

Investigating Psycholinguistic Sources of Complexity in Legal Texts: Corpus Evidence from Historical Corpora and Indian Judiciary Data

Akanksha Sagar¹(akisagar91@gmail.com), Gurujegan Murugesan¹, Sidharth Ranjan¹

¹Indian Institute of Technology Jodhpur

Legal documents are often difficult for laypeople to understand due to the persistent use of complex and archaic language, commonly referred to as *legalese*. This limited comprehensibility restricts individuals to interpret contractual obligations and effectively access justice or legal compliance. Following previous work in the literature [1,2,3,4], we investigate the *copy-paste hypothesis*, which proposes that legal professionals continue producing or writing complex registers either out of laziness or convenience or respect for tradition or extensive re-use of older legal templates that already contain convoluted terms. We further examine the extent to which the complexity in the legal texts can be explained through psycholinguistic features[5] viz., (i) usage of non-standard capitalization, (ii) low frequency jargons, (iii) passive voice, and (iv) center embedded words.

Method: We first extracted a set of legal archaic expressions comprising compound deictic and pronominal adverbs (viz., *herein*, *thereof*, *whereby*, *aforementioned*; see Table 1) from publicly available legal contract templates (<https://lawrato.com/legal-documents>) obtained from the Indian Judiciary System. The database consists of 56 contract templates with 76,155 tokens, 60,399 total words and 3,087 total sentences. We then compared archaic set frequency distributions across two historical corpora: (i) Corpus of Old English (OE; 650-1150 AD) originally sourced from Universal Dependency Treebank [6], (ii) Corpus of Late Modern English (CLMET-3.1) [7] spanning approximately 15-19th century. For psycholinguistic analysis, we quantified the four aforementioned features in our contract template dataset and subsequently performed count analyses.

Results: Our analysis on the Indian Judiciary System revealed that the identified archaic expressions were virtually absent in the OE corpus. Instead, their usage emerged prominently in the Late Modern English period, with peak frequencies observed around the 1760s and trend continued in the 19th century [Fig. 1]. These constructions continue to persist extensively within contemporary Indian legal contract templates from where our Archaic words are sourced. This persistence provides corpus evidence for the *copy-paste hypothesis*, suggesting that legal drafting practices inherit linguistic structures from earlier legal traditions through template reuse and institutional standardization. From a linguistic point of view, many of these archaic expressions (e.g., *herein*, *thereof*, *whereunder*) function as discourse-deictic expressions or pronominal adverbs within Pragmatics. They do not refer to physical locations in space but rather to parts of the document itself. As a result, this helps in co-reference resolution or reducing ambiguity and avoiding repeated usage of noun phrases in the legal texts. This also signifies the efficiency-driven tradeoff between linguistic accessibility and legal specificity. The percentage of non-standardized capitalization counted as 2.95% [Table 2] and archaic words as 1.93% of total words [Table 3]. Passive constructions accounted for 32.6% of the analyzed words compared to 67.4% active construction [Fig. 2]. Additionally, center embedded words accounted for 10.46% of the analyzed words [Fig. 3] suggesting potentially higher working-memory demands in *legalese*.

Conclusion: Our findings suggest that legal complexity emerges from an interaction between historical inheritance, institutional standardization, and psycholinguistic efficiency principles. The continued reuse of archaic expressions supports the *copy-paste hypothesis* while simultaneously revealing certain functional roles such as referential precision and ambiguity resolution in legal discourse. At the same time, psycholinguistic features viz., non-standardized capitalization, passive voice, low frequency jargons and center embedded words likely increase cognitive difficulty for non-expert readers. Future work will extend this analysis to large-scale Indian legal corpora such as English [8] and Hindi [9] Legal Documents Corpus, with a goal of developing a cognitively informed paraphrasing system that transforms complex legal sentences into simplified sentences.

Figure: Figure 1: Evolution of Legal Archaisms

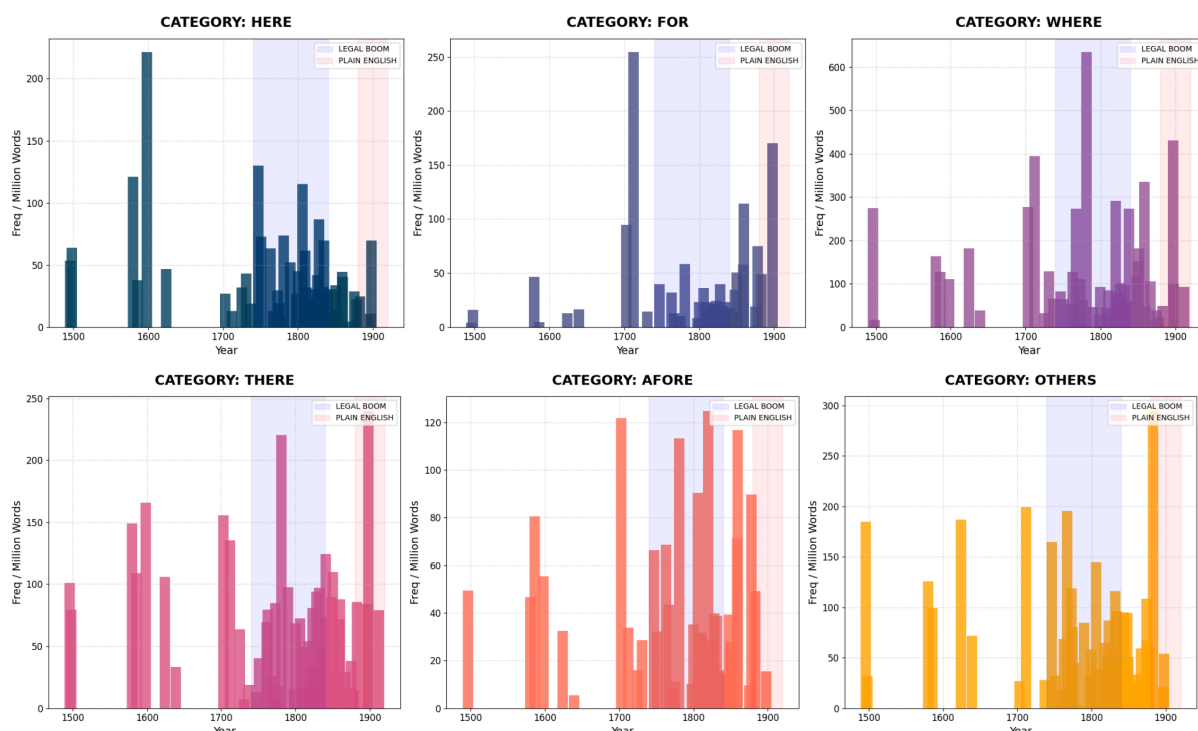


Figure 2: Passive & active voice

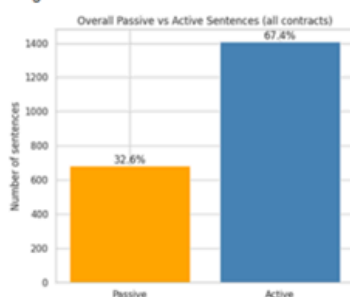
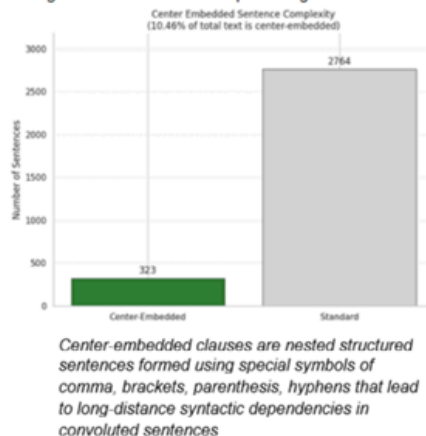


Figure 3: Center embedded percentage



Table

Table 1: Archaic Set	
Category	Examples
AFORE-	aforesaid, aforementioned, foregoing
FOR- / FORE-	forthwith, foregoing, forasmuch
HERE-	hereafter, hereby, hereunto
THERE-	thereafter, thereat, thereby
WHERE-	whereabout, whereabouts, whereas,
Others	witnesseth, notwithstanding

Table 2: Top 10 Capitalization (Total Capital 2.95%)			Table 3: Top 10 Archaics (Total Archaic 1.93%)		
Capitalization	Freq	%	Archaic	Freq	%
AND	112	0.19	hereto	162	0.27
WHEREAS	90	0.15	thereof	128	0.21
OF	80	0.13	hereinafter	126	0.21
AGREEMENT	58	0.1	hereby	120	0.2
THE	48	0.08	whereas	102	0.17
IN	47	0.08	herein	86	0.14
WITNESS	41	0.07	aforesaid	66	0.11
WHEREOF	38	0.06	hereunder	55	0.09
THIS	37	0.06	whereof	43	0.07
NOW	31	0.05	thereon	38	0.06

References

- [1] E. Martínez, F. Mollica, & E. Gibson, *Even lawyers do not like legalese*, Proc. Natl. Acad. Sci. U.S.A. 120 (23) e2302672120
- [2] K. A. Adams, *A Manual of Style for Contract Drafting*, American Bar Association, ed. 4, 2017
- [3] Tiersma, Peter M., *Some Myths About Legal Language*. Journal of Law, Culture and Humanities, Forthcoming, Loyola-LA Legal Studies Paper No. 2005-26
- [4] E. Martínez, F. Mollica, & E. Gibson, *Even laypeople use legalese*, Proc. Natl. Acad. Sci. U.S.A. 121 (35) e2405564121
- [5] Martínez E, Mollica F, Gibson E. *Poor writing, not specialized concepts, drives processing difficulty in legal language*. Cognition. 2022 Jul;224:105070. doi: 10.1016/j.cognition.2022.105070. Epub 2022 Mar 4. PMID: 35257980
- [6] Corpus of Old English, Universal Dependency Treebank. Link: <https://universaldependencies.org/ang/>
- [7] Diller, H., De Smet, H., Tyrkkö, J., *A European database of descriptors of English electronic texts*, The European English Messenger 19.2, 2010
- [8] Vijit Malik et al., *ILDC for CJPE: Indian Legal Documents Corpus for Court Judgment Prediction and Explanation*, Association for Computational Linguistics, 2021
- [9] Arnab Kapoor et al., *HLDC: Hindi Legal Documents Corpus*. In *Findings of the Association for Computational Linguistics*: Association for Computational Linguistics, 2022