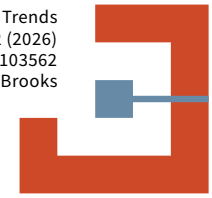
Trends

JALTCALL Trends
ISSN 2982-3749, 2(1), 103562 (2026)
https://doi.org/10.29140/jct.v2n1.103562



Castledown

Streamlining learner corpus development with LLMs and NLP

Gavin Brooks
Kyoto University of Foreign Studies, Japan
gavinbrooks@gmail.com

Learner corpus projects increasingly require scalable, reliable cleaning and annotation in order to develop corpora that are ecologically valid and span more learner contexts. This paper evaluates a hybrid pipeline that combines deterministic NLP preprocessing with a multi-LLM consensus classifier for context-dependent decisions. Using a pilot of 25 IELTS essays sampled from a 1,329-essay corpus, algorithms were used to handle document matching, anonymisation, segmentation, and mechanical typo correction, while three commercial LLMs independently classified off-list tokens (spelling errors, proper nouns, technical terms, and related categories). Automatic corrections were applied only when the models agreed, while disagreements triggered human-in-the-loop review. Deterministic plus targeted LLM use improved document matching and segmentation, while the LLMs were able to successfully clean 97.4% of the off-list words in the corpus. Results indicate that a pipeline that employs a multi-LLM consensus, coupled with algorithmic preprocessing and selective human oversight, can provide a practical, cost-conscious path to scaling learner corpus development while preserving methodological transparency.

Keywords: Corpus linguistics, learner corpora, LLMs, NLP

Introduction

The process of compiling learner corpora has changed significantly since the International Corpus of Learner English (ICLE) was compiled in the early 1990s (Granger, 2003). While first-generation corpora like ICLE provided important insights, they were labour-intensive to compile. Granger (2024) traces the field's progression from first-generation manual methods through second-generation digital infrastructures to an emerging third generation characterised by increased diversity, interdisciplinarity, and automation. However, despite these improvements, as modern learner corpus designs have grown more ambitious, compilation methods have struggled to keep pace (Paquot, 2024).

Recent studies have shown that large language models (LLMs) demonstrate

promise for assisting with corpus annotation and cleaning tasks. Fonteyn et al. (2025) demonstrated the benefit of using machine learning to annotate corpora by using transformer-based models to automate complex linguistic annotation when classifying Early Modern English *-ing* forms. Mahmoudi-Dehaki and Nasr-Esfahani (2025) extended this to the field of learner corpora, demonstrating that ChatGPT-4 could outperform human annotators in evaluating pragmatic competence in EFL learners' verbal production. They used focus groups to show that human annotations suffered from bias, fatigue, and inconsistency, issues that have been documented by other researchers (e.g., Díaz-Negrillo & Fernández-Domínguez, 2006).

Research by Creutz (2024) further demonstrates the benefits of using LLMs with learner text. Using multiple LLMs to correct grammatical errors in Finnish learner texts, he found that while single-model accuracy ranged from 74.5% (GPT-3.5) to 83.3% (GPT-4), he was able to achieve 84.3% accuracy by using multiple models working together. In doing so, he found that the different models exhibited complementary strengths: Claude produced conservative corrections, GPT-4 generated more fluent output, and GPT-3.5 offered cost-effective processing.

However, despite the promise demonstrated by this research, there are still questions about how LLM-based approaches can be best leveraged to clean and compile corpora. This study addresses this issue by testing a hybrid pipeline that combines deterministic NLP preprocessing with multi-LLM consensus validation, requiring unanimous agreement across three models before applying automated corrections, with human review for disagreements.

Methodology and results

Data

The pilot corpus used in this study consists of 25 IELTS essays selected from a larger collection of 1,329 argumentative essays written by pre-sessional students at a British university (Table 1). All writers demonstrated proficiency between IELTS 6.5 and 8.0, corresponding to upper-intermediate to advanced English competence. While it accounted for only about 2% of the total corpus, this sample size allowed the researcher to test the pipeline and verify all automated decisions.

Table 1. Descriptive stats for essays (total essays *N* = 1,329)

IELTS level	Essays	Mean words	Min words	Max words	Std dev
Overall	**1,329**	**1,946**	**269**	**9,791**	**297.62**
6.5	199	1,864	269	2,305	284.05
7.0	636	1,936	480	2,700	197.99
7.5	425	1,971	752	2,512	164.50
8.0	69	2,117	1,527	9,791	951.76
Pilot corpus	25	2,267	1,768	3,138	281

Pipeline architecture

The corpus cleaning pipeline comprised six sequential stages (Figure 1): (1) document matching and conversion; (2) anonymisation; (3) segmentation; (4) algorithmic typo correction; (5) multi-LLM spelling classification; and (6) metadata generation.

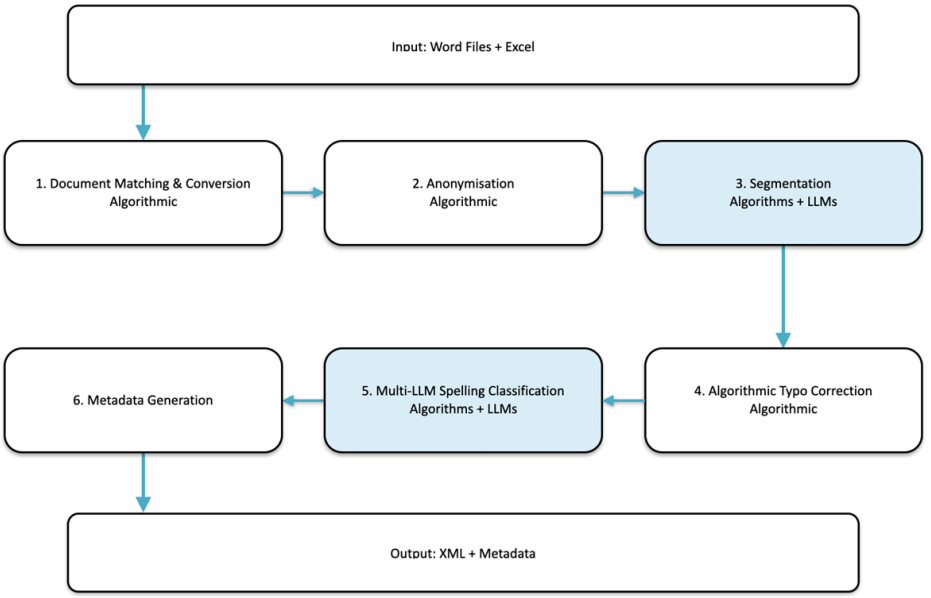


Figure 1. Pipeline cleaning architecture

All of the processing was done in Python 3.12 with the assistance of pandas, spaCy, RapidFuzz, and the APIs for Claude Sonnet 4.0, GPT-4.5, and Gemini 2.5 Flash. The appendix includes the full Claude system prompt and few-shot examples. The GPT-4.5 and Gemini 2.5 Flash prompts, per-rater classifications for the 93-token validation subset, and analysis scripts, including the Fleiss' κ script, are publicly available in the supplementary materials (Brooks, 2026). At each stage, a structured JSON output was created documenting all processing decisions. These stages are described in more detail below.

Deterministic NLP components (stages 1–4)

The first four pipeline stages employed deterministic algorithms to handle structured preprocessing tasks (Table 2). Stage 1 used fuzzy string matching (RapidFuzz, threshold = 85) to match Word documents against student database entries, comparing filename and embedded metadata. Stage 2 applied regular expressions and fuzzy matching to remove identifying information (names, IDs, emails) and tag documents with SHA-256 hashes. Stage 3 employed regex-based rules to classify paragraphs (front matter, body, back matter) and document components (titles, headers, lists); LLM assistance (Gemini 2.5 Flash) was used to resolve ambiguous cases. Stage 4 corrected mechanical typing errors such as duplicated punctuation, missing spaces, and inconsistent capitalisation using systematic regex rules. Context-dependent spelling decisions were left for Stage 5.

Table 2. Steps 1–4 deterministic and LLM pipeline performance

Stage	Method	Accuracy (%)
1a. Document matching (pilot)	Fuzzy string matching (RapidFuzz)	100
1b. Document matching (full corpus)	Fuzzy string matching (RapidFuzz)	94.56
2. Anonymisation	Regex + fuzzy matching + SHA-256	100
3a. Paragraph classification	Regex (baseline)	92.67
3b. Paragraph classification	Regex + LLM for ambiguous cases	100
3c. Component identification	Regex (baseline)	82.24
3d. Component identification	Regex + LLM for ambiguous cases	94.98
4. Typo correction	Regex rules for mechanical errors	NA

JALTCALL Trends
Vol. 2 No. 1
2025 JALTCALL Conference Papers
Special Issue

Multi-LLM consensus classification (stage 5)

The final preprocessing stage addressed context-dependent decisions: determining whether off-list tokens represented spelling errors requiring correction, or legitimate usages that should be preserved. Words were first lemmatised and compared against BNC-COCA wordlists (Nation, 2020); this pre-filtering reduced the number of words the LLMs needed to process. Off-list words were checked by three commercial LLM APIs (Claude Sonnet 4.0, GPT-4.5, Gemini 2.5 Flash). The LLMs received the target word with the surrounding sentences for context and part-of-speech tags. System prompts set the models up as "expert corpus linguists" and used few-shot examples and temperature 0.1, building on Creutz's (2024) study. Models were asked to classify tokens into eight categories:

1. **spelling_error:** Unintentional misspelling of an English word (e.g., "recieve" instead of "receive")
2. **proper_noun:** Personal names, place names, organisation names
3. **technical_term:** Discipline-specific vocabulary, jargon
4. **foreign_word:** Non-English lexical items
5. **compound_word:** Hyphenated or closed compounds not in wordlists
6. **informal_word:** Colloquialisms, slang, non-standard varieties
7. **regional_variant:** British/American spelling variants (e.g., "colour"/"color")
8. **valid_word**: Legitimately used word absent from wordlists

Corrections were applied only when all three LLMs agreed on both the category and correction, prioritising precision over automaticity. Disagreements triggered a human-in-the-loop (HITL) review, with all decisions logged for later analysis.

Metadata generation (stage 6)

The final stage synthesised outputs from all preceding steps into structured metadata following TEI P5 XML encoding guidelines (TEI Consortium, 2025) and Paquot et al.'s (2024) LC-meta schema. XML files containing corpus metadata were accompanied by stage-specific JSON files documenting all processing decisions, enabling both future integration with other standardised learner corpora and full methodological transparency.

Discussion

Across the full 25-essay pilot, all three models agreed on whether correction was needed for 97.4% of off-list tokens, aligning with evidence that LLMs can match or exceed human inter-annotator reliability benchmarks (Larsson et al., 2020; Yu et al., 2024; Mahmoudi-Dehaki & Nasr-Esfahani, 2025). The remaining 2.6% triggered HITL review, where disagreements were resolved through human judgment rather than forced consensus. To assess chance-corrected reliability and address the risk of shared bias or collective hallucination, a validation subset of 93 off-list tokens drawn from three essays was analysed in detail. On this subset, Fleiss' κ across the full 8-way category schema that allowed the model to identify what type of correction was needed was 0.79, indicating substantial agreement under Landis and Koch's (1977) commonly used interpretation scale, and binary unanimity was 100%. All 83 unanimous category-level classifications were manually audited against a human gold standard, with all 83 confirmed correct. The remaining 10 cases (10.75%) split across three category boundaries: technical_term vs proper_noun, technical_term vs valid_word, and valid_word vs informal_word (see Table 3). All 10 were resolved through a HITL review.

Table 3. Examples of disagreement cases

Token	Context (excerpt)	Claude	ChatGPT	Gemini	Category boundary
Concentrators	Cleaning Solar Panels and Solar Concentrators.	technical_term	technical_term	proper_noun	technical_term vs proper_noun
preprocessed	game map must be preprocessed before the A* algorithm…	technical_term	valid_word	valid_word	technical_term vs valid_word
pricy	the cost of the system is pricy…	valid_word	informal_word	informal_word	valid_word vs informal_word

Note: All 10 disagreement cases are available in the supplementary material (Brooks, 2026).

Furthermore, the pipeline's strategic task allocation, using deterministic methods for structured operations (document matching, anonymisation, typo correction) and LLMs for context-dependent decisions, was able to provide cost and speed improvements over LLM-only approaches. This pattern reflects Archer and Culpeper's (2018) observation that some tasks require complex contextual synthesis while simpler orthographic tasks are possible to solve algorithmically. The approach contrasts with supervised methods like Fonteyn et al.'s (2025) MacBERTh, which require training data. While few-shot LLM methods favour pilot studies and diverse corpora that lack the resources for large training sets, the JSON documentation at each stage enables a smooth transition to trained models for large-scale projects.

Limitations and implications

This pilot study examined only 25 essays from a single genre (IELTS argumentative essays) at advanced proficiency levels. Further studies are needed to determine if the pipeline is able to perform both at scale as well as across genres, proficiency bands, and modalities. The proprietary nature and rapid speed of change in commercial LLMs could also make replication hard. Following Paquot et al.'s (2024) LC-meta framework and FAIR principles, it is important for researchers using LLMs to note (1) exact model versions and API parameters, (2) system prompts and few-shot examples, (3) all automated classifications, and (4) version control for traceability. For future research, I intend to validate the pipeline across the full 1,329-essay corpus, test effectiveness on spoken transcripts and different genres of corpora, as well as develop GUI tools to allow for broader adoption by researchers lacking programming expertise.

Conclusion

This pilot study demonstrates that a multi-LLM approach offers a viable approach to scalable learner corpus cleaning. The 97.4% unanimous agreement rate achieved across three commercial models (Claude, ChatGPT, Gemini) is similar to the levels typically seen on human inter-annotator reliability benchmarks. Allowing a human to review the 2.6% of decisions that the LLMs disagree on may actually improve these human rating scores, as there will be fewer corrections to make, leading to less fatigue.

While these results are promising, they require validation at scale across the complete 1,329-essay corpus, genres, and proficiency levels. This is important because, as Paquot (2024) argues, improving our ability to reliably compile learner corpora is essential for corpus research to remain relevant in SLA. As LLMs become better and more accessible, hybrid human-LLM workflows can help to improve corpus development, enabling specialised corpora and longitudinal studies previously constrained by the challenges of manual compilation.

References

Archer, D., & Culpeper, J. (2018). Corpus annotation. In A. H. Jucker, K. P. Schneider, & W. Bublitz (Eds.), *Methods in pragmatics* (pp. 495–526). De Gruyter. https://doi.org/10.1515/9783110424928-020

Brooks, G. (2026). *Supplementary materials for "Streamlining learner corpus development with LLMs and NLP"* [Data set]. Zenodo. https://doi.org/10.5281/zenodo.19922960

Creutz, M. (2024). Correcting challenging Finnish learner texts with Claude, GPT-3.5 and GPT-4 large language models. *Proceedings of the Ninth Workshop on Noisy and User-Generated Text,* 1–10. https://aclanthology.org/2024.wnut-1.1/

Díaz-Negrillo, A., & Fernández-Domínguez, J. (2006). Error tagging systems for learner corpora. *Revista Espanola De Linguistica Aplicada, 19,* 83–102. http://www4.ujaen.es/~svalera/Research/informatizacion/RESLA 2006.pdf

Fonteyn, L., Manjavacas, E., & De Regt, J. (2025). Using machine learning to automate data annotation in corpus linguistics: A case study with MacBERTh. *International Journal of Corpus Linguistics.* https://doi.org/10.1075/ijcl.22088.fon

Granger, S. (2003). The International Corpus of Learner English: A new resource for foreign language learning and teaching and second language acquisition research. *TESOL Quarterly, 37*(3), 538–546. https://doi.org/10.2307/3588404

Granger, S. (2024). From early to future learner corpus research. *International Journal of Learner Corpus Research, 10*(2), 247–279. https://doi.org/10.1075/ijlcr.22034.gra

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), 159–174. https://doi.org/10.2307/2529310

Larsson, T., Paquot, M., & Plonsky, L. (2020). Inter-rater reliability in learner corpus research: Insights from a collaborative study on adverb placement. *International Journal of Learner Corpus Research, 6*(2), 237–251. https://doi.org/10.1075/ijlcr.20001.lar

Mahmoudi-Dehaki, M., & Nasr-Esfahani, N. (2025). Automated vs. manual linguistic annotation for assessing pragmatic competence in English classes. *Research Methods in Applied Linguistics, 4*(3), 100253. https://doi.org/10.1016/j.rmal.2025.100253

Nation, P. (2020). *The BNC/COCA word family lists.* https://www.wgtn.ac.nz/__data/assets/pdf_file/0005/1857641/about-bnc-coca-vocabulary-list.pdf

Paquot, M. (2024). Learner corpus research: A critical appraisal and roadmap for contributing (more) to SLA research agendas. *Corpus Linguistics and Linguistic Theory, 20*(3), 567–590. https://doi.org/10.1515/cllt-2024-0014

Paquot, M., König, A., Stemle, E. W., & Frey, J.-C. (2024). The core metadata schema for learner corpora (LC-meta): Collaborative efforts to advance data discoverability, metadata quality and study comparability in L2 research. *International Journal of Learner Corpus Research, 10*(2), 280–300. https://doi.org/10.1075/ijlcr.23033.paq

TEI Consortium. (2025). *TEI P5: Guidelines for electronic text encoding and interc*hange (Version 4.8.1). Text Encoding Initiative Consortium. https://tei-c.org

Yu, D., Li, L., Su, H., & Fuoli, M. (2024). Assessing the potential of LLM-assisted annotation for corpus-based pragmatics and discourse analysis: The case of apology. *International Journal of Corpus Linguistics, 29*(4), 534–561. https://doi.org/10.1075/ijcl.23087.yu

Author biodata

Gavin Brooks is an Associate Professor at Kyoto University of Foreign Studies. He researches vocabulary and corpus linguistics, focusing on EAL students' needs, lexical diversity in learner texts, and the development and use of learner corpora in EFL and ESL contexts.

Appendix

Prompt architecture and API parameters

The Stage 5 multi-LLM consensus classifier used three commercial LLM APIs in parallel to classify the off-list words. Each model received a structurally similar system prompt that gave it the eight category labels, described the input fields (word, lemma, POS tag, surrounding context), and requested a JSON-array response. The prompts were written specifically for each of the LLMs based on best practices at the time. The Claude prompt is reproduced below in full. The GPT-4.5 and Gemini 2.5 Flash prompts, per-rater classifications for the 93-token validation subset, and analysis scripts are available from the supplementary materials (Brooks, 2026).

System prompt:

```
You are an expert linguist analyzing unknown words in learner English
writing.

Your task is to classify each unknown word into one of these categories:
- spelling_error: Clear misspelling that needs correction
- proper_noun: Person, place, organization, or publication name
- technical_term: Domain-specific vocabulary (academic, scientific, etc.)
- foreign_word: Non-English word (borrowed terms, transliterations)
- compound_word: Hyphenated or merged words
- informal_word: Slang, colloquial, or very informal language
- regional_variant: British vs American spelling differences
- valid_word: Legitimate word missing from standard wordlists

Consider the part of speech (pos), context, and lemma when making
classifications.

EXAMPLES:
{few_shot_examples}

Classify these words and return a JSON array with this exact structure:

{word_data}

For each word, provide:
- "category": one of the categories above
- "confidence": float from 0.0 to 1.0
- "suggested_correction": corrected spelling if category is
  "spelling_error", otherwise empty string
- "explanation": brief reason for your classification

Return only the JSON array with the same structure but filled in.
```

Few-shot examples block:

```
Input: {“word”: “recieve”, “context”: “I will recieve the package tomorrow”}
Output: {“category”: “spelling_error”, “confidence”: 0.95,
         “suggested_correction”: “receive”,
         “explanation”: “Common misspelling of ‘receive’”}

Input: {“word”: “LinkedIn”, “context”: “I found his profile on LinkedIn”}
Output: {“category”: “proper_noun”, “confidence”: 0.9,
         “suggested_correction”: “”,
         “explanation”: “Social media platform name”}

Input: {“word”: “photosynthesis”,
        “context”: “Plants use photosynthesis to make energy”}
Output: {“category”: “technical_term”, “confidence”: 0.85,
         “suggested_correction”: “”,
         “explanation”: “Scientific/biological term”}

Input: {“word”: “colour”, “context”: “What colour is your car?”}
Output: {“category”: “regional_variant”, “confidence”: 0.8,
         “suggested_correction”: “”,
         “explanation”: “British spelling of ‘color’”}
```

JALTCALL Trends
Vol. 2 No. 1
2025 JALTCALL Conference Papers
Special Issue