Supplemental Information for: So much for plain language: An analysis of the accessibility of United States federal laws over time

Eric Martínez[1†], Francis Mollica[2], Edward Gibson[1]

[1]Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139

[2]Institute for Language, Cognition and Computation, School of Informatics, University of Edinburgh, UK

[†] Correspondence to ericmart@mit.edu

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 1.

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

Table 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: Davies, 2009), 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 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 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 3.

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

Table 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 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 (American Law Institute and National Conference of Commissioners on Uniform State Laws, 2002).

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

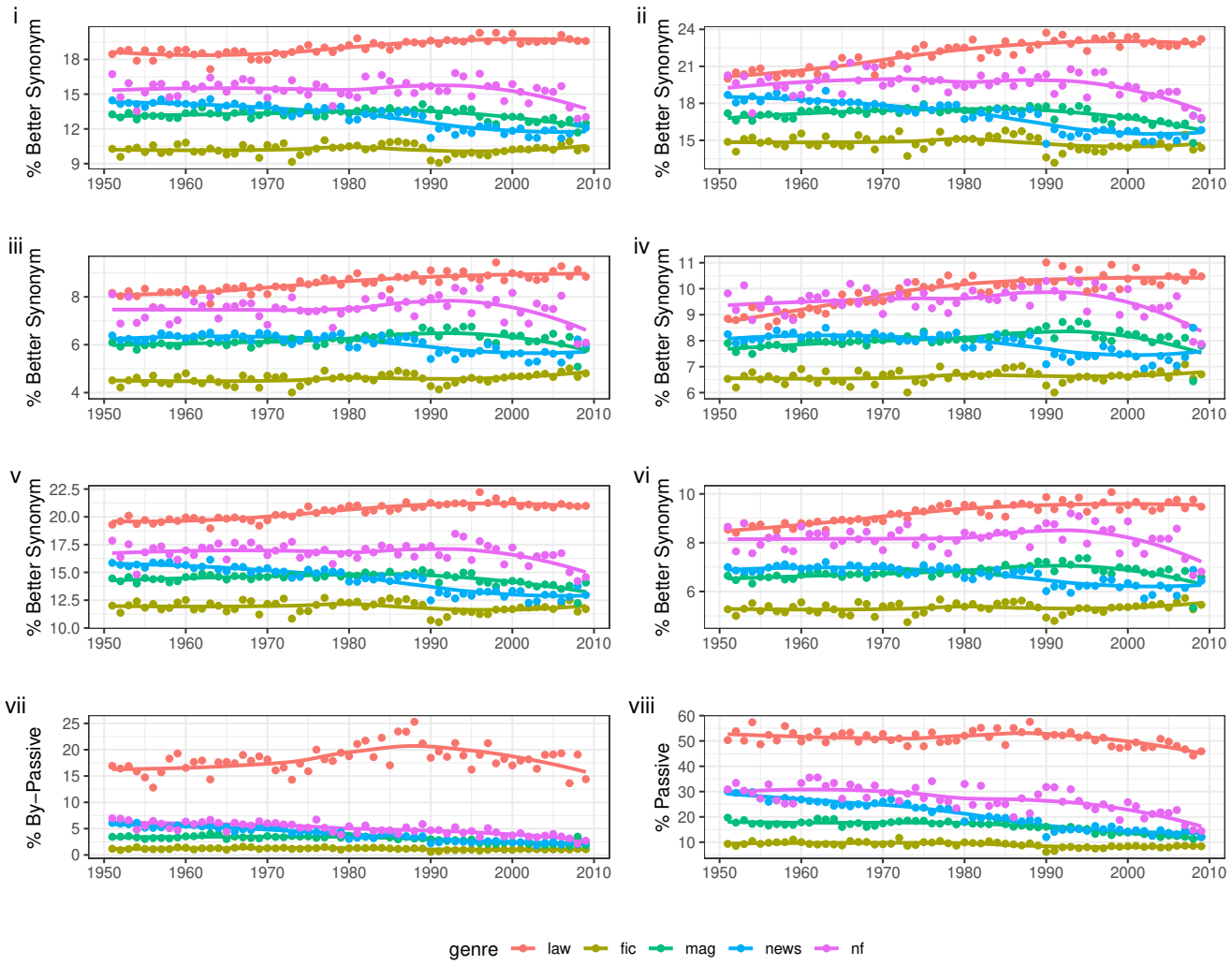

Figure 1: Comparison of supplemental indices of linguistic processing difficulty in federal laws vs four genres of standard English, including fiction books, magazine articles, newspaper articles, and non-fiction books (1951-2009).

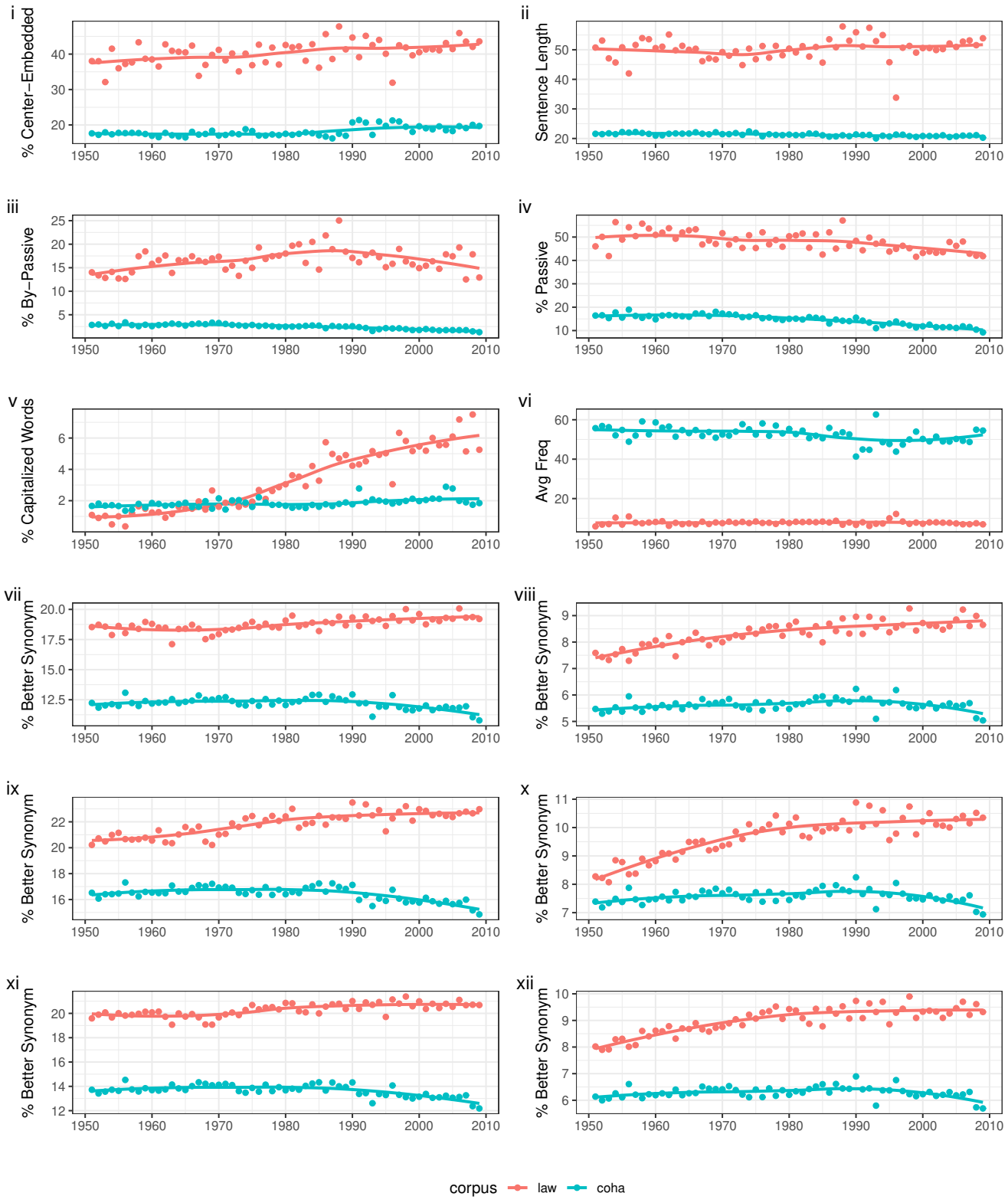

Figure 2: Comparison of indices of linguistic processing difficulty in federal laws vs Corpus of Historical American English (1951-2009).

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 1 i (for all content words) and iii (for all alphabetic words), as well as in 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 1 i (for all content words) and iii (for all alphabetic words), as well as in 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 1 v (for all content words) and vi (for all alphabetic words), as well as in 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 between federal laws and the COHA corpus of by-passives and passives overall are visualized in Figure 2 iii and iv, respectively. Genre-by-genre comparisons of by-passives and passives are visualized in Figures 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 2 i.

References

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

Davies, M. (2009). The 385+ million word corpus of contemporary american english (1990–2008+): Design, architecture, and linguistic insights. *International journal of corpus linguistics*, *14*(2), 159–190.