Writing a Book with Claude and a Corpus

Michael Barlow
University of Auckland
83 Symonds St. Auckland, NZ
mi.barlow@auckland.ac.nz

Abstract

The *Writing the Research Article* series consists of corpus-based guides to disciplinary writing, produced via a collaboration between a corpus linguist and a large language model (Claude). This paper traces that collaboration through four phases. In the first, prompt-driven phase, the books were drafted quickly, with the model performing most of the compositional work against discipline-specific corpora of research articles. A crisis defined the second phase: some "corpus-attested" examples were discovered to have been fabricated by the model, compromising the series central claim and prompting a program of verification governed by a corpus-first rule. The third phase involved an analysis environment in which category definitions are held in human-editable YAML files, and Python scripts, which compute every published statistic, with methods notes embedded in the results themselves. The fourth phase adds verification processes. An audit tool now classifies every quoted example in a manuscript as attested, truncated, silently edited, or invented. The corpora, definitions, code and results are published as a companion site that also lets a writer compare draft writing against the corpus.

Keywords: large language models, corpus-based description, academic writing, human-AI collaboration

1. Introduction

The *Writing the Research Article* series sets out to describe the characteristics of research articles in a given field by using a corpus of published articles from the relevant field. This allows for generalisations and examples to be based on empirical data. The titles published or in preparation include Business, Psychology, Social Science, Humanities, Economics, Engineering, Chemistry, Environmental Studies, and Medicine. The series is the product of a linguist-LLM collaboration: the linguist supplies the corpora, the analytical framework, and editorial judgement or feedback, and a large language model (Claude) works on drafting, analysis, and programming. The collaboration trajectory is not straightforward, but the aim is to start with a corpus analysis, followed by the creation of an outline of units, and then the drafting of content. Given the heavy involvement of AI, the books cannot be claimed as academic outputs, but there are obviously implications for scholarship and questions about what language models can and cannot do.

2. Prompt-Driven Authorship

In the first phase, the books were written quickly, with the model doing most of the initial draft following the directions given by the linguist. The method was essentially prompt-driven: the human author specified the design of each volume: the unit structure covering the architecture of the research article, along with features to examine, such as distribution of tenses, citation practice, hedging/boosting, move analysis (Swales, 1990) and stance/metadiscourse (Hyland, 2005). The workflow was rapid. A new disciplinary volume could be produced by systematic adaptation of an existing one.

In this phase, the chapters were combined into Word documents with a controlled inventory of paragraph styles (body text, corpus examples, writer's notes, etc.); tables were standardised and exported as uniformly sized PDFs; and the resulting files were imported into *inDesign*. The strengths of the arrangement were speed, scale, and consistency across the series. The weakness was that the central claim of the series, that the examples and figures were corpus-based, depended on the LLM's report of its behaviour.

3. Crisis and Verification

The second phase began with the suspicious presence of an em-dash in an example. Querying Claude led to the admission that while the first half of the example came from the corpus, the second half was fabricated. And then a further check revealed that some examples had been invented by the model. While the examples were all plausible and appropriate for the genre, they were not present in the corpus.

The response was a program of systematic double-checking. However, it was still prompt-based: asking for verification of the examples and the frequency data. This phase involved a considerable amount of back and forth to make the analysis more legitimate. Nevertheless, the same concerns remained.

4. A Workspace for Corpus Analysis

Rather than prompting a model to characterise a corpus and then policing its output, the next move was to construct an analysis environment in which the corpus characterisations are defined and computed. The architecture enforces a strict division between definition and procedure. Declarative YAML files state what counts: the wordlists, patterns, and categories that define hedges, boosters, connectives, rhetorical moves, formulaic phrases, reporting verbs, etc. These definition files work in tandem with Python scripts. For example, the Python script might check for negation occurring with a booster so that *no clear evidence* is not counted as boosting. Each measure writes the output to a results file together with method notes.

One small part of a YAML file is shown in Figure 1. The labels are not very intuitive. Measures, for example, refers to relevant Python scripts. The Codings show a classification of article types that involved intervention by the linguist. (See below.) The human also had to request that the YAML be fully commented.

This working method uncovered a variety of issues: for example, in the Economics corpus finding a genre distinction that had been invisible was revealed. The short-format papers (letters, brief communications) had been included with full-length research articles and this distorted both the macrostructure analysis and general statistics. The

short-form articles (Notes) were separated into a companion corpus and the format was described separately.

An automated analysis of move structure is always going to present some difficulties. The linguist questioned the use of *however* as an indicator of a gap strategy for a research article. The automatic procedure was replaced with a collaborative one. The regexes being used were made explicit, and the LLM gathered a batch of data from 5 or 10 articles and prompted the linguist to look at the evidence and agree with or override the initial categorisation made by the AI.

This coding produced a major correction in the coding of Economics RAs and it turned out that a continuation strategy rather than gap-filling was the most common way of organising the argument structure in an article.

```yaml
codings:
  # Unit 1's six structural families (applied_empirical, experimental,
  # pure_theory, theory_empirical, hybrid, thematic). 104 articles,
  # coded MB 2026-07-17. Read by measures/families.py.
  families: econ_structural_families.csv
  # Introduction niche strategies. 99 articles, coded MB 2026-07-17.
  # Ruling: continuation 34, gap_filling 27, counter_claiming 14,
  # mixed_findings 12, real_world_problem 11. Read by measures/niche.py.
  niche: econ_niche_strategies.csv
# One module per line, run in order by run_all.py from measures/<name>.py.
# Each returns {stats, examples, notes} into results/econ/results.json, the
# single source for every number in the book.
measures:
  - structure       # Unit 1: macrostructure - section frequency, lengths, skeleton
  - families        # Unit 1: the six structural families, from the codings CSV
  - intros          # Unit 2: openings, gap language, roadmap, contribution claims
  - niche           # Unit 2: niche strategies, from the codings CSV
  - conclusions     # Unit 8: length, openings, limitations, last sentences
  - abstracts       # abstract length, openings, finding-preview rates
```

Figure 1: Part of a YAML file for the Economics volume.

For other coding cases, such as the type of article (theoretical, applied empirical, experimental, etc.) the LLM output a CSV with all the data and its recommendation, and the linguist added a column with the final decision on the coding.

The workspace for each title is governed by a manifest naming its corpora, its definition layer, and the measures (scripts) to run; the measures cover macro-structure, section openings and closings, abstracts, rhetorical moves, hedging and boosting, tense and voice. The scripts run on spaCy dependency parses rather than surface strings.

While some features such as hedges/boosters are examined across all the corpora, discipline-specificity means that each corpus has to be analysed independently of other corpora with respect to the structure. For example, Psychology has a method section in three out of four articles, but it has a top-level heading only in one in twelve articles. It is nested within sections such as Study and Experiment. The procedure that emerged after some setbacks aimed to establish the macrostructure of each new corpus as the initial step. This might seem obvious, but working with an LLM is something of a black box with specific procedures running out of sight.

5. Verification as an Instrument

The fabrications of examples that precipitated the crisis now (hopefully) have a permanent countermeasure: a quote-audit tool which reads the typeset manuscript, extracts every quotation, and classifies it against the corpus as attested verbatim, truncated (a real sentence quoted as though complete), silently edited (sharing long runs with a corpus sentence but altered), or invented (no trace).

Each claim in the manuscript is compared with the recomputed value and is recorded in a ledger, with various notations regarding the process and decisions made.

The audit also had to be extended from the numbers to the text. A figure can be correct while the wording in the sentence around it overclaims in some way. The phrase "the corpus suggests" had to be changed to "as a rule of thumb," or something similar, in some cases. Another example is that the text lists common mistakes, but there is no corpus of mistakes and so, again, the wording has to be adjusted to acknowledge this fact. The mistakes are essentially substantial deviations from the corpus norms.

The LLM view of the process is as follows: verification tools are organised as a staged gate. Six stages run in order: corpus well-formedness; definitions (including a check that no taxonomy row is empty of corpus support, a signature of an imported category); measurement; document structure; claims; and production. A stage that passes can be locked, which records a hash of every input it depended on; if an input later changes, the lock is reported stale and names the file. Exceptions are registered rather than suppressed: a known corpus defect is recorded with its reason and an assessment of what it affects, and a figure with no reproducible source is listed with the analysis that produced it, so the register of unreproducible numbers is itself a document that can be read and challenged.

A cross-volume plausibility check compares nine statistics across all corpora and flags any figure more than four times from the median. This is the instrument that catches a measure which runs cleanly but measures the wrong thing: the citation counter, written for author-year styles, reported 0.9 citations per 10,000 words in chemistry against a cross-volume median of 98.7, exposing the fact that chemistry, engineering and medicine cite numerically and the measure did not apply to them at all.

The procedures listed above represent an improvement. However, the issues are not all fully resolved. The linguist requests the audit be run several times to make sure that nothing has been mis-characterised or overlooked. Another problem is that the LLM is good at writing .md files, but doesn't always read them and so there is the potential to introduce some error that should have been avoided.

6. General Access

One aim arising out of this iteration of an approach to analysing the RA corpora is to make as much as possible of the entire analytical apparatus accessible to anyone interested. Since the corpora are derived from openly licensed research articles (Kershaw and Koeling, 2020), they can be redistributed in full with per-article attribution, alongside the definitions, analysis code, coding sheets and criteria, etc.. A companion website (WritingTheResearchArticle.com) presents these as usable instruments: a concordance search over the corpus with collocates and section distribution; a browser for every category that the books count, showing the frequency and its examples; the definitions and counting rules in full.

In addition, there is a tool that takes a writer's own draft and reports its hedging, connective density, introduction share, conclusion length and opening conventions against the distribution in the corpus of published articles. The

draft tool runs entirely in the reader's browser, so that the user's writing is not uploaded.

The idea is to have the website data generated by the same programs, from the same definition files, that produce the statistics in the books. The corpus is indexed so that it can be searched using JavaScript to match the Python files used on the original corpus. The interface for Economics is shown in Figure 2. And Figure 3 shows the search interface with the categories that were used in the analysis displayed as options for the search.

The website also includes the workspace, the Python scripts, corpus, and so on. It is easiest to simply search the corpus via the website, but the procedural data and outputs are listed for anyone interested. There is no automatic link between the workspace and website contents and so it will be necessary to make any updates needed manually.

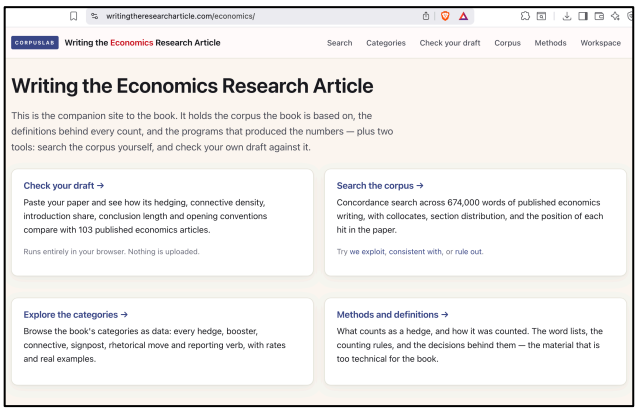

Figure 2: The concordance interface for the Economics corpus.

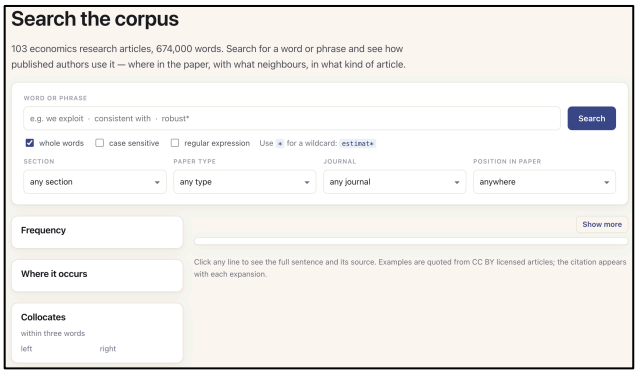

Figure 3: The search interface, with the categories used in the analysis available as search options.

7. Conclusion

This article lists the development in ways of working with Claude over time and indicates the shifts in the nature of the human-LLM collaboration. In Phase 1, the optimistic stage, the model did most of the work and the linguist set the general direction and gave feedback on the model output. Phase 2 involved the linguist probing the model output more thoroughly. This led to improvements, but it was *ad hoc* and based on where the attention of the linguist was focussed. The thrust of Phase 3 was the setting up of a workspace created and used by the LLM and inspected by the linguist. The workspace is itself quite complex because it includes, among other things, definitions, Python scripts, lists of procedures, and results. As an issue arises, something is added to the workspace to deal with it or, at least, note it. A streamlining and simplification of the workspace would likely improve the output. Finally, the externalisation via a website makes the corpus, definitions, results, etc. accessible to all.

8. Bibliographical References

Hyland, K. (2005). *Metadiscourse: Exploring Interaction in Writing*. London: Continuum.

Swales, J. M. (1990). *Genre Analysis: English in Academic and Research Settings*. Cambridge: Cambridge University Press.

9. Language Resource References

Kershaw, D. and Koeling, R. (2020). *Elsevier OA CC-BY Corpus* (Version 2) [Data set]. Mendeley Data. https://doi.org/10.17632/zm33cdndxs.2