Introduction: Using Corpus Evidence to Write Research Articles

This book takes a different approach to teaching academic writing. Rather than prescribing rules or offering model paragraphs to imitate, it opens up a large collection of published research articles and asks: *what do writers actually do?*

The answer, drawn from a corpus of 3,335 social science articles comprising over 2.2 million words, reveals patterns that no single article could show. When we examine thousands of Introductions, we find that writers create research gaps using a remarkably consistent set of constructions. Examples such as *however, little research has examined* appear with predictable regularity. When we examine thousands of Conclusions, we find that *contributes to* is the default way of claiming significance, appearing roughly once every 200 words. When we compare sections, we discover that the Discussion is hedged almost twice as heavily as Results — not because Discussion writers are less confident, but because interpretation carries greater uncertainty than reporting.

These are not rules. They are tendencies due to the accumulated choices of thousands of writers navigating the same rhetorical challenges you face. This book makes those tendencies visible so that you can make informed decisions about your own writing.

The Corpus

Feature	Detail
Articles	3,335
Total words	2,206,539
Sections	7 (Introduction, Literature Review, Methods, Results, Results-Discussion, Discussion, Conclusion)
Disciplines	Across the social sciences

The analyses in this book draw on a purpose-built corpus of social science research articles:

The corpus was divided into its component IMRDC sections, allowing us to compare the same linguistic features across different parts of the paper. This is what makes the analysis powerful: we can show not just that hedging is common in academic writing, but that it peaks in the Discussion at 156.8 instances per 10,000 words — more than double the Methods rate of 65.5. We can show not just that citations matter, but that the Introduction is six times more citation-dense than Results.

Every frequency, every pattern, and every example sentence in this book comes directly from this corpus.

Why Corpus Evidence Helps Writers

Most academic writing textbooks work from the top down: they present a model of what good writing looks like and ask students to reproduce it. Corpus analysis works from the bottom up: it examines what published writers actually produce and identifies the patterns that emerge.

This matters for three reasons.

First, corpus evidence replaces intuition with frequency. Experienced writers often cannot articulate why they make particular choices; they "just know" that a Discussion sounds wrong without hedging or that a Methods section needs some passive verbs. Corpus data makes this tacit knowledge explicit. When we show that passive past participles appear at 22.2 per 10,000 words in Methods — roughly once every 45 words — we are quantifying a pattern that experienced writers know but cannot make explicit.

Second, corpus evidence reveals the architecture of the paper. Individual articles look different from one another: different topics, different methods, different findings. But at the level of rhetorical structure, they are remarkably similar. Introductions create gaps. Methods justify procedures. Results report data. Discussions interpret. Conclusions claim contributions. The corpus makes this

shared architecture visible by showing which constructions cluster in which sections, and at what frequency.

Third, corpus evidence shows range, not just norms. For any given rhetorical move, the corpus reveals multiple ways of accomplishing it. A research gap can be signalled through absence (*"little research has examined..."*), contradiction (*"we argue that this view fails to account for..."*), unresolved debate (*"findings have been mixed..."*), or real-world urgency (*"despite policy efforts, the problem persists..."*). Seeing this range helps writers choose the strategy that best fits their study, rather than defaulting to the most common formula.

How to Use This Book: Conceptualise First, Then Find Language

A corpus-based writing book carries a risk: writers may treat it as a phrase bank, copying constructions wholesale without understanding the rhetorical work those constructions perform. This would be a mistake. The examples in this book are not templates to be filled in; they are evidence of how published writers have solved particular rhetorical problems.

Before consulting any section of this book, you should have already done the conceptual work:

Before writing your Introduction, you should be able to answer: *What is the specific gap, problem, or question my study addresses? Why does this gap matter? What exactly does my study do about it?* Only then should you look at how other writers have signalled gaps and announced their studies — not to copy their words, but to find constructions that fit your conceptualisation.

Before writing your Methods, you should be able to answer: *What did I do, in what order, and why did I make each choice?* The corpus shows you how other writers sequence and justify their procedures, but the logic of your own methodology must come first.

Before writing your Discussion, you should be able to answer: *What do my findings mean? How do they relate to what others have found?*

What are the limits of my claims? The corpus shows you how other writers hedge interpretations, compare with literature, and acknowledge limitations; but your interpretation of your own data must be original.

The process, then, is: **conceptualise > consult > compose.** Decide what you want to say, look at how others have said similar things, then write in your own voice using structures that serve your argument. The corpus gives you a repertoire of options. Your study determines which options to choose.

What the Corpus Reveals: A First Look at Section Proportions

Before examining specific structures, it is worth seeing the paper as a whole. Our corpus reveals that the five IMRDC sections are not equal in length:

Section	Share of paper	Words in a 6,000 word paper	Avg. sentence length
Introduction	32.6%	~1,960	31.2 words (longest)
Methods	15.5%	~930	28.5 words (shortest)
Results	10.2%	~610	29.8 words
Discussion	24.1%	~1,450	30.3 words
Conclusion	17.6%	~1,050	30.3 words

Three observations stand out. First, the Introduction is the longest section; roughly a third of the paper. Students who write a one-paragraph Introduction and a ten-page Results section are inverting the expected proportions. Second, Results is the shortest section at just 10.2%. Data presentation is compressed; interpretation (Discussion, 24.1%) takes more space than reporting. Third, Introduction and Conclusion together account for just over half the paper (50.2%) — the argumentative frame (why the study matters, what it contributes)

takes more space than the empirical content (Methods + Results = 25.7%).

These proportions are not prescriptive. Your paper may differ for good reasons; a complex methodology, an unusually rich dataset, a contested theoretical landscape. But they provide a useful benchmark. If your Results section is three times longer than your Discussion, it is worth asking whether you are reporting more than you are interpreting.

How This Book Is Organised

The structures in this book are presented in three groups, moving from the internal architecture of individual sections to features that arc across the whole paper.

Part 1: Rhetorical Moves — What writers do within each section

These structures describe the characteristic rhetorical "moves" — the recurring steps — that writers take within each IMRDC section. They show, for example, that Introduction writers create a research territory before identifying a gap, that Discussion writers cycle through restating, comparing, and explaining for each finding, and that Conclusion writers claim contribution before stating implications. Each move is illustrated with corpus frequencies and example sentences.

> Unit 1: Introduction Moves (CARS model and alternative niche strategies)
> Unit 2: Methods Moves
> Unit 3: Results Moves
> Unit 4: Discussion Moves
> Unit 5: Conclusion Moves

Part 2: Opening and Closing Strategies — How writers begin and end sections

> Unit 6: First and Last Sentences of Each Section

Part 3: Cross-Section Arcs — Features that change systematically across the paper

These structures examine features that shift from section to section, creating trajectories that run through the whole paper. Tense shifts from present to past and back. Hedging dips in Methods and peaks in Discussion. Citations cluster in Introductions and thin in Results. Together, these arcs reveal the paper's underlying argumentative logic.

Unit 7: Tense Distribution Across Sections
Unit 8: Citation Distribution Across Sections
Unit 9: The Hedging and Boosting Arc

Capstone: The Paper as a Single Argument

The final structure synthesises all previous analyses to show how the entire paper serves a single macro-argument: *there is a gap in knowledge, and we fill it.* It draws together the move structures, the tense arc, the citation arc, and the hedging arc to show how every linguistic choice contributes to this overarching case.

Unit 10: The Paper as a Single Argument

Throughout the book, remember: the corpus shows you what is common, not what is compulsory. Your task is not to reproduce these patterns mechanically but to understand the rhetorical problems they solve; and then to solve those problems in your own way, for your own study.