Poor writing, not specialized concepts, drives processing difficulty in legal language

Eric Martínez a, *, Francis Mollica b, Edward Gibson a

a Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139
b Institute for Language, Cognition and Computation, School of Informatics, University of Edinburgh, UK

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 (Charrow & Charrow, 1979; Diamond, Murphy, & Rose, 2012; Elwork, Sales, & Alfini, 1982; Heuer & Penrod, 1989) and Miranda warnings (Goldstein, Condie, Kalbeitzer, Osman, & Geier, 2003; Rogers, Harrison, Shuman, Sewell, & Hazelwood, 2007). 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; Rakoff, Daumier, & Case, 2014; Reif, 2004). 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 (Gibson, 1998; Miller & Chomsky, 1963). 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: Rayner, Ashby, Pollatsek, & Reichle, 2004), passive-voice constructions...
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 properties associated with processing difficulty relative to standard English texts. To do so, we expanded the corpus of contracts used by Gozdz-Roszkowski (n ≈ 1 million words; Gozdz-Roszkowski, 2011) with an additional set of contracts (n ≈ 2.5 million words) drawn from Westlaw’s database of court documents between 2018 and 2020. For our standard English corpus, we compiled: (a) a sample of Wall Street Journal articles (n ≈ 5 million words published in 1996, as part of the CSR Wall Street Journal corpus (Paul & Baker, 1992)); and (b) a broad sample (n ≈ 10 million words) of TV/Movie scripts, spoken language, newspaper articles, blogs, magazine articles and web pages from the Corpus of Contemporary American English (COCO: Davies, 2009). 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.\(^1\)

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.

---

\(^1\) 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.
2.2. Results

Results are visualized in Fig. 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% CI 2.48 – 2.63), passive voice (OR = 4.35; 95% CI 4.11 – 4.61), and non-standard capitalization (OR = 1.78; 95% CI 1.75 – 1.82).

2.2.1. Capitalization

In our analysis (Fig. 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).

2.2.2. Word frequency

As seen in Fig. 1 ii, 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 standard-English corpus (95% CI: 450.244 to 452.567).

2.2.3. 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 standard-English corpus (95% CI: 16.090 to 16.165). We observe similar results in the anti-conservative analysis (see SI).

2.2.4. 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 (Fig. 1 iii), 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).

2.2.5. 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 (Fig. 1 iv)–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).

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.

3.1. Methods

To do so, we developed a paradigm building on Masson and Waldron (1994), 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; Stanovich & West, 1989; Moore & Gordon, 2015) 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.

3.1.1. Materials

Our primary materials consisted of 12 pairs of short contract excerpts of roughly 150 words each (see Supplemental Fig. 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 (American Law Institute, 1981; American Law Institute and National Conference of Commissioners on Uniform State Laws, 2002).

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; Moore & Gordon, 2015) as a proxy for language experience.

All experiment code, data and analysis scripts are available on OSF: https://osf.io/xcqd9/.

3.1.2. 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 the quality of the data we received. Participants 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 pseudorandomly 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 (Acheson, Wells, & MacDonald, 2008). 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.

3.1.3. 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 (Cohen, 1960; McHugh, 2012). 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).

3.2. Experimental results

3.2.1. Comprehension

Fig. 2i illustrates the comprehension accuracy across registers and Fig. 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 Masson and Waldron (1994), 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 ($\beta=-0.121, SE=0.049, p<0.05$), 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 ($\beta=0.229, SE=0.080, p<0.05$).

\footnote{Our results do not qualitatively change if we use the full item set}
3.2. 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.

Fig. 2 iii 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 (Bergman & Roediger, 1999). 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 using Bayesian mixed effect logistic regression. We conducted a replication experiment with frequency, and no effect of capitalization and voice. In the SI, we find convergent results in a similar exploratory analysis of the comprehension data.

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 with frequency, and no effect of capitalization and voice. In the SI, we find convergent results in a similar exploratory analysis of the comprehension data.

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

![Fig. 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 (iv). Posterior distribution over logistic regression coefficients reflecting the influence of condition and each surface property on recall (v). Negative coefficient values reflect a decrease in recall performance. Points reflect the median; the outer line range reflects the 95% credible intervals and the inner line range reflect the 80% credible intervals.](image)
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 (Gozdz-Roszkowska, 2011) and non-standard capitalization (Arbel & Toler, 2020) 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 (Tobia, 2019). 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 (Tobia, 2019; Tobia, 2021). 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 underscores 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 (Bradley & Corwyn, 2002; Kieffer, 2010), and those of lower SES face greater disenfranchisement from the legal system (Legal Services Corporation, 2017), 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., Masson & Waldron, 1994). 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” (Mellinkoff, 2004).

Lastly, lawyers may not choose to write in an esoteric manner. Similar to the “curse of knowledge” (Hinds, 1999; Nickerson, 1999), they may not realize that their language is too complicated for the average reader to understand (Azuelos-Atias, 2018). 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 (Cain, Oakhill, & Lemmon, 2004; Kendeou & Van Den Broek, 2007; Long, Pratt, Johns, Morris, & Jonathan, 2008; Noordman & Vonk, 1992; Ozuru, Dempsey, & McNamara, 2009). 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.

Author statement

Eric Martinez: design of experimental materials, collection of corpus materials, implementation and analysis of corpus scripts, writing and editing manuscript. Frank Mollica: implementation and analysis of experimental study, supervision of corpus analysis, writing and editing manuscript. Edward Gibson: Supervision of experimental study and corpus analysis, writing and editing of manuscript.
Cognition 224 (2022) 105070

E. Martínez et al.

Fundings
We thank the Brain and Cognitive Science department at MIT for funding this research.

Acknowledgements
We are grateful to Anita Podrug and Yufei Liu for coding the recall responses in the experimental data, and Kinan Martín for help drafting the scripts for the corpus analyses. We would like to thank (alphabetical by first name) Alex Paunov, Anya Ivanova, Ashley Thomas, Ev Fedorenko, Giuseppe Ricciardi, Greta Tuckute, Halie Olsen, Heather Kosakowski, Rachel Ryskin, Ray Jackendorf, Rebecca Saxe, Saima Malik-Moraleda, and Shari Liu for their helpful comments at various stages of this paper.

Appendix A. Supplementary data
Supplementary data to this article can be found online at https://doi.org/10.1016/j.cognition.2022.105070.

References


