1. ABOUT THE DATASET

------------

Title: Content list of a corpus of texts from the English language GCSE exams

Creator(s): Beverley Jennings <https://orcid.org/0000-0002-3645-7411>

Organisation(s): Institute of Education, University of Reading

Rights-holder(s): Beverley Jennings

Publication Year: 2025

Description: This is a list of the separate documents used to create a corpus of texts from the English language GCSE exams. The list in includes titles of sources, authors, dates of publications, exam series, exam boards and the exam board websites from which they were sourced. These texts were used to create a corpus that was analysed as part of a PhD project that examined the language in the exams and the comprehension challenge that this language could present to students.

Changelog: n/a

Cite as: Jennings, Beverley (2025): Content list of a corpus of texts from the English language GCSE exams. University of Reading. Dataset. http://dx.doi.org/10.17864/1947.001403

Related publication: Jennings, B., Powell, D., Jaworska, S. & Joseph, H. (2024). A Corpus Study of English Language Exam Texts: Vocabulary Difficulty and the Impact on Students' Wider Reading (or Should Students be Reading More Texts by Dead White Men?). Journal of Adolescent& Adult Literacy, 67, 303–316. https://doi.org/10.1002/jaal.1331A PPE N D I X AA.1

Contact: n/a

Acknowledgements: n/a

2. TERMS OF USE

------------

This dataset is licensed under a Creative Commons Attribution 4.0 International Licence: https://creativecommons.org/licenses/by/4.0/.

3. PROJECT AND FUNDING INFORMATION

------------

Title: The vocabulary challenge of the English Language GCSE Exam and the implications for adolescent reading experience

Dates: 2018-2025

Funding organisation: n/a

Grant no.: n/a

This dataset was part of a PhD project

4. CONTENTS

------------

File listing

Exam\_texts\_corpus\_content.csv

5. METHODS

-----------

The extracts were sourced from the sample assessment materials and the three sets of past papers that were publicly available when the data was collected (June 2017, November 2017 and June 2018) from the four awarding exam boards in England: AQA, Edexcel, Educas and OCR. In total there were 59 extracts available from the exam board websites. The ETC contains 36,585 words, of which 6,854 are unique. [This section should describe how the dataset was generated. The following information should be included as relevant.

Further details available in Jennings et al. (2024)