The Cambridge Law Corpus: A Corpus for Legal AI Research

Andreas Östling[1], Holli Sargeant[2], Huiyuan Xie[2], Ludwig Bull[3], Alexander Terenin[2], Leif Jonsson[4], Måns Magnusson[1], Felix Steffek[2]

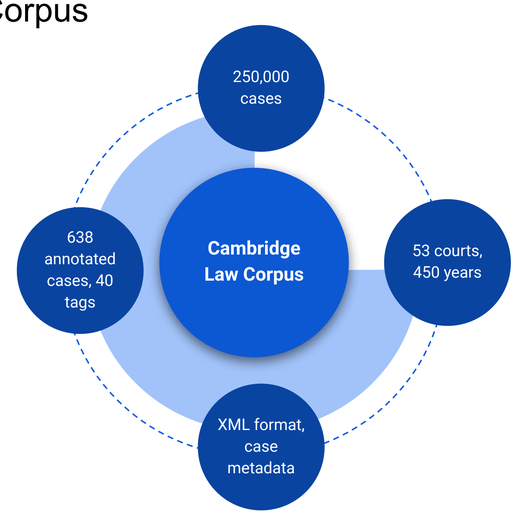
[1]Uppsala University [2]University of Cambridge [3]CourtCorrect [4]Sudden Impact AB

NEURAL INFORMATION PROCESSING SYSTEMS

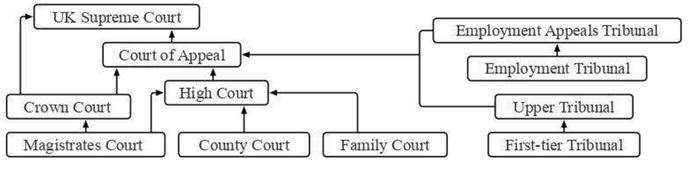
Summary

We introduce the Cambridge Law Corpus (CLC), a research corpus for use in legal AI research. It consists of over 250 000 court cases from the UK ranging from the 16th to the 21st century. We presents the first release, containing the raw text and expert annotations on case outcomes. We intend to continuously enrich the corpus with additional legal data, metadata and new annotations to create a research infrastructure for UK law. Due to the sensitive nature of the material, we offer an extensive legal and ethical discussion. As a consequence, the corpus will only be released under specific conditions.

The Corpus

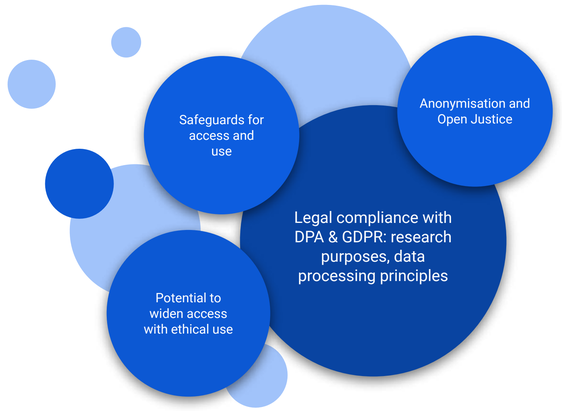


In addition to the court cases currently available, we plan to add future cases as well as legal statutes. A more diverse corpus with additional text opens up further research possibilities. One such possibility would be to train a language model to not only contain information about general legal text but also law and the structure and relationship between and within cases and statutes.



A simplified view of the UK court and tribunal structure.

Legal and Ethical Considerations



The legal framework for data collection and processing includes the Data Protection Act 2018 (DPA) and UK General Data Protection Regulation (GDPR). We rely on research purposes for the legality of the dataset (DPA sch 2 paras 26, 27; GDPR art 89).

We balance open justice and public availability of legal data with the protection of individuals and the sensitive nature of this data. We achieve this with the following safeguards:

- Appropriate technical and organisational safeguards exist to protect personal data (DPA s 19; GDPR art 89).
- Processing will not result in measures being taken in respect of individuals and no automated decision-making takes place (DPA s 14; GDPR art 22).
- Users who access the corpus must agree to comply with the DPA and the UK GDPR in addition to any local jurisdiction.
- Any individual may request the removal of a case or certain information which will be immediately removed (GDPR art 17; DPA s 47).
- The corpus will not pose any risks to people's rights, freedoms or legitimate interests (DPA s 19; GDPR art 89).
- Access to the corpus will be restricted to researchers based at a university or other research institution whose Faculty Dean (or equivalent authority) confirms, inter alia, that ethical clearance for the research envisaged has been received.
- Researchers using the corpus must agree to not undertake research that identifies natural persons, legal persons or similar entities. They must also guarantee that they will remove any data if requested to do so.

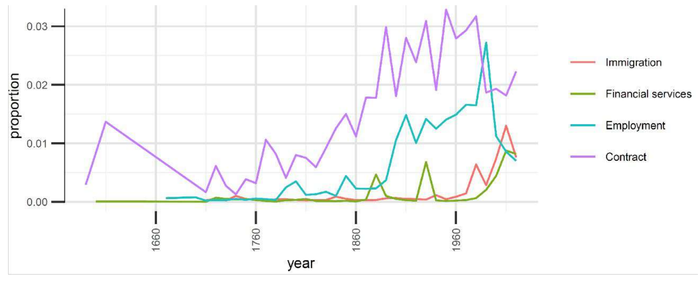
The purpose, design and method of the corpus avoid risks to individuals and prioritise positive outcomes through improving access to and research of case law.

Experimental Results

GPT-3.5-turbo	GPT-4	Annotator Reference
Lord Justice: I also agree.	*I would dismiss these appeals.*	I would dismiss these appeals.
I would answer the issue: *The Claimant did comply with the service requirements of the Share Purchase Agreement.* I will hear submissions as to the next steps to be taken in relation to this claim, and as to whether it is appropriate to embody the answers to the preliminary issues (disputed and undisputed) in formal declarations.	*The Claimant did comply with the service requirements of the Share Purchase Agreement.*	The Claimant did comply with the service requirements of the Share Purchase Agreement.
For the reasons I have given on each of the questions charterers are unsuccessful in their appeal. Here too it is unnecessary to address procedural questions and I consider it undesirable to seek to do so. It follows from my analysis above that charterers' appeal on questions 2 and 3 fails on the merits.	*For the reasons I have given on each of the questions charterers are unsuccessful in their appeal.*	For the reasons I have given on each of the questions charterers are unsuccessful in their appeal.

Comparing case outcome extraction by GPT-3.5-turbo and GPT-4 with the annotator reference.

Green: model output identical with manual annotation.



Proportion of words in documents belonging to one of four topics. The model is an LDA topic model.

Model	Acc	F1	WER	BERTScore	BLEU	ROUGE$_L$
End-to-end RoBERTa	0.997	0.185	0.536	0.207	0.010	0.185
RoBERTa pipeline	0.739	0.012	0.776	0.030	0.015	0.028
GPT-3.5	-	-	4.281	0.788	0.217	0.300
GPT-4	-	-	3.396	0.840	0.282	0.517

Evaluation results for the case outcome extraction task.

Acknowledgements

Funding by the UK Economic and Social Research Council (ESRC) and JST (Japan Science and Technology Agency) is gratefully acknowledged.

UNIVERSITY OF CAMBRIDGE UPPSALA UNIVERSITET courtcorrect