentence or clause (*"all such payments and benefits, including the payments and benefits under Section 3(a) hereof, being hereinafter referred to as the 'Total Payments' "*). Center-embedded structures cause processing difficulty for readers [43], [44] and have been shown to inhibit recall of legal content relative to clauses of equivalent meaning that have been un-embedded into separate sentences [101], [102]. Here we calculated the percentage of sentences in each corpus containing a center-embedded clause.

We coded a sentence as containing a center-embedded clause if a predicate dependent clause as parsed by Stanza (i.e. clausal subjects, clausal complements, open clausal complements, adjectival clauses, and adverbial clauses) was followed by a word as opposed to an end-of-sentence punctuation mark. As noted above, this method was 88.95% accurate in identifying these structures in a given sentence (95% CI: 86.98 to 90.73).

Passive-voice structures. For each corpus we calculated the percentage of sentences containing a reversible passive-voice structure. Federal Plain Language Guidelines advocate for using the active voice instead of the passive voice. Passive-voice structures are acquired later than active-voice structures and have been shown to pose comprehension difficulties for adults in certain circumstances, particularly in the context of implausible sentences *e.g. "the girl was kicked by the ball"* [46]. Although [101] recently found evidence that passive voice structures did not inhibit recall of legal content relative to active-voice structures in contracts, it may be that the stimuli used in [101] did not span the circumstances shown to induce the comprehension errors seen in adult experiments. To err with caution, we include passive voice structures in our analysis, particularly reversible passives or *by*-passives (e.g. *"the information shall be maintained by the Federal Government"* as opposed to *"the information shall be maintained"*), which can be more easily replaced by active-voice structures without a loss or distortion in meaning.

We coded a sentence as containing a reversible passive voice structure if a word was marked with the passive voice features by Stanza and had the word *by* in the same head according to the Stanza parse. As noted above, this method was 97.93% accurate in identifying by-passive structures in a given sentence (95% CI: 97.04 to 98.82).

For robustness purposes, we also computed the percentage of sentences containing non-reversible passives. These results are reported in the SI.

3.2.3 Analysis Plan

To evaluate the influence of the plain-language movement, for each of our six indices of processing difficulty we conducted both a classical and a break-point Bayesian regression limited to the legal-corpus data. Whereas a classical regression estimates a single slope for predictors across their range, a break-point regression assumes there is a fixed "break" in the range of a predictor (in this case time) and estimates two sets of slopes: one set for before the break-point and one set for after the break-point. For word frequency and sentence length, we used a linear regression to predict the mean value of these metrics per sentence and report standard deviation as a measure of variance. For all other indices, we used a beta-binomial logistic regression, which estimates an over-dispersion parameter, as a measure of variance, in addition to fixed effects. In the case of our sentence-level metrics (center-embedding and passive voice), the regression estimated the influence of our predictors on whether a sentence had a given metric. In the case of our word-level metrics (capitalization and word choice), the regression estimated the influence of our predictors on whether a word had a given metric. For all break-point regression models, we used normal priors, with a mean of 0 and a standard deviation of 3 (primarily for increased computational speed). Breakpoint models were compared with a baseline model containing a fixed effect of time—i.e., one set of slopes, using Bayes Factors. All

analyses were conducted using the brms package in R [116].

To more generally evaluate the accessibility of federal laws over time and as compared to plain English, we conducted separate Bayesian regression models (linear for frequency and sentence length; logistic for remaining indices) that included both of our corpora. For each index, we first considered two models: one with a main effect of Corpus (Legal vs baseline) and Year, and one with an additional interaction term between Corpus and Year. We used the default priors in brms and stan, which are flat priors. A Bayes-factor comparison for each index except average word frequency suggested the model with an interaction is a better explanation of the data (BF>10) than the model without an interaction. For word frequency, both models perform equally well. We therefore only report the results of the model with an interaction term in Table 3.3.

3.2.4 Transparency and Openness

The methods of this paper comply with the TOP guidelines of the Journal of Experimental Psychology: General. In particular, all original data and code for this project can be cited as [117] and is available at the link https://osf.io/ambp4/?view_only=b4ab367e4cfb4f83acd2c51a000cfa68.

3.3 Results

3.3.1 Efficacy of the Plain Language Movement

Were the plain-language movement to have been effective, one would expect (a) the prevalence of difficult-to-process features to have meaningfully decreased over time, and (b) the decrease to coincide with the onset of the plain-language movement. To evaluate this prediction, for each of our six indices of processing difficulty we conducted classical and break-point Bayesian regressions limited to

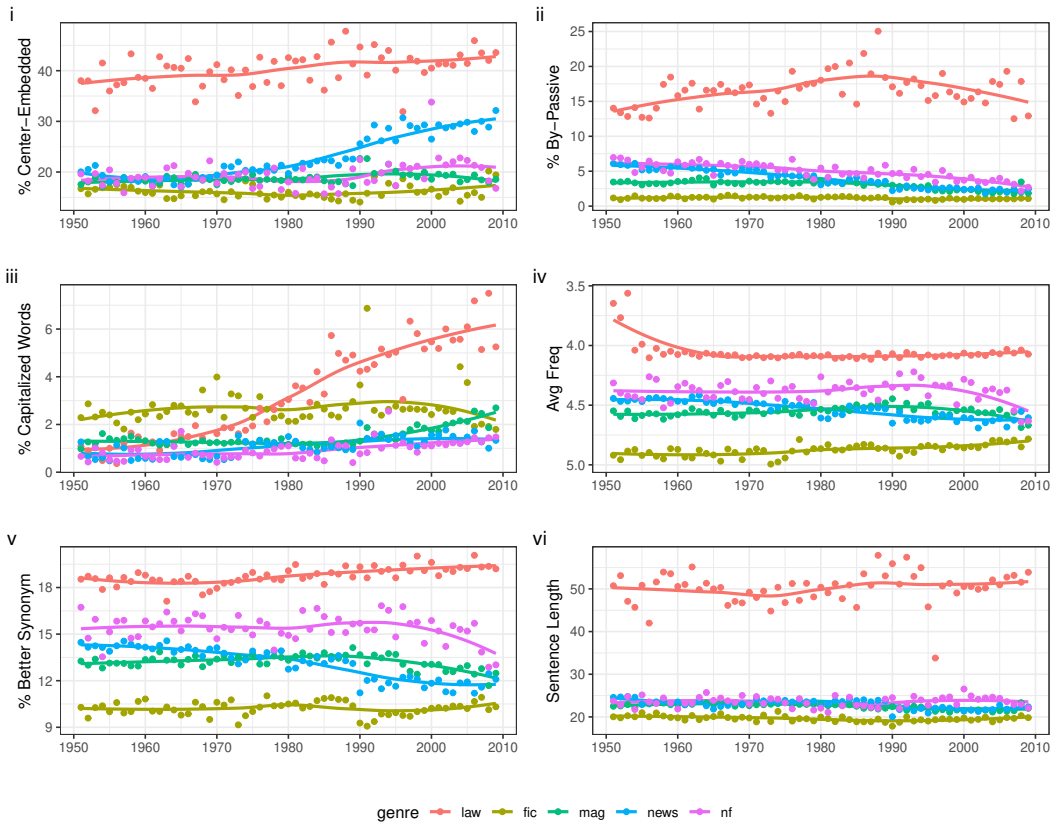

Figure 3.1: Comparison of indices of linguistic processing difficulty in federal laws vs four genres of English, including fiction books, magazine articles, newspaper articles, and non-fiction books (1951-2009). For any given year, most, if not all texts indices were vastly more prevalent in laws than any of the baseline genres. Individual points reflect mean values of an index within a genre. Lines reflect LOESS regression lines capturing the year-by-year trend of the prevalence of an index within each genre. Baseline texts were taken from the Corpus of Historical American English.

the legal corpus. We used two break-points: 1972, a plausible year for the plain language movement's call-to-arms, and 2010, the year of the passage of the Plain Language Act. If the plain-language movement overall had a simplifying effect, one would expect the slope of the regression line after the 1972 breakpoint (i.e., 1972-2022) to be both negative (positive for word frequency) and less (greater for word frequency) than the slope of the regression line before the breakpoint (i.e. 1951-1972). If the Plain Writing Act of 2010 led to a decrease in features, one would expect the slope of regression line for after the 2010 breakpoint (i.e., 2010-2022) to be both negative and less (positive/greater for word frequency) than the slope of the regression line before the breakpoint (i.e. 1951-2010). Regression coefficients for all indices can be found in Table 3.1 for 1972 and Table 3.2 for 2010.

For both time points, a classical regression model—i.e, a single linear effect of time, explains the data better than a breakpoint effect (BFs>10) for five out of six features: center-embedding, by-passives, capitalization, word choice, and word frequency, suggesting no significant impact of plain language movements on decreasing the prevalence of these features. For the remaining feature (sentence length), a classical regression model also explains the data better than a breakpoint effect for 1972 but not for 2010. That said, it should be noted that the effect size, while significant, is negligibly small and likely driven by the large amount of data and higher variability in values on the sparser side of the breakpoint (i.e. after 2010).[4] Visually, the single slope linear effect of time is clear even from the posterior draws of the breakpoint model. Due to the increased variance in the 2010 models, the posterior draws appear as if there might be

[4]As a follow up, we conducted heteroskedastic break-point regressions—i.e., we fit the variance before and after the breakpoint separately, for 2010. This analysis provides a little evidence (1<BFs<5) to suggest that the data are more variable post 2010. This extra variance is partly due to the fewer timepoints sampled post 2010. Regardless, models with a linear effect of time are still better explanations of the data than the heteroskedastic break-point regressions.

differences in slope across the break; however, this is less clear looking at the data points themselves. Further, for frequency, by-passive and word choice, the slopes post the 2010 breakpoint point are trending towards worse language processing outcomes contra the plain language movement. For capitalization, the slope post the 2010 breakpoint is trending towards better language processing outcomes, which while not statistically significant, suggests improvement.

	Intercept	Before 1972	After 1972	Dispersion	BF
Word Frequency	4.614 [4.613, 4.615]	-0.002 [-0.002, -0.002]	-0.002 [-0.002, -0.002]	0.582 [0.582, 0.582]	Inf
Word Choice	-1.58 [-1.63, -1.54]	-0.00 [-0.01, 0.00]	0.00 [-0.00, 0.00]	200.55 [165.56, 240.29]	5440
Capitalization	-3.60 [-3.74, -3.46]	0.04 [0.02, 0.05]	0.02 [0.01, 0.02]	83.20 [68.13, 99.59]	311
Center-embedding	-0.75 [-0.89, -0.62]	0.01 [-0.01, 0.02]	0.00 [-0.00, 0.01]	14.46 [12.03, 17.15]	2576
Sentence Length	50.92 [50.71, 51.13]	0.7 [0.04, 0.09]	-0.01 [-0.02, -0.00]	49.63 [49.56, 49.70]	Inf
Passive Voice	-1.93 [-2.09, -1.78]	0.02 [0.00, 0.03]	-0.00 [-0.01, 0.00]	21.05 [17.26, 25.46]	403

Table 3.1: Estimates and 95% confidence intervals for the intercept and slopes of the breakpoint regression models at 1972. Bayes Factors reflect the evidence for a linear trend across years over a non-linear (breakpoint) model.

	Intercept	Before 2010	After 2010	Dispersion	BF
Word Frequency	4.562 [4.561, 4.563]	-0.001 [-0.002, -0.001]	-0.058 [-0.058, -0.057]	0.580 [0.580, 0.580]	Inf
Word Choice	-1.59 [-1.64, -1.54]	-0.00 [-0.00, 0.00]	0.02 [0.01, 0.03]	207.06 [170.81, 247.55]	86
Capitalization	-2.86 [-2.99, -2.73]	0.02 [0.02, 0.03]	-0.04 [-0.10, 0.00]	84.19 [68.97, 101.28]	22
Center-embedding	-0.61 [-0.75, -0.47]	0.00 [-0.00, 0.01]	0.02 [-0.03, 0.06]	14.56 [12.04, 17.32]	847
Sentence Length	51.49 [51.30, 51.68]	0.03 [0.02, 0.04]	-0.68 [-0.75, -0.61]	49.62 [49.55, 49.69]	≈ 0
Passive Voice	-1.96 [-2.12, -1.80]	0.00 [-0.00, 0.01]	0.02 [-0.04, .06]	20.80 [17.12, 24.89]	712

Table 3.2: Estimates and 95% confidence intervals for the intercept and slopes of the breakpoint regression models at 2010. Bayes Factors reflect the evidence for a linear trend across years over a non-linear (breakpoint) model.

3.3.2 General Trends in Accessibility of Legal and Non-Legal Language

Even if plain language efforts have not coincided with a decrease in difficulty-inducing structures in legal texts, it may be the case that (a) difficulty-inducing structures became more prevalent in other texts relative to or as well as legal language, or that (b) legal language was not filled with very high indices of difficulty-inducing structures to begin with. To evaluate these alternative accounts, as well as to obtain a more general systematic account of the accessibility of federal laws–both temporally and relative to other genres of English–we

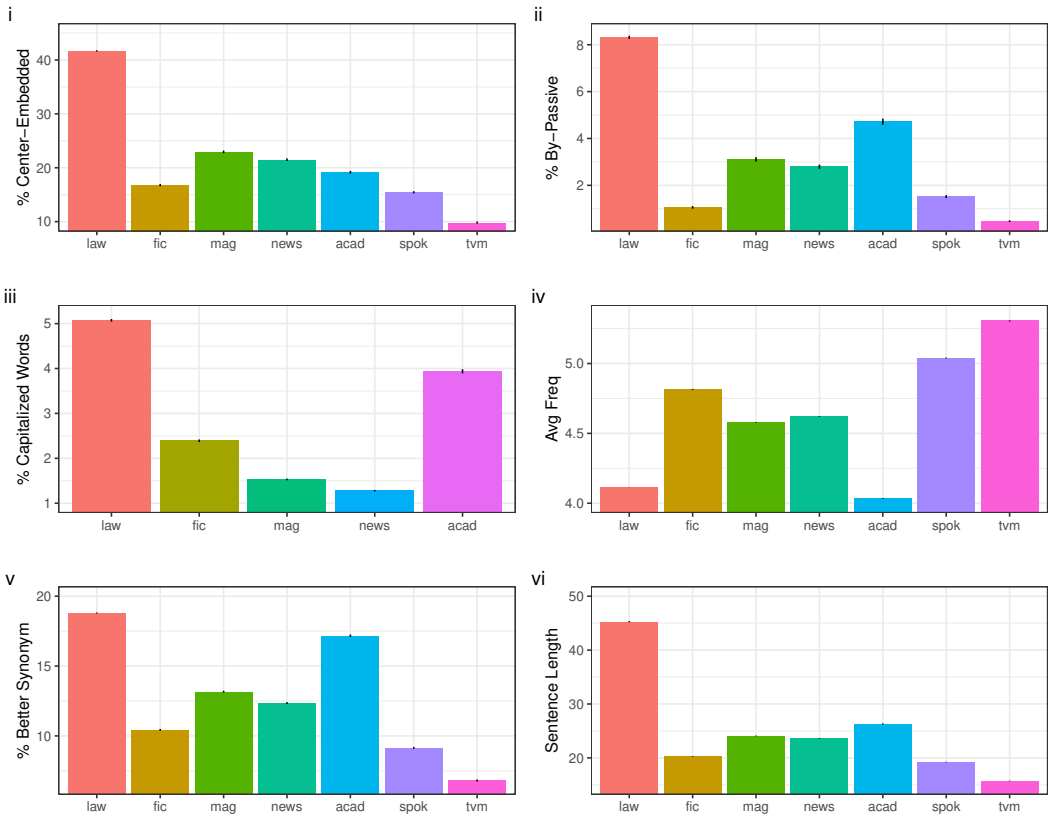

Figure 3.2: Comparison of indices of linguistic processing difficulty in contemporary federal laws vs five genres of English, including academic articles, fiction books, magazine articles, newspaper articles, and spoken language transcripts. Federal laws were taken from the 2021 edition of the United States Code. Baseline texts were taken from the Corpus of Contemporary American English. Height of bars reflects mean of index within a given genre, whereas error bars reflect 95% bootstrapped confidence intervals of the mean. With one exception (word frequency in academic texts), indices remain more prevalent in laws than any of the baseline genres,

first computed the descriptive statistics of each of index within the corpora over time.[5] We found that for each year, the prevalence of virtually every metric was higher in federal laws than in any of the four genres of the plain-language corpus (in most cases, the difference was striking). These results are visualized

[5]Because the Corpus of Historical American English only extends to 2009, we did not directly compare laws from 2010 to 2022 with our baseline corpus.

	Intercept	Corpus	Year	Corpus:Year	BF
Word Freq.	4.69	-0.60	-0.00	0.00	1.28
	[4.69-4.70]	[-0.60- -0.60]	[-0.00- -0.00]	[0.00-0.00]	
Word Choice	-1.72	-0.24	0.00	-0.00	Inf
	[-1.7- -1.72]	[-0.25-0.24]	[0.00-0.00]	[-0.00- -0.00]	
ALL CAPS	-3.88	-0.13	0.02	-0.01	Inf
	[-3.88 - -3.88]	[-0.13- -0.13]	[0.02-0.02]	[-0.01- -0.01]	
Embedding	-0.97	-0.57	0.00	-0.00	12.13
	[-0.97- -0.97]	[-0.57- -0.56]	[0.00- 0.00]	[-0.00- -0.00]	
Sent. Length	35.99	-14.73	.02	-0.04	Inf
	[35.60-36.30]	[-15.03- -14.34]	[-0.00-0.04]	[-0.06- -0.02]	
Passive Voice	-2.61	-1.00	-0.00	-0.01	Inf
	[-2.61 - -2.60]	[-1.01- -1.00]	[-0.01 –0.00]	[-0.01- -0.01]	

Table 3.3: Estimates and 95% confidence intervals for the intercept and slopes of the Bayesian regression models, as well as the Bayes Factor (BF) estimates in favor of these models over models without an interaction term.

in Figure 5.1.

We then used Bayesian regression methods to estimate the influence of corpus (legal vs baseline) over time (in years) for each of our indices of processing difficulty (results in Table 3.3). For every metric, our models revealed federal laws to contain more difficult to process structures than our baseline texts. While the credible intervals for our estimates of the main effect and interaction with time do not include zero, the parameters reflect very small effects (e.g., 20 years for change of 1%). If anything the tightness of the credible intervals is likely over-estimated due to the large size of the corpora. Therefore, we interpret the results to suggest no meaningful influence of time on the prevalence of a given metric, nor of the interaction between time and corpus.

3.3.3 Accessibility of Contemporary Legal vs Baseline Texts

Even if plain-language efforts failed to reduce the prevalence of complex psycholinguistic features in legal texts to the level of those in everyday text and speech, it is conceivable that this failure is a natural result of the higher con-

ceptual complexity of legal texts relative to other texts. If so, one would predict that texts of similar conceptual complexity (such as academic articles) would have the same rate of psycholinguistically complex features as legal texts.

In order to account for this possibility, as well as to more generally compare laws to contemporary baseline texts, we conducted an additional comparison between (a) all United States federal laws in force as of 2021 [118] and (b) a comparably sized sample of academic texts, fiction books, magazine articles, newspaper articles, TV/movie scripts, and spoken language transcripts published in 2019 [119].

For all baseline genres except academic texts, each index of processing difficulty was disproportionately common in legal texts relative to the baseline texts. For academic texts, each index of processing difficulty except one (word frequency) was disproportionately more common in legal texts. In most cases, the difference was striking. Full results reported in Figure 3.2.

3.4 Discussion

The present study first set out to investigate whether the plain-language movement succeeded in reducing certain features (a) that are associated with psycholinguistic complexity, (b) whose use is discouraged by plain-language advocates, and (c) that have been attested to be common in legal documents. According to our regression models, the slope of the line after 1972 suggested no change or harmful change, indicating that laws on balance had not gotten meaningfully simpler by our metrics since the onset of the plain-language movement. With regard to 2010, most (but not all) of our regression models did not reveal a positive change, indicating that the Plain Language Act of 2010 may have induced some modest improvements but did not coincide with a meaningful reduction of most of these features, either.

To further contextualize these findings, the present study next sought out to investigate to what extent federal laws have deviated from baseline texts with respect to the presence of these features, both (a) over time between 1951 and 2009; and (b) at present. With regard to (a), as visualized and documented above, all of the metrics we looked at were startlingly more prevalent in federal laws than each of our baseline texts, with the relative prevalence failing to decrease over the examined time interval. With regard to (b), with one exception, all of the features we looked at were startlingly more prevalent in the United States Code than each of our baseline contemporary texts. Insofar as these features are accurate proxies for processing difficulty, then, in line with common intuition and plain-language advocates and consistent with recent findings regarding private legal documents [101], this suggests that United States laws have been and continue to be more difficult to understand than other genres of English, including documents of comparable conceptual complexity, such as academic texts.

Our study provides the first systematic large-scale account of the accessibility of public legal language–both longitudinally and compared to more standard forms of English–substantiating previous anecdotal accounts of the efficacy of plain-language efforts made by plain-language advocates, who have described progress as "way slow" and acknowledged that "much remains to be done to improve" [49].

Having documented the profile of public legal language over the last 70 years and demonstrated the inefficacy of plain-language efforts over the same time period, further extensions to this study–both with regard to academic scholarship and government advocacy–should seek to confirm the extent to which these findings hold for other types of government documents, such as federal regulations and informational pages on government websites. For example, it may be the case that the plain-language movement led to a simplification not of laws them-

selves, but of supplemental supporting documents that provided a layperson's explanation of the content contained in those laws.

In addition, future work could also seek to understand the cause of the complexity of legal language. In other words, not only how lawyers and lawmakers write but why they choose to write they way that they do. One possibility is that the style in which laws are currently written is necessary to maintain communicative precision. Prior to our study, this hypothesis had been undercut by previous findings showing comprehension of legal content with a simplified register [41], [101], [120]. Our results further undercut this possibility, as our analysis focused on features that are known to have simpler alternatives with equivalent meaning (e.g. "mala fides" versus "bad faith"). While it seems entirely plausible that certain legal jargon is inevitable, our results suggest that in many instances such jargon can be replaced with simpler alternatives that preserve meaning.

Moreover, to the extent that legal jargon is inevitable, the inaccessibility of legal language would still be problematic even according to the law's own aims, as much of legal doctrine either assumes or requires that laws be accessible to the typical layperson. For example, in United States constitutional law, the fair notice doctrine requires "that laws give the person of ordinary intelligence a reasonable opportunity to know what is prohibited, so that he may act accordingly" [1], [3], [103], [104]. Insofar as laws are incomprehensible to the typical layperson, this would arguably imply that laws are not giving laypeople fair notice, which would in turn undermine both the constitutionality of those laws and the legitimacy of the fair notice doctrine.

Meanwhile, the ordinary meaning doctrine, which has been referred to as "the most fundamental principle of legal interpretation," not only of United States law but of jurisdictions across the world, requires that words in legal documents typically be interpreted according to how they are ordinarily under-

stood by laypeople [2], [7], [10], [11], [15]. However, insofar as legal documents are not ordinarily understood by laypeople, the coherence and legitimacy of this doctrine would also be undermined.

Aside from the inevitability of legal jargon, another possibility for why lawmakers write the way that they do is that esoteric text arises out a mismatch between the priorities of the writer and reader of a law. If lawmakers' priorities differ from the reader's priorities they may even do this implicitly as opposed to engaging in an outright "conspiracy of gobbledegook" [67]. This possibility seems to have been undercut by recent findings indicating that lawyers, like laypeople, disprefer complicated legalese to simplified legal language when tasked with reading and evaluating legal documents [102].

Another alternative, similar to what has been dubbed the "curse of knowledge" [61], [62], is that lawyers may not realize that their language is too complicated for the average reader to understand [90]. Although this hypothesis appears to be supported by previous findings that show an effect of features such as prior knowledge and reading skill on the processing of specialized texts [91]–[95], recent evidence in legal contexts has undercut this hypothesis. In particular, lawyer subjects in [102], like laypeople, were found to struggle to understand convoluted legal documents, and were not found to be disproportionately better at understanding convoluted legal documents relative to simplified legal documents compared to laypeople, nor were they found to underestimate the difficulty of convoluted legal documents relative to simplified legal documents.

An additional possibility is that legalese is a result of an iterative drafting process, in which conditions are often thought of after the creation of an initial draft and are more easily embedded within the center of existing sentences as opposed to separated out into a subsequent sentence. If so, this would predict that the complexity of legal language could be alleviated by thinking through the conceptual complexity of a legal document prior to writing as opposed to

copying and iteratively editing documents over time.

A final possibility is that lawyers and lawmakers write in a convoluted manner in order to lend official legal documents a ritualistic, spell-like element of authority (cf. Tiersma, 2008; Hart, 2012). If true, this could explain why the plain-language movement might have succeeded in spurring efforts to create unofficial descriptions of laws but not in the simplification of official legal documents such as legislation.

Further work into the plausibility of these hypotheses could yield insight into how best to persuade lawmakers to integrate the findings of our and similar studies and help alleviate the mismatch between the ubiquity and impenetrability of legal texts in the modern era.

3.4.1 Acknowledgements

Special thanks to George Nathaniel for helping collect materials. Special thanks to Kinan Martin, Aalok Sathe, Colton Casto, Evelina Fedorenko and Greta Tuckute for help and feedback on corpus analysis pipeline.

3.4.2 Constraints on Generality

The research question of this study related to the accessibility of the federal laws of the United States of America (a) over time since the onset of the plain-language movement, and (b) relative to other texts applicable to the general population of the United States of America. The legal materials that we used were an exhaustive set of (a) all federal laws passed by Congress since before the onset of the plain-language movement and May 2022, and (b) all federal laws currently in effect as of 2021. Therefore, we can be confident that our results generalize to the target set of legal documents identified by our research question. Our results also converge with recent findings of the same complex features in other types of legal documents relative to non-legal documents.

Although the complexity of legalese has been attested in other countries beyond the United States, and although other countries have had similar plain-language efforts, it is unclear whether our results would generalize to laws of other countries and other languages. It is also unclear to what extent these findings generalize to layperson summaries of legal documents within and beyond the United States.

With regard to our non-legal materials, our sample included a large and wide-ranging set of baseline genres of English that varied in their intended audience and formality. We therefore expect that our findings would hold were we to compare legal texts with other baseline genres according to our metrics.

The metrics we looked at are generally considered by plain-language advocates, as well as within the psycholinguistics and reading literatures, as valid proxies for accessibility, and the tools we used to measure those metrics have been validated as accurate beyond the present study. We can therefore be confident that our analyses reliably assessed the efficacy of the plain-language movement according to its own aims.

That said, it is possible that there are some indices of processing difficulty that we missed. There may be other ways in which laws are more complex than non-laws, and there may be some ways in which laws are less complex than non-laws.

Similarly, it is unclear to what extent the psycholinguistic complexity of laws can be dissociated with their conceptual complexity. Previous studies we ran have found that both lawyers and non-lawyers recall and understand more content in legal documents drafted without these features compared to legal documents of equivalent meaning drafted with these features [101], [120]. Similarly, in the present study we found that laws had higher indices of complex psycholinguistic features than texts of plausibly similar levels of conceptual complexity. However, it is plausible that some degree of psycholinguistic complexity in le-

gal texts is a result of conceptual complexity, and it is unclear to what degree complex psycholinguistic features in legal documents can be removed without leading to a loss or distortion of meaning.

Chapter 4

Even Lawyers Don't Like Legalese

This chapter is adapted from the following publication:

Eric Martínez, Frank Mollica & Edward Gibson, *Even Lawyers Do Not Like Legalese*, 120 Proc. Nat'l Acad. Sci. U.S. 1 (2023)

It is reproduced here with slight modifications.

Abstract

This chapter builds upon the previous two by investigating why lawyers tend to write in such a convoluted manner. Across two preregistered experiments, the paper evaluates five hypotheses proposed by scholars and commentators for why lawyers write in a complex manner. Experiment 1 revealed that lawyers, like laypeople, were less able to recall and comprehend legal content drafted in a complex "legalese" register than content of equivalent meaning drafted in a simplified register. In Experiment 2, lawyers rated simplified contracts as equally enforceable as legalese contracts, and rated simplified contracts as

preferable to legalese contracts on several dimensions–including overall quality, appropriateness of style, and likelihood of being signed by a client. These results suggest that lawyers who write in a convoluted manner do so as a matter of convenience and tradition as opposed to outright preference and that simplifying legal documents would be beneficial for lawyers and nonlawyers alike.

4.1 Introduction

There is a burgeoning psycholinguistics literature documenting the various domains in which efficiency shapes human language, such that successful communication can be achieved with minimal effort on average by the sender and receiver [23]–[34]. Two ways in which this efficiency manifests itself relate to word length and syntax. For example, words that are more frequent (such as "the") tend to be shorter than less frequent words (such as "accordion"), such that utterances tend not to be longer than necessary given one's communicative aims [35]. With regard to syntax, it has been observed across languages that words that depend on each other tend to be close together in an utterance [36], so as to (by hypothesis) avoid overloading working memory capacity when interpreting an utterance.

However, one domain in which this efficiency has been attested to not apply is in the context of the legal system, as the language in contracts, statutes, and other legal documents is often observed to be notoriously inaccessible to a typical layperson, such that legal content seems to not be understood by a listener with minimal effort (e.g. [41], [53]–[58], [64]). Recent empirical work has supported the longstanding anecdotal observation/intuition that legal language is complex. For example, on a syntactic level, the language in contracts [101] and legislation [121] has been found to be laden with center-embedded clauses (leading to long-distance syntactic dependencies) at a rate several times higher

than standard English texts, including academic articles and other texts aimed at an educated audience.

Meanwhile, on a word level, legal documents have also been found to be laden with words that are infrequently used in everyday speech. Previous research had long identified center-embedding [43], [122] and word frequency [74] to be reliable proxies for processing difficulty in normal texts. Recent work confirmed this to be true in legal documents, also, as contracts drafted with these features were recalled and comprehended at a lower rate than legal documents of equivalent meaning drafted without these features (and center-embedding in particular was found to inhibit recall to a greater degree than word frequency) [101].

While the above studies have shed insight into the question of how legal language is complicated to understand, it remains an open question why legal language is so complicated to understand–that is, why do lawyers write in such a convoluted manner in the first place? Answering this question is relevant not only to major questions in psycholinguistics, but to legal doctrine and public policy, as well.

Across modern civilization, societal norms and rules are established and communicated largely in the form of written laws. Because law is encoded in the form of natural language, it follows that an understanding of language is crucial to drafting, interpreting, and enforcing the rules and standards that comprise legal doctrine and underpin modern society. In particular, understanding why lawyers and lawmakers write in such a convoluted manner can help inform policy efforts to make laws more accessible–which have been advocated for decades [48], [49], [52], with little to no success [121]. Such efforts are crucial to ensuring the comprehension and compliance of societal norms, as well as upholding the legitimacy of legal doctrines that either expressly assert or implicitly assume that legal documents are or ought to be easily interpretable to laypeople, such

as *Ordinary Meaning* [7], [15], [86] and *Fair Notice* [1].

Here we conducted two well-powered, pre-registered experiments aimed at evaluating five hypotheses presented in the theoretical literature for why lawyers write the way that they do.[1] In Experiment 1, we found that lawyers, like laypeople, were less able to recall and comprehend legal content drafted in a complex "legalese" register than content of equivalent meaning drafted in a simplified register. In Experiment 2, we found that lawyers rated simplified contracts as equally enforceable as legalese contracts, and rated simplified contracts as preferable to legalese contracts on several dimensions–including overall quality, appropriateness of style, and likelihood of being signed by a client. These results suggest that lawyers who write in a convoluted manner do so as a matter of convenience and tradition as opposed to an outright preference, and that simplifying legal documents would be beneficial for lawyers and non-lawyers alike.

4.2 Hypotheses

In previous literature, scholars proposed several hypotheses for why lawyers write in a complicated manner. Here we briefly present each of these hypotheses in turn, as well as the associated predictions of these hypotheses that we pre-registered for our experiments.

Curse of Knowledge Hypothesis. Some scholars have speculated, in line with what has been dubbed the "curse of knowledge" in other disciplines [61], [62], that the reason legal language is so difficult to understand is because lawyers do not realize that they write in an esoteric manner [63]. If this were true, one would predict that lawyers would not show the same degree of difficulty as laypeople in understanding complicated legal texts relative to simplified legal

[1]Data and code for both experiments are available at the following link: https://osf.io/vtscj/?view_only=b29d7f40400646589eec651703534990

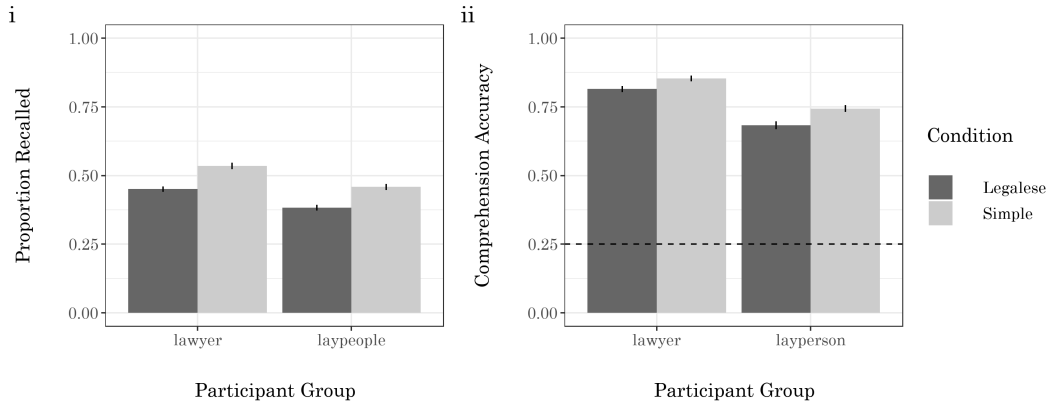

Figure 4.1: Proportion of legal content recalled (i) and comprehended (ii) in legalese and simple contracts by lawyer and non-lawyer participants. Error bars represent 95% bootstrapped confidence intervals. Dotted line in (ii) represents chance performance in comprehension task.

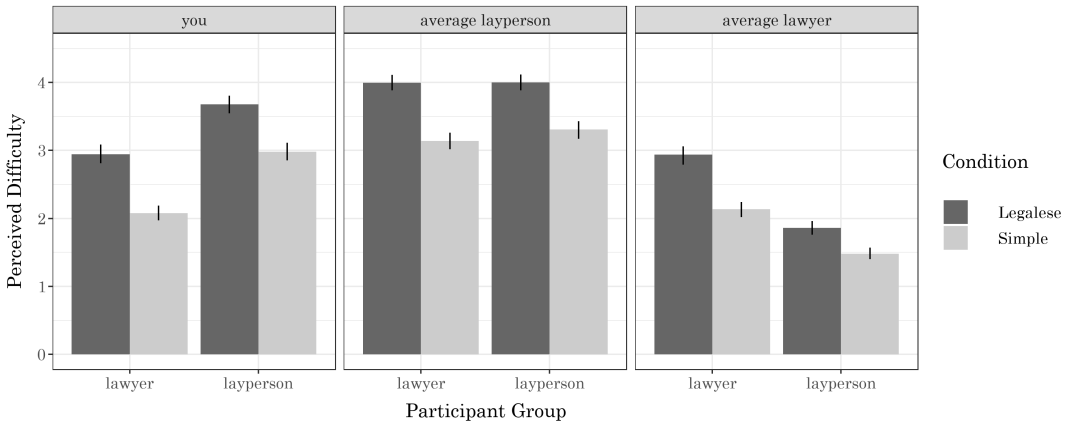

Figure 4.2: Subjective difficulty ratings by lawyer and lay participants regarding how difficult participants found a given text (a) for themselves (left panel); (b) for the average layperson (middle panel); and (c) the average lawyer (right panel).

texts, and that lawyers woud underestimate how difficult legalese texts are for laypeople.

Copy-and-Paste Hypothesis. Some commentators have speculated that lawyers simply write in a complex register out of "habit, laziness" [64] or respect for "tradition" [39], that they "copy and paste" [65] from existing templates with old, complicated terms because that's the "quickest and cheapest way to produce a contract"[66]. If this hypothesis were true, one would expect that lawyers would rate plain-English contracts as of equal quality as legalese contracts, and that lawyers would be equally likely to agree to sign off on a contract written in a simpler register written by someone else as they would for a contract written in a legal register.

In-Group Signaling Hypothesis. Some commentators have hypothesized that lawyers write in legalese to be accepted by their peers, to sound more "lawyerly," to "mark themselves as members of the profession" [64]. If so, one would predict that lawyers would rate contracts written in legalese as sounding more appropriate/suitable for a lawyer than those written in plain English, and would rate the author of that contract as more hireable than the author of a plain-English contract.

It's Just Business Hypothesis Some commentators have hypothesized that lawyers write in legalese as a way of "preserving their monopoly" [67] on legal services and "justifying fees" [64]. If this hypothesis were true, one would predict that lawyers would rate contracts written in legalese as being more likely to be signed by clients than contracts written in a simple register.

Complexity of Information Hypothesis. Some have speculated that legal language needs to be complex in order to satisfy certain communicative aims, such as conveying complex legal concepts in a way that "is far more precise than ordinary language" [39], to avoid ambiguity, and/or to ensure enforceability. To evaluate this hypothesis, we constructed a question that asked whether a

given contract excerpt was enforceable. If this hypothesis were true, one would predict that lawyers would rate simplified contracts as unenforceable or lower quality than complicated contracts.

4.3 Results

4.3.1 Experiment 1

In Experiment 1, we evaluated the curse of knowledge hypothesis.[2].

To evaluate the predictions of this hypothesis, we conducted a pre-registered experiment in which we evaluated lawyers' ($n = 105$) comprehension and recall of two types of legal contracts. The first set, "legalese" contracts, were written in a style containing linguistic features that have been shown to be disproportionately common in legal texts relative to non-legal texts, and which have also been shown to inhibit recall and comprehension of legal content relative to contracts without these features. The second set, "plain-English" contracts, were of equivalent meaning drafted without these difficult-to-process features. We analyzed lawyers' performance alongside a reanalysis of Martinez, Mollica & Gibson's [101] experiment of laypeople ($n = 108$) that used an identical set of materials and procedure.

Results are visualized in Figures 1 and 2. Contrary to the predictions of the curse of knowledge hypothesis, we observed a main effect of legal training and register on recall ($\beta = -.353$, SE=.159, p=.026) and comprehension ($\beta = -.808$, SE=.100, p<.001), but not an interaction between register and legal training on recall (p=.360) or comprehension (p=.638). That is, although lawyers were significantly better than laypeople at comprehending and recalling legal content overall in our materials, both lawyers and laypeople were better at compre-

[2]The pre-registration for Experiment 1 can be viewed at the following link: https://osf.io/y8xjd/?view_only=bf30ec08c7bd4f3c92d7c0024ce73eae

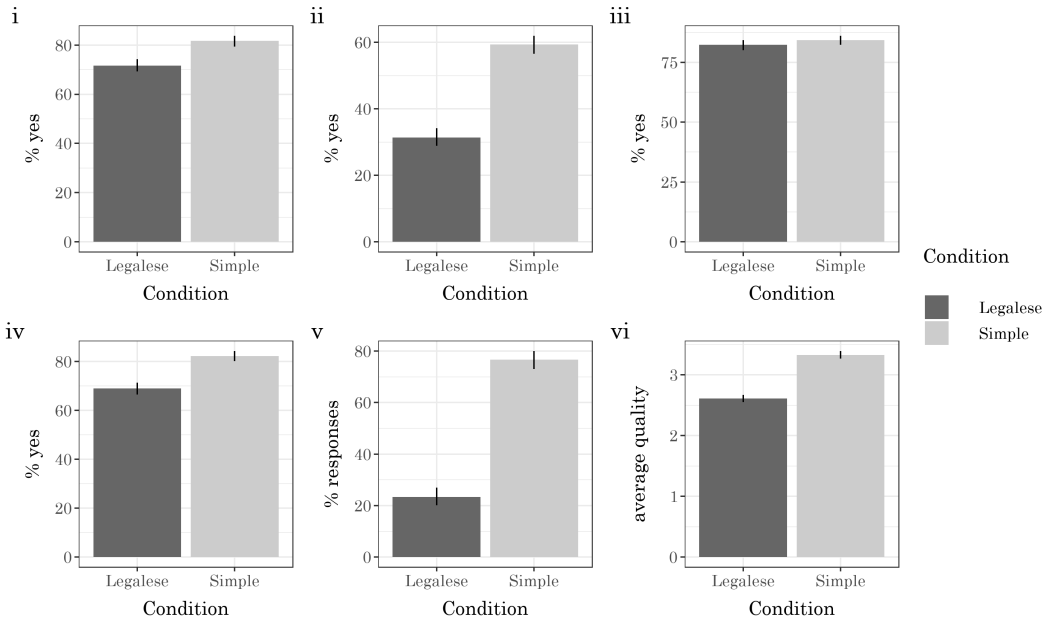

Figure 4.3: Results of lawyer responses to questions regarding the quality of legalese and simple contracts according to a series of desiderata, including (i) appropriateness of style, (ii) hireability of author, (iii) enforceability of document, (iv) likelihood of document being signed by client, (v) willingness to use document as written, and (vi) overall quality of document.

hending (β =.354, SE=.088, p<.001) and recalling (β =.360, SE=.121, p=.003) plain-English texts than legalese texts, there was no evidence that lawyers were disproportionately better than laypeople at comprehending (p=.638) or recalling (p=.360) legal content in legalese texts relative to plain-English.

We observed converging results when comparing lawyer and layperson's subjective difficulty ratings of each text, as lawyer participants' predictions of how difficult a text would be for the average layperson did not significantly differ from those of lay participants. See SI for details.

4.3.2 Experiment 2

In Experiment 2, we sought to evaluate the predictions associated with the four remaining hypotheses: The in-group signaling hypothesis, the it's just business

hypothesis, the complexity of information hypothesis, and the copy-and-paste hypothesis. To do so, we presented lawyers (n=102) with the same set of contracts used in Experiment 1, and asked them to rate the contracts on a variety of dimensions, including overall quality and enforceability of the contract, hireability of the author who wrote the contract, willingness to sign off on the contract as written, and likelihood that a client would agree to the contract's terms.[3].

Results of Experiment 2 are visualized in Figure 3. In line with all of the pre-registered predictions of the copy-and-paste hypothesis and against all of the pre-registered predictions of the in-group signaling, it's just business and complexity of information hypotheses, lawyers rated contracts written in plain-English as significantly higher quality (β =1.705, SE=.329, p<.001) and no less enforceable than legalese contracts (p=.717); rated the authors of plain-English contracts as significantly more hireable than those of legalese contracts ($\beta = 1.835$,$SE = .318$,$p < .001$); were significantly more likely to say that they would agree to use the contract as-written (β =1.432, SE=.270, p<.001); and predicted that clients would be significantly more likely to sign plain-English contracts than legalese contracts (β =1.232, SE=.338, p<.001).

The results of both experiments were robust to all measured demographic variables, including race, gender, age, years of practice experience, and "fanciness" of lawyer (see definition in methods). These results are reported and visualized in the SI.

4.4 Discussion

This study provides the first attempt to empirically investigate the long-puzzling question of why lawyers write the way that they do, undermining most prior accounts of the cognitive origins of legalese. For example, some commentators

[3]Pre-registration for Experiment 2 can be viewed here: https://osf.io/b98j5/?view_only=a9f9ba58bd114b5db0cfd820798344b1

have maintained that lawyers prefer or are otherwise forced to write in a complex manner in order to satisfy certain communicative aims, to sound more lawyerly, or to justify exorbitant fees to clients. Others have speculated that lawyers simply do not realize they are writing in a complicated manner due to how easy it is for them to understand. In contrast, the fact that lawyers in our studies rated plain-English contracts as higher quality, even more likely to be signed by clients and no less enforceable than legalese contracts, and rated the authors of plain-English contracts as more hireable than authors of legalese contracts undermines both of these sets of hypotheses, suggesting that in many instances lawyers both can and prefer to write in a more understandable manner as opposed to being bound by the nature of law, or engaging in a "conspiracy of gobbledygook."

Meanwhile, the fact that lawyers rated both contracts as enforceable and likely to be signed by clients but preferred plain-language contracts on several dimensions suggests, consistent with what we have dubbed the "copy-and-paste" hypothesis, that lawyers may simply draw from old, pre-existing templates laden with arcane and convoluted language due to that being easier and cheaper to produce than drafting a simpler contract from scratch. This finding is consistent with recent empirical work indicating that lawyers rely heavily on templates in drafting contracts, with future agreements only rarely deviating from previous ones even when deviations would apparently benefit the involved parties[123]. In addition to cost, said stickiness may also be borne out of lawyers' training in the importance of precedent, which overall might lead to an adherence to templates laden with old, archaic language by virtue of the fact (or assumption) that they worked before, and that the specific language may have been "defended in court" previously.

From a policy perspective, our results also provide insight into the long-standing question of how to make legal language more understandable. Al-

though for decades, the United States government has engaged in top-down efforts to simplify public legal documents for the benefit of society at large [49], [50], recent work has revealed these efforts to have failed, as laws, like contracts, remain laden with difficult to process features such as center-embedding and low-frequency words [121]. While this failure may lead some to conclude that simplifying legal language is an intractable affair, our results paint a more optimistic picture, suggesting that lawyers (a) believe legal documents can and should be simplified to better serve their communicative aims; and (b) like laypeople, struggle to comprehend complex legal language relative to a simpler alternative. Our results further suggest that the processing difficulty of legal texts may be alleviated as lawyers and lawmakers become more aware of both the ways in which public legal documents tend to be complex, as well as the alternatives available to them in order to make them less complex.

It is worth noting that our results do not imply that legal documents can be simplified limitlessly without sacrificing communicative aims, nor do we discount the role of formality in legal writing. Like other professionals, lawyers may use a more formal tone in legal documents in order to, for example: (a) demonstrate their status of members of the legal community, which may require convergence on a style that is identifiable and replicable, and (b) signal to a reader that a text should be taken seriously an official legal document as opposed to a form of casual, non-binding communication.

Instead, our results indicate that such formality is not necessarily synonymous with complexity. That is, in many cases, lawyers can and should adopt a simpler register in order to achieve a level of formality that best aligns with their communicative aims as opposed to burdening clients and themselves with obfuscatory legalese.

4.4.1 Constraints on Generality

Examining the participant sample, the stimuli and general experimental design suggests that the results of the present study would likely generalize to a broad array of relevant real-world scenarios.

With regard to participants, our sample included a large number of lawyers that, according to available estimates [21], [124], were broadly representative of the legal profession with regard to to a number of demographic factors, including age, ethnicity, gender, years of legal experience, and type of legal employment. Analyses further revealed that our results were the same when controlling for these demographic variables in our analysis, such that we expect the results to generalize to the broader population of United States lawyers. It is unclear whether they would generalize to the legal profession in other countries.

With regard to materials, our focus was on contracts, given that contract law is one of the most central areas of legal doctrine [21], [125], and because contracts are one of the most common types of legal documents encountered in everyday life. Our stimuli consisted of a diverse array of contract excerpts whose content mirrored the most common types of clauses found in contracts in the United States [126], [127]. Although our focus was on contracts, the linguistic features we looked at have been found to be disproportionately prevalent in both private legal documents (contracts) and public legal documents (e.g. legislation) relative to other forms of written and spoken English [101], [120]. Thus, we expect the results to generalize to other types of legal documents beyond those examined in the present study, though it is likely that some types of provisions will be less amenable to simplification than those used in the present study.

Regarding the ecological validity of the design, one might wonder whether lawyers' responses to questions in a hypothetical setting would generalize to real-world behavior. Given that an important role of a lawyer in the real world is to reason about hypothetical scenarios and engage in counterfactual reasoning, the

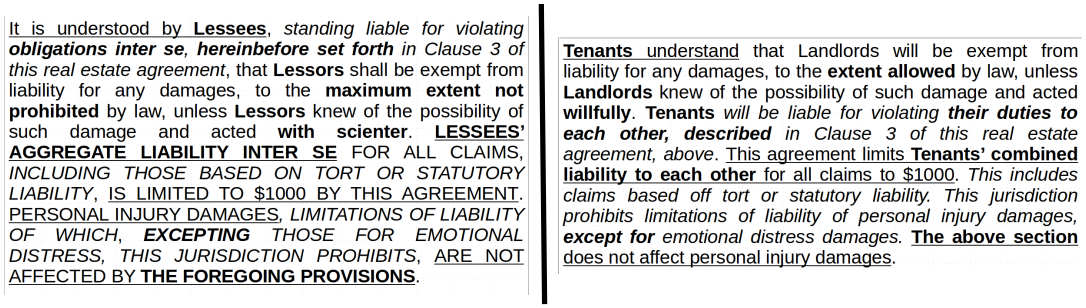

Figure 4.4: An example stimulus pair in legalese (left) and simple (right) register. The differences in surface properties across registers are depicted by font style. Bold denotes word frequency. Italic denotes embedded clauses. Underlined denotes voice. Unfortunately, we have run out of font styles to make differences in capitalization more apparent. Image reprinted from [101] SI.

fact that our experimental design asked lawyers to reason about hypothetical scenarios and engage in counterfactual reasoning would seem to imply that our study was well-aligned with the job of a lawyer in the real world. By extension, this would suggest that our design was an ecologically valid way to test our hypotheses.

A related concern relates to whether there was a performative element–if lawyers know they are subjects in an experiment and are being observed by scientists, maybe they will behave differently than in the real world. Although this is an important concern, we have no reason to expect that lawyers knew what result we were interested in, given that: (a) we did not give away the specific research question we were interested in when recruiting lawyers for our study; and (b) we ensured that lawyers were unaware of register manipulation during the experiment. Supposing that lawyers did not know what result we were interested in, we also have no reason to expect that their behavior was systematically influenced to help the researchers get a desired result. Thus, we have no reason to expect that a potential performative element drove our results.

4.5 Methods

4.5.1 Experiment 1

Materials

The primary materials consisted of 12 pairs of short contract excerpts of roughly 150 words each (see Supplemental Fig. 4.4). Each pair contained of (a) one excerpt drafted in a legalese register, containing features identified by previous studies to be strikingly more prevalent in legal texts relative to non-legal texts, including center-embedded clauses, low-frequency jargon, non-standard capitalization and passive-voice structures; and (b) one excerpt drafted in a simple register, identical in content to the other excerpt but without the above features.

For each contract pair, 12-15 comprehension questions were drafted in a "neutral" register. In addition to the main experimental materials, we also implemented the author recognition task (ART; [77], [78]) as a measure of individual differences in experience with language.

Participants and Procedure

United States attorneys (n=106) were recruited to participate as subjects in our experiment, through a combination of direct email invitations, word-of-mouth recruitment, and social media posts. Participants received $100 for their participation in the study. Participants were retained in our analysis as long as they were licensed to practice law in the United States. Participants were required to enter an official law school or law firm email, or provide their official bar number in order to help verify their attorney status. Of the 106 participants, 105 were verified to be attorneys and were retained in the final analysis.

With regard to demographics, the mean age of retained participants was 34 (median: 31). 60.8% of participants identified as male. 38.2% identified as

non-white. Participants had a mean of 5.9 years of practice experience. 50.9% of the sample were coded as "fancy" lawyers, meaning that they either (a) graduated from a top-25 law school according to U.S. News and World Report, or (b) worked at a top-200 law firm according to American Lawyer (AmLaw) magazine.[4]

Retained participants were psuedorandomly assigned to six trials (3 legalese; 3 simple). Participants did not see the same contract in both a simple and legal register. Assignment of stimuli to participant was pseudorandom fto ensure that across participants every trial was administered with approximately the same frequency. The order of trials was randomized for each participant.

A trial consisted of (a) reading an excerpt, (b) a subset of the ART, (c) recalling the excerpt, and (d) answering comprehension questions. For the reading component, participants were presented with exactly one excerpt, written in either legalese or plain English. They were asked to carefully read the text twice, and were given as much time as needed to do so. For the ART component, participants were given the names of 50 individuals and were asked to select which names corresponded to real authors. We expanded the ART task to 300 trials in order to keep the timing of a trial consistent. The original items from the published ART were presented first. For the remaining trials, the participants were administered novel items that looked virtually the same as authentic materials (half of the names corresponding to real authors, the other half corresponding to high-school track stars). We do not use these novel items in our analysis as they have not been validated [80]. After being shown the ART materials, participants were asked to recall as much of the excerpt they had read as possible. They were told that they could use their own words, but that their version should stay true to the original. Finally, each trial ended with the comprehension questions corresponding to the excerpt.

[4]This was determined based on the email participants provided when taking the study.

Analysis Plan

Following Martinez, Mollica Gibson [101], two trained research assistants coded whether a proposition was successfully recalled (see SI for details). Coders were unaware of whether a participant had seen or recalled the simple or legalese version of a text. Twenty percent of the retellings were coded by both coders so as to assess inter-rater reliability using Cohen's kappa coefficient [81], [82]. For our regression analyses, we perform both a conservative analysis and an anti-conservative analysis, with regard to ties. Our results do not qualitatively change, so we only report the conservative analysis in text (see SI for anti-conservative analysis).

4.5.2 Experiment 2

Materials

Our primary materials consisted of the same 12 pairs of short contract excerpts as those used in Study 1. In addition, we also constructed a series of questions aimed at testing specific hypotheses for why lawyers write the way that they do. Here we discuss each of these questions in turn. The full list of questions, as well as the experimental interface, is provided in the SI.

Copy and Paste Hypothesis. To test this hypothesis, we constructed a question that asked participants to rate the quality of a given contract excerpt (in plain English or legalese), as well as another question that asked participants whether they would agree to sign off on a given contract excerpt assuming it were written by someone else.

In-Group membership hypothesis. To test this hypothesis, we constructed two types of questions: one that asks whether the style of a particular excerpt sounds appropriate for a lawyer, and another that asks whether a participant would hire the author of the excerpt.

It's just business hypothesis. To evaluate this hypothesis, we constructed a question that asked participants to rate whether a client would be likely to sign a particular contract excerpt.

Complexity of Information hypothesis. To evaluate this hypothesis, we constructed a question that asked whether a given contract excerpt was enforceable. We constructed a question that asked participants to rate the quality of a given contract excerpt (in plain English or legalese)

Participants and Procedure

United States attorneys (n=105) were recruited to participate as subjects in our experiment through similar means as Study 1. Participants received $40 for their participation in the study, and were retained in the analysis using the same criteria as Study 1.

With regard to demographics, the mean age of retained participants was 35.7 (median: 33). 62.7% of participants identified as male. 38.2% identified as non-white. Participants had a mean of 8.3 years of practice experience (median: 5.5). 40.2% of the sample were coded as "fancy" lawyers.

With regard to procedure, Retained participants were psuedorandomly assigned to six trials. Assignment of stimuli to participant was pseudorandom to ensure that across participants every trial was administered with approximately the same frequency. The order of trials was randomized for each participant. Within each trial, participants were first presented with one version of a contract excerpt in either legalese or plain-English, and asked to answer several questions about it. Participants were then presented with the other version of the contract excerpt and asked to answer the same questions about it. Participants were then shown the two versions side-by-side and asked to answer several questions about the two versions in tandem.

Acknowledgements

Thanks to Sofie Cheung, Sofia Zhang, Euphy Liu and Anita Podrug for help with coding the recall responses. Thanks to Nancy Kanwisher, Roger Levy, Rebecca Saxe, Ev Fedorenko, and Laura Schulz for helpful feedback and discussions regarding the hypotheses and experimental design. Thanks to audiences at Tedlab and Evlab, the MIT Brain Cognitive Sciences "Cognitive Lunch" talk series, and the Human Sentence Processing conference in Pittsburgh, PA, March 2023. And finally, thanks to the editor, one anonymous reviewer, and Adele Goldberg for their comments on the submitted paper.

Chapter 5

Even Laypeople Use Legalese

This chapter is adapted from the following publication:

Eric Martínez, Frank Mollica & Edward Gibson, *Even Laypeople Use Legalese*, 121 Proc. Nat'l Acad. Sci. U.S. 35 (2024)

It is reproduced here with slight modifications.

Abstract

While the previous chapter lends support for the idea that lawyers copy and paste from existing templates laden with complex syntax and archaic legal jargon, this chapter investigates how convoluted legalese makes its way into legal documents in the first place. Here, a corpus analysis (n=59 million words) first replicated and extended prior work revealing laws to contain strikingly higher rates of complex syntactic structures relative to six baseline genres of English. Next, two pre-registered text generation experiments (n=280) tested two leading hypotheses regarding how these complex structures enter into legal documents during the drafting process. In line with the magic spell hypothesis, we found people tasked with writing official laws wrote in a more convoluted manner

than when tasked with writing unofficial legal texts of equivalent conceptual complexity. Contrary to the copy-and-edit hypothesis, we did not find evidence that people editing a legal document wrote in a more convoluted manner than when writing the same document from scratch. From a cognitive perspective, these results suggest law to be a potential rare exception to the general tendency in human language towards communicative efficiency. In particular, these findings indicate law's complexity to be derived from its performativity, whereby low-frequency structures may be inserted to signal law's authoritative, world-state-altering nature, at the cost of increased processing demands on readers. From a law and policy perspective, these results suggest that the tension between the ubiquity and impenetrability of the law is not an inherent one, and that laws can be simplified without a loss or distortion of communicative content.

5.1 Introduction

Since the dawn of modern civilization, humankind has codified and communicated societal norms and rules largely in the form of written laws. In order for people to understand and comply with social norms and rules, it follows that legal content must be drafted in a way such that people can ultimately understand and comply with it.

Indeed, the principle that law should provide such "fair notice" to the general public is a core tenet of modern legal doctrine, which mandates that laws provide proper warning of prohibited conduct "in language that the common world will understand," [1], [2] to "give the person of ordinary intelligence a reasonable opportunity to know what is prohibited, so that he may act accordingly." [3], [4].

In addition to legal doctrine, principles of communicative efficiency likewise

suggest that laws should be understandable. For example, a burgeoning psycholinguistics literature has uncovered various properties of human language that appear optimized for easing the communicative burden on speakers and listeners [23]–[34], such as (a) syntactic dependency length minimization [36], [37], and (b) a preference for shorter words over longer words in everyday speech [35].

These principles notwithstanding, legal documents have long been observed to be notoriously difficult to understand [38]–[41]. In particular, recent work has revealed legal documents, including both private contracts and federal legislation, to be laden with center-embedded clauses at a rate twice as high as other genres of texts, including those aimed at an educated audience [101], [121].

Moreover, legal documents containing these features have been shown to cause processing difficulty relative to legal documents without these features, even for lawyers and experienced lay readers [101], [102].

The mismatch between the ubiquity and impenetrability of legal documents has long been acknowledged not just by those tasked with reading legal documents but those tasked with promulgating them, as well [100]. In the United States, policy efforts to simplify laws have been advocated for decades [48], [49], [52], with little to no success [121].

And although recent work has revealed that even lawyers prefer simplified legal documents over complex legal documents [102], it remains an open question how complex features such as center-embedded syntax make their way into legal documents in the first place.

To answer this question, we conducted two well-powered pre-registered experiments testing two leading hypotheses for why lawyers write the way that they do, including: (a) the magic spell hypothesis, according to which lawyers and lawmakers write in a convoluted manner in order to lend legal documents a ritualistic, spell-like element; and (b) the copy-and-edit hypothesis, according

to which conditions and specifications are often considered only after the creation of an initial draft and are more easily embedded into the center of existing sentences as opposed to being written-out into separate sentences.

In line with the magic spell hypothesis [40], [121], we found that people tasked with writing laws wrote in a more convoluted manner (i.e. more center-embedded syntax) than when tasked with writing control texts of plausibly equivalent conceptual complexity. Contrary to the copy-and-edit hypothesis, we did not find evidence that people editing a legal document wrote in a more convoluted manner than when writing the document from scratch.

These findings suggest that lawyers and lawmakers write in a complex manner in order to confer legal documents a ritualistic, spell-like element, presenting broad-ranging implications for law, policy and cognitive science.

5.2 Law's Syntactic Complexity

Perhaps the most distinctive feature of *legalese* is center-embedded syntax, in which clausal content is embedded within the center of another clause as opposed to being edge-embedded or written as a separate sentence.

Consider the following example from a Massachusetts Drunk Driving Law:

> *Whoever, upon any way or in any place to which the public has a right of access, or upon any way or in any place to which members of the public have access as invitees or licensees, operates a motor vehicle with a percentage, by weight, of alcohol in their blood of eight one-hundredths or greater, or while under the influence of intoxicating liquor, or of marijuana, narcotic drugs, depressants or stimulant substances, all as defined in section one of chapter ninety-four C, or while under the influence from smelling or inhaling the fumes of any substance having the property of releasing toxic vapors as*

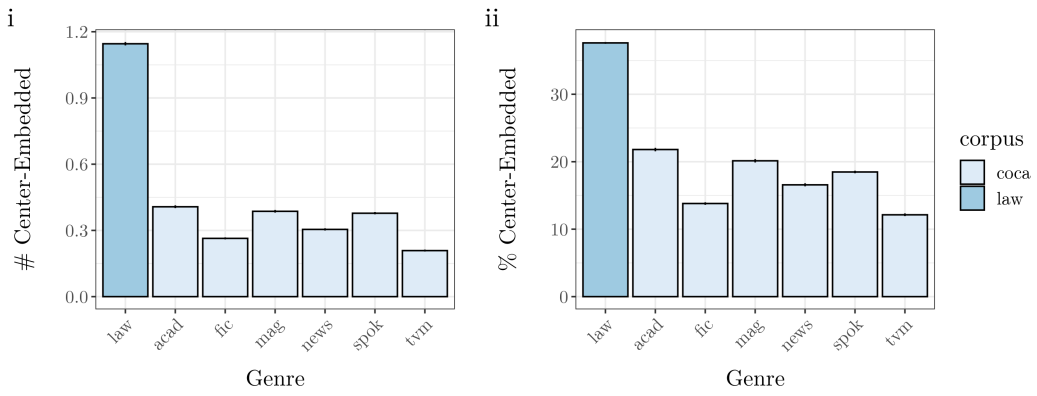

Figure 5.1: Number of center-embedded clauses per sentence (i) and percentage of sentences with center-embedded syntax (ii) in laws compared to six baseline genres of written and spoken English: academic texts, fiction, magazine articles, newspaper articles, and TV/Movies. Laws were taken from the 2021 edition of the United States Code, the official compilation of all federal laws currently in force. Baseline genres were taken from the most recent year (2019) of the Corpus of Contemporary American English. Error bars represent 95% bootstrapped confidence intervals.

> *defined in section 18 of chapter 270 shall be punished by a fine of not less than five hundred nor more than five thousand dollars or by imprisonment for not more than two and one-half years, or both such fine and imprisonment.* [128]

The clausal material in red is embedded into the center of the main clause, separating important words from each other and leading to a structure than is unusually difficult to process [37], [44], [122].

Prior work has indicated that this example is by no means unique, as legal documents have been found to contain strikingly higher rates of center-embedded syntax relative to other genres of English, including those aimed at an educated audience [101], [121].

For robustness purposes, here we first sought to replicate and extend these results using a more direct method of detecting center-embedded syntax compared to prior work (see methods), in which we used state-of-the-art natural

language processing tools to detect the number of center-embedded verbs in a sentence in (a) the United States Code [118]; and (b) six baseline genres in the Corpus of Contemporary American English [72]: academic texts, fiction, newspaper articles, magazine articles, spoken transcripts, and TV/Movie scripts.

Results are visualized in Figure 1. Consistent with prior work, laws contained several times more center-embedded clauses than any of the baseline genres of English. When looking at the percentage of sentences with center-embedded clauses, laws likewise contained strikingly higher rates than any other genre.

In addition, prior analyses have also indicated that center-embedded syntax disproportionately contributes to the higher difficulty in recalling *legalese* vs plain-English compared to other markers of *legalese*, such as passive voice and non-standard capitalization [101]. The increased processing difficulty associated with center-embedded syntax in legal texts and non-legal texts has been hypothesized to be associated with increased demands on working memory capacity resulting from long-distance syntactic dependencies [43], [101]. However, it remains an open question to what extent legal texts have longer syntactic dependencies relative to baseline texts.

To answer this question, we also compared the syntactic dependency length in our legal vs non-legal corpora. As with center-embedded syntax, and consistent with the predictions of the theoretical literature, laws contained strikingly longer dependencies than any of the other baseline genres. Full results reported in SI.

5.3 Hypotheses

Having replicated and extended prior work demonstrating the prevalence of complex syntactic structures in legal texts, we next turned to testing two lead-

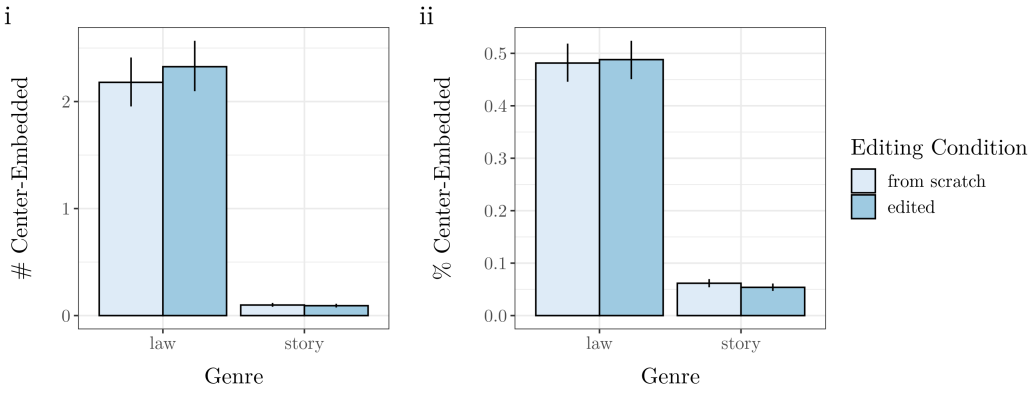

Figure 5.2: Number of center-embedded clauses per sentence (i) and percentage of sentences with center-embedded syntax (ii) in criminal laws versus crime stories. Error bars represent 95% bootstrapped confidence intervals.

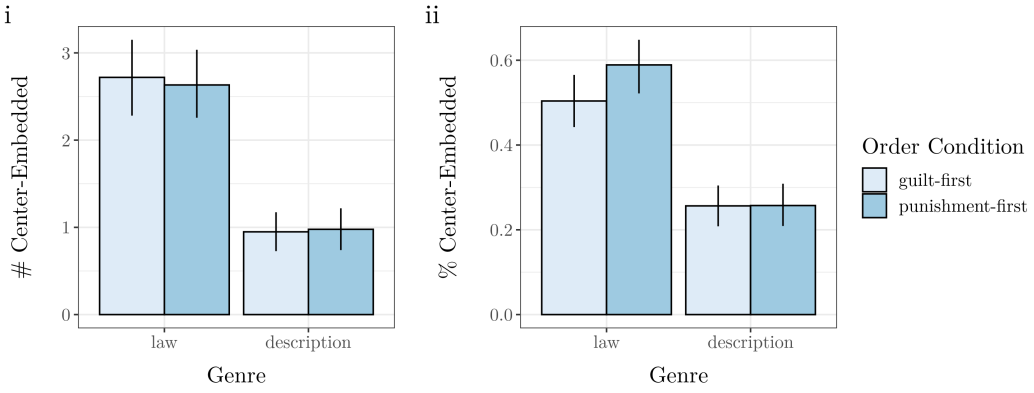

Figure 5.3: Number of center-embedded clauses per sentence (i) and percentage of sentences with center-embedded syntax (ii) in participant-drafted laws versus unofficial descriptions of laws. Error bars represent 95% bootstrapped confidence intervals.

ing hypotheses proposed in previous literature for how such features enter into legal documents in the first place. Below we briefly present each of these hypotheses in turn, as well as the associated predictions of these hypotheses that we preregistered for our experiments.

Magic Spell Hypothesis. Some have posited that lawyers and lawmakers write in a convoluted manner in order to lend legal documents a ritualistic, spell-like element [40], [121]. These ritualistic types of language are often referred to as performative utterances [129], which unlike descriptive utterances, not only describe the state of the world but change the state of the world they are describing.

In order to effectively convey performativity, such utterances have been attested to contain distinctive, low-frequency structures, as in the case of magic spells, which are characterized by such features as rhyming (e.g. "Double, double toil and trouble; Fire burn, and cauldron bubble": [130]) and foreign-sounding jargon ("wingardium leviosa" [131]). Indeed, in a pilot experiment we found that participants tasked with writing a magic spell rhymed in 58.8% of sentences, as compared to 1.8% of sentences when tasked with writing a mere recollection of a fantastical event involving a magic spell (see SI).

Given that legal documents, like spells and other performative utterances: (a) have been shown to possess low-frequency structures (such as center-embedded syntax), at several times the rate of standard texts [101], [121], and (b) are meant not only to describe the state of the world but also change the state of the world (by establishing, eliminating and/or modifying legally binding social rules), one might similarly hypothesize that such low-frequency structures are inserted so as to signal a legal document's authoritative nature.

If this hypothesis were true, one would predict that people tasked with writing an official legal document would write in a more convoluted manner (including more center-embedded syntax) than when writing a non-performative

law-related document of equivalent conceptual complexity.

Copy-and-Edit Hypothesis. Recent work has speculated that convoluted legal language may be a result of an iterative drafting process, in which conditions and specifications are often thought of only after the creation of an initial draft or template and are more easily embedded into the center of existing sentences as opposed to being written-out into separate sentences [121]. Given the observed reliance of lawyers and lawmakers on templates and "boilerplate provisions" in the drafting of legal documents [123], [132], this would explain why the prevalence of structures such as center-embedded syntax are so much higher in legal documents than other genres where the drafting process is less path-dependent and drawn-out [101], [121]. If this hypothesis were true, one would predict that people tasked with editing a legal document will write in a more convoluted manner (including more center-embedded clauses) than when tasked with writing a legal document of equivalent semantic content from scratch.

5.4 Results

5.4.1 Experiment 1

In Experiment 1, we evaluated both the magic spell Hypothesis and copy-and-edit hypothesis.[1] To evaluate the predictions of these hypotheses, we conducted a preregistered experiment in which we asked participants (n=200) to write either a (a) legal provision prohibiting a crime; or (b) a story describing someone committing that crime.

In half of the trials (*from-scratch condition*), participants were initially given all of the details of the crime and were tasked with writing their law or story all at once. In the other half of trials (*editing condition*), participants were first

[1]All data, code and pre-registrations for this paper can be viewed at the following OSF repository link: link.

given details of a paired-down version of the crime and were asked to write their law or story based on that version. After completing their draft, participants in these trials were then presented with additional details of the crime and were asked to revise their draft to incorporate these additional details.

Results are visualized in Figure 5.2. In line with the predictions of the magic spell hypothesis, participants' responses contained a higher percentage of sentences with center-embedded syntax in the law condition (48.1%; 95% CI: 46.0 to 51.1) compared to responses in the story condition (5.8%; 95% CI: 5.2 to 6.2). The difference was striking [OR: 8.3], and significant (β =2.859, SE=.113, p<.0001), and held true when looking at the number of center-embedded clauses per sentence (β =3.126, SE=.204, p<.0001) as opposed to just the percentage of sentences with center-embedded clauses.

Contrary to the predictions of the copy-and-edit hypothesis, participants in the editing condition were not significantly more likely to center-embed than in the from-scratch condition (p=.262), nor was there an interaction between genre and editing manipulations (p=.244). This was true both when looking at the number of center-embedded clauses per sentence and when examining the percentage of sentences with center-embedded clauses (p=.755 for editing manipulation; p=.165 for interaction between editing and genre manipulations).

5.4.2 Experiment 2

To further test the robustness of the magic spell hypothesis, we conducted a second experiment in which we asked participants (n=80) to write either (a) an official law prohibiting a crime (*law condition*); or (b) an unofficial description of a law prohibiting a crime (*description condition*), with the latter being a plausibly tighter control than a story for a text of similar conceptual complexity as a law.

To control for possible ordering effects of the materials, in half of the trials

the instructions described the requirements of guilt for the prohibited crime, followed by the punishment for the crime (*guilt-first condition*). In the other half of the trials, the instructions described the punishment for the crime followed by the requirements of guilt for the crime (*punishment-first condition*).

Results are visualized in Figure 5.3. As in Experiment 1, in line with the magic spell hypothesis, participants were more likely to produce sentences containing center-embedded clauses in the law condition (54.6%; 95% CI: 50.3 to 59.1) than in the control condition (25.7%; 95% CI: 22.5 to 28.9). The difference was striking [OR: 2.1], and was significant both when looking at the number of center-embedded clauses per sentence (β =1.391, SE=.184, p<.0001) as well as the percentage of sentences with center-embedded clauses (β =1.552, SE=.227, p<.0001).

The results of the ordering manipulation were also consistent with the magic spell hypothesis, as participants were not significantly more likely to produce sentences with center-embedded syntax in the guilt-first condition (p=.613) than in the punishment-first condition, nor was there an interaction between genre and ordering manipulation (p=.414). Converging results were found when analyzing the the number of center-embedded clauses per sentence (p=.362 for ordering manipulation; p=.274 for interaction between ordering and genre manipulations).

To further test the robustness of the magic spell hypothesis and account for the possibility that participants responses in the two conditions were not matched for conceptual complexity, we conducted additional exploratory analyses where (a) responses were filtered if they did not include more than 80% of the propositions in the instructions; and (b) conceptual complexity (operationalized as proportion of propositions included in a participant's response) was included as a fixed-effect predictor in our regression models.

The results of these analyses were consistent with those reported in the main

text; genre remained a significant and strong predictor of participant's likelihood to center-embed. These analyses are reported in full in the SI.

5.5 Discussion

This paper has empirically investigated the long-puzzling question of why laws are written in a complex manner, testing two leading hypotheses across two well-powered, pre-registered experiments.

In line with the magic spell hypothesis, we found that people tasked with drafting laws wrote in a more convoluted manner than when tasked with drafting various control texts of plausibly equivalent conceptual complexity. Contrary to the copy-and-edit hypothesis, we did not find evidence that people editing a legal document wrote in a more convoluted manner than when writing the document from scratch.

These lines of evidence were robust to various control attempts, including (a) comparing laws to different genres (stories and descriptions of laws) to serve as control texts; and (b) manipulating the order in which specifications of a given law were presented (requirements of guilt first vs punishment first).

Answering this question is relevant to advancing longstanding questions of both cognitive science and legal doctrine / public policy.

On the cognitive science side, as documented above, there is a burgeoning psycholinguistics literature documenting the various domains in which communicative efficiency shapes human language [23]–[34]. Given that law stands as an attested exception to this observed efficiency, uncovering the cognitive factors giving rise to the processing difficulties of legal documents can help inform the degree and domains in which human language is optimized for communicative efficiency, as well as the factors giving rise to said (in)efficiency.

In particular, these results suggest law to be a type of performative utter-

ance [129], meant not just to communicate states of the world but to explicitly alter the state of the world. In such instances, distinctive low-frequency structures may be inserted in order to effectively signal the performative nature of the utterance, which in turn might increase processing demands on readers. In the case of other types of performative language, such as "actual" magic spells, such structures may include rhyming or foreign-sounding terminology. In the case of laws, this deviation may come largely in the form of altering the syntactic structure of the clausal material from right-branching to center-embedded, creating as a byproduct an overload on a reader's working memory capacity.

On the law and policy side, these results add to an emerging body of literature demonstrating that the language of legal documents can be simplified without a loss or distortion of legal content [101], [102], [121], which might provide a source of optimism to efforts to simplify legal documents (which have been advocated for for decades [50], to no avail [121]). These findings also shed insight into debates related to the aforementioned legal doctrines that expressly assert or implicitly assume that laws be understandable to the public at-large. Jurists have long acknowledged the tension between the doctrinal mandate that laws be understandable to the common person and the observation that laws are not understandable to the common person [1]–[4]. Whereas recent proposals to resolve this tension have taken for granted the necessity of law's complexity and have called for scaling back the mandate that laws be accessible to the common person [133], our results suggest such compromises may not be necessary. Instead, our results indicate that lawmakers can faithfully comply with this mandate while simultaneously preserving the desired level of conceptual complexity.

5.6 Materials

5.6.1 Corpus Analysis

Materials

Our primary materials consisted of two corpora. Our legal corpora consisted of the 2021 edition of the United States Code [118], the official compilation of all federal legislation currently in force.

Our baseline corpora consisted of academic texts, fiction texts, newspaper articles, magazine articles, spoken transcripts, and TV/Movie scripts from the Corpus of Contemporary American English [72]. In order to best match the legal corpora, we used only texts from the most recent year of the corpus (2019).

Procedure

To calculate the number of center-embedded clauses in each sentence, we first (a) tokenized each corpus into sentences; and (b) got a syntactic parse of the sentence using the Stanza package from the Stanford NLP group [107]. Following [101], [121] we then filtered out sentences that (i) contained fewer than 10 alphabetic words; (ii) did not end in a punctuation mark; or (iii) contained 3 or more punctuation characters in a row.

For each sentence, we then calculated the number of center-embedded verbs (operationalized as the number of main verbs between a noun and its root).

For validation purposes, we hand-coded a random sample of 300 sentences for the presence of center-embedded clauses. This revealed the parser to be 92.3% accurate in detecting whether a sentence contained a center-embedded clause (95% CI: 89.3 to 95.5).

5.6.2 Experiment 1

Materials

Table 5.1: Set of propositions of sample item from Experiment I. Propositions in red are those not initially presented to participants in the copy-and-edit condition.

Requirements of guilt for offense:

- Any person
  - Who trafficks in marijuana
  - By knowingly or intentionally:
    - manufacturing, distributing, dispensing, or cultivating; or
    - possessing with intent to manufacture, distribute, dispense, or cultivate; or
    - By bringing into the commonwealth
  - A net weight of fifty pounds or more of any mixture containing marijuana

Punishment of offense:

- Not less than two and one-half nor more than fifteen years in prison

The primary materials consisted of eight items, with each item consisting of sets of instructions to write a passage relating to (respectively) the commission of a legally prohibited criminal offense (i.e. a crime), such as arson, bribery, or drunk driving. Each item consisted of 4 conditions (2 manipulations with 2 conditions each). The first manipulation was genre, which consisted of a *legal* condition and a *story* condition. In the legal condition, the materials consisted of instructions asking participants to write a law prohibiting a crime. In the story condition, participants were asked to write a story involving someone committing a crime. Both conditions had an associated cover story explaining the

motivation behind the task. In the legal condition, participants were told that they were a "lawmaker" who was "tasked with writing a law that prohibits a certain crime, and specifies the punishment for that crime if the crime is committed." In the story condition, participants were told that they were a "fiction writer" who was "tasked with writing a story about someone who commits a crime and is punished for committing the crime."

The second manipulation was sequencing, whose conditions consisted of a *from-scratch* condition and an *editing* condition. In the from-scratch condition, the details and specifications of the crime were presented all at once. In the editing condition, in contrast, the specifications were presented in two stages. In Stage 1, the version of the crime included within the instructions was paired-down and did not contain all of the specifications. In Stage 2, the version of the crime included all of the specifications, and the instructions directed participants to edit their text so as to include all of the additional instructions.

Participants and Procedure

Participants (n=200) were recruited via the online platform Prolific. This sample size was based on a power analysis, which determined the number of participants that would give us an 80% chance to detect an effect size that was at least half as large as the effect of the interaction between genre + sequencing obtained in a pilot experiment (this was smallest effect of any predictor variable in our pilot experiment). Participants were eligible if they resided in the United States, were 18 years or older, and native speakers of English. Each participant completed 8 trials of the same series of tasks.

On a given trial, participants would be presented with materials in one of the four conditions, and asked to write either a law or story in accordance with the material's instructions. As noted above, when in the from-scratch condition, participants were asked to draft their text all-at-once, whereas in the

editing condition, participants were first asked to write an initial draft based on a paired-down version of the crime described, and then subsequently presented with the full version of the crime and asked to edit their draft to incorporate the additional details associated with that version. Across the 8 trials, each participant was presented with 2 items in each of the 4 conditions, never seeing the same item more than once.

Prior to each trial, participants were given a comprehension check question where they were (a) told which of the two genres they would be asked to write (a law or a story), and (b) asked to confirm which of the two genres they would be asked to write. Participants were not allowed to proceed to the trial until answering the comprehension check correctly.

Prior to completing the first trial, participants were asked to promise that they would not use a language model (such as GPT) to complete the task. After completing the last trial, they were prompted with a similar message asking to promise that they did not use a language model (such as GPT) to complete the task.

Participants were retained in the analysis if they completed all trials and were determined not to use a language model in their responses.

Analysis Plan

To evaluate participant responses, responses were separated into sentences using an automatic parser—in particular, the tokenizers package in R. The tokenized sentences were spot-checked by a human and corrected for errors. After tokenization, sentences were hand-coded for center-embedded syntax, both in terms of (a) the degree of center-embedded syntax (defined as the number of center-embedded verbs); and (b) the binary presence of center-embedded syntax (i.e. were any verbs in the sentence center-embedded).

Following our preregistration, we then analyzed the effect of our two manipu-

lations on the prevalence of center-embedded syntax by conducting two separate regressions for each of the two operationalizations of center-embedded syntax, including (a) a mixed-effects binary logistic regression with the binary presence of center-embedded syntax (in a given sentence) as the outcome variable; and (b) a mixed-effects possoin regression with degree of center-embedded syntax as the outcome variable. Both regressions featured (a) genre, sequencing condition and their interaction as fixed-effects; and (b) genre, sequencing condition, item and participant as random effects. Results did not qualitatively change for either regression. We report both in the text.

5.6.3 Experiment 2

Materials

Table 5.2: Instructions in Experiment II for law condition and description condition

Law Condition	Description Condition
You are a lawmaker. You are tasked with writing a law that prohibits a certain crime, and specifies the punishment for that crime if the crime is committed. Below are the preconditions and punishment for the crime. Please write the law ensuring that it sounds authoritative and legally binding.	You are a tour guide, working in a country with strict crime laws. In order to raise awareness among your foreign customers of the crime laws, you are tasked with writing a description of the precondition for a particular crime in your country, as well as the punishment for committing that crime. Below are the preconditions and punishment for the crime. Please write the description, ensuring that they are comprehensive and accurate.

Similar to Experiment 1, the primary materials of Experiment 2 consisted

of eight items, each of which consisted of 4 conditions (2 manipulations with 2 conditions each). The first manipulation was genre, which consisted of a *law* condition and a *description* condition. The law condition was identical to the law condition in Experiment 1, and consisted of instructions asking participants to write a law prohibiting a crime. In the description condition, participants were asked to write an unofficial description of of a law prohibiting a crime.

As in Experiment 1, both conditions had an associated cover story explaining the motivation behind the task. As in Experiment 1, participants in the law condition were told that they were a "lawmaker" who was "tasked with writing a law that prohibits a certain crime, and specifies the punishment for that crime if the crime is committed." In the description condition, participants were told that they were a "tour guide" working in a country with strict crime laws. In order to raise awareness among foreign customers of the crime laws, they were "tasked with writing a description of the preconditions for a particular crime in your country, as well as the punishment for committing that crime."

In order to control for potential order effects, the second manipulation was ordering, whose conditions consisted of a *guilt-first* condition and a *punishment-first* condition. In the guilt-first condition, the details of the crime in question were presented such that the the requirements of guilt for the offense were presented first, followed by the punishment of the offense. In the punishment-first condition, the ordering was reversed, such that the punishment of the offense was presented first, followed by the requirements of guilt.

Unlike Experiment 1, there was no sequencing manipulation—across all conditions, the materials asked participants to write their law or description all-at-once from scratch instead of in stages.

Participants and Procedure

Participants (n=80) were recruited via the online platform Prolific. This sample size was based on a power analysis, which determined the number of participants that would give us an approximately 80% chance to detect an effect size that was at least 1/5 as large as the effect of genre obtained in Experiment 1. Participants were eligible if they resided in the United States, were 18 years or older, and native speakers of English. Each participant completed 8 trials of the same series of tasks.

On a given trial, participants were presented presented with materials in one of the four conditions, and asked to write a text of the appropriate genre. Across the 8 trials, each participant was presented with 2 items in each of the 4 conditions, never seeing the same item more than once. As in Experiment 1, participants were given a comprehension check prior to each trial, were asked before and after the experiment to promise to not use / have used a language model to generate their responses, and were retained according to the same exclusion criteria.

Analysis Plan

Responses were tokenized and coded for center-embedded syntax following the same procedure as in Experiment 1. As in Experiment 1, we analyzed the effect of our two manipulations on the prevalence of center-embedded syntax by conducting two separate regressions for each of the two operationalizations of center-embedded syntax, including (a) a mixed-effects binary logistic regression with the binary presence of center-embedded syntax (in a given sentence) as the outcome variable; and (b) a mixed-effects Poisson regression with degree of center-embedded syntax as the outcome variable. Both regressions featured (a) genre, ordering and their interaction as fixed-effects; and (b) genre, ordering, item and participant as random effects. Results did not qualitatively change for

either regression. We therefore report both in the text.

Acknowledgements

Thanks to Goeun Kim for help drafting the materials. Thank you to Rebecca Saxe for the inspiration for testing and naming the magic spell hypothesis. Thanks to Nancy Kanwisher, Roger Levy, Ev Fedorenko, Laura Schulz, and Claire Hill for helpful feedback and discussions regarding the hypotheses and experimental design. Thanks to audiences at Tedlab and Evlab, the MIT Brain Cognitive Sciences "Cognitive Lunch" talk series, and the WEIRD conference in Minneapolis, MN October 2023.

Chapter 6

Discussion

Abstract

This thesis sought to answer three broad questions related to the cognitive underpinnings of legal complexity, including: (1) How is legal language complex; (2) Has legal language gotten less complex over time; (3) Why is legal language complex. This chapter recapitulates the extent to which this thesis has answered each of these three questions, as well as the implications of these findings for broader questions of cognitive science, law and policy, including: (4) Is language inefficient; (5) Why should legal language be simple; (6) How can legal language become simple.

Note that this discussion, as with the rest of the thesis, focuses primarily on United States law. Although prior work has found converging results on a smaller scale in Canadian legal documents [41], some commentators have speculated that the language of some legal systems may be simpler than that of North American jurisdictions [134]. It remains an open question whether the results, as well as the inferences derived from those results, would hold for legal documents of countries with other legal systems (whether civil-code or other common-law jurisdictions) and/or other languages. Future work should seek to

verify this empirically, whether via corpus analyses or behavioral experiments.

6.1 How is law complex?

This thesis first set out to document the cognitive and linguistic profile of legalese—that is: (A) What are the psycholinguistically complex features that are disproportionately prevalent in legal texts relative to non-legal texts; (B) To what extent do these features collectively and individually make legal documents hard to understand for lawyers and non-lawyers; and (C) What are the cognitive mechanisms giving rise to this difficulty.

With regard to (A), prior to the work described in this thesis, there had been long-standing speculation and anecdotal accounts of the presence of certain psycholinguistically complex features in legal texts, such as low-frequency jargon, center-embedded syntax, non-standard capitalization (ALL CAPS), passive voice structures, and unusually long sentences. The corpus analyses conducted in Chapter II provided quantitative empirical support for this speculation, finding that each of these features was strikingly more prevalent in contracts relative to 9 baseline genres of written and spoken English—including academic articles, blog posts, fiction texts, non-fiction texts, spoken transcripts, newspaper articles, magazine articles, wall street journal articles, and TV/movie scripts.

Chapter III further replicated and extended these results in a separate genre of legal texts, comparing the entirety of the United States Code with six baseline genres of English from the Corpus of Contemporary American English. With just one exception (average word frequency in academic texts), all of the aforementioned features were strikingly more common in the United States Code relative to each of the baseline genres of English.

With regard to (B), the experimental evidence presented in Chapter II revealed that contracts drafted with all of these features were more difficult to

both comprehend and recall than contracts of equivalent meaning drafted without all of these features. This finding replicated across multiple experimental paradigms and held true for participants of all reading levels. Chapter IV further revealed that this effect was true not just for lay participants but for lawyers, as well—and not just for lawyers overall but for lawyers across different demographic subgroups, such as age, years of legal experience and fanciness of legal training.

In addition, contrary to the implicit assumptions of prior research and advocacy efforts [40], [49], analyses of individual linguistic structures revealed that not all of these features impact processing difficulty in legal texts equivalently. In particular, center-embedding and word frequency were found to negatively inhibit recall of legal context to a greater degree than passive voice structures and non-standard capitalization.

With regard to (C), as stated in Chapter II, some legal theorists have taken the position that "law is a system built upon expert knowledge of technical concepts," such as *habeas corpus*, *promissory estoppel*, and *voir dire* [85]. Under this account, the processing difficulty of legal texts is simply a natural result of not knowing specialized legal concepts. Others have argued that "law is a system built upon ordinary concepts," such as *cause*, *consent*, and *best interest* [85], [86]. Under this account, processing difficulty could be explained by the presence of unnecessarily complicated psycholinguistic factors.

The work presented in this thesis better aligns with an ordinary concepts account of the law. Previous work in the general psycholinguistics literature has suggested that center-embedded syntax is difficult to process due to working memory constraints, given that (a) center-embedded syntax increases the length of syntactic dependencies; and (b) long-distance syntactic dependencies force a reader to hold words in memory for longer before they arrive at a given predicate [36], [37], [135], [136].

Correspondingly, the fact that center-embedded syntax was more than twice as prevalent per sentence in both laws and contracts relative to the standard-English corpora, and inhibited recall to a greater degree than other features in our experimental study suggests that the cause of the processing difficulty of legal texts may be largely related to working memory constraints as opposed to a mere lack of understanding of specialized legal concepts.

Furthermore, the results of the corpus analyses in Chapters II, III and V, all further undermine the notion that the cognitive demands of legalese are a natural byproduct of a lack of understanding of specialized legal concepts. After all, if certain concepts are not known by those without expert legal training, then one would not expect to find many words to describe those concepts aside from the low-frequency jargon used by legal experts (just as one might not expect to see higher-frequency synonyms for terms such as *quark* or *electron* in physics, for example). Contrary to this prediction, our corpus analyses revealed that contracts and laws both contained even more cases of words with high-frequency synonyms than standard English texts, thus undercutting the view that processing difficulty is driven merely by lack of specialized knowledge. Although it is plausible that specialized concepts contribute to the perceived processing difficulty of legal texts, the results suggest that insofar as low-frequency legal terminology presents processing difficulty for laypeople, this often results not from unfamiliarity with the concept underlying that terminology but with the terminology itself (such as the low-frequency phrases *ab initio* and *ex post facto*, which in many cases respectively refer to the same concepts as the high-frequency phrases "from the start" and "after the fact").

6.2 Has law gotten simpler over time?

Having documented the profile of legalese overall, the thesis next set out to investigate the profile of legalese over time. Over the last several decades, there have been efforts on behalf of the US government to simplify legal documents for society at-large. However, prior to the work documented in this thesis, there had been no systematic evaluation of (A) how effective this so-called "plain language movement" has been; nor (B) the accessibility of legal language compared to standard English across time more generally.

Analyzing every law passed by congress between 1951 and May 2022, the corpus analyses presented in Chapter III found that top-down efforts to simplify legal language in the United States have largely been ineffectual.

In particular, our analyses revealed no evidence of a decrease in the prevalence of psycholinguistically complex features such as low-frequency jargon, center-embedded syntax, passive-voice structures and non-standard capitalization as a result of the initial onset of the plain-language movement in 1972.

Although our analyses did reveal some evidence of a decrease in some features following the enactment of the plain writing act of 2010, this was not the case for most of our features. Furthermore, even for the features for which there was a decrease, the change appeared to be unstable as opposed to steady improvement.

Moreover, on a more general level, the work presented here suggests that laws have and continue to remain difficult-to-process relative to standard English texts. For example, comparisons between laws passed by congress with matched control texts of several baseline genres of English published between 1951 and 2009 revealed that for virtually any given year, laws contained strikingly higher rates of psycholinguistically complex structures relative to any of our baseline genres of English.

In addition, when comparing the current edition of the US Code with con-

temporary texts from six baseline genres, we found that as a general matter laws continue to remain laden with difficult to process features at strikingly higher rates than other genres.

Taken together, these results suggest that laws remain strikingly more difficult to process than standard English texts, substantiating previous anecdotal accounts of the efficacy of plain-language efforts made by plain-language advocates, who have described progress as "way slow" and acknowledged that "much remains to be done to improve" [49].

6.3 Why is law complex?

Having documented why legalese is difficult to understand on the compehender side, the thesis next sought to investigate this question from the producer side—that is, why do lawyers and lawmakers write this way in the first place.

Scholars and commentators had proposed approximately five sets of hypotheses/explanations for why legal language is so complicated to understand.

For example, some scholars had speculated, in line with what has been dubbed the "curse of knowledge" in other disciplines [61], [62], that the reason legal language is so difficult to understand is because lawyers do not realize that they write in an esoteric manner [63].

Under what we have dubbed the "copy-and-paste" hypothesis, other commentators speculated that lawyers simply write in a complex register out of "habit, laziness" [64] or respect for "tradition" [39], that they "copy and paste" [65] from existing templates with old, complicated terms because that's the "quickest and cheapest way to produce a contract"[66].

A third hypothesis, referred to in this thesis as the "in-group signaling" hypothesis, maintained that lawyers write in legalese to be accepted by their peers, to sound more "lawyerly," to "mark themselves as members of the profession"

[64].

A fourth hypothesis, labeled here as the "it's just business' hypothesis, argued that lawyers write in legalese as a way of "preserving their monopoly" [67] on legal services and "justifying fees" [64].

Finally, according to the *complexity of information* hypothesis, legal language needs to be complex in order to satisfy certain communicative aims, such as conveying complex legal concepts in a way that "is far more precise than ordinary language" [39], to avoid ambiguity, and/or to ensure enforceability.

Prior to this thesis, there remained no empirical evaluation of these hypotheses. The work presented in Chapter IV sought to rectify this gap via 2 pre-registered studies that tested the predictions of each of these five hypotheses.

In Study 1, contrary to the predictions of the curse of knowledge hypothesis, we found that lawyers, like laypeople, were less able to recall and comprehend legal content drafted in a complex "legalese" register than content of equivalent meaning drafted in a simplified register. In Study 2, contrary to the in-group signaling, it's just business, and complexity of information hypotheses, we found that lawyers rated simplified contracts as equally enforceable as legalese contracts, and rated simplified contracts as preferable to legalese contracts on several dimensions–including willingness of the lawyer to hire the contracts' author, appropriateness of the contract's style, and likelihood of the contract being signed by a client. In contrast, lawyers' preference for simplified contracts was consistent with the predictions of the copy-and-paste hypothesis.

Although the findings of this study lend support for the idea that lawyers copy and paste complex structures from existing templates, it still remained an open question how these complex structures enter into legal documents in the first place.

To further investigate this question, Chapter V proposed and presented two additional hypotheses for why lawyers write in a complex manner, including

(A) the magic spell hypothesis; and (B) the copy-and-edit hypothesis. Consistent with the magic spell hypothesis, two pre-registered experiments found that people tasked with writing official laws wrote in a more convoluted manner than when tasked with writing unofficial legal texts of equivalent conceptual complexity. Contrary to the *copy-and-edit hypothesis*, we did not find evidence that people editing a legal document wrote in a more convoluted manner than when writing the same document from scratch.

These findings indicate law's complexity to be derived from its performativity, whereby certain low-frequency structures, such as center-embedded syntax and low-frequency jargon, may be inserted to signal law's authoritative, world-state-altering nature, at the cost of increased processing demands on readers.

It remains an open question whether the presence of these low-frequency structures is a mere byproduct of tradition, or whether there's something about complexity of syntax and jargon that makes legal documents sound inherently more official or authoritative.

It is also conceivable that there are some hypotheses that we did not directly test here that might help to explain why law is more complicated than everyday speech. One such account, which we might refer to as the "adversarial interpreter hypothesis," is that the heightened linguistic complexity of law is a byproduct of the adversarial (as opposed to cooperative) nature in which law is interpreted relative to ordinary language. Under this account, to prevent parties from exploiting ambiguities in the text, legal actors may insert complex structures (so as either to remove the ambiguities or simply make them more difficult to detect).

Moreover, it is also plausible that some hypotheses which were undermined in this work may not be fully ruled out as at least partial accounts for the heightened complexity of the law. For example, in line with (a weaker form of) the complexity of information hypothesis, it is possible that some legal content

requires the use of complicated jargon and/or syntax at a higher rate than typical English communication, while in some cases—in line with the it's just business hypothesis—it might be in a lawyer's best financial interest to make legal content overly complicated for the reader.

Furthermore, it is possible that different hypotheses might serve as plausible accounts for the presence of some features of legalese (e.g. complicated jargon) but not others (e.g. center-embedded syntax). Future work might seek to more precisely disentangle these accounts.

6.4 Is legal language inefficient?

From a cognitive standpoint, one implication of this work is in informing the degree to and domains in which human language is optimized for communicative efficiency. In particular, this thesis suggests that legal language may be a rare exception to the general trend of human language towards communicative efficiency.

As mentioned in the introduction, there is a burgeoning psycholinguistics literature documenting the various ways in which language is optimized for easing the cognitive burden on both producers and comprehenders during the communication of an utterance. One of the hallmarks of communicative efficiency is syntactic dependency length minimization, whereby words in a sentence that depend on each other, both in their interpretation and in their statistical distribution, tend to be close to each other in linear order [136]. For example, in prior work, evidence from dozens of languages across disparate families has indicated that utterances in human language generally have lower-than-expected syntactic dependencies relative to different baselines [36].

Contrary to this tendency, the work presented in this thesis discovered that words in legal documents have unusually long syntactic dependencies. For ex-

ample, as revealed in Chapters II and III, legal documents of various genres are laden with center-embedded syntax—long observed to be associated with long-distance syntactic dependencies—at several times the rate of other baseline genres of English, including those aimed at an educated audience. Chapter V presented more direct evidence of long-distance syntactic dependencies, finding that laws had strikingly longer dependencies on average than six baseline genres of English, even when adjusting for sentence length.

Moreover, Chapter V also presented evidence that this tendency was not merely a natural consequence of heightened conceptual complexity, as center-embedded structures were several times more prevalent in official legal documents than unofficial legal documents of equivalent conceptual complexity. The fact that the only difference between the unofficial legal documents and official legal documents was with respect to their performativity suggests that performativity is the key driver of law's long syntactic dependencies and, by extension, of its communicative inefficiency. In particular, given that the key distinction between performative utterances and descriptive utterances is that performative utterances are meant to alter the state of the world instead of communicate states of the world, this suggests that distinctive low-frequency structures such as center-embedded syntax and esoteric jargon may be inserted so as to effectively signal that a given legal document is not only communicating but altering legal rights and obligations.

In exchange, this may impose disproportionately high processing demands on a reader relative to a non-performative document otherwise conveying the same relevant legal content.

In addition to law, this thesis suggests other domains in which language may be less optimized for communicative efficiency. In particular, given that other other types of performative utterances (such as magic spells) similarly contain low-frequency structures compared to non-performative control texts, this sug-

gests that performative utterances more generally may stand as an exception to the general tendency of human language towards communicative efficiency.

That said, as noted in Chapter IV, these results do not suggest that legal documents can be simplified limitlessly without sacrificing communicative aims. For example, the prevalence of center-embedding in unofficial legal documents in Experiment II of Chapter V, though twice as low as in official legal documents, was several times higher than in crime stories. One explanation for this discrepancy is that the legal content in both official and unofficial legal documents may be of higher conceptual complexity than that of crime stories, and that this heightened conceptual complexity necessitates heightened syntactic complexity.

Moreover, it is also worth noting that some notions of communicative efficiency might be broad enough to encapsulate performativity. If so, then insofar as certain low-frequency structures are necessary to satisfy certain performative aims, this would imply that the inclusion of such structures within certain performative utterances would not violate communicative efficiency per se, even if those structures led to increased processing demands over non-performative utterances of otherwise equivalent conceptual complexity.

However, given that performative utterances are by definition generally contrasted with utterances that merely communicate states of the world, it seems reasonable to adopt a narrower definition of communicative efficiency that excludes or distinguishes performative aims from communicative aims. If so, utterances with difficult-to-process structures whose inclusion over easier-to-process structures is necessary only to advance performative aims would be appropriately characterized as inefficient.

By extension, given the evidence presented in this thesis, it follows that one might appropriately characterize law as inefficient under this account.

6.5 Why should law be simplified?

Before turning to the question of how best to integrate these findings, it is worth reviewing why lawyers and lawmakers should care to integrate these findings.

One answer is to point out that there are several principles of modern legal doctrines that mandate that they be comprehensible to the reader.

For example, under the *fair notice* principle of criminal and constitutional law, laws are mandated to provide proper warning of prohibited conduct "in language that the common world will understand," [1], [2] to "give the person of ordinary intelligence a reasonable opportunity to know what is prohibited, so that he may act accordingly." [3], [4]. Jurists have recently argued that such a fair notice principle may plausibly be satisfied only if ordinary people are able to "read and understand the law for themselves, without need to absorb distinctively legal training" [5], [6]. If so, then given that this work has revealed that ordinary people are often not able to understand the law for themselves, without need to absorb distinctively legal training, this indicates that many such laws ought to be invalid according to existing legal doctrine. It follows that in order to restore the validity of these laws and the legitimacy of the fair notice doctrine, laws ought to be simplified to be more in line with baseline genres of English.

Similar implications of this work derive from the *ordinary meaning doctrine*, which requires that words in legal documents typically be interpreted according to how they are ordinarily understood by laypeople [2], [7], [10], [11], [15]. As revealed in this work words in legal documents are often not ordinarily understood by laypeople (nor, in some cases, by lawyers), thereby undermining the coherence of the ordinary meaning doctrine. To the extent that legal actors are committed to the coherence of legal doctrine, it follows that the legal documents ought to be simplified so as to make them ordinarily understood by laypeople.

A third example relates to the philosophy of textualism, which has not only become the dominant interpretive approach of the United States Supreme Court [6], but has also become increasingly prevalent in the American judiciary writ-large [20]; and is even widely endorsed by legal academics, as well [21]. According to Justice Amy Coney Barrett, textualists "view themselves as agents of the people rather than of Congress" and "approach language from the perspective of an ordinary English speaker." [22] Given that ordinary English speakers appear unable to comprehend legal documents relative to non legal documents, it follows that this would likewise undermine the practice of legal interpretation as exercised by contemporary legal officials. It therefore follows, both from a theoretical standpoint and as a practical matter, that legal officials ought to be in favor of simplifying the encoding of legal content.

Indeed, work from this thesis has revealed that legal actors not only should care about making legal documents understandable from a doctrinal standpoint but in fact do care from a professional standpoint. The experimental evidence in Chapter IV revealed that lawyers, like laypeople, struggle to understand legal content written in a complex register relative to the same content drafted in a simplified register. Chapter IV also revealed that lawyers dispreferred complex legal documents to simplified legal documents on virtually every dimension, including overall quality, appropriateness of style, and likelihood of being signed by a client. Finally, this work also revealed that simplified legal documents were no less enforceable than complex legal documents, indicating that lawyers not only preferred simplified legal documents in the abstract but found them desirable from a practical standpoint as well. If even lawyers disprefer legalese, then it seems it is in their interest to dispense with it.

Given these reasons, it is perhaps no surprise that lawmakers have in fact attempted to make legal language easier to understand. These efforts in the United States, referred to as the "plain-language movement'', date back at least

as early as the 1970s, when Richard Nixon mandated that the Federal Registry be drafted in "layman's terms" and Jimmy Carter issued Executive Orders intended to make government regulations "easy-to-understand by those who were required to comply with them" [48], [49]. One of the more recent call-to-arms, the Plain Writing Act of 2010, attempted to establish formal guidelines regarding how to write government documents clearly for a lay audience [52].

Of course, these efforts not only benefit legal actors tasked with drafting and interpreting laws but also the laypeople tasked with complying with them. As mentioned previously, written laws are the means through which societal rules and norms are established and communicated across modern society.

Although the work presented in this thesis suggests that efforts to simplify legal language have failed, the existence of these efforts demonstrates the recognized importance on behalf of legal actors to make law comprehensible to those who are required to comply with it.

6.6 How can law be simplified?

Assuming one is on board with the aforementioned reasons for why law should be easier to comprehend, the next question is how to make law easier to comprehend. The work presented in this thesis has informed these efforts in several ways.

The first way derives from advancing our understanding of which features, cognitively and linguistically, are most likely to contribute to the difficulty faced by readers when reading legal documents. Prior to this work, plain-language guidelines had implicitly assumed that the stereotypical markers of legalese—passive voice, jargon, non-standard capitalization, and center-embedded syntax—contributed more-or-less equivalently to the processing difficulty of legal texts [49].

Contrary to this assumption, this thesis has revealed that some features, such as center-embedded syntax and jargon, disproportionately inhibit understanding and recall of legal content relative to other features, such as passive voice and non-standard capitalization. Correspondingly, these insights suggest that, all-else-equal, plain-language efforts should disproportionately focus on eliminating center-embedded syntax and low-frequency jargon as opposed to focusing equivalently on the prevalence of passive voice and non-standard capitalization.

Moreover, by revealing, contrary to previous speculation [67], that these features inhibit recall and comprehension of lawyers and laypeople of all reading levels, this work has underscored that simplifying legal documents would be beneficial for all as opposed to merely a subset of the lay population.

A second way in which this work can inform efforts to simplify legal language derives from advancing our understanding of the cognitive pressures leading legal experts to write this way in the first place.

For example, prior to this work, many commentators had speculated or assumed that lawyers prefer or are otherwise forced to write in a complex manner in order to satisfy certain communicative aims, to sound more lawyerly, or to justify exorbitant fees to clients. Were this to be true, one might pessimistically infer that simplifying legal language is an intractable affair, especially given the failures of many top-down efforts to simplify legal language in the past. The results of the this thesis work instead paint a more optimistic picture, indicating that lawyers (a) believe legal documents can and should be simplified to better serve their communicative aims; and (b) like laypeople, struggle to comprehend complex legal language relative to a simpler alternative. The results suggest that the processing difficulty of legal texts may be alleviated as lawyers and lawmakers become more aware of both the ways in which public legal documents tend to be complex, as well as the alternatives available to them in order to make

them less complex.

In particular, this suggests that the creation and adoption of plain-language templates may be a plausible avenue for change, given that (a) the results indicate that legal texts are laden with overly complex structures that can feasibly be replaced with easier-to-process alternatives; and (b) there exist several proofs of concept for the idea that large, real-world legal documents can be simplified without a loss or distortion of meaning [137], [138].

Although the idea of creating plain-language templates at scale may seem like a daunting affair, the rise of AI presents a potentially promising method of automating the practice of creating simplified legal texts, given that AI models have been shown to (a) possess high levels of legal knowledge [139], [140]; and (b) flexibly change the register of an inputted text [141], [142]. Although recent evidence suggests that even the most cutting-edge AI models remain below the level of an average barred attorney when it comes to generative tasks such as essay writing [140], it seems plausible to assume given the rapid advancement of AI that this possibility will become more feasible within the relatively near future.

Note that some might be tempted to draw a different conclusion from the magic spell hypothesis results of Chapter V. After all, if (a) low-frequency structures such as center-embedded syntax are necessary to signal law's authoritative, world-altering nature; and (b) the authoritative, world-altering nature is a necessary component of the law; then one might conclude that (c) low-frequency structures such as center-embedded syntax are a necessary component of legal documents.

The response to this objection is twofold. First, although low-frequency structures may sound more authoritative on the producer side, they do not appear to have equivalent or similar effect on the comprehender side. As revealed in Chapter V, lawyers rated plain contracts as better or equivalent to legalese

contracts on several desiderata, including a client's likelihood to agree to sign the agreement (i.e. agree to comply with its world-state-altering terms). To the extent that authoritativeness is a necessary condition for law on the comprehender side as opposed to the production side, it would follow that low-frequency structures are not a necessary condition to sufficiently signal law's authority to the reader.

Second, not all low-frequency structures have equivalent impact on processing difficulty. As revealed in this thesis, structures such as passive voice, ALL-CAPS, jargon and center-embedding were all strikingly infrequent in baseline English relative to legal texts. Yet they did not all inhibit comprehension or recall to the same degree. Similarly, other forms of performative utterances, such as magic spells, do not seem to be intuitively more difficult to process than their descriptive counterparts. Thus, to the extent that low-frequency structures are in fact necessary to sufficiently signal law's authority to the reader, the results suggest that their inclusion does not necessarily have to result in increased processing demands on readers.

Appendix A

Supplemental Information for *Poor Writing, Not Specialized Concepts, Drives Processing Difficulty in Legal Language*

A.1 Corpus Analysis

A.1.1 Methods

Documents were first tokenized into separate sentences using the Stanza natural language package. The number of tokenized sentences in the law and standard-English corpora was 115,287 and 884,221, respectively. Afterwards, we filtered out sentences as described in Table B.1.

As discussed in the main text, the standard-English corpus consisted of (a) a sample of Wall Street Journal articles published in 1996, as part of the CSR Wall Street Journal corpus [71]; and (b) a broad sample ($n \approx 10 million words$) of TV/Movie scripts, spoken language, newspaper articles, blogs, magazine articles

Filter	Legal Corpus	COCA
sentences without punctuation	73,486	756,452
sentences with 3+ consecutive punctuation marks	69,282	754,381
duplicate sentences	56,888	646,669
sentences with fewer than 10 words	44,687	455,488

Table A.1: Filtering Processes and number of remaining sentences for each corpus.

and web pages from the Corpus of Contemporary American English (COCA: davies2009385+). The number of sentences in each of these subgenres post-filtering is given in Table B.2.

Academic articles (COCA)	34,385
Blog posts (COCA)	43,905
Fiction (COCA)	38,503
Magazine articles (COCA)	43,587
Newspaper articles (COCA)	37,259
Spoken transcripts (COCA)	28,211
TV/Movies (COCA)	28,685
Wall Street Journal articles (CSR)	106,865
Web pages (COCA)	39,242

Table A.2: Sentences in each subgenre post-filtering

With regard to the filtering steps, we removed sentences without punctuation, as well as those with fewer than 10 words so as to remove headings in the contract corpus, which are not really sentences but would otherwise be counted as such without this filter. The removal of 3 consecutive punctuation marks was added as a filter so as to get rid of more non-sentences in both corpora. The duplicate sentences filter was added to remove the high number of repeat sentences in the standard-english corpus.

Word frequency. To perform this calculation, we looked at all the words in the corpora marked as a verb, noun, adjective or adverb according to Stanza. We then looked at how frequently each of these words appeared in the SUBTLEX word frequency dictionary, a corpus of American film subtitles commonly used

as a proxy for standard-English word frequency. Proper nouns and other words that did not appear in the corpus received a score of "NA." Main results are reported in the main text. Within each subgenre, the average frequency value is given in Table A.3.

Academic articles (COCA)	24,780.19 (95% CI: 24,508.08 to 25,069.94)
Blog posts (COCA)	58,073.56 (95% CI: 57,691.94 to 58,513.68)
Fiction (COCA)	56,350.24 (95% CI: 55,888.12 to 56,808.25)
Magazine articles (COCA)	38,116.42 (95% CI: 37,787.74 to 38,432.42)
Newspaper articles (COCA)	39,599.17 (95% CI: 39,242.96 to 39,962.11)
Spoken transcripts (COCA)	86,742.98 (95% CI: 86,068.52 to 87,411.30)
TV/Movies (COCA)	103,500.46 (95% CI: 102,689.78 to 104,314.84)
Wall Street Journal articles (CSR)	50,159.95 (95% CI: 49,752.73 to 50,537.04)
Web pages (COCA)	29,991.84 (95% CI: 29,818.63 to 30,169.14)
Contracts	18,753.08 (95% CI: 18,557.23 to 18,962.70)

Table A.3: Average Frequency

Word choice. We performed this calculation using two separate methods under the assumption that 1) legal register word choice is not restricted by precision and 2) legal concepts are restricted by precision—as it is often claimed that legal terms often resemble common words in form but have a more specialized meaning, such as the concept of "consideration" in contract law [73].

Under the first assumption, we looked at the same group of words included in the word-frequency analysis and for each of those words: (a) looked at the most common meaning/sense of that word according to WordNet (so as to determine what the authors most likely intended to say by using that word); (b) using the meaning/sense obtained in (a), looked at all possible synonyms of that word according to WordNet (i.e. assuming the authors meant to use the most common meaning, what other words could they have used instead); (c) computed the SUBTLEX frequency value of each of these other synonyms; and (d) coded whether the SUBTLEX frequency value of any of the synonyms was higher than that of the actual word used in the text (if yes, we coded that word

as having a 'better synonym' / 'higher-frequency synonym')

Using this method, we find that 13.747% of the words in the contract corpus were determined to have a higher-frequency synonym (95% CI: 13.701 to 13.792), as compared to 10.978% in the standard-English corpus (95% CI: 10.958 to 10.999). Within the standard-English corpus, the percentage of words with a higher-frequency synonym in each subgenre is given in Table A.4.

Academic articles (COCA)	14.707 (95% CI: 14.625 to 14.784)
Blog posts (COCA)	10.789 (95% CI: 10.722 to 10.856)
Fiction (COCA)	9.883 (95% CI: 9.813 to 9.957)
Magazine articles (COCA)	12.078 (95% CI: 12.011 to 12.151)
Newspaper articles (COCA)	11.275 (95% CI: 11.206 to 11.345)
Spoken transcripts (COCA)	8.941 (95% CI: 8.855 to 9.021)
TV/Movies (COCA)	7.167 (95% CI: 7.082 to 7.246)
Wall Street Journal articles (CSR)	13.592 (95% CI: 13.547 to 13.637)
Web pages (COCA)	11.333 (95% CI: 11.263 to 11.402)
Contracts	17.253 (95% CI: 17.188 to 17.314)

Table A.4: Percentage of words with a higher-frequency synonym in each subgenre (first method)

When considering only content words, the proportion of words with a higher-frequency synonym was 29.040% in the contract corpus (95% CI: 28.957 to 29.120) and 25.732% in the standard-English corpus (95% CI: 25.650 to 25.816). Within the standard-English corpus, the percentage within each subgenre is given in Table A.5.

Under the assumption that legal concepts are constrained by precision, we followed the same steps and words except that for Step (a), we looked at the least common meaning/sense of a given word instead of the most common word. Results for the contracts versus standard-English corpus reported in main text. Within the standard-English corpus, the percentage of words with a higher-frequency synonym in each subgenre using this assumption is given in Table A.6.

Looking only at improper words, the percentage within each subgenre ac-

Academic articles (COCA)	26.452 (95% CI: 26.319 to 26.586)
Blog posts (COCA)	21.262 (95% CI: 121.141 to 21.387)
Fiction (COCA)	19.956 (95% CI: 19.819 to 20.092)
Magazine articles (COCA)	22.800 (95% CI: 22.683 to 22.932)
Newspaper articles (COCA)	22.225 (95% CI: 22.118 to 22.381)
Spoken transcripts (COCA)	19.598 (95% CI: 15.906 to 16.241)
TV/Movies (COCA)	16.069 (95% CI: 15.906 to 16.241)
Wall Street Journal articles (CSR)	26.102 (95% CI: 26.017 to 26.176)
Web pages (COCA)	22.156 (95% CI: 22.032 to 22.291)
Contracts	30.245 (95% CI: 30.143 to 30.341)

Table A.5: Percentage of content words with a higher-frequency synonym in each subgenre (first method)

Academic articles (COCA)	11.042 (95% CI: 10.967 to 11.114)
Blog posts (COCA)	7.505 (95% CI: 7.453 to 7.569)
Fiction (COCA)	6.669 (95% CI: 6.630 to 6.747)
Magazine articles (COCA)	8.805 (95% CI: 8.754 to 8.864)
Newspaper articles (COCA)	7.645 (95% CI: 7.585 to 7.703)
Spoken transcripts	5.521 (95% CI: 5.458 to 5.581)
TV/Movies (COCA)	4.496 (95% CI: 4.429 to 4.559)
Wall Street Journal articles (CSR)	9.416 (95% CI: 9.378 to 9.454)
Web pages (COCA)	8.022 (95% CI: 7.964 to 8.083)
Contracts	13.437 (95% CI: 13.380 to 13.491)

Table A.6: Percentage of words with a higher-frequency synonym in each subgenre (first method)

cording to this assumption is given in Table A.7.

Academic articles (COCA)	19.860 (95% CI: 19.738 to 19.986)
Blog posts (COCA)	14.796 (95% CI: 14.690 to 14.904)
Fiction (COCA)	13.508 (95% CI: 13.399 to 13.622)
Magazine articles (COCA)	16.622 (95% CI: 16.512 to 16.729)
Newspaper articles (COCA)	15.087 (95% CI: 14.977 to 15.192)
Spoken transcripts (COCA)	12.101 (95% CI: 11.965 to 12.224)
TV/Movies (COCA)	10.081 (95% CI: 9.946 to 10.229)
Wall Street Journal articles (CSR)	18.082 (95% CI: 18.009 to 18.151)
Web pages (COCA)	15.684 (95% CI: 15.566 to 15.811)
Contracts	23.556 (95% CI: 23.461 to 23.638)

Table A.7: Percentage of content words with a higher-frequency synonym in each subgenre (first method)

Capitalization. Here we sought to determine what percentage of words in contracts were in ALL CAPS relative to standard English. To do so, we looked at all of the alphabetic words in each of our corpora and calculated the proportion of words in each corpus that were marked by Stanza as being entirely in uppercase letters. Main results reported in the main text. Within the standard-English corpus, the percentage within each subgenre is given in Table A.8.

Academic articles (COCA)	1.500 (95% CI: 1.474 to 1.529)
Blog posts (COCA)	2.346 (95% CI: 2.314 to 2.380)
Fiction (COCA)	2.252 (95% CI: 2.217 to 2.286)
Magazine articles (COCA)	1.805 (95% CI: 1.774 to 1.833)
Newspaper articles (COCA)	1.290 (95% CI: 1.265 to 1.315)
Wall Street Journal articles (CSR)	2.021 (1.991 to 2.051)
Web pages (COCA)	1.310 (95% CI: 1.295 to 1.326)
Contracts	2.780 (95% CI: 2.753 to 2.806)

Table A.8: Percent Capitalization

Passive-voice structures. To compute the prevalence of passive voice structures as a whole in both corpora, we calculated the number of words marked with the passive voice features in Stanza. To compute the prevalence of by-passive structures, we performed the same calculation and then looked at the number of passives that had the word *by* in the same head according to Stanza. Main results of each of these are reported in the main text. Within the standard-English corpus, the percentage of passive structures within each subgenre is given in Table A.9. The number of by-passive structures in each subgenre is given in Table A.10:

Center-embedded clauses. To determine the number of embedded clauses (both center-embedded and right-branching) as a whole, for every sentence in each corpus we looked at the number of predicate dependent clauses (i.e. clausal subjects, clausal complements, open clausal complements, adjectival clauses, and adverbial clauses). To determine the number of center-embedded clauses,

Academic articles (COCA)	29.920 (95% CI: 29.358 to 30.532)
Blog posts (COCA)	16.861 (95% CI: 16.453 to 17.241)
Fiction (COCA)	8.728 (95% CI: 8.430 to 9.045)
Magazine articles (COCA)	15.238 (95% CI: 14.839 to 15.613)
Newspaper articles (COCA)	16.745 (95% CI: 16.319 to 17.200)
Spoken transcripts (COCA)	12.735 (95% CI: 12.305 to 13.183)
TV/Movies (COCA)	6.230 (95% CI: 6.006 to 6.593)
Wall Street Journal articles (CSR)	18.382 (95% CI: 17.939 to 18.844)
Web pages (COCA)	17.393 (95% CI: 17.148 to 17.633)
Contracts	75.762 (95% CI: 74.676 to 76.822)

Table A.9: Percent Passive Structures

Academic articles (COCA)	4.395 (95% CI: 4.186 to 4.611)
Blog posts (COCA)	2.229 (95% CI: 2.084 to 2.376)
Fiction (COCA)	1.011 (95% CI: .909 to 1.109)
Magazine articles (COCA)	2.191 (95% CI: 2.065 to 2.325)
Newspaper articles (COCA)	2.396 (95% CI: 2.240 to 2.550)
Spoken transcripts (COCA)	1.626 (95% CI: 1.466 to 1.765)
TV/Movies (COCA)	.566 (95% CI: .471 to .646)
Wall Street Journal articles (CSR)	2.655 (95% CI: 2.493 to 2.820)
Web pages (COCA)	3.372 (95% CI: 3.258 to 3.484)
Contracts	14.076 (95% CI: 13.643 to 14.464)

Table A.10: Percent By-Passive Structures

we performed the above calculation and then looked at whether the clause was followed by a word as opposed to an end-of-sentence punctuation mark. Within each sub-genre of standard-English, the number of embedded clauses is given in Table A.11. The number of center-embedded clauses within each subgenre is given in Table A.12:

Experiment Methods

All experiment code and data can be found at OSF:
https://osf.io/xcqd9/?view_only=325a9567b2f54dc99eff8e8d5683e1bf

Academic articles (COCA)	81.741 (95% CI: 80.683 to 82.832)
Blog posts (COCA)	93.209 (95% CI: 92.234 to 94.148)
Fiction (COCA)	72.321 (95% CI: 71.427 to 73.280)
Magazine articles (COCA)	81.940 (95% CI: 81.006 to 82.827)
Newspaper articles (COCA)	89.293 (95% CI: 88.298 to 90.301)
Spoken transcripts (COCA)	90.244 (95% CI: 89.033 to 91.508)
TV/Movies (COCA)	57.218 (95% CI: 56.317 to 58.084)
Wall Street Journal articles (CSR)	86.556 (95% CI: 85.514 to 87.524)
Web pages (COCA)	94.880 (95% CI: 94.369 to 95.426)
Contracts	177.636 (95% CI: 175.495 to 179.912)

Table A.11: Embedded Clauses

Academic articles (COCA)	28.689 (95% CI: 28.054 to 29.344)
Blog posts (COCA)	38.180 (95% CI: 37.522 to 38.821)
Fiction (COCA)	20.748 (95% CI: 20.254 to 21.278)
Magazine articles (COCA)	27.667 (95% CI: 27.078 to 28.240)
Newspaper articles (COCA)	36.999 (95% CI: 36.323 to 37.690)
Spoken transcripts (COCA)	19.666 (95% CI: 19.048 to 20.244)
TV/Movies (COCA)	10.420 (95% CI: 10.022 to 10.825)
Wall Street Journal articles (CSR)	22.064 (95% CI: 21.778 to 22.347)
Web pages (COCA)	34.700 (95% CI: 33.995 to 35.356)
Contracts	72.903 (95% CI: 71.492 to 74.326)

Table A.12: Center-Embedded Clauses

A.1.2 Annotation details

Two trained research assistants coded whether a proposition was successfully recalled. In doing so, they were presented with a participant's retelling of a passage and then asked whether each legally relevant proposition of the passage was (a) fully recalled; (b) partially recalled; or (c) not recalled. Coders were told that for a response to count as "fully recalled," it did not have to be recalled verbatim (i.e. they can use their own words or syntax), so long as they were confident that the meaning of what subject wrote is the same as the proposition.

For example, suppose the original text said "A court in Boston will resolve the dispute," and the participant wrote "something will be resolved by a court."

When coding responses, a coder might see three propositions that say: (i) "A court in Boston," (ii) "will resolve," and (iii) "the dispute."

For (i), the coder would put a 0.5 for "partially recalled" (since "in Boston" was missing from "a court"); for (ii), the coder would put a 1 for "fully recalled" (since "will be resolved" means basically the same thing as "will resolve"), and for (iii), the coder would put a 0 for "not recalled" (since "the dispute" was not in the response).

To reduce potential bias, coders were unaware of whether a participant had seen or recalled the simple or legalese version of a text. Moreover, the rubric that the coders used to score participants' responses was the legalese version: for each proposition, they were given the language of that proposition in legalese, and were told to score a participants' response as having recalled that proposition if it had language that had the same meaning as that proposition. Thus, any differences in recall favoring the Plain English version would arise in spite of the coding bias, which was towards the Legalese version.

Of the roughly 650 retellings, each coder was responsible for coding roughly 60 percent ($\approx$ 390) of the retellings, such that (a) every retelling/proposition would be coded at least once, and (b) 20% of the retellings would be coded by both coders so as to assess inter-rater reliability. Coder reliability was assessed with Cohen's kappa coefficient [81], [82].

We adjudicated ties as follows: (i) a tie between one "fully recalled" judgment and one "not recalled" judgment resulted in a final "partially recalled" judgment; (ii) a tie between one "fully recalled" judgment and one "partially recalled" judgment resulted in a final "fully recalled" judgment for a given proposition; and (iii) a tie between one "partially recalled" judgment and one "not recalled" judgment resulted in a final "not recalled" judgment. For our regression analyses, we perform both a conservative analysis (recoding "partially recalled" as "not recalled') and an anti-conservative analysis (recoding "partially

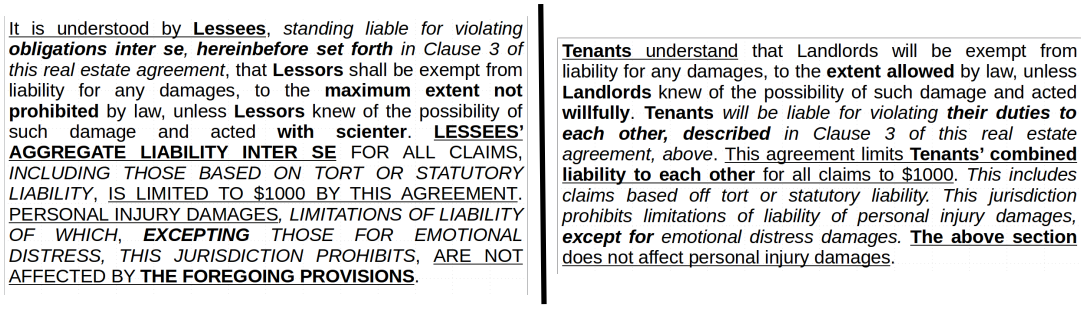

Figure A.1: An example stimulus pair in legalese (left) and simple (right) register. The differences in surface properties across registers are depicted by font style. Bold denotes word frequency. Italic denotes embedded clauses. Underlined denotes voice. Unfortunately, we have run out of font styles to make differences in capitalization more apparent.

recalled" as "fully recalled"). Our results do not qualitatively change, so we will only report the conservative analysis here.

A.1.3 Anti-conservative recall analysis

Again following masson1994comprehension, we expect participants to recall fewer legal propositions when the text is presented in a legal register compared to a simple register. We further predicted an interaction with language experience, such that recall would be worse for people with less language experience. Descriptively, propositions from excerpts in a simple register (49.3%) were recalled more than propositions presented in a legalese register (41.4%).

We conducted a mixed effect logistic regression with register (sum coded), standardized ART score and their interaction as fixed effects and excerpt and participant as random effects with register as a random slope for each. Using likelihood ratio tests, we fail to find a significant interaction and, thus, report here a simpler model without the interaction term. Again replicating masson1994comprehension, fewer legal propositions were recalled when they were presented in a legalese register compared to a simple register ($\beta = -0.18523$, SE$= 0.050, p < 0.05$). We do not find an effect of language experience on recall.

A.2 Pilot Power Analysis

Our primary concern was ensuring we had enough power to detect a reliable difference in recall across registers as previous experience suggests that there is an upper bound limit to human performance on recall of propositional content [83]. Therefore, we ran a pilot study focusing on the recall task. We recruited 32 participants for the pilot. Three participants were removed before recall was hand-coded as they copied the text word for word.

The same coders used in the main experiment hand-coded recall. For the power analysis, disagreements were coded as not being recalled. We fit a mixed effect logistic regression model predicting recall as a function of register with random slopes and intercepts for both participant and contract. Using the effect size estimated under this model ($\beta = 0.327$), we used the simr package [143] in R to conduct a power analysis to verify that 100 participants (our feasible sample size) would be sufficiently powered to detect a main effect of register on recall. We used the default z-test at an $\alpha = 0.05$ for 1000 simulations. Based on our simulations, we estimated the power to be 87.9% (CI: 85.72 - 89.86).

A.3 Additional Analyses of Comprehension Data

A.3.1 Robustness Check

In our original analysis, for 94.5% of question items, mean accuracy was above chance (25%) in both versions. To further ensure that our main effect of register was not driven by certain items that were systematically interpreted by participants as containing a different meaning in the simple register versus the legalese register, we conducted an additional regression model in which we filtered out question items in which participants' overall comprehension in either version was below 25%. In this model, we still find a main effect of register ($\beta = .165$,

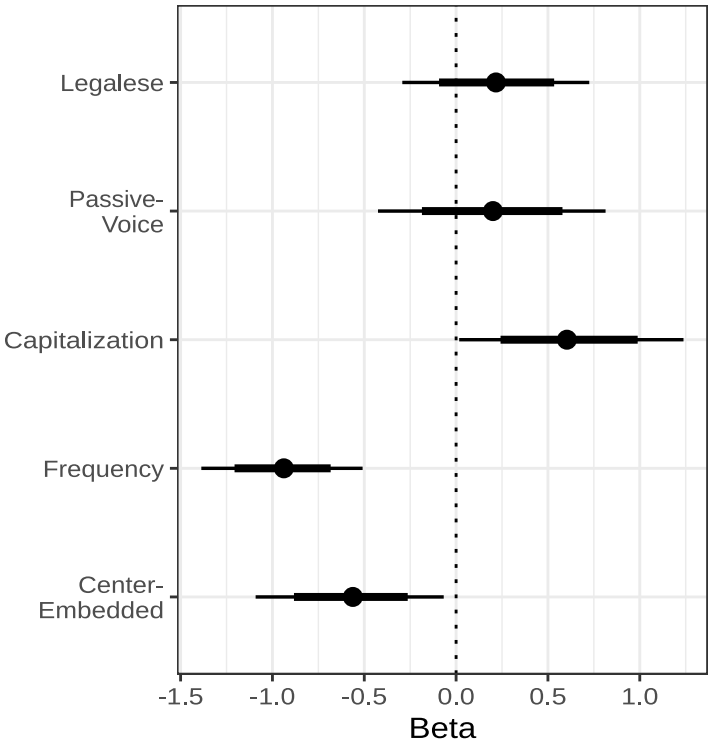

Figure A.2: Effects of linguistic structures on comprehension. Outer line range reflects the 95% credible intervals over the interaction term, inner line range reflect the 80% credible intervals over the interaction term and points reflect medians.

SE$= 0.049, p < 0.05$).

A.3.2 Exploratory Analyses of Linguistic Features

To explore the influence of our surface properties on comprehension behavior, we conducted a similar exploratory analysis. Of the 170 comprehension questions, 106 of the comprehension questions tested content that differed in surface features across registers. For each of these questions, we included a main effect of condition and coded the interactions between condition and center-embedding, voice, word frequency or capitalization. We conducted a mixed effect logistic regression predicting comprehension accuracy with condition and surface form properties as a fixed effects and random intercepts for excerpt and participant with complete random slopes. Our fixed effect was coded so that each coefficient reflects either an increase or decrease in comprehension accuracy for a legalese register relative to the average recall rate of the simple register. Figure A.2 represents the 95% and 80% credible intervals over the regression coefficient for each surface property. As in recall, we find an effect of center-embedding such that comprehension questions that were presented with center-embedding (i.e., in the Legalese condition) were responded to less accurately than when un-embedded (i.e., in the Simple register). We also find a main effect of frequency such that comprehension questions that were presented with lower frequency words (i.e., in the Legalese condition) were responded to less accurately than with higher frequency words (i.e., in the Simple register). Further, we find a main effect of capitalization such that comprehension questions whose content was presented capitalized (as in the Legalese register) were answered more accurately than when presented in standard case (as in simple register). Overall, these results suggest that the same surface features (frequency and center-embedding) hinder comprehension and recall of legal texts. While we do not find an ifluence of capitalization on recall, we find some evidence that capitalization might aid

comprehension. It is important to note that in these exploratory analyses, we should not make strong conclusions, especially as this experiment was not designed to test this exact question, but rather use these analyses to suggest future confirmatory experiments.

A.4 Replication Study

A.4.1 Methods

Materials

Primary materials for the replication study consisted of the same 12 pairs of contract excerpts used in the main experimental study, as well as the same comprehension questions associated with those excerpts. The author recognition test was not included among the experimental materials.

Participants and Procedure

Subjects (87) were recruited via Prolific to participate in the study. Only those with a comprehension score above chance were included in the final analysis, resulting in a final sample of 76 participants.

With regard to procedure, participants were first given a legal text (written either in legalese or simple English), along with all of the comprehension questions associated with that text, and were asked to read the text and answer all of the comprehension questions. participants were given as much time as desired to read the text and answer the comprehension questions. As with the main experiment, participants were psuedo-randomly assigned to six trials (3 legalese; 3 simple). Participants did not see the same contract in both a simple and legal register. Assignment of stimuli to participant was pseudo-random to ensure that across participants every trial was administered with approximately

the same frequency. The order of trials was randomized for each participant.

Participants were also given an additional monetary incentive if they correctly answered at least 90% of comprehension questions correctly across each of their six trials.

The analysis for this study was the same as the comprehension analysis for the main experiment except that, because participants did not complete the author recognition test, we did not include the author recognition test as a fixed effect predictor.

A.4.2 Results

Figure **??** illustrates the comprehension accuracy across registers. Descriptively, participants were more accurate in the simple register (72.3%) than in the legalese register (67.5%). As with our original experiment, we find a main effect of register on comprehension, with participants scoring significantly lower on texts written in legalese as compared with texts written in a simple register ($\beta = -.2223$, SE$= 0.0690, p < 0.05$). We also find that this main effect holds when removing items with below chance (25%) accuracy ($\beta = .4568$, SE$= 0.1362, p < 0.05$).

A.5 Additional Language Processing Metrics

There is no gold standard measure of processing difficulty for texts; however, being able to gauge the difficulty of a text is important for several applications, most notably education. Therefore, researchers have derived several measures of linguistic and discourse representations that are thought to correlate with processing difficulty. As stated in the text, we focused our indices on easily changeable linguistic structures/properties that were supposedly present in le-

gal texts. In this section, we present the metrics used by Coh-Metrix 3.0[1], a standardized set of processing related indices [144]. For a full discussion of the Coh-Metrix indices, see the documentation. We provide mean estimates for our texts in Tables A.13-A.16 and discuss differences across the conditions below. We discuss the measures where we find a large difference between our simple and legalese registers below.

	Label	leg.mean	leg.low	leg.high	sim.mean	sim.low	sim.high
1	DESPC	1.58	1.33	1.83	1.50	1.25	1.75
2	DESPL	2.42	1.79	3.17	4.17	3.08	5.21
3	DESPLd	0.12	0.00	0.29	0.47	0.12	0.88
4	**DESSC**	3.33	2.75	4.00	5.58	4.75	6.25
5	**DESSL**	31.87	27.13	36.62	17.39	14.86	21.39
6	DESSLd	9.55	6.46	12.59	6.25	4.93	7.32
7	DESWC	96.58	90.41	103.17	89.67	83.50	95.84
8	**DESWLlt**	5.07	5.00	5.15	5.25	5.18	5.33
9	DESWLltd	2.80	2.74	2.86	2.75	2.67	2.86
10	DESWLsy	1.72	1.68	1.77	1.76	1.71	1.80
11	DESWLsyd	0.98	0.93	1.04	0.97	0.92	1.02
12	PCCNCp	60.75	40.31	80.34	48.22	30.21	65.60
13	PCCNCz	0.50	-0.30	1.27	-0.03	-0.69	0.62
14	PCCONNp	22.43	11.63	34.09	21.44	9.45	35.01
15	PCCONNz	-1.13	-1.84	-0.55	-1.30	-2.12	-0.63
16	PCDCp	76.38	59.45	90.78	54.54	34.99	73.77
17	PCDCz	1.33	0.57	2.06	0.06	-0.68	0.70
18	PCNARp	11.57	7.57	15.41	17.70	13.00	22.69
19	PCNARz	-1.31	-1.56	-1.06	-1.00	-1.21	-0.78
20	PCREFp	70.76	55.24	85.09	76.56	67.40	84.80
21	PCREFz	1.40	0.45	2.49	1.00	0.56	1.55
22	**PCSYNp**	11.65	5.51	18.50	63.80	48.67	75.35
23	**PCSYNz**	-1.50	-1.98	-1.12	0.34	-0.20	0.73
24	PCTEMPp	41.14	17.96	63.88	31.85	14.62	53.29
25	PCTEMPz	-1.04	-2.54	0.41	-0.82	-1.63	0.01
26	PCVERBp	18.23	8.18	29.78	5.76	2.29	9.75
27	PCVERBz	-1.50	-2.24	-0.78	-2.18	-2.84	-1.60

Table A.13: Descriptive and Text Easability Coh-Metrics analysis of our stimuli. Means and bootstrapped 95% confidence intervals.

[1] http://cohmetrix.com/

	Label	leg.mean	leg.low	leg.high	sim.mean	sim.low	sim.high
28	CRFAO1	0.83	0.69	0.94	0.83	0.72	0.92
29	CRFAOa	0.81	0.70	0.92	0.67	0.55	0.79
30	CRFCWO1	0.14	0.11	0.17	0.17	0.13	0.22
31	**CRFCWO1d**	0.06	0.03	0.08	0.13	0.09	0.16
32	CRFCWOa	0.13	0.09	0.17	0.12	0.09	0.16
33	**CRFCWOad**	0.04	0.03	0.06	0.11	0.09	0.13
34	CRFNO1	0.83	0.69	0.94	0.75	0.60	0.88
35	CRFNOa	0.79	0.66	0.92	0.59	0.49	0.71
36	CRFSO1	0.90	0.81	0.97	0.86	0.78	0.94
37	CRFSOa	0.84	0.72	0.95	0.68	0.58	0.80
38	**LSAGN**	0.24	0.22	0.26	0.31	0.27	0.35
39	LSAGNd	0.23	0.20	0.27	0.19	0.18	0.21
40	LSAPP1	0.33	0.16	0.48	0.29	0.12	0.46
41	LSAPP1d	0.00	0.00	0.00	0.00	0.00	0.00
42	LSASS1	0.45	0.37	0.53	0.43	0.37	0.50
43	LSASS1d	0.11	0.05	0.15	0.15	0.11	0.19
44	LSASSp	0.30	0.18	0.41	0.40	0.33	0.46
45	LSASSpd	0.09	0.04	0.15	0.15	0.11	0.18
46	LDMTLD	89.14	77.32	99.52	69.85	59.67	80.88
47	LDTTRa	0.72	0.69	0.74	0.69	0.67	0.72
48	**LDTTRc**	0.86	0.84	0.89	0.79	0.76	0.82
49	LDVOCD	37.25	11.84	62.64	11.17	0.00	27.61

Table A.14: Co-reference Cohesion, Latent Semantic Analysis and Lexical Diversity Coh-Metrics analysis of our stimuli. Means and bootstrapped 95% confidence intervals.

There are five sets of differences between out legalese and simple stimulus conditions. First, there are several differences in Coh-Metrix that serve as a manipulation check. The legalese condition is more syntactically complex (PCSYNp/z) than the simple condition. The simple condition has greater syntactic structure overlap between both adjacent sentences (SYNSTRUTa) and all sentences within the paragraph (SYNSTRUTt) compared to the legalese condition. There are more agentless passive voice structures in the legalese than the simple condition (DRPVAL). The legalese condition is more left-embedded (measured by the number of words before the main verb) than the simple condition (SYNLE). The simple condition has higher frequency words than the

	Label	leg.mean	leg.low	leg.high	sim.mean	sim.low	sim.high
50	CNCADC	14.72	7.59	22.80	17.92	8.75	28.29
51	CNCAdd	34.23	28.84	39.45	35.59	27.35	43.89
52	**CNCAll**	102.14	92.61	112.85	78.94	68.33	89.69
53	**CNCCaus**	61.80	53.24	71.11	34.66	22.42	46.51
54	CNCLogic	30.66	20.68	40.39	31.07	17.55	44.09
55	CNCNeg	0.00	0.00	0.00	0.00	0.00	0.00
56	CNCPos	0.00	0.00	0.00	0.00	0.00	0.00
57	CNCTemp	9.42	4.40	14.36	11.42	5.99	17.19
58	CNCTempx	15.09	9.83	21.34	15.38	9.11	21.93
59	SMCAUSlsa	0.10	0.08	0.12	0.06	0.04	0.08
60	SMCAUSr	0.26	0.12	0.38	0.26	0.09	0.45
61	**SMCAUSv**	22.43	19.89	25.46	42.88	34.18	50.57
62	**SMCAUSvp**	31.02	25.10	36.56	53.28	46.26	59.62
63	**SMCAUSwn**	0.29	0.27	0.32	0.36	0.33	0.39
64	SMINTEp	9.34	4.24	15.09	15.28	9.16	20.50
65	**SMINTEr**	3.32	2.44	4.26	1.19	0.74	1.67
66	SMTEMP	0.73	0.60	0.87	0.76	0.70	0.84
67	**SYNLE**	11.00	7.97	14.25	4.49	3.25	5.78
68	SYNMEDlem	0.65	0.43	0.86	0.84	0.82	0.86
69	SYNMEDpos	0.44	0.28	0.58	0.59	0.56	0.62
70	SYNMEDwrd	0.68	0.45	0.91	0.88	0.84	0.91
71	SYNNP	1.11	1.01	1.22	1.15	1.04	1.27
72	**SYNSTRUTa**	0.08	0.06	0.10	0.13	0.10	0.16
73	**SYNSTRUTt**	0.08	0.06	0.09	0.13	0.11	0.15

Table A.15: Connectives, Situation Model and Syntactic Complexity CohMetrics analysis of our stimuli. Means and bootstrapped 95% confidence intervals.

legalese condition (WRDFRQmc). The verbs used in the simple condition have greater WordNet similarity overlap than the verbs used in the legalese condition (SMCAUSwn).

The second set of differences reflect causal coherence. It should be noted that these metrics are unreliable for short texts like our stimuli, yet we report them for transparency. There are more connectives (including causal connectives) in the legalese than the simple condition (CNCAll, CNCCaus). That being said, the simple condition has more causal verbs and particles than the legalese condition (SMCAUSv, SMCAUSvp). Th legalese condition has a higher ratio of

intentional particles to intentional verbs than the simple condition (SMINTEr). Taken together, it would appear that the simple and legalese conditions are both causally coherent but rely on different strategies to support causal coherence.

The third set of differences reflect co-referential coherence. The simple condition has greater content word overlap across adjacent (CRFWO1d) and all (CRFCWOad) sentences than the legalese condition, suggesting more co-reference. Similarly, there is greater givenness in the simple condition than the legalese condition (LSAGN) suggesting that there is more co-referential coherence. Also, the legalese condition has a higher type-token ratio than the simple condition (LDTTRc). Taken together, these measures suggest that the simple condition has greater co-referential coherence than the legalese condition. This is not a confound to our experimental manipulation though. These differences are really just a byproduct of using active voice and turning embedded clauses into separate sentences.

The penultimate set of differences reflect composite readability scores. The simple condition is more easily readable than the legalese register as measured by a lower Flesch-Kincaid Grade level (RDFKGL), a higher Flesh Reading Ease Score (RDFRE) and a higher L2 Readability score (RDL2) for the simple condition than the legalese condition. This is unsurprising because these measures take into account our manipulated surface features.

The last set of differences are merely descriptive. There are more words (DESWLlt) and sentences (DESSC) in the simple than the legalese condition; however, there are less words per sentence (DESSL) in the simple condition than the legalese. This is a result of splitting center-embedded clauses into two sentences. There are more nouns in the simple condition than the legalese condition (WRDNOUN), which is also a result of removing passive and center-embedded structures. While there are more 3rd person singular pronouns in the legalese condition (WRDPRP3s); there are more 3rd person plural pronouns in

the simple condition (WRDPRP3p). As our stimuli are relatively short, these differences are likely not stable markers of any significant processing difficulty.

	Label	leg.mean	leg.low	leg.high	sim.mean	sim.low	sim.high
74	DRAP	16.39	10.38	22.94	19.42	13.87	24.96
75	DRGERUND	16.21	7.83	24.86	12.58	5.46	20.35
76	DRINF	5.24	1.63	10.22	4.59	0.85	8.86
77	DRNEG	10.08	5.44	15.31	11.60	6.54	17.57
78	DRNP	373.90	351.07	395.67	380.25	359.23	400.92
79	DRPP	171.02	159.53	183.10	148.71	134.83	161.75
80	**DRPVAL**	32.97	26.48	39.20	8.05	1.81	16.02
81	DRVP	181.36	164.60	197.09	171.75	158.57	184.46
82	WRDADJ	57.46	41.60	75.22	59.11	43.69	76.71
83	WRDADV	29.46	23.86	35.16	35.50	30.71	40.11
84	WRDAOAc	423.41	404.03	437.58	417.43	400.74	433.17
85	WRDCNCc	383.46	364.37	401.15	380.26	364.19	397.08
86	WRDFAMc	542.48	534.93	550.07	547.94	542.99	552.73
87	WRDFRQa	2.86	2.78	2.93	2.71	2.64	2.79
88	WRDFRQc	1.81	1.72	1.89	1.89	1.83	1.94
89	**WRDFRQmc**	0.12	0.00	0.29	0.91	0.72	1.10
90	WRDHYPn	6.58	6.28	6.92	6.61	6.17	7.05
91	WRDHYPnv	2.11	1.98	2.24	2.35	2.18	2.54
92	WRDHYPv	1.56	1.37	1.80	1.84	1.59	2.06
93	WRDIMGc	408.45	388.97	428.27	406.17	388.32	424.19
94	WRDMEAc	417.54	403.45	431.50	421.94	405.03	441.08
95	**WRDNOUN**	314.70	296.75	334.95	350.52	338.24	362.40
96	WRDPOLc	2.99	2.87	3.10	3.08	2.93	3.23
97	WRDPRO	12.76	9.38	16.08	7.04	2.78	11.34
98	WRDPRP1p	0.00	0.00	0.00	0.00	0.00	0.00
99	WRDPRP1s	0.00	0.00	0.00	0.00	0.00	0.00
100	WRDPRP2	2.84	0.00	7.04	1.49	0.00	4.46
101	**WRDPRP3p**	0.00	0.00	0.00	2.68	0.00	5.69
102	**WRDPRP3s**	0.93	0.00	2.78	0.00	0.00	0.00
103	WRDVERB	111.18	100.98	122.67	122.52	107.99	139.27
104	**RDFKGL**	17.18	15.19	19.29	11.92	10.69	13.66
105	**RDFRE**	28.62	22.74	34.40	40.58	34.82	45.25
106	**RDL2**	7.28	5.44	9.35	13.84	10.80	17.65

Table A.16: Syntactic Pattern Densities, Word Information and Readability Coh-Metrics analysis of our stimuli. Means and bootstrapped 95% confidence intervals.

Appendix B

Supplemental Information for *So Much for Plain Language: An Analysis of the Accessibility of United States Laws Over Time*

B.1 Methods

Documents were first tokenized into separate sentences using the Stanza natural language package. Afterwards, we filtered out sentences through a series of five steps, to remove (a) sentences without punctuation, (b) sentences with more than 3 consectutive punctuation marks, (c) duplicate sentences, (d) sentences with more than 3 consecutive "@" symbols, and (e) sentences with fewer than 10 words.

With regard to these filtering steps, we removed sentences without punctuation, as well as those with fewer than 10 words so as to remove headings in the contract corpus, which are not really sentences but would otherwise be counted as such without this filter. The removal of 3 consecutive punctuation marks and

'@' symbols was added as a filter so as to get rid of more non-sentences in both corpora. The duplicate sentences filter was added to remove the high number of repeat sentences in the standard-english corpus.

The number of sentences remaining after each step in our primary corpora are described in Table B.1.

Filter	**Legal Corpus**	**COHA**
sentences without punctuation	1,907,203	8,482,309
sentences with 3+ consecutive punctuation marks	1,836,271	8,472,319
duplicate sentences	1,472,735	8,325,379
sentences with 3+ consecutive @ symbols	1,301,314	7,467,771
sentences with fewer than 10 words	848,555	4,835,240

Table B.1: Filtering Processes and number of remaining sentences for each corpus.

As discussed in the main text, the standard-English corpus consisted of a broad sample of fiction, non-fiction, popular magazines, and newspapers from the Corpus of Historical American from the years 1951 and 2009 English (COHA: davies2009385+), while the legal corpus consisted of every public law, private law, concurrent resolution, and proclamation issued by the United States federal government between 1951 and 2009. The number of sentences in each of these subgenres post-filtering is given in Table B.2.

Concurrent Resolutions (LAW)	16,794
Private Laws (LAW)	17,424
Proclamations (LAW)	76,554
Public Laws (LAW)	737,783
Fiction (COHA)	2,282,312
Non-Fiction (COHA)	552,501
Magazine articles (COHA)	1,219,045
Newspaper articles (COHA)	681,382

Table B.2: Sentences in each COHA subgenre post-filtering

As shown in the table, the corpus featured a comparatively small number of sentences in the concurrent resolution, private law, and proclamation subgenres.

These subgenres were also not represented in every year of our corpus; that is, there were years in which Congress did not pass any private laws, concurrent resolutions and/or proclamations. Because of this, we do not perform separate longitudinal genre-by-genre analyses of these subgenres and instead report the comparisons between the legal corpus and the different subgenres of the COHA.

The number of sentences remaining after each step in our secondary corpora (United States Code and Academic texts from the Corpus of Contemporary American English) are described in Table B.3.

Filter	**U.S. Code**	**Academic (COCA)**
sentences without punctuation	1,057,821	3,435,037
sentences with 3+ consecutive punctuation	1,055,101	3,429,513
duplicate sentences	893,303	3,369,139
sentences with 3+ consecutive @ symbols	893,303	2,897,611
sentences with fewer than 10 words	569,993	2,465,573

Table B.3: Filtering Processes and number of remaining sentences for each corpus.

Word frequency. To perform this calculation, we looked at all the words in the corpora marked as a verb, noun, adjective or adverb according to Stanza. We then looked at how frequently each of these words appeared in the SUBTLEX word frequency dictionary, a corpus of American film subtitles commonly used as a proxy for standard-English word frequency. Values were Zipf-adjusted. Proper nouns and other words that did not appear in the corpus received a score of "NA." Analyses and genre-by-genre visualizations are reported in the main text. Comparisons between the legal and COHA corpus are visualized in Figure B.2.

Word choice. We performed this calculation using three separate methods. Under the first method, we operated under the assumption that legal concepts are not restricted by precision—as it is often claimed that legal terms often resemble common words in form but have a more specialized meaning, such as the concept of "consideration" in contract law [73].

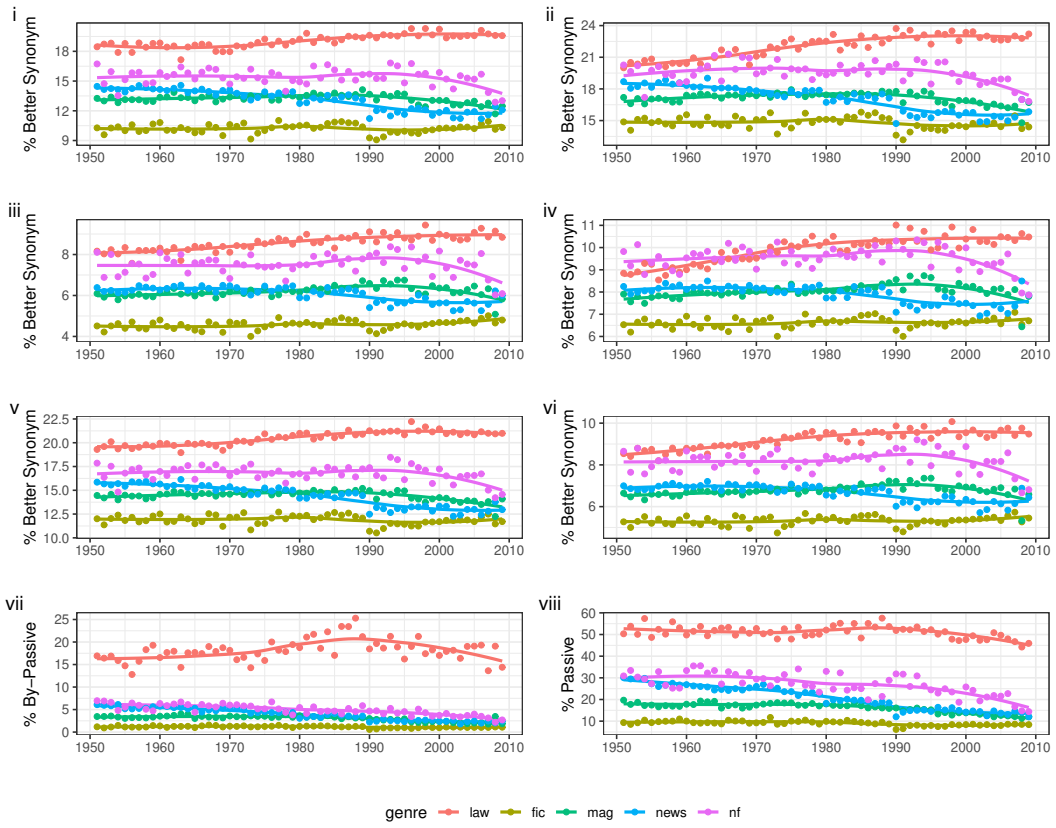

Figure B.1: Comparison of supplemental indices of linguistic processing difficulty in federal laws vs four genres of standard English, including fiction books, magazine articles, newspaper articles, and non-fiction books (1951-2009).

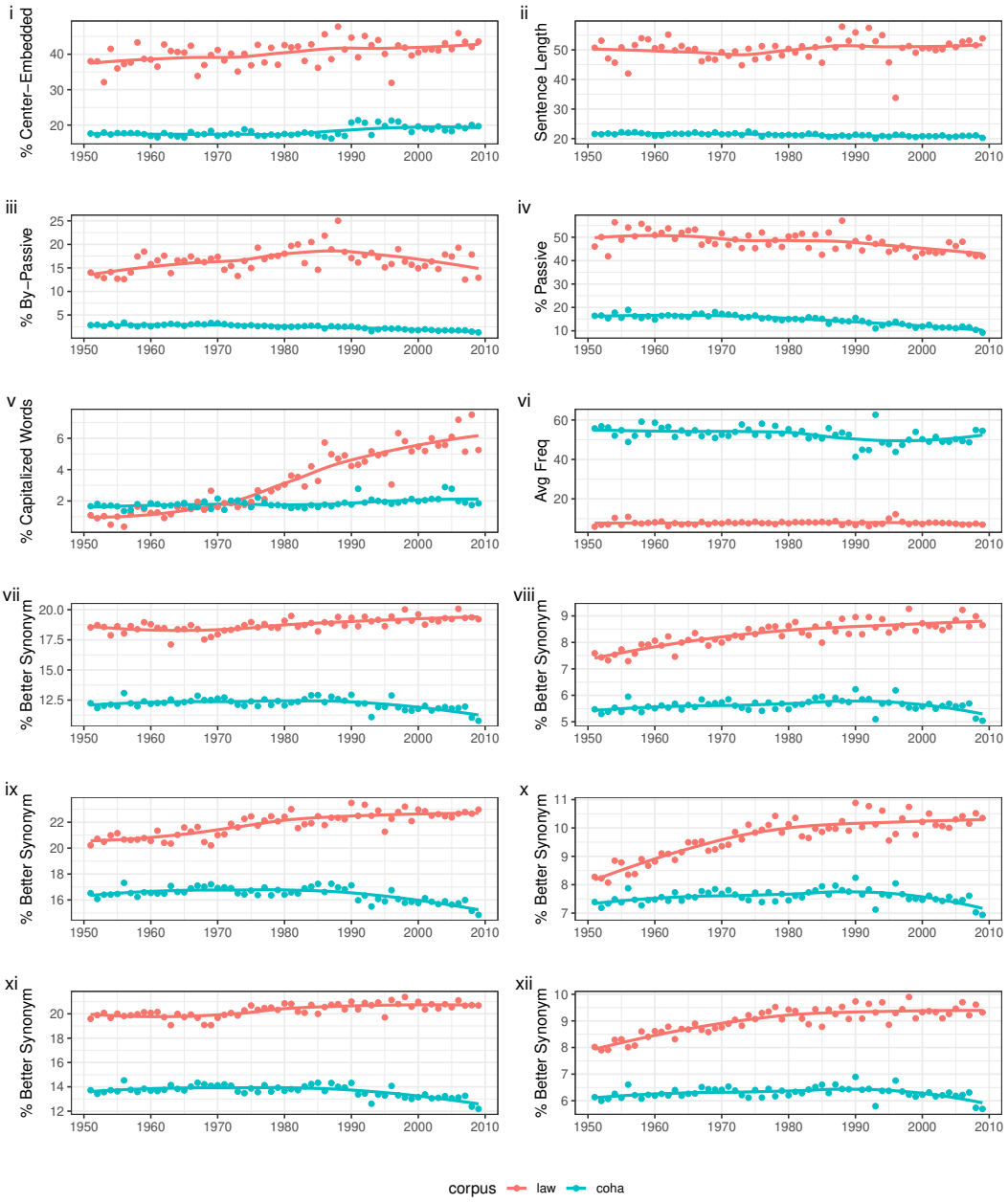

Figure B.2: Comparison of indices of linguistic processing difficulty in federal laws vs Corpus of Historical American English (1951-2009).

Under this method, we looked at the same group of words included in the word-frequency analysis and for each of those words: (a) looked at the least common meaning/sense of that word according to WordNet; (b) using the meaning/sense obtained in (a), looked at all possible synonyms of that word according to WordNet (i.e. assuming the authors meant to use the least common meaning, what other words could they have used instead); (c) computed the SUBTLEX frequency value of each of these other synonyms; and (d) coded whether the SUBTLEX frequency value of any of the synonyms was higher than that of the actual word used in the text (if yes, we coded that word as having a 'better synonym' / 'higher-frequency synonym')

Under our second method, we operated under the assumption that legal terms are not constrained by precision, and that the intended meaning of words in legal texts is the most common meaning of that word. Accordingly, we followed the same steps and words except that for Step (a), we looked at the most common meaning/sense of a given word instead of the most common word. For both assumptions, we calculated the percentage of words with a higher-frequency synonym both as a proportion of (a) all alphabetic words and (b) all content-words, not including proper nouns.

Under the third method, we operated under the assumption that the intended meaning of a word in legal texts tends to neither be the most common more least common sense of that word and is instead random. Accordingly, we followed the same steps and words except that for Step (a), we chose a random sense of that word. For each of the three methods, we calculated the percentage of words with a higher-frequency synonym both as a proportion of (a) all alphabetic words and (b) all content-words, not including proper nouns.

Note that WordNet's determinations of sense frequencies come from the SemCor semantically annotated SemCor corpus (Miller et al., 1993). Because of the sparsity of this corpus, the sense frequencies are less reliable for less

common words and senses of those words. In those cases, one may view our conservative method as assuming the author meant an uncommon sense of the word (as opposed to the absolutely least common sense), and view our anti-conservative method as assuming the author meant a more common sense of the word (as opposed to the absolutely most common sense).

The results of the first method are reported in the main text. Comparisons between laws and COHA corpus overall for the first method are visualized in Figure S1 B.1 i (for all content words) and iii (for all alphabetic words), as well as in B.2 vii (for all content words) and viii (for all alphabetic words).

With regard to the second method, comparisons between laws and COHA corpus are visualized in Figure S1 B.1 i (for all content words) and iii (for all alphabetic words), as well as in B.2 ix (for all content words) and x (for all alphabetic words).

With regard to the third method, comparisons between laws and COHA corpus are visualized in Figure S1 B.1 v (for all content words) and vi (for all alphabetic words), as well as in B.2 xi (for all content words) and xii (for all alphabetic words).

Capitalization. Here we sought to determine what percentage of words in contracts were in ALL CAPS relative to standard English. To do so, we looked at all of the alphabetic words in each of our corpora and calculated the proportion of words in each corpus that were marked by Stanza as being entirely in uppercase letters.

Passive-voice structures. To compute the prevalence of passive voice structures as a whole in both corpora, we calculated the number of words marked with the passive voice features in Stanza. To compute the prevalence of by-passive structures, we performed the same calculation and then looked at the number of passives that had the word *by* in the same head according to Stanza. Main results of each of these are reported in the main text. Comparisons be-

tween federal laws and the COHA corpus of by-passives and passives overall are visualized in Figure B.2 iii and iv, respectively. Genre-by-genre comparisons of by-passives and passives are visualized in Figures B.1 v and vi, respectively.

Center-embedded clauses. To determine the number of embedded clauses (both center-embedded and right-branching) as a whole, for every sentence in each corpus we looked at the number of predicate dependent clauses (i.e. clausal subjects, clausal complements, open clausal complements, adjectival clauses, and adverbial clauses). To determine the number of center-embedded clauses, we performed the above calculation and then looked at whether the clause was followed by a word as opposed to an end-of-sentence punctuation mark. Main results reported in main text. Comparisons between federal laws and COHA corpus are visualized in Figure B.2 i.

Appendix C

Supplemental Information for *Even Lawyers Don't Like Legalese*

C.1 Experiment 1

C.1.1 Author Recognition Test Analyses

As noted in the main text, for Experiment 1 we administered the Author Recognition Test (a validated proxy for reading ability) to participants as part of each trial. Although this was mainly used as a filler task, and we made no pre-registered predictions regarding the results, for transparency we report the results of its potential effect on comprehension and recall data. The results of all of these analyses were convergent with the findings in Martinez, Mollica & Gibson's sample of laypeople [101].

When adding ART score as a fixed-effect predictor to our comprehension model, we found a main effect of ART score on comprehension (β =.221, SE=.078, p=.005), such that those who had higher ART scores had higher comprehension

scores. However, we did not find a significant interaction between ART and register (p=.075). When adding ART score as a fixed-effect predictor to our recall model, we did not find a main effect of ART score on recall (p=.787).

C.1.2 Recall annotation details

Two trained research assistants coded whether a proposition was successfully recalled, using the same method as [101]. In particular, they were given a participant's retelling of a passage and then asked whether each legally relevant proposition of the passage was (a) fully recalled; (b) partially recalled; or (c) not recalled. Coders were told that for a response to count as "fully recalled," it did not have to be recalled verbatim (i.e. they can use their own words or syntax), so long as they were confident that the meaning of what subject wrote is the same as the proposition.

For example, suppose the original text said "A court in Boston will resolve the dispute," and the participant wrote "something will be resolved by a court." When coding responses, a coder might see three propositions that say: (i) "A court in Boston," (ii) "will resolve," and (iii) "the dispute."

For (i), the coder would put a 0.5 for "partially recalled" (since "in Boston" was missing from "a court"); for (ii), the coder would put a 1 for "fully recalled" (since "will be resolved" means basically the same thing as "will resolve"), and for (iii), the coder would put a 0 for "not recalled" (since "the dispute" was not in the response).

To reduce potential bias, coders were unaware of whether a participant had seen or recalled the simple or legalese version of a text. Moreover, the rubric that the coders used to score participants' responses was the legalese version: for each proposition, they were given the language of that proposition in legalese, and were told to score a participants' response as having recalled that proposition if it had language that had the same meaning as that proposition. Thus, any

differences in recall favoring the Plain English version would arise in spite of the coding bias, which was towards the Legalese version.

Of the roughly 650 retellings within the lawyer data, one coder was responsible for coding 100 percent of the retellings, while the second coder was responsible for coding a random subset (20%) of the retellings, so as to assess inter-rater reliability. Coder reliability was assessed with Cohen's kappa coefficient [81], [82].

We adjudicated ties as follows: (i) a tie between one "fully recalled" judgment and one "not recalled" judgment resulted in a final "partially recalled" judgment; (ii) a tie between one "fully recalled" judgment and one "partially recalled" judgment resulted in a final "fully recalled" judgment for a given proposition; and (iii) a tie between one "partially recalled" judgment and one "not recalled" judgment resulted in a final "not recalled" judgment. For our regression analyses, we perform both a conservative analysis (recoding "partially recalled" as "not recalled') and an anti-conservative analysis (recoding "partially recalled" as "fully recalled"). Our results do not qualitatively change, so we will only report the conservative analysis in the main writeup.

C.1.3 Anti-conservative recall analysis

As noted in the main writeup, for our recall analyses, we performed both a conservative analysis and an anti-conservative analysis. For the conservative analysis, As our results did not qualitatively change, so we only reported the conservative analysis in text.

For both the conservative and anti-conservative analyses, we conducted a mixed effect logistic regression with register, legal training and the interaction between the two as fixed effects, and participant as random effects, with register as a random slope for each. As in the conservative analysis, for the anti-conservative analysis, we found a main effect of register (β = .225, SE=.098,

$p = .022$) and legal training ($\beta = -.340$, SE=.153, $p = .026$) on recall. As in the conservative analysis, we did not find a main effect of the interaction between register and condition ($p = .174$).

C.1.4 Subjective rating analyses

As noted in the main writeup, in addition to recall and comprehension analyses, we also asked participants to rate how difficult a text they found the text (a) for themselves; (b) for the average layperson; and (c) for the average lawyer.

Below is the wording of each of the prompts:

- (you): "How complex/difficult do you find this text to understand?"
- (lawyer):"How complex/difficult do you think the average layperson/non-lawyer would find this text to understand?"
- (layperson): "How complex/difficult do you think the average lawyer would find this text to understand?"

The 5 answer choices for each prompt were as follows:

- extremely simple/easy
- somewhat simple/easy
- neither complex/difficult nor simple/easy
- somewhat complex/difficult
- extremely complex/difficult

Results are visualized in the main text.
To analyze these results, we ran three different models.

First, we conducted a model that compared how difficult lawyers and laypeople predicted texts would be for the average layperson.

Second, we conducted a model that compared (a) how difficult lawyers predicted texts would be for the average layperson compared with (b) how difficult lay participants perceived the texts to be for themselves.

Third, we conducted a model that compared (a) how difficult laypeople predicted texts would be for the average lawyer compared with (b) how difficult lay participants perceived the texts to be for themselves.

This model's predictions are less relevant for the curse of knowledge hypothesis, but for robustness purposes we report it anyway.

The model for all three models was as follows:

- clmm(as.factor(Response) condition*training + (1 + condition | subject) + (1 + condition | item), data =.,)

The only difference among the three models was that the "response" variable was filtered to include a different subset of the data (to include the relevant conditions)

For the first model, we found a main effect of condition (β =-1.784, SE=.161, p<.001), but not training (p=.974), nor the interaction between condition and training (p=.127).

That is, contrary to the predictions of the curse of knowledge hypothesis, we did not find evidence that lawyers underestimated the difficulty of legal texts for non-lawyers, nor did they particularly underestimate the difficulty of legal texts written in legalese.

For the second model, we found a main effect of condition (β =-1.784, SE=.161, p<.001). We also found an effect of training (β =-.938, SE=.293, p=.001), with laypeople's own subjective ratings being significantly easier than lawyers' predictions of the average layperson's subjective ratings.

We also found an interaction between training and condition (β =.562, SE=.250, p=.025), such that laypeople's ratings of simple texts were disproportionately high relatively to lawyers' predictions of those texts relative to the groups' legalese ratings.

For the third model, we found a main effect of condition (β =-2.235, SE=-8.718, p<.0001). We also found an effect of training (β =-3.094, SE=.339, p<.0001), with laypeople's predictions of lawyers' ratings being significantly easier than lawyers' subjective ratings of the texts.

We also found an interaction between training and condition (β =.792, SE=.266, p=.003), such that laypeople's predictions of lawyers' ratings of simple texts were disproportionately high relatively to lawyers' ratings of those texts relative to the legalese ratings.

C.2 Experiment 2

Hypotheses and Predictions

In Experiment 2, we aimed to test the following hypotheses and associated predictions. All of these were pre-registered on OSF.

Hypothesis I: Lawyers simply write in a complex register out of "habit, laziness" [39] or "tradition" [40]; they "copy and paste" (Adams, 2022) from existing templates with old, complicated terms because that's the "quickest and cheapest way to produce a contract" [66], not out of any preference.

- Prediction 1: Lawyers will rate plain English contracts as of equal quality as legalese contracts.
- Prediction 2: Lawyers will agree to sign off on a contract written in Plain English.

Hypothesis II: Lawyers write in legalese in order to be accepted by peers. The legalese signals in-group membership [40].

- Prediction 1: Lawyers will rate contracts written in legalese as sounding more "lawyerly" (more appropriate/suitable for a lawyer) than those written in plain English.
- Prediction 2: Lawyers will rate authors of contracts written in legalese more hirable than authors of contracts written in plain English.

Hypothesis III: Lawyers write in legalese as a way of "preserving their monopoly" [67] on legal services and "justifying fees" [40]

- Prediction: Lawyers will predict contracts written in legalese as being more likely to be signed by clients than contracts written in plain English.

Hypothesis IV: Contractual language needs to be complex in order to convey complex legal concepts in a way that "is far more precise than ordinary language" [39] and/or to be enforceable

- Prediction: Plain English contracts will be rated as unenforceable by legal experts

Materials

To evaluate our predictions, we measured two sets of outcome variables. The first set of outcome variables were measured individually for each text and were as follows:

- text quality ("How would you rate the overall quality of the above contract excerpt?")
- enforceability ("Suppose two parties signed a contract that included the above excerpt. Would the excerpt likely be legally enforceable (assuming the rest of the contract was enforceable)?")

- useability ("Suppose someone at your firm drafted a contract that included the above excerpt. Would you and your firm agree to execute it as currently written (assuming the rest of the contract was okay)?")

- hireability ("Suppose the excerpt was drafted by someone outside your firm. Would your firm be likely to hire them to draft future contracts, all else equal?")

- lawyerliness ("Does the style/tone of the excerpt sound appropriate for a lawyer?")

- likelihood of being signed ("Suppose you drafted this excerpt for a client as part of a larger contract. Would a client be likely to sign this contract (assuming the rest of the contract was written in a similar style)?")

Text quality was measured on a scale of 1-5 (1 being "extremely low-quality" and 5 being "extremely high-quality"). All other outcome variables in this set were measured on a yes-no scale.

The second set of outcome variables were measured for each contract pair as opposed to each individual contract. These variables were as follows:

- more usable ("Which of the two versions would you be more likely to execute, given the choice?")

- more likely to be signed ("Which of the two versions would a client be more likely to agree to sign?")

- more lawyerly ("Which of the two versions sounds more appropriate for a lawyer?")

- more hireable ("Suppose the two versions were drafted by two different authors. Which of the two would your firm be more likely to hire to draft future contracts, all else equal?")

All outcome variables in this set were measured on a two-point scale (version 1 or version 2).

C.2.1 Analysis Plan

To evaluate Hypothesis I, we conducted two regressions.

- An ordinal regression with the following syntax: `clmm(text quality ~ condition + (1 + condition | item) + (1 + condition | subject), data = .)`
- A logistic regression with the following syntax: `glmer(is usable ~ condition + (1 + condition | item) + (1 + condition | subject), data = ., family = binomial(link = "logit"))`

For Hypothesis II we we conducted exact binomial tests for the more lawyerly and more hireable variables.

For Hypothesis III we conducted the following logistic regression, as well as an exact binomial test:

- `glmer(client would sign ~ condition + (1 + condition | item) + (1 + condition | subject), data = ., family = binomial(link = "logit"))`

For Hypothesis IV, we conducted the following regression:

- `glmer(is enforceable ~ condition + (1 + condition | item) + (1 + condition | subject), data = ., family = binomial(link = "logit"))`

In our pre-registration, we stated that if we encountered issues fitting models, we would use Bayesian regression techniques with similar syntax. We did not encounter issues fitting models, and therefore will report our pre-registered models.

C.2.2 Supplementary Results

Results are visualized in Table S1 and Figure S1. As noted in the main text, all of the predictions of hypotheses 1-3 were disconfirmed, and all of the predictions of hypothesis 4 were confirmed.

With regard to hypothesis 1, contrary to the first prediction, we found that lawyers were more likely to say that they would use simple contracts over legalese contracts (β =1.432, SE=.270, p<.001), and rated simple texts as higher quality than legalese texts (β =1.705,SE=.329,p<.001).

With regard to hypothesis 2, contrary to both predictions, participants were more likely to rate the authors of simple texts as hireable compared to the authors of legalese texts ($\beta = 1.835$,$SE = .318$,$p < .001$), and we did not find participants to be more likely to rate legalese texts as sounding more lawyerly than simple texts (p=.692).

With regard to hypothesis 3, contrary to the pre-registered prediction, we found that participants were more likely to predict that clients would sign a simple contract compared to a legalese contract ($\beta = 1.232, SE = .338, p <$ .001).

With regard to hypothesis 4, in line with the predictions, we found that participants were more likely to say that they would agree to use the simple contracts as written (β =1.705,SE=.329,p<.001), and we did not find a significant difference in how enforceable the different contracts were rated as (p=.717).

C.2.3 Exploratory Demographic Analyses

Our main analyses, predictions, and conclusions drawn on the basis of our analyses were limited to those that we included in our pre-registration on OSF.

Although our pre-registered statistical models did not include any demographic analyses, one might wonder whether our results may have been driven

by the demographic composition of our lawyer sample.

Below is a description of each of these results as applied to our hypotheses. To help lend a visual sense of the robustness of the results, results for Experiment 2 limited to "experienced" attorneys (those with 10 or more years of practice experience) are visualized in Figure S4. Results limited to "fancy" attorneys (those who attended a top-25 law school or work at a top-200 law firm) are visualized in Figure S5.

To more rigorously account for the possibility of results being driven by demographic factors, we conducted additional versions of our pre-registered analyses, adding each of our demographic variables as fixed-effect predictors. Doing so did not alter our results.

Specifically, in cases where our pre-registered models found a main effect of a given predictor variable, the same was true when adding all of the demographic variables as additional fixed-effect predictors.

Conversely, in cases where our pre-registered models did not find a main effect of a given predictor variable, the same was true when adding all of the demographic variables as additional fixed-effect predictors.

Curse of Knowledge

With regard to the curse-of-knowledge hypothesis, We added the demographic factors to a modified model of comprehension and recall models that (a) were limited to lawyers; and (b) did not contain fixed-effects related to legal training. These models were as follows:

- `glmer( comprehension ~ condition + age + is fancy + gender + ethnicity + practice experience + (1 + condition | item) + (1 + condition | subject), data = ., family = binomial(link = "logit"))`
- `glmer( recall ~ condition + age + is fancy + gender + ethnicity`

```
+ practice experience + (1 + condition | item) + (1 + condition
| subject), data = ., family = binomial(link = "logit"))
```

As with our pre-registered models, these models revealed a main effect of register, such that lawyers were significantly more likely to comprehend (β =.373,SE=.094,p<.0001) and recall (β =.376,SE=.132,p<.0001) legal content written in simple contracts relative to legalese contracts.

The comprehension model revealed a main effect of "fanciness," such that fancy lawyers had significantly higher comprehension overall than non-fancy lawyers (β =.433,SE=.158,p=.006). There were no other main effects in the two models.

In-Group Signaling

With regard to the in-group signaling hypothesis, We added the demographic factors to our model of hireability, as follows:

- `clmm( hireability ~ condition + age + is fancy + gender + ethnicity + practice experience + (1 + condition | item) + (1 + condition | subject), data = ., family = binomial(link = "logit"))`

As with our pre-registered model, this model revealed a main effect of register, such that lawyers rated authors of simple contracts as significantly more hireable than authors of legalese contracts (β =1.900,SE=.322,p<.0001).

It's Just Business

With regard to the it's just business hypothesis, We added the demographic factors to our model of willingness to sign, as follows:

- `clmm( client would sign ~ condition + age + is fancy + gender + ethnicity + practice experience + (1 + condition | item) + (1 + condition | subject), data = ., family = binomial(link = "logit"))`

As with our pre-registered model, this model revealed a main effect of register, such that lawyers predicted that clients would be significantly more likely to sign simple contracts than legalese contracts (β =1.208,SE=.370,p=.001).

Complexity of Information

With regard to the complexity of information hypothesis, We added the demographic factors to our model of enforceability, as follows:

- `clmm( enforceability ~ condition + age + is fancy + gender + ethnicity + practice experience + (1 + condition | item) + (1 + condition | subject), data = ., family = binomial(link = "logit"))`

As with our pre-registered model, this model revealed no effect of register. That is, we did not find evidence that lawyers rated legalese contracts as more enforceable than legalese contracts (p=.156).

Copy and Paste

We added the demographic factors to our models of quality and usability, as follows:

- `clmm(quality ~ condition + age + is fancy + gender + ethnicity + practice experience + (1 + condition | item) + (1 + condition | subject), data = ., family = binomial(link = "logit"))`
- `clmm(would use ~ condition + age + is fancy + gender + ethnicity + practice experience + (1 + condition | item) + (1 + condition | subject), data = ., family = binomial(link = "logit"))`

In both cases, we still found a main effect of register on responses, such that lawyers were significantly more likely to say that they would use simple contracts over legalese contracts (β =1.533,SE=.285,p<.0001), and rated simple contracts as significantly higher quality than legalese contracts (β =1.807,SE=.338,p<.0001).

Table C.1: Endorsement rates by desiderata

Desiderata	Legalese			Simple		
	endorsement %	lower CI	upper CI	endorsement %	lower CI	upper CI
hireability	31.4	28.8	33.9	59.4	56.7	62.1
likelihood of being signed	69.0	66.4	71.4	82.2	80.2	84.4
enforceability	82.3	80.2	84.4	84.3	82.2	86.2
quality	2.61	2.55	2.67	3.33	3.27	3.39

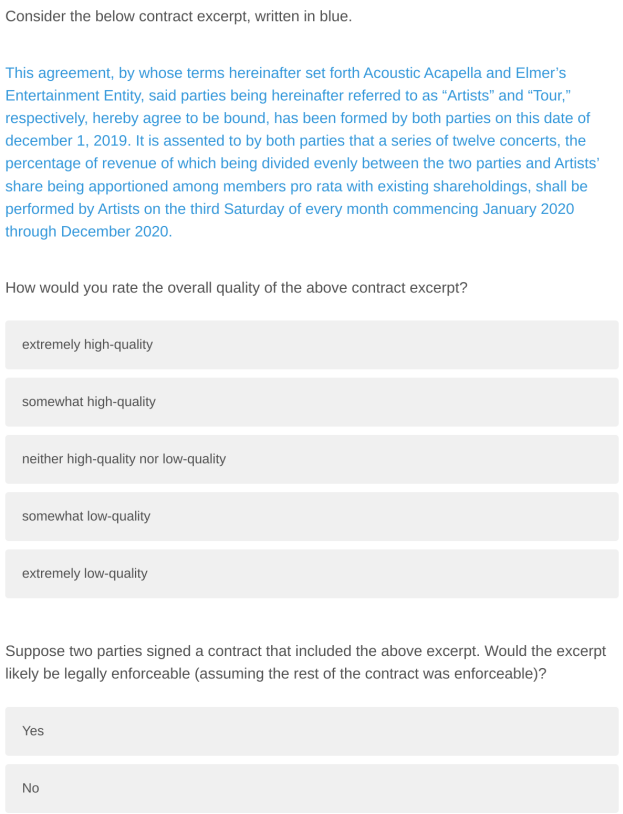

Figure C.1: Interface of Experiment II.

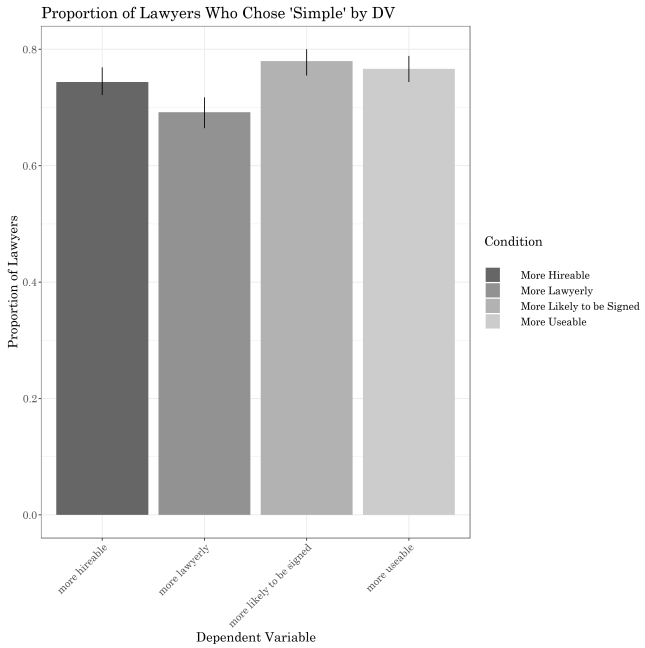

Figure C.2: Proportion of lawyers who endorsed simple version over legalese version according to different desiderata.

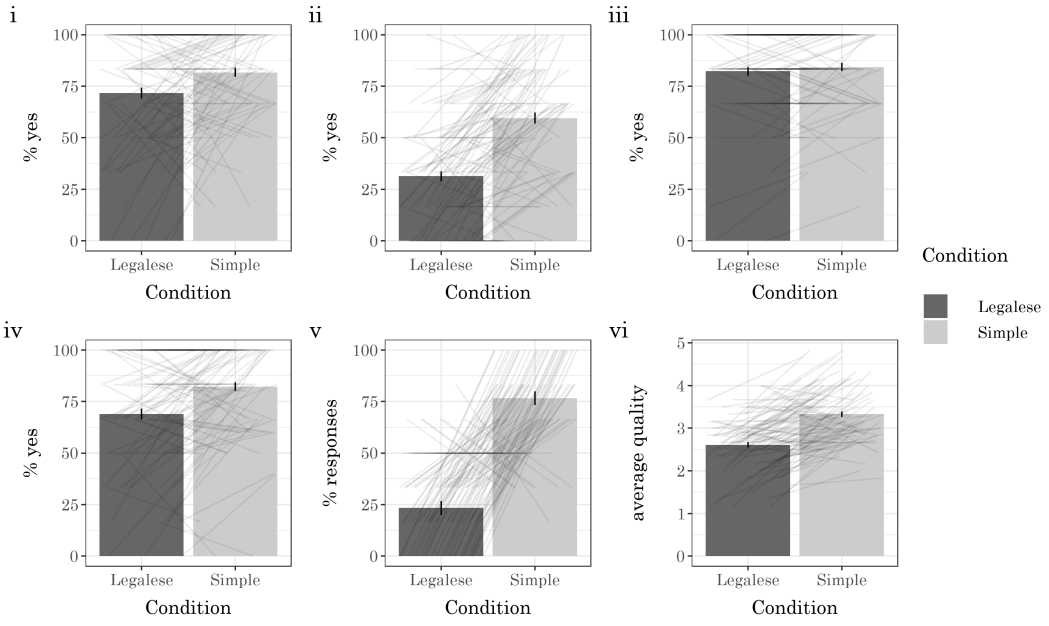

Figure C.3: Proportion of lawyers who endorsed simple and legalese contracts according to different desiderata.

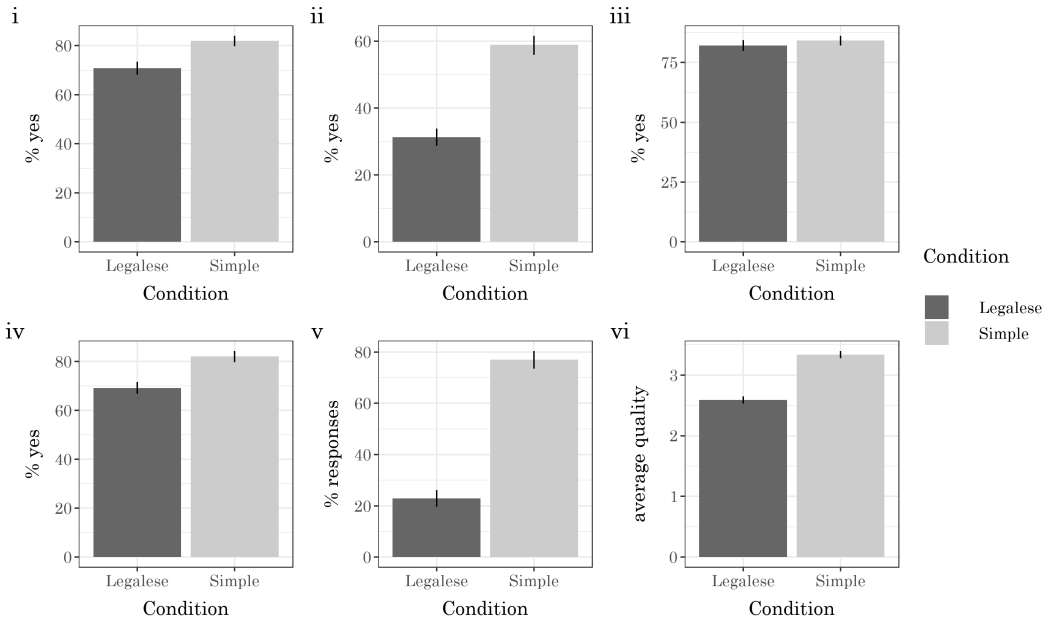

Figure C.4: Proportion of experienced lawyers who endorsed simple and legalese contracts according to different desiderata.

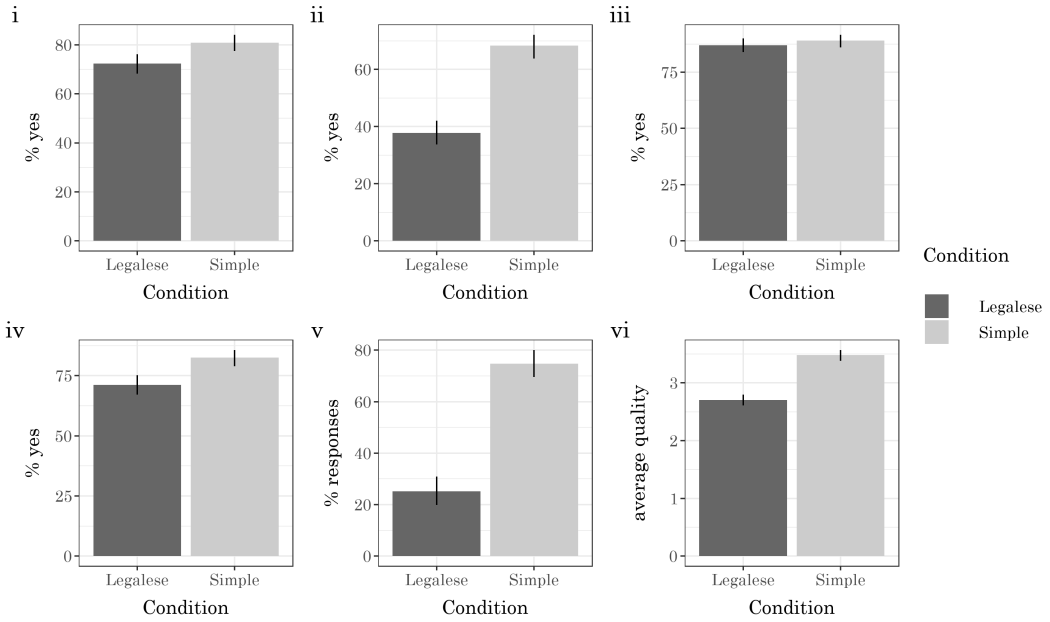

Figure C.5: Proportion of fancy lawyers who endorsed simple and legalese contracts according to different desiderata.

Appendix D

Supplemental Information for *Even Laypeople Use Legalese*

D.1 Corpus Analysis

D.1.1 Syntactic Dependency Length

The processing difficulty associated with center-embedded clauses has been hypothesized to be associated with increased demands on working memory capacity resulting from long-distance syntactic dependencies. However, it remains an open question to what extent legal texts have longer syntactic dependencies relative to baseline texts.

To answer this question, we conducted an additional analysis of the same two corpora as in our center-embedding analyses (United States Code and Corpus of Contemporary American English) where we computed syntactic dependency lengths per sentence.

As with the center-embedding analysis, we used the Stanza package to dependency parse each sentence. Following [121], we filtered out (i) sentences with fewer than 10 alphabetic words; (ii) sentences without end-of-sentence punctu-

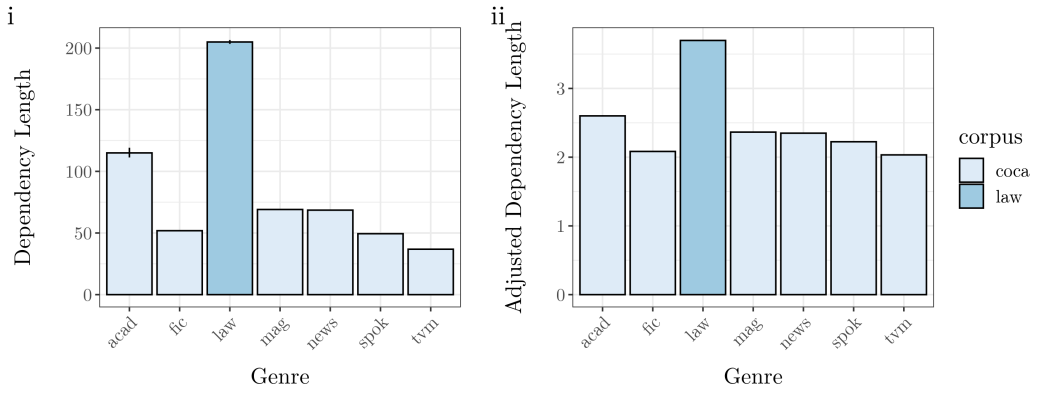

Figure D.1: Dependency length and adjusted dependency length in laws vs academic texts, fiction texts, newspaper articles, magazine articles, spoken transcripts, and TV/Movie scripts.

ation marks and (iii) sentences with 3 or more consecutive punctuation marks.

With the remaining sentences, for each word we calculated the distance between that word and its head word (defined as the difference in ordinal position/index between the word and its head word).

For each sentence, we then calculated (a) the total distance across words; and (b) the total distance across words divided by the total number of words.

Results are visualized in Figure D.1.

D.2 Pilot Experiment

As noted in the main text, to validate the common observation that magic spells are laden with rhyming, we conducted a pilot experiment.

D.2.1 Methods

For this pilot experiment, we constructed 8 sets of instructions to write a magic spell or description of a performed magic spell. Each item consisted of 2 conditions. In the magic spell condition, participants were asked to write, as part

of an authoritative textbook on magic spells, a spell/incantation that accomplishes a certain task. In the control/description condition, participants were asked to write their recollection of having observed the same task having been accomplished through the use of a magic spell.

Description was used as a baseline for a non-performative text with similar conceptual complexity. If the "magic spell" hypothesis were true (i.e. if magic spells contain more rhyming than non-performative texts of similar conceptual complexity), the prediction is that people will rhyme more in the magic spell condition than in control conditions, since the spells are authoritative/ritualistic.

D.2.2 Participants and Procedure

Participants (n=20) were recruited via the online platform prolific. Participants were eligible if they resided in the United States, were 18 years or older, and native speakers of English. Participants performed 8 trials, seeing 8 items exactly once, 4 in each condition.

As in the main experiments, participants were given a comprehension check prior to each trial, were asked before and after the experiment to promise to not use / have used a language model to generate their responses, and were retained according to the same exclusion criteria.

D.2.3 Results

Results shown in D.2. Participants tasked with writing a magic spell rhymed in 58.8% of sentences (95% CI: 54.2 to 63.2), as compared to 1.8% of sentences when tasked with writing a mere recollection of a fantastical event involving a magic spell (95% CI: .9 to 2.9).

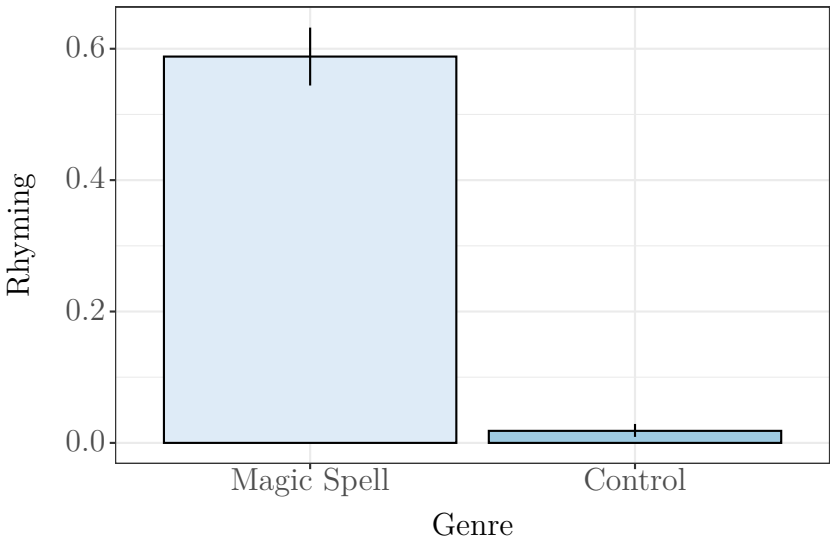

Figure D.2: Prevalence of rhyming per sentence in magic spells vs descriptions of fantastical event involving a magic spell.

Experiment 1 Cover Stories and Items

D.2.4 Cover Stories

Story Condition

You are a **fiction writer**. You are tasked with writing a short story about someone who commits a crime and is punished for committing that crime.

Below are the specifications for the crime. Please write the story, ensuring that it sounds authentic and engaging.

Law Condition

You are a **lawmaker**. You are tasked with writing a law that prohibits a certain crime, and specifies the punishment for that crime if the crime is committed.

Below are the specifications for the law. Please write the law, ensuring that it sounds authoritative and legally binding.

D.2.5 Items

Below are the propositions and source for each item.

Item	Description	Source	Propositions

1	Bank Robbery.	Compare with 18 U.S.C. § 2113	**Requirements of guilt for offense:**
			• Whoever
			  • Entering or attempting to enter
			  • A bank, credit union, or any savings and loan association
			  • With intent to commit
			  • in such bank, credit union, or such savings and loan association
			  • Any felony
			  • affecting such bank, credit union, or such savings and loan association
			  • And in violation of any statute of the United States, or any larceny
			Punishment of offense:
			• 20 years in prison

Continued on next page

Table D.1 – continued from previous page

Item	Description	Source	Propositions

2	Driving under the Influence	Compare with Mass Gen. c.90 § 24	**Requirements of guilt for offense:**
			• Any person who
			  • Upon any way or in any place where the public has a right of access
			  • Operates a vehicle or directs the operation of a vehicle
			  • While under the influence of marijuana; or
			  • With a percentage, by weight, of alcohol of 8/100 or greater
			Punishment of offense:
			• Fine of $5000 and 2.5 years in prison

Continued on next page

Table D.1 – continued from previous page

Item	Description	Source	Propositions

3	Drug Trafficking	Compare with Mass Gen. c.94C § 32E	**Requirements of guilt for offense:**
			• Any person
			  • Who trafficks in marijuana
			  • By knowingly or intentionally:
			  • manufacturing, distributing, dispensing, or cultivating; or
			  • possessing with intent to manufacture, distribute, dispense, or cultivate; or
			  • By bringing into the commonwealth
			  • A net weight of fifty pounds or more of any mixture containing marijuana
			Punishment of offense:
			• Not less than two and one-half nor more than fifteen years in prison

Continued on next page

Table D.1 – continued from previous page

Item	Description	Source	Propositions

4	Money Laundering	Compare with 18 U.S.C. § 1956	**Requirements of guilt for offense:**
			• Any person
			  • Knowing that the property involved in a financial transaction represents the proceeds of some form of unlawful activity
			  • Conducts a financial transaction which involves the proceeds of unlawful activity
			  • With the intent to carry out specified unlawful activity;
			  • With intent to engage in conduct constituting a violation of the Internal Revenue Code; and
			  • Knowing that the transaction is designed to conceal the proceeds of unlawful activity
			Punishment of offense:
			• Fine of $500,000; or
			• Up to 20 years in prison; or
			• Both

Continued on next page

Table D.1 – continued from previous page

Item	Description	Source	Propositions

5	Arson	Compare with 18 U.S.C. § 81; Florida Title XLVI, C. 806.01	**Requirements of guilt for offense:**
			• Any person who
			  • Within the territorial jurisdiction of the United States
			  • Willfully and maliciously; or while in the commission of any felony
			  • Sets fire to or burns
			  • Any dwelling
			  • Whether occupied or not
			  • OR Any structure where persons are normally present
			  • OR Any other structure known to be occupied
			Punishment of offense:
			• Up to 25 years in prison

Continued on next page

Table D.1 – continued from previous page

Item	Description	Source	Propositions

6	Tax Evasion	Compare with 26 U.S.C. § 7202	**Requirements of guilt for offense:**
			• Any person
			  • Required under Title 26 to pay tax
			  • Who willfully fails to
			  • pay tax
			  • Make tax returns
			  • Keep tax records
			  • Or supply such information
			  • At the times required by law or regulations
			Punishment of offense:
			• Fine of $25,000 and 1 year in prison

Continued on next page

Table D.1 – continued from previous page

Item	Description	Source	Propositions

7	Perjury	Compare with 18 U.S.C. § 1621	**Requirements of guilt for offense:**
			• Whoever
			  • Taking an oath before a competent tribunal
			  • In any case in which a law of the United States authorizes an oath to be administered
			  • That he will testify, declare, depose, or certify truly, or that any written testimony, declaration, deposition, or certificate by him subscribed is true
			  • Willfully and contrary to such oath
			  • States or subscribes any material matter which he does not believe to be true
			Punishment of offense:
			• Up to 5 years in prison

Continued on next page

Table D.1 – continued from previous page

Item	Description	Source	Propositions

8	Bribery	Compare with 18 U.S.C. § 201	**Requirements of guilt for offense:**
			• Whoever
			  • Directly or indirectly
			  • Corruptly
			  • Gives, offers or promises
			  • Anything of value
			  • To any public official
			  • With the intent
			  • To influence any official act; OR
			  • Induce the public official to act in violation of their official duties
			Punishment of offense:
			• Imprisonment of up to 2 years

Experiment 2 Control Analysis

Although the description condition in Experiment 2 was intended to as a control condition of equivalent conceptual complexity as laws, it is conceivable that the results from the main analysis of Experiment 2 were in fact driven by differences in conceptual complexity—that is, perhaps participants simply included fewer propositions in their descriptions of laws compared to official laws, resulting in fewer propositions to center-embed.

D.2.6 Methods

To account for this possibility, we first hand-coded the number and proportion of propositions (for a given item) included in a participant's response for each trial. For example, consider the case of item 8, which has 10 propositions. If a participant's draft law included all except 1 of the propositions, we (a) coded that response as having included 9 total propositions and .9 proportion of propositions for that trial; and (b) proposition-by-proposition, coded 9 propositions as 1 (recalled) and 1 proposition as 0 (not recalled).

To determine whether participants' responses were more conceptually complex in the legal condition as compared to the description condition, we conducted a mixed-effects logistic regression, where (a) "was recalled" (0 or 1) was the outcome variable; (b) genre (legal vs description), ordering condition, and the interaction between the two were fixed-effects predictors; and (c) proposition, item and participant as random intercepts.

Next, to assess the effect of conceptual complexity on participants' propensity to center-embed (and whether this eliminated the effect of register), we then conducted two separate regression models similar to those in the main writeup.

The first model was the same as that reported in the main text, with the exception that participant responses were removed if the proportion of propositions included was not greater than 80% for a given trial.

The second model was identical to the first model, with the exception that conceptual complexity (operationalized as proportion of propositions recalled for a given trial) was added as a fixed-effect predictor in the regression.

D.2.7 Results

Number of Propositions

Descriptively, propositions were included in laws at a higher rate (.856) than descriptions (.837).

This difference was statistically significant according to our model (p=.031).

Effect of Propositions on Center-Embedding

Our main results were robust to the two additional control models.

First, after filtering out all trials with 80% or fewer propositions included, there remained a significant main effect of register in terms of amount of center-embedded clauses (β =-1.514,SE=.242,p<.0001) and binary presence of center-embedding (β =-1.525,SE=.261,p<.0001) .

Although our second analyses revealed a main effect of conceptual complexity on amount of center-embedding (β =2.419,SE=.832,p=.004), they did not reveal a main effect of conceptual complexity on the binary presence of center-embedding (p=.066).

Moreover, these models likewise still revealed a significant main effect of register in terms of amount of center-embedded clauses (β (β =-1.447,SE=.264,p<.0001) and binary presence of center-embedding (β =-1.514,SE=.242,p<.0001).

Results after filtering trials without greater than 80% of propositions included are visualized in D.3

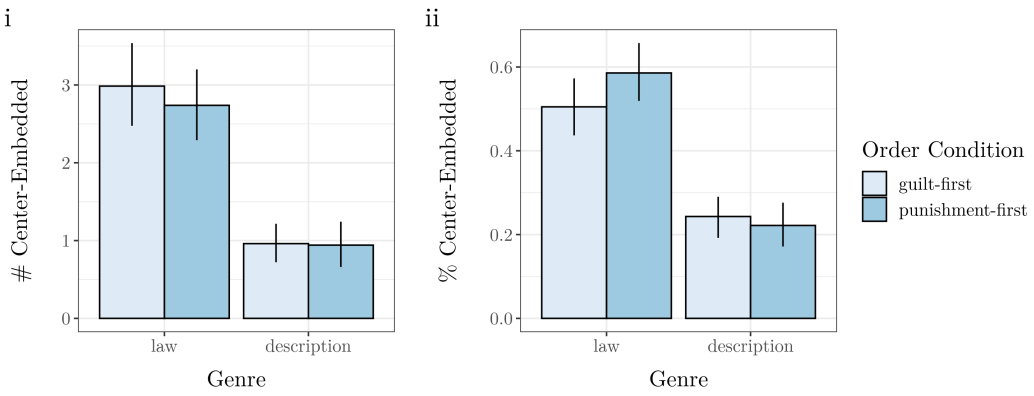

Figure D.3: Prevalence of center-embedding after filtering responses without greater than 80% of propositions for a given item.

References

[1] *Mcboyle v. united states*, 1931.

[2] *Moskal v. united states*, 1990.

[3] *Grayned v. city of rockford*, 1972.

[4] *Hoffman estates v. flipside, hoffman estates, inc.* 1982.

[5] R. H. Fallon Jr, "The statutory interpretation muddle," *Nw. UL Rev.*, vol. 114, p. 269, 2019.

[6] K. Tobia, B. G. Slocum, and V. Nourse, "Ordinary meaning and ordinary people," *U. Pa. L. Rev.*, vol. 171, p. 365, 2022.

[7] B. G. Slocum, "Ordinary meaning," in *Ordinary Meaning*, University of Chicago Press, 2015.

[8] K. P. Tobia, "Testing ordinary meaning," *Harv. L. Rev.*, vol. 134, p. 726, 2020.

[9] *California civil code, part 2. contracts, § 1549–1701 (2018)*, Accessed: 2024-05-28, 2018. URL: https://leginfo.legislature.ca.gov/faces/codesTOCSelected.xhtml?tocCode=CIV.

[10] *Richards v. united states*, 1962.

[11] *United states v. turkette*, 1981.

[12] *Venter v. r*, TS 910, 1907.

[13] *River wear commissioners v adamson*, 2 App Cas 743, 1877.

[14] *Interpretation act (cap 1) s 9a*, 1993.

[15] B. G. Slocum and J. Wong, "The vienna convention and ordinary meaning in international law," *Yale J. Int'l L.*, vol. 46, p. 191, 2021.

[16] *Vienna convention on the law of treaties art. 31, opened for signature may 23, 1969, 1155 u.n.t.s. 331*, United Nations Treaty Series, 1969.

[17] *Levin by levin v. desert palace, inc.* 1983.

[18] *Jowett, inc. v. us*, 2000.

[19] *Harris v. department of veterans affairs*, 1998.

[20] A. Peters, "Are they all textualists now?" *Northwestern University Law Review*, vol. 118, no. 5, pp. 1201–1276, 2024.

[21] E. Martínez and K. Tobia, "What do law professors believe about law and the legal academy?" *Geo. L. J.*, vol. 112, 2023.

[22] A. C. Barrett, "Congressional insiders and outsiders," *The University of Chicago Law Review*, pp. 2193–2211, 2017.

[23] E. Gibson, R. Futrell, S. P. Piantadosi, I. Dautriche, K. Mahowald, L. Bergen, and R. Levy, "How efficiency shapes human language," *Trends in cognitive sciences*, vol. 23, no. 5, pp. 389–407, 2019.

[24] S. T. Piantadosi, H. Tily, and E. Gibson, "Word lengths are optimized for efficient communication," *Proceedings of the National Academy of Sciences*, vol. 108, no. 9, pp. 3526–3529, 2011.

[25] S. T. Piantadosi, H. Tily, and E. Gibson, "The communicative function of ambiguity in language," *Cognition*, vol. 122, no. 3, pp. 280–291, 2012.

[26] S. T. Piantadosi, "Zipf's word frequency law in natural language: A critical review and future directions," *Psychonomic bulletin & review*, vol. 21, no. 5, pp. 1112–1130, 2014.

[27] K. Mahowald, E. Fedorenko, S. T. Piantadosi, and E. Gibson, "Info/information theory: Speakers choose shorter words in predictive contexts," *Cognition*, vol. 126, no. 2, pp. 313–318, 2013.

[28] K. Mahowald, I. Dautriche, E. Gibson, and S. T. Piantadosi, "Word forms are structured for efficient use," *Cognitive science*, vol. 42, no. 8, pp. 3116–3134, 2018.

[29] R. Futrell and E. Gibson, "L2 processing as noisy channel language comprehension," *Bilingualism: Language and Cognition*, vol. 20, no. 4, pp. 683–684, 2017.

[30] R. Futrell and R. Levy, "Noisy-context surprisal as a human sentence processing cost model," in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, vol. 1, 2017, pp. 688–698.

[31] R. Ryskin, R. Futrell, S. Kiran, and E. Gibson, "Comprehenders model the nature of noise in the environment," *Cognition*, vol. 181, pp. 141–150, 2018.

[32] Y. Zhang, R. Ryskin, and E. Gibson, "A noisy-channel approach to depth-charge illusions," *Cognition*, vol. 232, p. 105 346, 2023.

[33] E. Gibson, S. T. Piantadosi, K. Brink, L. Bergen, E. Lim, and R. Saxe, "A noisy-channel account of crosslinguistic word-order variation," *Psychological science*, vol. 24, no. 7, pp. 1079–1088, 2013.

[34] E. Gibson, L. Bergen, and S. T. Piantadosi, "Rational integration of noisy evidence and prior semantic expectations in sentence interpretation," *Proceedings of the National Academy of Sciences*, p. 201 216 438, 2013.

[35] G. K. Zipf, "Human behavior and the principle of least effort: An introd. to human ecology," 1949.

[36] R. Futrell, K. Mahowald, and E. Gibson, "Large-scale evidence of dependency length minimization in 37 languages," *Proceedings of the National Academy of Sciences*, vol. 112, no. 33, pp. 10 336–10 341, 2015.

[37] E. Gibson and N. J. Pearlmutter, "Constraints on sentence comprehension," *Trends in cognitive sciences*, vol. 2, no. 7, pp. 262–268, 1998.

[38] P. M. Tiersma, "Reforming the language of jury instructions," *Hofstra L. Rev.*, vol. 22, p. 37, 1993.

[39] P. Tiersma, "Some myths about legal language," *Journal of Law, Culture and Humanities, Forthcoming, Loyola-LA Legal Studies Paper*, no. 2005-26, 2005.

[40] P. Tiersma, "The nature of legal language," in *AILA applied linguistics series: Vol. 5. Dimensions of forensic linguistics*, John Benjamins Publishing Company, 2008, pp. 7–25.

[41] M. E. Masson and M. A. Waldron, "Comprehension of legal contracts by non-experts: Effectiveness of plain language redrafting," *Applied cognitive psychology*, vol. 8, no. 1, pp. 67–85, 1994.

[42] *Act of june 12, 1913, p.l. 481, 37 p.s. § 61*, Public Law, 1913.

[43] E. Gibson, "Linguistic complexity: Locality of syntactic dependencies," *Cognition*, vol. 68, no. 1, pp. 1–76, 1998.

[44] G. A. Miller and N. Chomsky, "Finitary models of language users," in *Handbook of Mathematical Psychology*, John Wiley & Sons, 1963, pp. 2–419.

[45] K. Rayner, J. Ashby, A. Pollatsek, and E. D. Reichle, "The effects of frequency and predictability on eye fixations in reading: Implications for the ez reader model.," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 30, no. 4, p. 720, 2004.

[46] F. Ferreira, "The misinterpretation of noncanonical sentences," *Cognitive psychology*, vol. 47, no. 2, pp. 164–203, 2003.

[47] Y. A. Arbel and A. Toler, "All-caps," *Journal of Empirical Legal Studies*, vol. 17, no. 4, pp. 862–896, 2020.

[48] Exec. Order No. 13648, *44 fr 69609*, 1979.

[49] Plain Language Action Information Network, *Federal plain language guidelines*, 2011.

[50] M. A. Blasie, "The rise of plain language laws," *U. Miami L. Rev.*, vol. 76, p. 447, 2021.

[51] M. A. Blasie, "Regulating plain language," *Wis. L. Rev.*, p. 687, 2023.

[52] Plain Writing Act of 2010. ().

[53] R. P. Charrow and V. R. Charrow, "Making legal language understandable: A psycholinguistic study of jury instructions," *Columbia law review*, vol. 79, no. 7, pp. 1306–1374, 1979.

[54] A. Elwork, B. D. Sales, and J. J. Alfini, *Making jury instructions understandable*. Michie Company, 1982.

[55] L. Heuer and S. D. Penrod, "Instructing jurors," *Law and Human Behavior*, vol. 13, no. 4, pp. 409–430, 1989.

[56] S. S. Diamond, B. Murphy, and M. R. Rose, "The kettleful of law in real jury deliberations: Successes, failures, and next steps," *Nw. UL Rev.*, vol. 106, p. 1537, 2012.

[57] N. E. S. Goldstein, L. O. Condie, R. Kalbeitzer, D. Osman, and J. L. Geier, "Juvenile offenders' miranda rights comprehension and self-reported likelihood of offering false confessions," *Assessment*, vol. 10, no. 4, pp. 359–369, 2003.

[58] R. Rogers, K. S. Harrison, D. W. Shuman, K. W. Sewell, and L. L. Hazelwood, "An analysis of miranda warnings and waivers: Comprehension and coverage," *Law and human behavior*, vol. 31, no. 2, pp. 177–192, 2007.

[59] P. L. Refo, "The vanishing trial," *Journal of Empirical Legal Studies*, vol. 1, no. 3, pp. v–vii, 2004.

[60] J. S. Rakoff, H. Daumier, and A. C. Case, "Why innocent people plead guilty," *The New York Review of Books*, vol. 20, 2014.

[61] R. S. Nickerson, "How we know—and sometimes misjudge—what others know: Imputing one's own knowledge to others.," *Psychological bulletin*, vol. 125, no. 6, p. 737, 1999.

[62] P. J. Hinds, "The curse of expertise: The effects of expertise and debiasing methods on prediction of novice performance.," *Journal of Experimental Psychology: Applied*, vol. 5, no. 2, p. 205, 1999.

[63] S. Pinker, *The sense of style: The thinking person's guide to writing in the 21st century*. Penguin Books, 2015.

[64] P. Tiersma, "The nature of legal language," *Dimensions of forensic linguistics*, pp. 7–25, 2008.

[65] K. A. Adams, *A Manual of Style for Contract Drafting*, 4th ed. American Bar Association, 2017, ISBN: 978-1634259644.

[66] C. A. Hill, "A comment on language and norms in conplex business contracting," *Chi.-Kent L. Rev.*, vol. 77, p. 29, 2001.

[67] D. Mellinkoff, *The language of the law*. Wipf and Stock Publishers, 2004.

[68] T. Samples, K. Ireland, and C. Kraczon, "Tl; dr: The law and linguistics of social platform terms-of-use," *Berkeley Technology Law Journal (forthcoming 2023)*, 2023.

[69] J. A. Obar and A. Oeldorf-Hirsch, "The biggest lie on the internet: Ignoring the privacy policies and terms of service policies of social networking services," *Information, Communication & Society*, vol. 23, no. 1, pp. 128–147, 2020.

[70] S. Goźdź-Roszkowski, *Patterns of linguistic variation in American legal English: A corpus-based study*. Peter Lang Frankfurt am Main, 2011.

[71] D. B. Paul and J. Baker, "The design for the wall street journal-based csr corpus," in *Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992*, 1992.

[72] M. Davies, "The 385+ million word corpus of contemporary american english (1990–2008+): Design, architecture, and linguistic insights," *International journal of corpus linguistics*, vol. 14, no. 2, pp. 159–190, 2009.

[73] American Law Institute and National Conference of Commissioners on Uniform State Laws, *Uniform commercial code (U.C.C.) s 2-216(2))*, 2002.

[74] C. B. Marks, M. J. Doctorow, and M. C. Wittrock, "Word frequency and reading comprehensiony1," *The Journal of Educational Research*, vol. 67, no. 6, pp. 259–262, 1974.

[75] K. P. Tobia, "Legal concepts and legal expertise," *Social Science Research Network*, vol. 0, 2020.

[76] B. J. Baldie, "The acquisition of the passive voice," *Journal of Child Language*, vol. 3, no. 3, pp. 331–348, 1976.

[77] K. E. Stanovich and R. F. West, "Exposure to print and orthographic processing," *Reading Research Quarterly*, vol. 0, pp. 402–433, 1989.

[78] M. Moore and P. C. Gordon, "Reading ability and print exposure: Item response theory analysis of the author recognition test," *Behavior research methods*, vol. 47, no. 4, pp. 1095–1109, 2015.

[79] American Law Institute, *Second restatement of contracts*, 1981.

[80] D. J. Acheson, J. B. Wells, and M. C. MacDonald, "New and updated tests of print exposure and reading abilities in college students," *Behavior research methods*, vol. 40, no. 1, pp. 278–289, 2008.

[81] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and psychological measurement*, vol. 20, no. 1, pp. 37–46, 1960.

[82] M. L. McHugh, "Interrater reliability: The kappa statistic," *Biochemia medica*, vol. 22, no. 3, pp. 276–282, 2012.

[83] E. T. Bergman and H. L. Roediger, "Can bartlett's repeated reproduction experiments be replicated?" *Memory & Cognition*, vol. 27, no. 6, pp. 937–947, 1999.

[84] J. S. Fisher and G. A. Radvansky, "Patterns of forgetting," *Journal of Memory and Language*, vol. 102, pp. 130–141, 2018.

[85] K. P. Tobia, "Essays in experimental jurisprudence," Ph.D. dissertation, Yale University, 2019.

[86] K. Tobia, "Law and the cognitive science of ordinary concepts," *Law and Mind: A Survey of Law and the Cognitive Sciences*, 2021.

[87] M. J. Kieffer, "Socioeconomic status, english proficiency, and late-emerging reading difficulties," *Educational Researcher*, vol. 39, no. 6, pp. 484–486, 2010.

[88] R. H. Bradley and R. F. Corwyn, "Socioeconomic status and child development," *Annual review of psychology*, vol. 53, no. 1, pp. 371–399, 2002.

[89] Legal Services Corporation, *Justice gap report: Measuring the civil needs of low-income americans*, 2017.

[90] S. Azuelos-Atias, "Making legal language clear to legal laypersons," *Legal Pragmatics*, vol. 288, p. 101, 2018.

[91] P. Kendeou and P. Van Den Broek, "The effects of prior knowledge and text structure on comprehension processes during reading of scientific texts," *Memory & cognition*, vol. 35, no. 7, pp. 1567–1577, 2007.

[92] K. Cain, J. Oakhill, and K. Lemmon, "Individual differences in the inference of word meanings from context: The influence of reading comprehension, vocabulary knowledge, and memory capacity.," *Journal of educational psychology*, vol. 96, no. 4, p. 671, 2004.

[93] D. L. Long, C. Prat, C. Johns, P. Morris, and E. Jonathan, "The importance of knowledge in vivid text memory: An individual-differences investigation of recollection and familiarity," *Psychonomic Bulletin & Review*, vol. 15, no. 3, pp. 604–609, 2008.

[94] L. G. Noordman and W. Vonk, "Readers' knowledge and the control of inferences in reading," *Language and Cognitive Processes*, vol. 7, no. 3-4, pp. 373–391, 1992.

[95] Y. Ozuru, K. Dempsey, and D. S. McNamara, "Prior knowledge, reading skill, and text cohesion in the comprehension of science texts," *Learning and instruction*, vol. 19, no. 3, pp. 228–242, 2009.

[96] B. A. Garner *et al.*, "Black's law dictionary," 2004.

[97] A. L. Institute, *Model Penal Code*. 1984.

[98] M. H. Arsanjani, "The rome statute of the international criminal court," *American Journal of International Law*, vol. 93, no. 1, pp. 22–43, 1999.

[99] H. E. Mattila, *Comparative legal linguistics: language of law, Latin and modern lingua Francas*. Routledge, 2016.

[100] M. Adler, "The plain language movement," in *The Oxford handbook of language and law*, 2012.

[101] E. Martínez, F. Mollica, and E. Gibson, "Poor writing, not specialized concepts, drives processing difficulty in legal language," *Cognition*, vol. 224, p. 105 070, 2022.

[102] E. Martínez, F. Mollica, and E. Gibson, "Even lawyers do not like legalese," *Proceedings of the National Academy of Sciences*, vol. 120, no. 23, e2302672120, 2023.

[103] J. A. Love, "Fair notice about fair notice," *Yale LJ*, vol. 121, p. 2395, 2011.

[104] P. H. Robinson, "Fair notice and fair adjudication: Two kinds of legality," *U. Pa. L. Rev.*, vol. 154, p. 335, 2005.

[105] Library of Congress. "United states statutes at large." (2021).

[106] M. Davies, "Expanding horizons in historical linguistics with the 400-million word corpus of historical american english," *Corpora*, vol. 7, no. 2, pp. 121–157, 2012.

[107] P. Qi, Y. Zhang, Y. Zhang, J. Bolton, and C. D. Manning, "Stanza: A python natural language processing toolkit for many human languages," *arXiv preprint arXiv:2003.07082*, vol. 0, 2020.

[108] M. Brysbaert and B. New, "Moving beyond kučera and francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for american english," *Behavior research methods*, vol. 41, no. 4, pp. 977–990, 2009.

[109] G. A. Miller, "Wordnet: A lexical database for english," *Communications of the ACM*, vol. 38, no. 11, pp. 39–41, 1995.

[110] W. J. Van Heuven, P. Mandera, E. Keuleers, and M. Brysbaert, "Subtlex-uk: A new and improved word frequency database for british english," *Quarterly journal of experimental psychology*, vol. 67, no. 6, pp. 1176–1190, 2014.

[111] R. Hiltunen, "The grammar and structure of legal texts," in *The Oxford handbook of language and law*, 2012.

[112] D. Kurzon, "'legal language': Varieties, genres, registers, discourses," *International Journal of Applied Linguistics*, vol. 7, no. 2, pp. 119–139, 1997.

[113] K. Marton and R. G. Schwartz, "Working memory capacity and language processes in children with specific language impairment," 2003.

[114] R. Flesch, "Flesch-kincaid readability test," *Retrieved October*, vol. 26, no. 3, p. 2007, 2007.

[115] M. Solnyshkina, R. Zamaletdinov, L. Gorodetskaya, and A. Gabitov, "Evaluating text complexity and flesch-kincaid grade level," *Journal of social studies education research*, vol. 8, no. 3, pp. 238–248, 2017.

[116] P.-C. Bürkner, "Brms: An r package for bayesian multilevel models using stan," *Journal of statistical software*, vol. 80, pp. 1–28, 2017.

[117] E. Martinez, F. Mollica, and E. Gibson, *Accessibility of legal texts over time*, Mar. 2023. URL: osf.io/ambp4.

[118] *United States Code*. Washington, D.C.: U.S. Government Publishing Office, 2021. URL: https://www.govinfo.gov/app/collection/uscode/2021.

[119] M. Davies, "The 385+ million word corpus of contemporary american english (1990-2008+): Design, architecture, and linguistic insights," *International Journal of Corpus Linguistics*, vol. 14, no. 2, pp. 159–190, 2009. DOI: 10.1075/ijcl.14.2.02dav.

[120] E. Martinez, F. Mollica, Y. Liu, A. Podrug, and E. Gibson, "What did i sign? a study of the impenetrability of legalese in contracts," in *Proceedings of the Annual Meeting of the Cognitive Science Society*, vol. 43, 2021.

[121] E. Martinez, F. Mollica, and E. Gibson, "So much for plain language: An analysis of the accessibility of united states federal laws over time," *Journal of Experimental Psychology: General*, in press.

[122] N. Chomsky, "On certain formal properties of grammars," *Information and control*, vol. 2, no. 2, pp. 137–167, 1959.

[123] J. Nyarko, "Stickiness and incomplete contracts," *U. Chi. L. Rev.*, vol. 88, p. 1, 2021.

[124] A. B. Association. "Aba profile of the legal profession." (2020), URL: https://www.americanbar.org/content/dam/aba/administrative/news/2020/07/potlp2020.pdf (visited on 03/28/2023).

[125] E. Martínez, "Measuring legal concepts," *Available at SSRN 4715691*, 2024.

[126] A. L. Institute, *Restatement (Second) of Contracts*. St. Paul, MN: American Law Institute, 1981.

[127] A. L. Institute and N. C. of Commissioners on Uniform State Laws, *Uniform Commercial Code*, Official Text. Philadelphia, PA: American Law Institute, 2020. URL: https://www.ali.org/publications/show/uniform-commercial-code/.

[128] *Massachusetts General Laws Chapter 90, Section 24*.

[129] J. L. Austin, "Performative utterances," in *Philosophical Papers*, J. O. Urmson and G. J. Warnock, Eds., Clarendon Press, 1961.

[130] W. Shakespeare, *Macbeth*. E-Kitap Projesi & Cheapest Books, 2024.

[131] J. K. Rowling, *Harry Potter and the philosopher's stone*. Bloomsbury Publishing, 2015, vol. 1.

[132] R. Anderson IV, "Path dependence, information, and contracting in business law and economics," *Wis. L. Rev.*, p. 553, 2020.

[133] J. Cross, "The fair notice fiction," *Alabama Law Review*, vol. 75, no. 2, 2023.

[134] C. A. Hill and C. King, "How do german contracts do as much with fewer words," *Chi.-Kent L. Rev.*, vol. 79, p. 889, 2004.

[135] E. Gibson, "The dependency locality theory: A distance-based theory of linguistic complexity," *Image, language, brain*, pp. 95–126, 2000.

[136] R. L. J. Futrell, "Memory and locality in natural language," Ph.D. dissertation, Massachusetts Institute of Technology, 2017.

[137] H. Spamann, *Simplified DGCL: including a guide to the federal proxy rules*. 2021.

[138] C. Coupette, J. Singh, and H. Spamann, "Simplify your law: Using information theory to deduplicate legal documents," in *2021 International Conference on Data Mining Workshops (ICDMW)*, IEEE, 2021, pp. 631–638.

[139] D. M. Katz, M. J. Bommarito, S. Gao, and P. Arredondo, "Gpt-4 passes the bar exam," *Philosophical Transactions of the Royal Society A*, vol. 382, no. 2270, p. 20 230 254, 2024.

[140] E. Martínez, "Re-evaluating gpt-4's bar exam performance," *Artificial Intelligence and Law*, pp. 1–24, 2024.

[141] J. Huang and M. Tan, "The role of chatgpt in scientific communication: Writing better scientific review articles," *American journal of cancer research*, vol. 13, no. 4, p. 1148, 2023.

[142] M. de Rivero, C. Tirado, and W. Ugarte, "Formalstyler: Gpt based model for formal style transfer based on formality and meaning preservation.," in *KDIR*, 2021, pp. 48–56.

[143] P. Green and C. J. MacLeod, "Simr: An r package for power analysis of generalized linear mixed models by simulation," *Methods in Ecology and Evolution*, vol. 7, no. 4, pp. 493–498, 2016.

[144] A. C. Graesser, D. S. McNamara, M. M. Louwerse, and Z. Cai, "Coh-metrix: Analysis of text on cohesion and language," *Behavior research methods, instruments, & computers*, vol. 36, no. 2, pp. 193–202, 2004.