Meta Data for Annotated Dataset:

**Description:**

This repository contains two datasets derived from Hansard transcripts of UK parliamentary debates on labour-law remedies. The first, **MANUALLY\_ANNOTATED\_DATA.xlsx,** is the core dataset of 862 annotated statements. Each statement is linked to an actor, organisation, concept, right, and stance label following a specialist codebook. The file is colour coded, indicating the assigned annotator. The codebook was developed by Prof. Aristea Koukiadaki and may be added to the repository subject to approval.

The second, **512\_each\_statements\_nonstatements\_chunks.xlsx**, is a balanced subset of 1,024 entries with 512 statements (label = 1) and 512 non-statements (label = 0). This smaller dataset was constructed to support stable training of the statement segmentation classifier, ensuring equal representation of positive and negative cases. Non-statement examples include procedural remarks, greetings, and quotations, allowing the classifier to learn meaningful distinctions beyond trivial features such as length.

This dataset forms the basis for the training and evaluation of the segmentation, concept classification, and right classification modules of the *dnaPipelineR* pipeline reported in Tunmore (2025), MSc Extended Research Project. They underpin the reproducibility of the dissertation experiments and may also serve as a resource for researchers in computational political discourse analysis, particularly for studies of UK parliamentary debates.

**Ethics & Licensing:**

All transcripts are sourced from Hansard, the official public record of UK parliamentary proceedings. Although the dataset is not anonymous (speeches are attributed to named MPs), it is not considered private or sensitive, as it concerns public officials performing their duties. Hansard is freely accessible online, and its publication by Parliament constitutes permission for reuse.

The datasets are published as part of Tunmore’s (2025) MSc Extended Research Project, submitted to the University of Manchester for the degree of MSc in Data Science (Social Analytics) in the Faculty of Humanities.

**Source:**

**Primary source:** Debate texts were sourced from Hansard (<https://hansard.parliament.uk/>), the official repository of all UK Parliamentary debates, downloaded as TXT files.

**Debates included:** See the Table below for a full breakdown of debates used in the creation of this corpus, including assigned annotator, title of debates, number of statements, and Hansard links. All debates were downloaded in July 2025.

**Table 1:** Debates included in the annotated dataset with statement counts, assigned annotator, and Hansard links.

|  |  |  |
| --- | --- | --- |
| **Debate Title and Assigned Annotator** | **Statement**  **Count** | **Link to Hansard Debate** |
| **Annotator 1** | **182** |  |
| Draft Enterprise and Regulatory Reform Act 2013 (Consequential amendments) (Bankruptcy) and the Small Business, Enterprise and Employment Act 2015 (Consequential amendments) Regulations 2016 | 24 | <https://hansard.parliament.uk/Commons/2016-03-21/debates/730a7b57-e874-41f7-98f4-e1b4108a0a54/DraftEnterpriseAndRegulatoryReformAct2013(ConsequentialAmendments)(Bankruptcy)AndTheSmallBusinessEnterpriseAndEmploymentAct2015(ConsequentialAmendment> |
| Employment Rights Bill Volume 763: debated on Tuesday 11 March 2025 | 50 | <https://hansard.parliament.uk/Commons/2025-03-11/debates/5F064261-A587-4975-B120-2B7045B095D7/EmploymentRightsBill> |
| Employment Rights Bill Volume 763: debated on Tuesday 11 March 2025 - Debate Resumed | 50 | <https://hansard.parliament.uk/Commons/2025-03-11/debates/9193774F-AB84-4ACE-8AED-08BB980810B6/EmploymentRightsBill> |
| Employment Rights Bill Volume 763: debated on Wednesday 12 March 2025 | 50 | <https://hansard.parliament.uk/Commons/2025-03-12/debates/D532A0C0-8B26-4BE4-B61A-36908DB2CBA7/EmploymentRightsBill> |
| Employment Rights Bill: Employment Rates Volume 764: debated on Monday 17 March 2025 | 8 | <https://hansard.parliament.uk/Commons/2025-03-17/debates/8EBAF816-1C0E-47E5-84F2-FF0458C0BD89/EmploymentRightsBillEmploymentRates> |
| **Annotator 2** | **543** |  |
| Trade Union Bill Volume 599: debated on Monday 14 September 2015 | 543 | <https://hansard.parliament.uk/Commons/2015-09-14/debates/1509146000001/TradeUnionBill> |
| **Annotator 3** | **137** |  |
| Strikes (Minimum Service Levels) Bill Volume 726: debated on Monday 16 January 2023 | 137 | <https://hansard.parliament.uk/commons/2023-01-16/debates/A08B1E2A-B712-4C30-9A71-57DAC922C4A5/Strikes(MinimumServiceLevels)Bill> |
| **Grand Total of Statements** | **862** |  |

**Annotation Process:**

**Annotators:** Three MSc Data Science students.Annotator 1: sampled multiple debates (~50 statements per debate).Annotators 2 and 3: fully annotated one lengthy debate each.

**Guidance:** Codebook supplied by Prof. Aristea Koukiadaki

**Software used:** Discourse Network Analyzer (<https://github.com/leifeld/dna>)

**Export format:** CSV and Excel (training-ready).

**Timescale:** Annotation of datasets began in July 2025 and concluded in August 2025 with the integration of annotations into a single CSV for use in modelling.

**Variables:**

Each row in the dataset represents one annotated statement. Annotations were guided by a codebook provided by Columns include:

* **Actor**: Speaker name from Hansard.
* **Organisation**: Affiliation (party or institutional).
* **Statement**: Minimal contiguous text span representing the claim.
* **Concept**: One of 21 categories (ACCESS, ADEQ, ADR, AUTONOMY, BALANC, COLLECT, COMPAR, COMPL, COURTS, CREATE, DETER, ECONOMIC, ENFORCE, JUST, LEGAL, PHYSICAL, POLIT, POWER, RELATE, REPLIC, RIGHTS).
* **Right**: One of 6 labour standards (EQUAL, GENERAL, SECURITY, STATUS, STRIKE, WAGES).
* **Agreement**: Binary (agree=1, disagree=0). *Note: excluded from modelling due to annotator inconsistency.*