

Lexical Bundles in Graded News

Effects of Language Simplification on Frequency, Structure, and Function

Huy Khang Ninh Nguyen

Nottingham Trent University, School of Arts and Humanities — MA TESOL
Submission ID: 2281 | khangninh92@gmail.com

INTRODUCTION

- Graded news is widely used in ELT to support extensive reading, offering level-appropriate input for learners.
- Lexical bundles — frequent four-word sequences — support fluent processing and signal discourse functions (Biber et al. 1999; Biber, Conrad & Cortes 2004).
- Simplification may prune or replace the bundles that make authentic news efficient to process, yet few studies compare graded and authentic versions of the same stories.
- This study compares four-word lexical bundles in graded (B1/B2) news from Breaking News English with their original source articles.

METHODOLOGY

- Corpus-driven comparative design: three corpora built from Breaking News English (B1, B2 graded) and their linked original sources (non-graded).
- 411 articles per corpus across 8 topics (arts, business, crime, environment, health, politics, science, sports), published 2015–2025.

Corpus	Word count	Texts
B1 (graded)	77,614	411
B2 (graded)	94,566	411
Non-graded	253,169	411

- Four-word bundles extracted with corpus-sensitive frequency thresholds (40 per million words for graded corpora; 20 per million for non-graded) and a dispersion criterion of ≥ 3 texts.
- Manual refinement removed punctuation-disrupted and context-bound sequences and merged overlapping bundles.
- Final bundle counts: B1 = 84 types / 307 tokens; B2 = 59 types / 316 tokens; Non-graded = 121 types / 892 tokens.

RESEARCH QUESTIONS

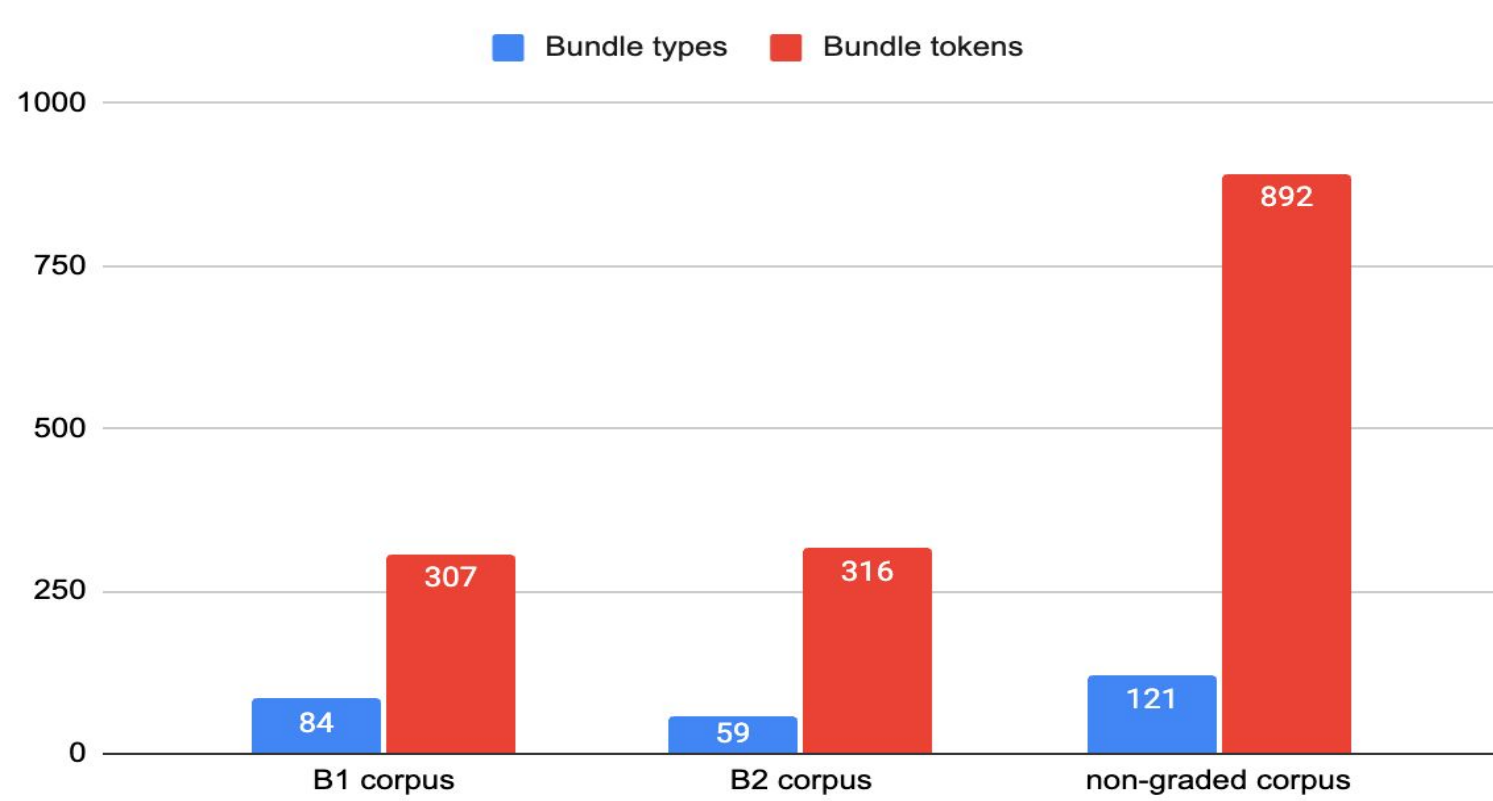
To what extent do four-word lexical bundles in graded news differ from those in original news across frequency, structure, and function?

- RQ1 — Frequency: How do frequency distributions of bundles differ between graded and authentic news?
- RQ2 — Structure: How do structural patterns of bundles differ between graded and authentic news?
- RQ3 — Function: How do functional patterns of bundles differ between graded and authentic news?

THEORETICAL FRAMEWORK

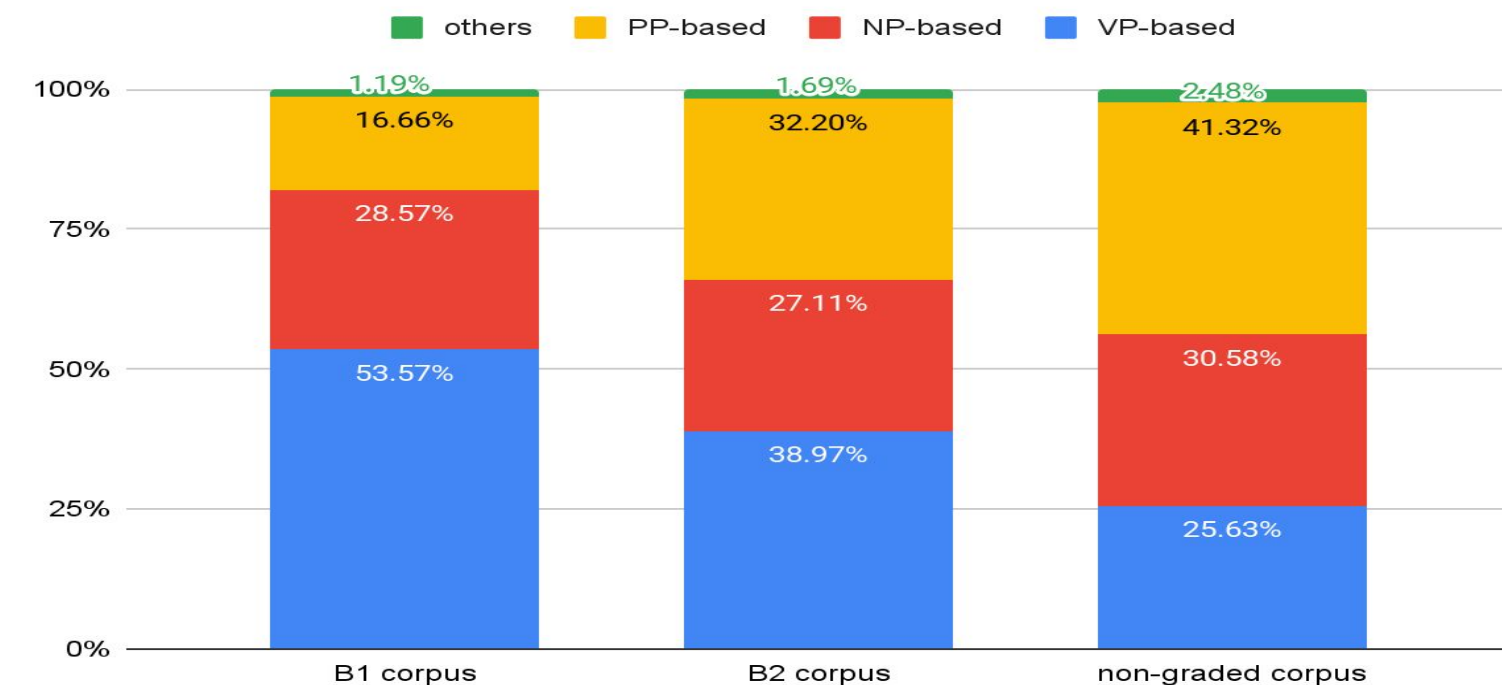
- Lexical bundles are identified using frequency and dispersion thresholds (Biber et al. 1999; Cortes 2004).
- Structural taxonomy (Chen & Baker 2010): VP-based, NP-based, and PP-based bundles.
- Functional taxonomy (Biber, Conrad & Cortes 2004): referential, stance, and discourse-organising bundles.
- Text simplification operates through lexical control, syntactic trimming, and discourse pruning (Siddharthan 2014; Nation & Waring 2020), which may reshape bundle distributions.

KEY FINDINGS — FREQUENCY



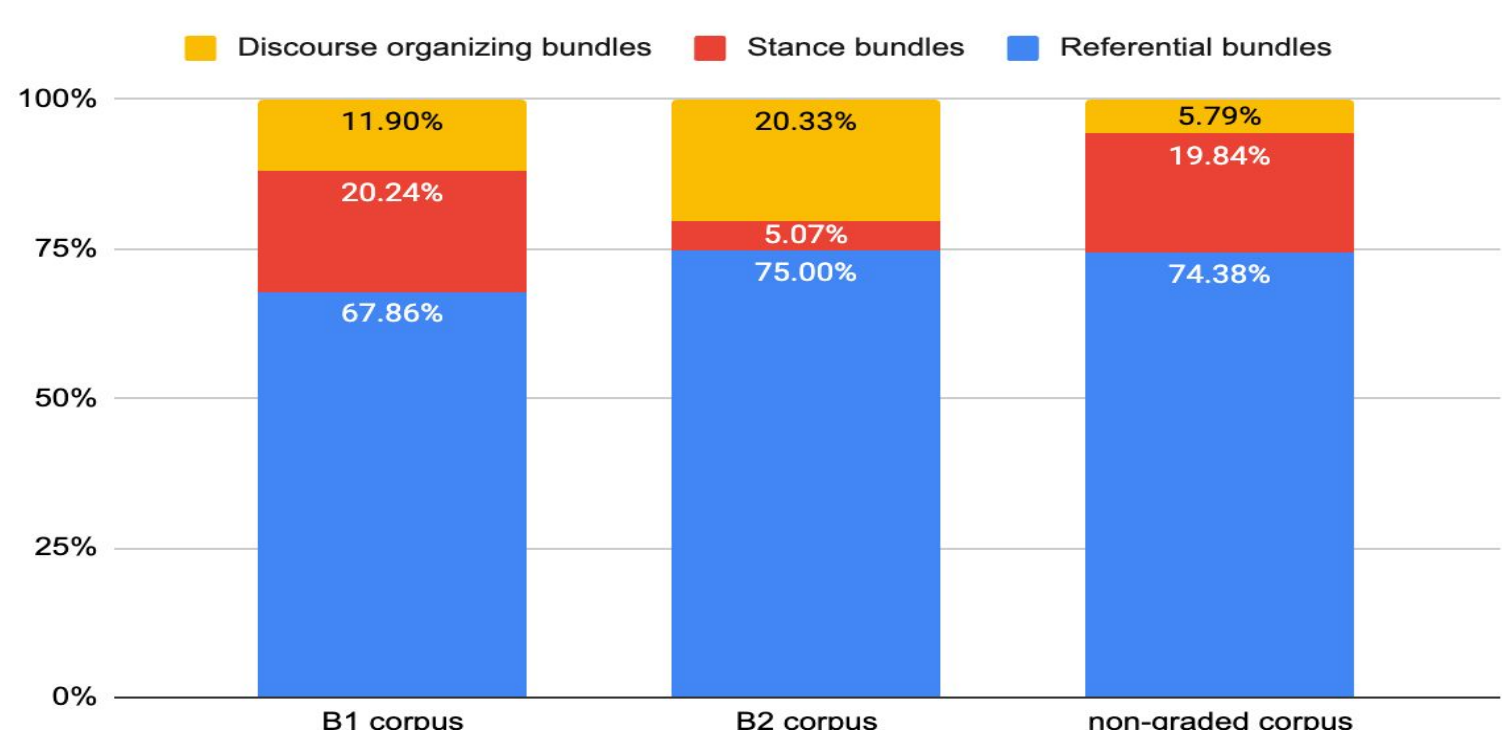
- The non-graded corpus contains the most bundle types (121) and tokens (892) — expected, given its larger size.
- B2 shows fewer bundle types than B1 (59 vs. 84) despite being the larger corpus: editors trim transparent quotative/existential frames (e.g., there has been a) when moving from B1 to B2.
- Graded corpora recycle a narrower set of high-frequency templates (e.g., per cent of the, one of the most); authentic news spreads its frequency across a wider variety of bundles.

KEY FINDINGS — STRUCTURE



- VP-based bundles dominate B1 (53.6%) but fall in B2 (39.0%) and further in authentic news (25.6%).
- PP-based bundles rise steadily — 16.7% (B1) → 32.2% (B2) → 41.3% (authentic) — becoming the largest category in authentic news.
- NP-based bundles stay fairly stable (27–31%) across all corpora; simplification favours clause-led VP frames, while authentic news favours the denser NP/PP packaging typical of professional journalism.

KEY FINDINGS — FUNCTION



- Referential bundles dominate all three corpora (68–75%), reflecting news writing's reliance on nominal reference and qualification.
- B1 shows more stance marking (20.2%) than B2 (5.1%); discourse-organising bundles rise from B1 (11.9%) to B2 (20.3%).
- Authentic news shows a more even mix — stance (19.8%) and discourse-organising (5.8%) — reflecting newsroom attribution and topic-shifting routines.

IMPLICATIONS

- Repetition of core bundles in graded texts supports fluency-building for intermediate learners.
- A narrower bundle range means learners risk missing authentic newsroom phraseology — especially attribution and time/place framing.
- Recommend bridging: recycle under-represented PP/attribution bundles at B2 through noticing, input enhancement, and controlled production tasks.
- Materials writers can sequence bundles across levels to phase in authentic phraseology as learners progress.

LIMITATIONS & FUTURE RESEARCH

- Relatively small, single-source corpora limit generalisability.
- Extraction and classification relied on a single rater's judgement.
- Future work: larger, multi-platform corpora; inter-rater reliability checks; learner-based experiments on graded vs. authentic exposure.

REFERENCES

- Allan, R. (2016). Lexical bundles in graded readers. *System*, 59, 61–72.
- Bednarek, M. and Caple, H. (2017). *The Discourse of News Values*. Oxford: Oxford University Press.
- Biber, D., Conrad, S. and Cortes, V. (2004). If you look at...: Lexical bundles in university teaching and textbooks. *Applied Linguistics*, 25(3), 371–405.
- Biber, D., Johansson, S., Leech, G., Conrad, S. and Finegan, E. (1999). *Longman Grammar of Spoken and Written English*. Essex: Pearson.
- Chen, Y.H. and Baker, P. (2016). Investigating criterial discourse features across L2 development. *Applied Linguistics*, 37(6), 849–880.
- Cortes, V. (2004). Lexical bundles in published and student disciplinary writing. *English for Specific Purposes*, 23, 397–423.
- Nation, I.S.P. and Waring, R. (2020). *Teaching Extensive Reading in Another Language*. New York: Routledge.
- Siddharthan, A. (2014). A survey of research on text simplification. *ITL*, 165(2), 259–298.
- Zakaria, A., Renandya, W.A. and Aryadoust, V. (2023). A corpus study of language simplification and grammar in graded readers. *LEARN Journal*, 16(2), 130–153.