

The English Classroom Corpus

Dataset Card — CEFR-Graded English Language Teaching Corpus (A1–C2)

Rights holder: The English Classroom AB, Kalmar, Sweden — 100% owned by the founder and sole author. No affiliated entities hold any rights in the corpus. **Contact:** Jennifer Gardner, Founder & CEO — jennifer@the-englishclassroom.com **Status:** Available for non-exclusive licensing (AI training, fine-tuning, evaluation, and product integration)

What this is

A 600-hour corpus of original English language teaching content, authored 2019–2026, graded across all six CEFR levels (A1–C2) and thematically mapped to all 17 UN Sustainable Development Goals. The corpus was built as a complete, sequenced curriculum — not an aggregation — meaning level progression, recycling of language items, and pedagogical scaffolding are internally consistent across the full set.

This structure cannot be scraped from the open web. Graded instructional sequences with aligned exercises, answers, and audio are precisely the data class that is scarce in web-crawled corpora.

Contents

Component	Detail
Instructional hours	600 hours of classroom-ready lessons
CEFR coverage	A1, A2, B1, B2, C1, C2 — internationally recognised grading
Lesson structure	Pre-planned lessons with step-by-step teacher instructions and paired student resources
Exercise data	Comprehension questions, activities, and answer keys (<i>confirm: answer keys exist for all levels?</i>)
Audio	2,500+ native-speaker English sound clips paired to lesson content, recorded by the author (British English native speaker)
Thematic mapping	All 17 UN SDGs covered at every level; each lesson tagged by SDG topic

Component	Detail
Formats	Exportable as PDF and SCORM packages; conversion to structured text/JSONL available on request
Language variety	British English, single authorial voice

Why it is valuable as training data

- **Graded difficulty labels.** Every text carries a reliable CEFR level assigned by the author, a qualified UK teacher — usable for controllable-generation training, readability modelling, and learner-level adaptation.
- **Instructional pairs.** Teacher instructions ↔ student tasks ↔ answers form natural input/output structures for instruction-tuning in educational contexts.
- **Pedagogical sequencing.** Lessons are ordered within a real curriculum, providing progression signal absent from shuffled web text.
- **Text-audio alignment.** Native-speaker audio paired with graded text supports speech and pronunciation applications.
- **Thematic breadth with consistency.** SDG topics (climate, health, equality, economics, justice) treated at six proficiency levels each — the same concept expressed at A1 and C2 is a rare parallel-simplification signal.

Provenance and rights

- **Authorship:** All 600 hours of instructional content authored solely by Jennifer Gardner, qualified UK teacher and founder. Single-author, single-voice corpus with no contributing writers and no rights held by any third party.
- **No third-party copyrighted material** incorporated (*confirm: no reproduced texts, images, or extracts from published works within lessons*).
- **No user-generated content.** The corpus contains only authored instructional material. No student work, no forum content, no submissions.
- **No personal data.** The corpus contains no names, records, or information relating to platform users or any identifiable individuals.
- **Audio:** All 2,500+ clips recorded by the author. Text and speech share a single origin — full rights held by The English Classroom AB.
- **Prior licensing:** The corpus has been commercially licensed for AI training under a non-exclusive agreement with a US-based AI data company (*subject to confidentiality review — confirm what may be disclosed*).

- **Clean separation:** The corpus is independent of, and outside the scope of, all other existing agreements of the rights holder.

Real-world validation

- Deployed via The English Classroom platform to 7,000+ registered teachers across 80+ countries since 2020.
- In active classroom use in humanitarian and low-resource education contexts, including deployment with UNRWA and active users in Palestine, Lebanon, Syria, Ukraine.
- User-rated 4.8/5. Teacher testimonials available from the UK, Pakistan, and other markets.
- Total audience across platform, blog subscribers, and social channels exceeds 50,000.

Licence types available

Licence	Scope	Notes
AI training licence (non-exclusive)	Training, fine-tuning, evaluation of ML models	Term-limited or perpetual-snapshot options
Product integration licence	Use of curriculum as content spine within an ed-tech product	Per-product scope, white-label available
Institutional licence	Direct classroom use by schools, NGOs, ministries	Existing channel

Exclusive arrangements considered by sector carve-out only. Sample set available under NDA on request.

Prepared July 2026. Figures accurate at time of writing.