Measuring Left-Embeddedness for the Accessible Simplification of Legal & Medical English

Georgina Willoughby
Supervised by Prof. Constantin Orăsan & Dr. Diptesh Kanojia
Centre for Translation Studies • Institute for People-Centred AI • University of Surrey

THE PROBLEM

Legal and medical English routinely place the subject far from its verb, loading reader working memory.

WHO IT AFFECTS

L2 readers and non-specialists confronting high-stakes specialised text in daily life.

THE APPROACH

Measure left-embeddedness across legal and medical corpora, then evaluate how simplification handles it.

THE QUESTION

Can a linguistically informed approach simplify without losing the meaning that matters?

1 | Why this matters

Specialised English is unreadable for many of the people who depend on it.

Legal judgements, statutes, and patient information leaflets routinely contain sentences in which the grammatical subject is separated from its main verb by long chains of qualifying material.

For **L2 readers**, this is not a stylistic preference, it is a comprehension barrier:

- Readers must hold the subject in working memory across the entire gap before the predicate resolves.
- On the BLaRC legal corpus, gaps are several times longer than in texts written for advanced (CEFR C2) readers.
- Legal and medical texts have direct material consequences for the people who must understand them.

This PhD asks: *how can a linguistically informed approach to simplification reduce this cognitive load while preserving the meaning that matters?*

2 | What is left-embeddedness?

The distance between the grammatical **subject** and the **main verb**: everything a reader must hold in working memory before the sentence resolves.

The Board (SUBJECT) *of Directors, in light of the statutory requirements set out in section 14(2)(b),* **resolved** (VERB)

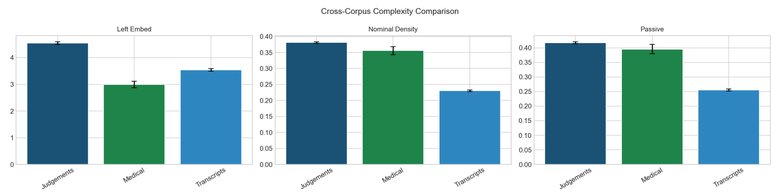
← GAP: 13 words

What fills the gap. Prepositional chains, embedded subclauses, statutory citations, parentheticals. In legal corpora only ∼5% of gap tokens are content-bearing (nouns/adjectives), the rest is structural padding.

How we measure it. A spaCy dependency pipeline identifies the `nsubj` and the root verb; the gap is the token distance between them.

3 | Complexity across registers

Preliminary analysis across three corpora using a spaCy `nsubj`-anchored measurement.



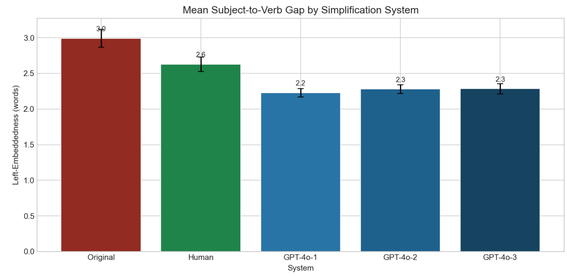
Sentence-level means across BAILII Judgements, EMC Patient Information Leaflets, and court Transcripts. Error bars: 95% CI.

Legal judgements exhibit the highest values across all measured syntactic metrics.

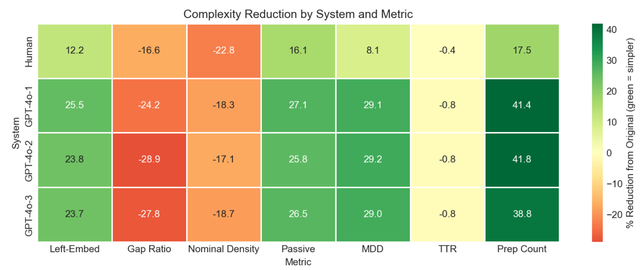
Medical leaflets are structurally simpler but share the passive-voice and nominal-density profile of legal text, suggesting accessibility issues that surface-level readability scores miss.

4 | Initial observations from simplification data

A pilot comparison of human and GPT-4o simplifications shows how each handles subject-to-verb distance.



Mean subject-to-verb gap across systems. The three GPT-4o runs converge at a lower value than the human reference.



Each cell shows percentage change relative to the original sentence. Green: the metric decreased after simplification (simpler). Red: the metric increased (more complex)

Initial results on medical paired data suggest that reductions in left-embeddedness do not necessarily align with reductions in other complexity measures.

Understanding how these metrics interact will form an important part of the next stage of the project.

5 | The measurement pipeline

A single **spaCy** pipeline tags every sentence in every corpus. For each sentence we record seven metrics:

Metric	Captures
Left-embed gap	subject–verb distance
Gap ratio	gap ÷ sentence length
Lexical density	content ÷ total words
Nominal density	nouns + adj. share of tokens
Passive rate	passive `nsubjpass`
MDD	mean dependency distance
Prep. count	`ADP` chains in the gap

Work so far: complexity analysis on BLaRC; the full pipeline applied to BAILII Judgements, court Transcripts, and EMC leaflets; pilot comparison of human and GPT-4o simplification on medical paired data.

6 | The next stages of the PhD

Evaluation of left-embeddedness

Assess whether reducing left-embeddedness leads to more accessible and faithful simplifications, using automatic metrics and human evaluation on legal data.

- **Does resolving left-embeddedness improve simplification quality?**

Beyond left-embeddedness

Expand the framework to additional linguistic phenomena that contribute to processing difficulty.

- **Which linguistic features create the greatest barriers to accessibility?**

Implicit meaning and discourse

Investigate how discourse structure, information packaging, and implicit meaning affect comprehension in legal communication.

- **What information remains difficult even when syntax is simplified?**

Towards linguistically-informed simplification

Develop simplification approaches that account for syntactic complexity, discourse structure, and implicit meaning.

- **What would accessibility-oriented simplification look like?**

7 | Contribution

This PhD aims to contribute:

- A reproducible measurement framework for linguistic complexity in legal English.
- An empirical account of how current LLM-based simplification systems handle syntactic complexity.
- New insights into the role of discourse structure and implicit meaning in specialised communication.
- A linguistically informed approach to accessibility-oriented text simplification, designed with L2 readers and non-specialists in mind.

Acknowledgements

This research is funded by the Leverhulme Trust through the **ADA Doctoral Scholarships Network** hosted by the University of Surrey. With thanks to my supervisors **Prof. Constantin Orăsan** and **Dr. Diptesh Kanojia**, and to the wider ADA community at the Centre for Translation Studies and the Surrey Institute for People-Centred AI.

Selected references

Ure, J. (1971). *Lexical density and register differentiation.*
Halliday, M.A.K. (1985). *Spoken and Written Language.*
Gibson, E. (1998). Linguistic complexity. *Cognition* 68.
BLaRC: Bilingual Legal Reference Corpus.
EMC: Electronic Medicines Compendium.

LEVERHULME TRUST — ADA — CENTRE FOR TRANSLATION STUDIES, UNIVERSITY OF SURREY

 | gw00697@surrey.ac.uk