The International Journal of
**Technology, Innovation,
& Education**

Beyond learning outcomes: Tracking Tunisian learners' actual use of phraseology in a data-driven learning course

Ben Amor Olfa [iD] [1] * **and Faiza Derbel** [iD] [2]

[1] Faculty of Arts and Human Sciences of Sousse, University of Sousse, Tunisia; Laboratory of Approaches to Discourse (LAD), University of Sfax, Tunisia.
[2] Faculty of Letters, Arts and Humanities, University of Manouba, Tunisia; Laboratory of Approaches to Discourse (LAD), University of Sfax, Tunisia.

BenAmor.Olfa@flsh.u-sousse.tn
* Corresponding author

Abstract

This study investigates how master students majoring in Business Studies engage with the corpus toolkit *AntConc* in a data-driven learning (DDL) intervention targeting non-finite phraseological patterns in academic writing. Thirty-five Tunisian EFL students participated in a five-week course combining guided practice with independent concordance searching. Six annotated patterns, drawn from a business-domain corpus, were deployed for structural and functional analysis. Learner activity during DDL tasks was tracked through screen recordings for a sub-sample of 20 students and supplemented by written reflection logs. Tracking captured query variety, search refinement, time on task, use of *AntConc* features, pattern documentation, and discourse function recognition, adapted from established frameworks (Pérez-Paredes et al., 2011, 2012). Learning outcomes were assessed through pre-, post-, and delayed tests. Results showed significant gains in the use of target structures, though retention varied across learners. Correlations indicated that varied querying and systematic notetaking were associated with higher gains. Thematic analysis of the reflection logs corroborated these findings, revealing increased phraseological awareness but persistent challenges in discourse function identification. Thereby, the study demonstrates the value of integrating behavioural monitoring into corpus-based instruction, supporting the design of scaffolded, tool-mediated environments that foster phraseological competence and learner autonomy.

1. Introduction

The integration of corpora in language teaching has revolutionised pedagogical approaches, notably through Data-Driven Learning (DDL) (Johns, 1991) which enables learners to explore language patterns and uncover linguistic rules. In DDL-based learning environments, students employ one of two methods: either based on examining printed corpus materials (hands-off) or direct, hands-on interaction with corpora. Various terminologies have emerged to characterise and distinguish these approaches, including "hard and soft version" (Gabrielatos, 2005), "hands-on DDL and paper-based materials" (Boulton, 2010), "direct and indirect consultation of corpora" (Chambers, 2007), and "deductive and inductive DDL" (Cresswell, 2007). This study focuses on a direct hands-on use of a curated business corpus as resource to teach non-finite clause phraseology to Business English (BE) students within the Tunisian higher education context.

While an expanding corpus of research exists (Boulton & Vyatkina, 2021, 2024; Li et al., 2025; Lusta et al., 2023; Pérez-Paredes, 2019), stressing the effectiveness of both direct and indirect approaches, a considerable number of teachers continue to perceive corpus-based methodologies as peripheral to language instruction. This cautious adoption can be attributed to limited familiarity with corpus tools, concerns over classroom time constraints, and uncertainty about the level of learner on-task engagement. Given this, while the potential benefits of incorporating corpora into teaching are widely recognized in the literature, they have yet to achieve widespread acceptance and integration within educational circles (Boulton, 2017; Pérez-Paredes, 2019). In the Maghreb region, research on DDL remains particularly limited, with only a few empirical studies to date—most notably Bouabida (2020) in Morocco and Chaalal (2024) in Algeria—underscoring the need for more context-specific research. The present study therefore contributes to addressing this gap by extending DDL inquiry to the Tunisian higher-education context.

While early DDL models promoted full learner autonomy (O'Sullivan, 2007; Sun & Wang, 2003), recent work demonstrates the need for scaffolded support to help students interpret concordance lines—that is, to read and infer recurrent lexical and grammatical patterns from keyword-in-context displays—effectively (O'Keeffe, 2021; Pérez-Paredes, 2019). Balancing teacher guidance with learner discovery is important in contexts where students lack prior experience with corpus tools, as it reduces cognitive overload while fostering autonomy. In line

with this, Boulton and Vyatkina (2024) call for research integrating usage-tracking instruments to capture learner behaviour during pedagogical use of corpus software and scaffolding types. They note that DDL studies rarely document students' search queries because of difficulties in capturing detailed interface interactions, leaving a gap in our understanding of how learners actually interact with corpora in practice (Crosthwaite et al., 2019). This responds to the need for further research by examining how learners use corpora during DDL-based instruction.

More specifically, this study investigates how EFL business master's students engage with *AntConc* during phraseology-focused academic writing activities. A comparative corpus analysis of expert journal articles and graduate dissertations informed the design of intervention materials. *AntConc* was selected for its accessibility and ability to use wildcards and filters to explore non-finite patterns efficiently. Learner search behaviour was systematically tracked through screen recordings and written logs, and this behavioural data was triangulated with pre-, post-, and delayed test scores and qualitative reflections to provide a comprehensive account of tool use and learning behaviour.

The instructional focus addresses non-finite clause structures, which are frequent in academic writing (Biber et al., 1999) yet challenging for EFL learners. Constructions such as *taken together* or *when analysing results* enhance information density, cohesion, and rhetorical precision. Six phraseological patterns were taught and annotated for both structure and semantic function in the corpus as part of the learners' BE writing curriculum. The goal was to help learners recognize, interpret, and apply these patterns in their own writing through tailored *AntConc*-mediated DDL tasks.

2. Reviewing research on DDL and phraseology

DDL posits that learners construct linguistic knowledge by investigating authentic texts and inducing patterns from concordance evidence (Johns, 1991). It aligns with the Noticing Theory (Schmidt, 1993) and usage-based grammar, both of which maintain that repeated exposure to context-bound patterns fosters entrenchment and facilitates productive language use (Bybee, 2010). From a phraseological perspective, pattern grammar (Hunston & Francis, 2000) and corpus-based descriptions stress that many grammatical choices are conventionalised lexical–grammatical clusters rather than abstract rules. Thus, learning grammar for academic writing largely involves recognising and using recurrent phraseological packages (Biber et al., 2004; Hunston & Francis, 2000). DDL enables learners to notice such packages in real texts and test

hypotheses about their form and function, thereby supporting both form-focused and function-oriented learning.

Non-finite clause constructions (e.g., *to investigate the effect*, *being confident in the results*, *data collected during…*) are pervasive in academic prose and serve key rhetorical functions: they condense information, signal stance, indicate purpose, and organise discourse (Biber et al., 1999; Hyland, 2005; Quirk et al., 1985). For many EFL learners, these structures pose both formal and functional challenges, requiring choices about voice, adjuncts, argument structure, and rhetorical intent (Green, 2017; Petrovitz, 2001). Research indicates that language acquisition depends not only on exposure but also on how learners interact with examples—through noticing, hypothesis-testing, and consolidation (Flowerdew, 2009; Yoon & Hirvela, 2004). Learner behaviour is thus a key factor influencing learning outcomes (Huang, 2022; Tall & Razali, 2006): strategic searching, repeated sampling, context inspection, and systematic documentation**,** all increase the likelihood of forming accurate generalisations from concordance lines (Crosthwaite et al., 2019; Pérez-Paredes et al., 2012).

In DDL research, learning behaviour is also viewed as a multidimensional construct encompassing cognitive strategies (*e.g., query selection, search refinement*), metacognitive processes (*e.g., monitoring, hypothesis testing*), and observable interactional actions (*e.g., time on task, tool use, documentation*) (Crosthwaite et al., 2019; Pérez-Paredes et al., 2012). Recent studies have operationalised these behaviours as measurable actions such as query variety, refinement sequences, time spent on searching and analysing, use of concordancer functions (file view, KWIC, collocates), and systematic note-taking (Crosthwaite et al., 2019; Gilquin, 2022; Pérez-Paredes et al., 2011, 2012; Qiu, 2024). Quantifying such behaviours enables researchers to investigate not just *whether* learning occurred but *how* it was achieved, linking engagement patterns to learning outcomes. Early DDL research relied heavily on manual logs (see Table 1 below) and learner questionnaires and interviews (Chambers & O'sullivan, 2004; Ma, 1994), which capture learner reflection but not without limitation of self-reporting since learners often verbalise their thoughts without analysing or explaining their actions, making such data incomplete and potentially biased (Cohen, 2013). To obtain fine-grained process data, researchers have adopted digital user-tracking methods including screen recordings, web-proxy logs, keystroke capture, and automated query logs (Cotos, 2014; Crosthwaite et al., 2019; Gaskell & Cobb, 2004; Pérez-Paredes et al., 2011).

These methods offer objective sequences of learner actions — exact queries, temporal patterns, and navigational choices — enabling researchers to reconstruct search trajectories and to link them to immediate task success or later retention (Pérez-Paredes et al., 2012). Hybrid designs that triangulate digital logs with written reflections can capture both what learners did and what they believed they were doing — a combination that can reveal convergences and divergences between perceived and actual strategy use (Hafner & Candlin, 2007; Pérez-Paredes et al., 2011).

Table 1. Tracking learning behaviour in previous DDL studies

Study	Tracking type	Participants	Focus	Target linguistic feature
Ma (1994)	Manual log	ESL learners	Learner concordance diaries, reflections	Mixed lexico grammatical
Chambers & O'Sullivan (2004)	Manual log	8 university French learners	Logs on revision and corpus consultation in writing	Grammar and Lexico-grammatical patterning
Hafner & Candlin (2007)	Manual + digital	Law students	Search log analysis, interviews, writing samples	Legal discourse grammar
Gaskell & Cobb (2004)	Digital (IP logs)	ESL learners	Error correction through concordance information	Word and sentence level
Park & Kinginger (2010)	Manual + digital	L2 writers	IP logging, screen recording, and reflections	Writing process
Pérez-Paredes et al. (2011)	Digital log	EFL learners	Guided vs. unguided DDL, search behaviour	It-cleft sentences and inversion
Pérez-Paredes et al. (2012)	Digital log	EFL learners	Search sequence patterns, strategy classification	it-clefts
Cotos (2014)	Digital log	Graduate students	Search behaviour, task engagement	Functional meaning in academic writing
Crosthwaite et al. (2019)	Digital log	EFL learners	Query selection, learning outcomes	Lexico-grammatical patterns
Kotamjani et al. (2017)	Digital log	EFL learners	Tracking writing process	Grammar and vocabulary
Gilquin (2022)	Digital log + keystroke logging	ESL learners	Writing process	Grammar
Qiu (2024)	Digital log + stimulated recalls	Graduate EFL learners	Self-directed writing process	sentence-level linguistic features

A number of studies have applied these approaches in varied contexts, targeting collocations, grammar, or broader writing processes. Table 1 above summarises these tracker-based studies in corpus pedagogy showing that, while some focused on vocabulary and lexico-grammatical patterns (e.g., Ma, 1994; Crosthwaite et al., 2019), others delved into more complex areas such as functional meaning in academic writing (Cotos, 2014) or legal discourse grammar (Hafner & Candlin, 2007). Collectively, these studies highlight both the versatility of tracking methods and their potential to reveal the interaction between learner strategies, task design, and linguistic outcomes. At the same time, they underscore the need to consider the pedagogical conditions under which these behaviours emerge.

DDL is often promoted as a means to foster learner autonomy (Boulton, 2010; Charles & Hadley, 2022; Johns, 1991), yet research shows that unmediated autonomy can result in unproductive searching, confusion over discourse functions, and superficial pattern recognition (Charles, 2006; Hafner & Candlin, 2007). Contemporary DDL pedagogy therefore emphasises scaffolded autonomy combining teacher guidance, worked examples, or worksheets with opportunities for independent exploration (Braun, 2005; Flowerdew, 2009). Tool design plays a critical role in this balance: annotated systems such as *NooJ* reduce noise and provide targeted output, lowering cognitive load for novices (Crosthwaite et al., 2019), whereas raw concordancers like *AntConc* encourage exploratory discovery but demand greater query-crafting skill. This creates a trade-off: strong scaffolding tends to improve task completion and immediate understanding, while looser guidance supports broader exploration and potentially deeper generalisation. Effective pedagogy thus requires calibrating support — offering strategies such as node suggestions, model queries, and demonstrations of wildcard use — while creating opportunities for experimentation and hypothesis building (Flowerdew, 2009; Boulton & Vyatkina, 2024). The question is not so much whether to scaffold and more about determining the optimal amount and timing for support, relative to learners' proficiency and instructional goals.

3. Focus and scope of the study

Although several studies have used digital user-tracking to reveal search patterns and strategy sequences (Pérez-Paredes et al., 2011, 2012; Cotos, 2014; Crosthwaite et al., 2019), most have focused on general query behaviour or collocation discovery rather than grammar-centred phraseology such as non-finite clauses. Empirical work rarely links coded behavioural features

to gains on grammatical phraseology tests or triangulates screen-recordings with written reflections to compare perceived and actual strategies. Few studies report how annotated corpora (XML exports, tags) are used in classroom concordancing tasks — for instance, whether learners display tags in *AntConc* when verifying discourse function judgments.

In this study, we address these gaps by (i) operationalising learning behaviour (query variety, search refinement, time on task, tool use, pattern documentation, and discourse-function recognition), (ii) tracking these behaviours via screen recordings and written logs, and (iii) relating them to pre-, post-, and delayed test outcomes for non-finite patterns. We also examine the ways in which learners engage with annotated corpus feedback during DDL tasks. Together, these analyses contribute empirical evidence to debates on scaffolded autonomy, tool design, and the link between observable behaviours and learning gains.

Drawing on the theoretical insights outlined above and in light of the aims formulated for the study, two research questions are proposed:

1. To what extent do learners improve in their use of targeted non-finite phraseological patterns after engaging in *AntConc*-mediated DDL activities?
2. Which patterns of learner behaviour operationalised in the instructional plan emerge during *AntConc* tasks and contribute to test gains and/or retention?

4.   Research Methods and Materials

Participants in the study consisted of thirty-five first-year master's students (11 male and 24 female) in a French-medium Economics and Finance program at the Faculty of Economics and Management of Mahdia (FSEGM), Tunisia, with an average age of 23 years. English is a third language for these students but believed to be essential for their future careers. Their proficiency levels, based on the institution's records and program entry scores, ranged between B1 and B2 according to the Common European Framework of Reference (CEFR). Most reported moderate-to-frequent use of computers to support writing (e.g., consulting dictionaries, translation tools, grammar sites), but none had prior experience with corpora or concordancers. Given this was the students' first exposure to DDL, we provided an initiation session to familiarise them with *AntConc*. Convenience sampling was used in this exploratory investigation. This very group of BE students was selected to experiment with this innovative approach through a plan of a series of guided DDL sessions targeting non-finite clause phraseology. The main objective as mentioned in the introduction is to collect data about user

behaviour and response to the pedagogical strategy based on corpus exploration and the inference of rules related to phraseology in a corpus of BE texts.

'User-tracking' is known as a reliable way of tracing process (Pérez-Paredes et al., 2011, 2012) and observing how learners interact with the corpus tools while completing the programmed corpus-based instructional tasks. An instructional corpus was compiled comprising Business English texts from published journal articles (JA) and MA/PhD dissertations (TD). To ensure balance representation, 20 student dissertations and 49 research articles were included, reflecting differences in length and word count (see Appendix A), reaching for a total of 1,123,725 words.

Given the corpus size and complexity, *NooJ* software (Silberztein, 2020) was used to undertake automated parsing and annotation of non-finite clause constructions and their phraseology. Annotation relied on cascades of syntactic transducers implemented as automata within syntactic graphs, enabling recognition of recurrent phraseological and functional patterns. Part-of-speech tagging and functional parsing facilitated systematic extraction of target patterns. It is worth nothing that *NooJ*'s precision and recall have been previously assessed and validated (Ben Amor & Derbel, 2020).

The instructional focus of the study was placed on six non-finite clause phraseological patterns frequently used in academic writing (Biber et al., 1999). These structures were selected for their high communicative value and recurrent presence in expert academic prose, particularly within the genres analysed in the corpus. The six target patterns were:

1. it + be + adjective + to-infinitive (e.g., *It is important to consider...*)
2. it/this + verb + us + to-infinitive (e.g., *This allows us to explore...*)
3. there is a/the need + to-infinitive (e.g., *There is a need to revise...*)
4. adverb/adjective + verb-ing (e.g., *widely increasing, worth noting*)
5. adverb + verb-ed (e.g., *widely adopted, clearly defined*)
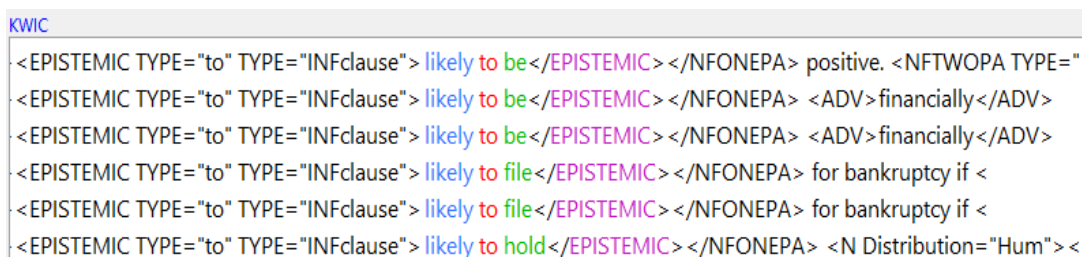6. noun + verb-ing/verb-ed (e.g., *data suggesting, issues raised*)

In addition to their surface-level grammatical structure, these patterns were analysed for their semantic and discourse functions, drawing on the functional taxonomy proposed by Biber et al. (2004). Each occurrence in the corpus was categorized under one of five function-based categories:

1. Stance (Epistemic and Attitude) e.g., *It is necessary to...*),

2. Referential (e.g., *data suggesting*...),

3. Discourse organizers (e.g., *This allows us to*...),

4. Goal-oriented (e.g., *to ensure that*...), and

5. Multi-functional, where more than one function overlapped.

This functional categorization forms part of a broader study reported elsewhere (Ben Amor, 2025). In this research, both the phraseological structure and semantic function of each non-finite clause pattern were encoded in the corpus using *NooJ*. Programmable graphs and syntactic transducers enabled the integration of structural tags with function labels, allowing precise extraction and analysis of each pattern in its rhetorical context. The annotated corpus was exported in XML format and imported into *AntConc* for use in DDL activities. Students could reveal these tags for feedback by enabling the "Show Tags" option in *AntConc*'s global settings. Figure 1 illustrates the structural and functional tags for the first pattern (adjective + to-infinitive), labelled "NFONEPA" with the functional tag "EPISTEMIC" displayed in concordance lines.


```
KWIC
<EPISTEMIC TYPE="to" TYPE="INFclause"> likely to be</EPISTEMIC></NFONEPA> positive. <NFTWOPA TYPE="
<EPISTEMIC TYPE="to" TYPE="INFclause"> likely to be</EPISTEMIC></NFONEPA> <ADV>financially</ADV>
<EPISTEMIC TYPE="to" TYPE="INFclause"> likely to be</EPISTEMIC></NFONEPA> <ADV>financially</ADV>
<EPISTEMIC TYPE="to" TYPE="INFclause"> likely to file</EPISTEMIC></NFONEPA> for bankruptcy if <
<EPISTEMIC TYPE="to" TYPE="INFclause"> likely to file</EPISTEMIC></NFONEPA> for bankruptcy if <
<EPISTEMIC TYPE="to" TYPE="INFclause"> likely to hold</EPISTEMIC></NFONEPA> <N Distribution="Hum"><
```

Figure 1. Screenshot of structural and functional tags in concordance lines

The instructional programme was implemented as part of an academic writing course covering five 90-minute sessions. Learners' activity consisted in exploring non-finite clause phraseological patterns and their semantic functions in a BE corpus using *AntConc* (version 3.5.8.0). Each session involved guided DDL tasks supported by worksheets designed to help learners recognise and generalise phraseological patterns (see Appendix C for a sample). Participants first examined sample sentences on the worksheet such as "...*it is still strong enough to dominate the accommodating and anchoring effects of imperfect credibility; it is worth noting that the algorithm allows for alternative specifications*." They were then asked to extract similar patterns in *AntConc* (see Figure. 2), entering the highlighted node words into the search bar and using wildcards for combinations like *worth *; * worth noting; a * enough * to-infinitive; * strong enough to-infinitive*. Results were sorted to group relevant patterns with preceding or following lexical slots including *worth mentioning; well worth; good enough; still*

strong enough. Learners recorded their findings on the worksheet, drew conclusions about the structures, compared their frequency across expert and student writing, and documented the discourse functions these constructions served.

Figure 2. Sample of phraseological patterns

The tasks were designed to encourage learners to observe how non-finite constructions function within real academic contexts, and to make inductive generalizations about form, frequency, and usage (Liu & Jiang, 2009). Students also completed tasks whereby they contrasted patterns across sub-corpora (e.g., dissertations vs. research articles), examining how grammatical choices and their functions varied in terms of sub-genre. User-tracking techniques were employed using screencast software (*Camtasia*, *OBS Studio*) to monitor task performance, enabling detailed observation and documentation of:

- Query formulation processes,
- Time spent on specific tasks or searches,
- Revisions or repeated attempts,
- On-screen interaction with *AntConc* features (e.g., sorting, keyword navigation).

The recordings were later analysed to investigate learners' strategies, difficulties, and degree of autonomy when working with *AntConc*. These screen-based inferences were corroborated with learners' responses on the worksheet and further triangulated with reference to post-task tests and log reflections. Based on these data sources it was possible to develop a process-oriented account of corpus engagement.

The following section describes the data collection instruments and the procedures used for learner 'tracking', along with the performance tests implemented to measure possible learning gains.

4.1. Performance Tests

To tease out any possible gains in the performance of the learners during this experiment, a pre-test, an immediate post-test, and a delayed post-test were administered to the group at different intervals over the five-week DDL implementation phase: the pre-test in session one, the immediate post-test in the fifth week, and the delayed post-test two months after the intervention. The objective of these tests was to assess the learners' progress, if any, and ability to identify the phraseology in non-finite clauses and their discourse functions.

The tests included five task types across sections, contributing to a total score of 40 points (see Appendix B):

1- Pattern Recognition and Function Identification – learners identified non-finite constructions in context and classified their discourse function.

2- Phraseology Comparison – learners produced and reflected on contrastive patterns expressing different discourse functions.

3- Rewriting Task – learners transformed finite structures into non-finite clause constructions.

4- Interpretation of opaque phraseological expressions – learners analysed figurative or less transparent non-finite patterns.

5- Production Task – learners wrote a short academic paragraph incorporating a range of non-finite clause constructions across semantic categories.

Test items were rotated in sequence to reduce the likelihood of practice effects, while the underlying structures remained constant across administrations. Each test was completed within a 45-minute session. Prior to the testing, three teachers validated the test content, and internal consistency analysis was performed yielding a Cronbach's alpha of 0.76, indicating acceptable reliability.

4.2. Student Logs and Screen Recordings

To track learners' engagement with the corpus tool during the DDL activities, combining screen recordings and learner written reflection logs were employed. The two sources of data enabled

the collection of both operational data (recordings of what learners *did*) and perceptual data (what learners *thought about* what they did) during the completion of the tasks.

From the whole group of 35 participants in the *AntConc* condition, 20 students were randomly selected for detailed 'tracking' using *Camtasia* and *OBS Studio* screen recording software. To ensure transparency and randomness in selection, all students were assigned an anonymous numeric identifier (e.g., S01 to S35), and 20 unique numbers were selected using a random number generator (via random.org) before the start of the intervention. This sampling strategy ensured representativeness while keeping the volume of video data manageable for systematic analysis.

4.3. Screen Recordings

Screen recordings captured learners' real-time interactions with *AntConc* and other resources during all DDL sessions (90 minutes per session, once per week). This type of process data, as advocated by Carol Chapelle (2003), enabled the researchers to reconstruct the learners' navigation paths, search strategies, and decision-making patterns during the completion of corpus-based activities. The six categories presented in Table 2 were used to code the recordings, adapted from Pérez-Paredes et al. (2011, 2012) and revised for phraseology-focused tasks, reflecting a framework based on detectable, observable learner actions that correspond to the decisions they make regarding the use of the materials.

Table 2. Screen recordings categories in the present study

Code	Category	Definition
SR1	Query Variety	Range of different search strings identified (variation in key terms, use of wildcards, etc.)
SR2	Search Refinement	Evidence of learners adjusting or rephrasing queries to improve results
SR3	Time on Task	Time spent actively interacting with AntConc (measured in minutes)
SR4	Tool Use	Use of AntConc's filtering or sorting features (e.g., sorting by L1/R1 context), File View, and Show Tags.
SR5	Pattern Documentation	Note-taking behaviours: copying, writing down, or organising extracted patterns
SR6	Recognition of discourse function	Search attempts: backtracking or repetition for discourse function patterns

Two trained coders were asked to independently analyse a subset of the recordings (20%), and inter-rater reliability was calculated using Cohen's kappa ($\kappa = 0.87$), indicating high agreement.

4.4. Written reflection logs

Alongside the screen recordings, the same 20 participants completed reflection logs after each DDL session (see Appendix D for the prompts), in which they summarised their interaction with *AntConc*, reflected on the usefulness of the corpus for understanding target phraseological patterns, evaluated difficulties encountered in searching, interpreting, or applying non-finite structures, and compared this method to traditional grammar learning approaches. The Logs were collected after each session, yielding a total of 100 individual written logs (5 sessions × 20 participants). Responses were analysed using a thematic coding framework, which included the five categories as illustrated in Table 3 below.

Table 3. Reflection log categories

Code	Theme	Sample Indicators
WL1	Learning pathways	Description of how they approached the task (e.g., "I searched for X first…")
WL2	Tool engagement	Attitudes toward using *AntConc* (e.g., "It was helpful/confusing/easy to use")
WL3	Pattern awareness	Comments on the structure and meaning of patterns (e.g., "I noticed that…")
WL4	Strategy reflection	Reflections on what worked or didn't (e.g., use of synonyms, trial and error)
WL5	Challenges	Specific issues with tool use or understanding corpus output

Thematic analysis followed an inductive approach, with emergent patterns identified and refined iteratively. To ensure analytic consistency, a second coder independently reviewed 25% of the logs. Inter-coder reliability reached $\kappa = 0.85$.

The two data sets — screen recordings and written logs — were analysed in parallel and then cross-compared to explore convergence or divergence in learner behaviour and self-reported perceptions. The combination of observational (screen) and introspective (written) data allowed the study to link observed corpus engagement to individual learner performance trends across the pre-test, post-test, and delayed post-test.

4.5. Data Analysis

To answer the first research question regarding learners' possible improvement in the use of non-finite phraseological patterns, a within-subjects repeated measures ANOVA was

conducted on test scores collected by means of three successive tests: pre-test, immediate post-test, and delayed post-test (two months later). This analysis aimed to identify statistically significant gains over time and to evaluate the long-term retention of the targeted patterns. Normality and sphericity assumptions were checked, and post-hoc comparisons with Bonferroni correction were applied to locate differences between time points.

In response to the second research question, which explored how learners interacted with *AntConc* and how their behaviour related to learning outcomes, a mixed-methods approach was adopted. First, quantitative behaviour coding was conducted based on the screen recordings from a randomly selected subgroup of 20 participants. The raw data was categorized under six behaviour types (SR1–SR6) based on search strategies, interaction levels, and task completion. Descriptive and inferential statistics (ANOVA) were then used to examine relationships between behaviour categories and individual test gains. Second, qualitative analysis of written learner logs was performed using thematic coding to explore students' self-reported strategies, challenges, and reflections on corpus use. These logs were also triangulated with screen data to detect convergence or divergence between stated and observed behaviours. This combined analysis allowed for a deeper understanding of how corpus search behaviours influenced language learning outcomes and provided interpretive insights into the variation in learner engagement while performing DDL activities.

5. Results

The description of the results obtained from the various sources of data are organized in three main sections. The first section reports on learners' performance across the pre-test, immediate post-test, and delayed post-test, providing evidence for their progress in mastering the targeted non-finite phraseological patterns upon completing the *AntConc*-based DDL activities. The second section presents findings from the screen recordings, examining how learners interacted with the corpus tool, identifying distinct behavioural profiles, and exploring how these behaviours correlate with individual learning outcomes. The final section provides an analysis of written reflection logs and delves into learners' self-perceived strategies and experiences.

5.1. Learner Performance on Phraseological Pattern Tests

A repeated measures ANOVA was conducted to examine differences in learners' performance across three time points; the pre-test, post-test, and delayed post-test. The results indicated a statistically significant effect of time on test scores, $F_{(2, 68)} = 231.26$, $p < .001$, with a large

effect size (partial η² = .872), suggesting substantial improvement over time. Descriptive statistics showed that mean scores increased from M = 8.94 (SD = 2.08) on the pre-test to M = 12.83 (SD = 2.08) on the post-test and remained stable on the delayed test (M = 12.90, SD = 1.89) (see Figure 3).
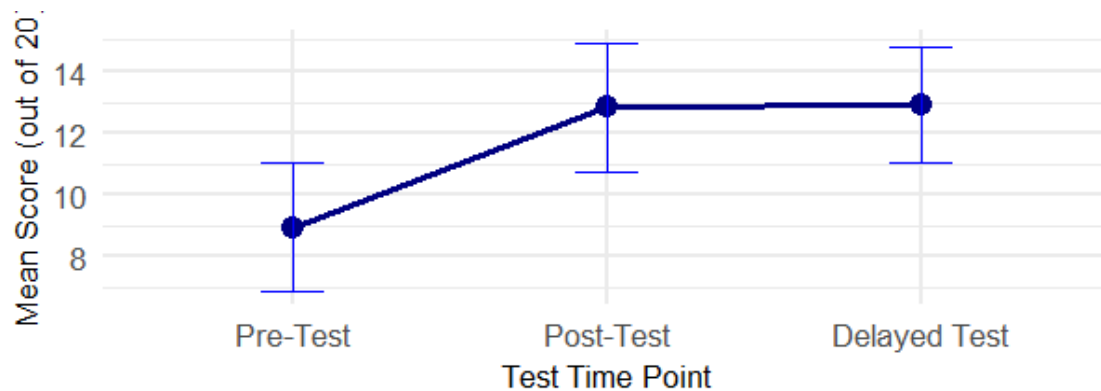
Figure 3. Learners' test performance across test time points

Post-hoc pairwise comparisons with Bonferroni adjustment revealed that both the pre-test to post-test (mean difference = 3.89, p < .001) and pre-test to delayed test (mean difference = 3.96, p < .001) improvements were statistically significant. However, the difference between the post-test and delayed test was not significant (mean difference = -0.07, p = 1.00), indicating that the learning gains were largely retained over time. These results suggest that the *AntConc*-based DDL intervention had a significant and sustained positive impact on the learners' ability to use non-finite phraseological patterns.

5.2.   Corpus Interaction Behaviours and their Relationship to Learning Outcomes

A Pearson correlation analysis was conducted to explore the relationship between learners' behavioural engagement (as coded from screen recording logs) and their performance outcomes. Results in Table 4 revealed a significant positive correlation between total behavioural engagement score and gain scores, r(18) = .467, p = .038, indicating that learners who interacted more actively with the *AntConc*-based tasks showed greater improvement in their use of non-finite phraseological patterns.

Among the individual behaviour categories, Query Variety (SR1) had a strong and significant correlation with both total engagement (r = .850, p < .001) and gain scores (r = .663, p = .001). Other observable behaviours strongly correlated with overall engagement included Search Refinement (SR2) (r = .830, p < .001), Pattern Documentation (SR5) (r = .778, p < .001), and Tool Use (SR4) (r = .757, p < .001) (see Table 4).

Table 4. Correlation results

Variable	Gain score	Retention score	Total behavioural score
SR1 – Query Variety	**.663** (**)	–.280	**.850** (**)
SR2 – Search Refinement	.340	–.043	**.830** (**)
SR3 – Time on Task	.345	–.054	**.729** (**)
SR4 – Tool Use	.311	–.258	**.757** (**)
SR5 – Pattern Documentation	.352	–.090	**.778** (**)
SR6 – Discourse Function	.059	.188	**.665** (*)
Total observable behavioural Score	**.467** (*)	–.133	—

Interestingly, no significant correlations were found between observable behavioural scores and retention scores, suggesting that while behavioural engagement predicts short-term learning gains, it may not be as strongly associated with long-term retention of the patterns. These patterns are visually summarised in Figure 4, which illustrates the strength and direction of correlations between each observable behaviour and the two performance outcomes. The blue bars in Figure 4 represent Pearson correlation coefficients between each behaviour and gain scores, whereas the grey bars represent correlations with retention scores. A dashed horizontal line indicates zero correlation, which distinguishes positive from negative associations.
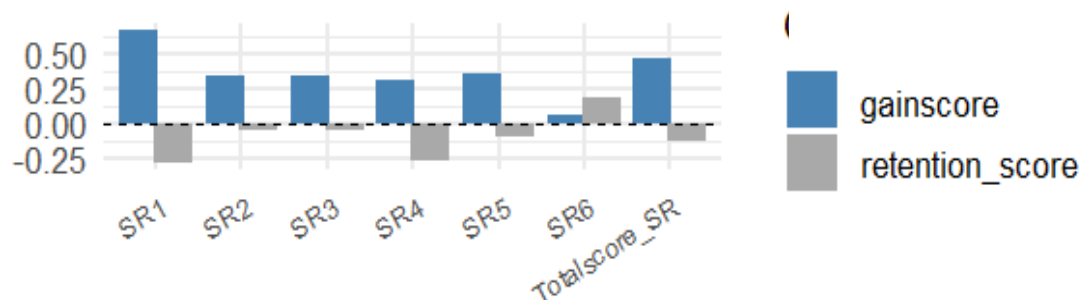
Figure 4. Correlations between screen recording categories and performance outcomes

As illustrated, Query Variety (SR1) stands out with the strongest positive correlation with gain score (r = .663), suggesting that learners who explored a wider range of corpus queries tended to benefit more from the intervention. Other behaviours such as Search Refinement (SR2), Time on Task (SR3), and Pattern Documentation (SR5) also show moderate positive correlation with gain scores, albeit not statistically significant. In contrast, the correlations with retention scores

were generally weaker and, in some cases, negative, indicating that immediate learning gains were more strongly associated with behavioural engagement than long-term retention.

Overall, Figure 4 above supports the conclusion that certain types of behavioural engagement — particularly query variety and strategic documentation of patterns — contributed more substantially to learners' immediate performance improvements. The lack of strong associations with retention suggests that other factors may influence long-term uptake, warranting further investigation.

5.3. Written Reflection Logs

To complement performance results and observable user behaviour, the written logs of participants were analysed to better characterise their experience with corpus-based activities. Thematic coding of 20 written logs provided evidence of five recurring themes: learning pathways, tool interaction, pattern interaction, strategy reflection, and challenges. These qualitative data were triangulated with screen recording data to ascertain areas of convergence and divergence, if at all, between learners' self-reported strategies and actual performance during the *AntConc* activities.

The participants described varied pathways for navigating the tasks. For example, in one task with adverbs followed by past participles (e.g., *widely used, broadly defined*), the participants were led to identify corresponding patterns in the corpus. Although 15 participants identified the complexity of this structure in their logs, video recordings indicated that they employed Wildcard searches (e.g., **ly *en; *ly *ed*) to discover suitable sequences, continually modifying search queries and scrolling through concordance lines. This activity signals exploratory interaction and is supported by the reported high correlation between query variety and learning success. As another participant explained: "*I kept switching the verb to achieve more combinations with various adverbs*" (P11). Nonetheless, as can be inferred from the close examination of the videos, query variation did not always reflect an understanding of patterns, with five participants simply altered surface features without inferring the underlying functional or structural principles.

Another group of five participants (P03, P08, P14, P16, P19) had a more target-oriented approach. They initially queried adverbs in general (e.g., **ly*) before targeting verb participles. They also spent far longer finishing the exercise. For example, P08 used instances such as *significantly *ed, when *ed, and commonly *ed*, which reflect active pattern targeting. This type

of learners brought their written comments into alignment with screen-checked behaviour, stressing overlap between reported learning paths and actual use.

Although statistical correlation between test results and tool usage was insignificant, 11 participants had higher interaction with *AntConc*'s advanced features. Their logs recorded activities such as full context analysis through the File View and the selection of show tags in global settings. Screen recordings confirmed these activities. For instance, P13 said: "*I looked at how the phrase showed [emerged] with the whole code [tags] in the whole text and watched out for the context telling me whether I used it in writing*". This is in line with observed interaction patterns—clicking on concordance lines to see file view, scrolling over broader textual contexts and clicking on show tags to check for the discourse functions. These students appeared to employ corpus functionality to acquire an appreciation of usage, but not everyone achieved matching test score gains.

Looking at their interaction with the target non-finite phraseological patterns, some learners produced rich comments on their perception of structure and function. Participant 3 (P03) said: "*I can use lots of patterns — I recall attitude, discourse organizers, and referential expressions.*" P04 commented on genre variation: "*I recall the JA corpus where non-finites occur more frequently… especially with adjectives plus to-infinitive, and also "enough to-infinitive.*" Other participants also commented on how they learned to construct longer and more complex search strings using wildcards — a technique that was unfamiliar to them prior to the intervention. As Participant 6 noted, *"I didn't know I could use stars to search for full patterns. It helped me to see the structure more clearly and try my own versions."* This ability to formulate extended pattern searches was corroborated by the screen recordings, which revealed that several learners attempted to query more advanced sequences such as *"* it * more likely to *"* and "*this * * us to *"*. These longer strings indicate an increased tool familiarity and a developing awareness of the phraseological nature of language, particularly as learners tried to adapt core structures to retrieve broader and more varied examples from the corpus. These observations confirm that not all the learners were able to capture both the structures and their phraseological patterns. This is in line with the earlier correlations between test results and pattern documentation (SR5) in Section 5.2.

A recurring theme throughout the logs was the perception that corpus use may offer advantages over learning grammar in the traditional way. The participants referred to it as "*motivating*," "*more creative*," and providing access to "*more examples*." The following comment from P04

was especially evocative: "*I can gather texts and use my own corpus…it helps a lot and it makes your vision clearer than the traditional method everything is on the screen*." Another participant (P13) added: "*I learned more with AntConc because there are more examples. I can play around in my structures when I write, and I can look at the codes to check my understanding*." These comments, as the learners put them, suggest a sense of ownership of learning and openness to data-driven discovery, although such metacognitive awareness did not necessarily translate into high performance outcomes, as explained in Section 5.2.

Despite favourable attitudes, the logs contained expressions of the learners experiencing some difficulties with the DDL method. Five participants (P01, P05, P10, P12, P20) struggled with identifying and comparing patterns within sub-corpora. They also mentioned uncertainty about how to assign discourse functions, particularly when patterns could plausibly fit more than one category. This was reflected in the screen recordings, where Students 1, 5, 10, and 20 did not activate the "Show Tags" option in *AntConc* to verify their answers regarding the discourse function of the patterns. It is likely that this omission limited their ability to confirm the accuracy of their categorisations, which may have contributed to lower performance on test items requiring precise identification of discourse functions.

Additionally, screen captures confirm other difficulties. For instance, Participants frequently typed incorrect search terms or misused wildcards, yielding zero hits or unrelated output. In these cases, learners engaged in "frenzied" activity, clicking randomly across features (e.g., P05 during session 3), or waited patiently for significant amounts of time with minimal activity (e.g., P01 and P12 during session 4). These findings reveal a disconnect between some learners' self-reporting of confidence and their proven expertise in corpus tools.

The written reflections provide rich insight into students' active engagement with corpus-based writing instruction. In the majority of cases, reflections underlined inferred behaviour (e.g., in question range, context analysis), while in others, they were inconsistent, indicating a gap between intention and enactment. Together, the logs and recordings emphasise the merit of coupling learner voice and behavioural evidence in improving measurement of the merits and difficulties of data-driven learning.

6. Discussion

The key insight emerging from this study is that tracking learners' behaviour while performing *AntConc*-based DDL activities provided important insights into their engagement and

challenges. Both screen recordings and written logs contributed to students' learning of non-finite patterns in English as revealed by the test scores, aligning with previous studies which reported similar results confirming the positive effects of DDL (Crosthwaite & Cheung, 2019; Granger et al., 2002; Meunier & Reppen, 2015). Statistical analysis revealed a substantial improvement of approximately 43.5% from pre-test to post-test, with performance sustained at delayed post-test. These findings corroborate earlier research on the long-term benefits of DDL (Elsherbini, 2017; Shivaraju et al., 2017). The gains reported by Elsherbini and Shivaraju — around 40% — suggest that our learners' progress falls within the same range observed in comparable studies.

Screen-recording analysis confirmed that behaviours, namely, Query Variety and Pattern Documentation, were strongly correlated to short-term learning gains. This implies that students who experimented with more diverse search queries and systematically documented patterns tended to internalise the target structures. This is consistent with previous research that active corpus data manipulation aids deeper noticing and retention (Flowerdew, 2012; Yoon & Hirvela, 2004). Surprisingly, no behavioural measure proved to be a good predictor of long-term retention, illustrating that variables other than observable search behaviour— such as follow-up practice or integration into writing — may play a greater role in sustaining knowledge over the long term (Boulton, 2012).

The triangulation of written reflection logs with user-tracking recordings uncovered convergence and divergence between reported user strategies and observed tool use. More than 80% of the participants accurately reported use of wildcards and show tags option. This was substantiated by screen output showing advanced query construction. About one-third of the learners reported overuse or understanding of discourse functions, e.g., five students failed to click on the "Show Tags" button in order to verify functional categorisation but reported checking functions. Such discrepancies are also reflected in learner strategy research, where learners' perceptions and actual behaviour do not match (Godwin-Jones, 2017; Lai & Chen, 2015).

This misalignment underlines one of the important implications for corpus-based pedagogy: self-reported logs cannot be considered reliable sources of information for capturing the complexity of DDL processes. While reflective accounts are indeed informative about the intentions and interpretations of learners, they do not reliably pinpoint the procedural steps that learners follow while interacting with corpus tools and materials. It is thus necessary to resort

to objective means of 'tracking' which could yield screen recordings or automated logs, which can reveal what learners actually do, identify the strategies they use, and detect unnoticed gaps in their use of the tools. In light of the information in "operational data", the teacher can subsequently design effective scaffolding and introduce technical control that ensure students complete each required move before proceeding. For instance, in this study, the fact that the use of the "Show Tags" feature was overlooked by five learners might mean that learners misunderstood the teacher's intention or lesson objective. They may be under the impression that they are engaging with discourse-functional information when they are not. The possibility of such mismatch to occur has implications for teacher practice: How he/she ought to introduce, model, and monitor specific corpus functions to line them up for potential learning situations.

The findings underscore the importance of teacher scaffolding not only in how to use the corpus effectively but also in how to interpret the functions of the patterns they find. While learners quickly developed technical skills (e.g., wildcard searches), they needed more explicit guidance in interpreting discourse functions and using them in writing tasks. For example, the activity in Appendix C (Activity 1.4) requires students to analyse the discourse functions of two different non-finite constructions—*Adj + to-infinitive* and *too + Adj + to-infinitive*. Students classify adjectives in the first pattern according to epistemic or attitudinal stance and compare their distribution across corpora and then examine instances of the second pattern to determine whether the construction expresses impossibility, unwillingness, or evaluation. Therefore, embedding reflective prompts (e.g., Appendix D, Question 3: *"How did you recognise their function in context?"*) alongside real-time tool feedback could help bridge the gap between exploratory search behaviour and accurate functional understanding (Charles, 2014; Frankenberg-Garcia, 2014).

These findings support a two-track pedagogical approach in which minimal class time is devoted to procedural tool mechanics, that learners grasp rapidly, and greater emphasis is placed on analysis-driven tasks that demand deeper engagement with annotated texts. In particular, dissecting the functional meaning of non-finite structures encourages learners to think beyond structural patterns and attend to both macro- and micro-level features of disciplinary discourse (Cotos et al., 2017), thereby helping business students align their academic writing with disciplinary conventions (Hyland, 2005).

The present study was conducted with a relatively small number (specialist cohort) of Tunisian master's students, which puts limits on generalisability. Additionally, while behavioural

engagement predicted short-term gains, continued inquiry is required to investigate the interplay between corpus use, follow-up writing practice, and longer-term retention. Analysis of writing quality should be included in future research in a bid to determine whether test performance gains are mirrored in natural academic writing production.

The results obtained from the varied data instruments used in this study present a convergent view of the effect of the *AntConc*-based DDL intervention on the learning of non-finite phraseological patterns. The quantitative data provide evidence that in-depth corpus work can produce measurable short-term gain. Conversely, the behavioural analysis brings to focus specific strategies — such as constant use of different questions and systematic logging of patterns — that appear to be the basis for such benefits. Parallel to these findings, qualitative introspection underscores persistent challenges in interpreting discourse function and intermittent divergence between perceived and actual engagement. These convergent and divergent results suggest that while technical mastery of corpus tools can be gained quickly, developing an understanding of the functional behaviour of phraseological patterns requires sustained practice, targeted feedback, and systematic reflection on the meanings and functions of phraseological patterns.

7. Conclusion

This study examined the integration of *AntConc*-mediated DDL tasks into a BE writing course to enhance learners' recognition, interpretation, and use of non-finite clause phraseology. By combining screen recordings, written logs, and pre-/post-/delayed tests, it showed that such pedagogy leads to measurable gains and sustained awareness. The study contributes to DDL research by providing a replicable methodology for tracing corpus use and linking learner–tool interaction with writing development, offering both researchers and teachers a model for designing scaffolded, data-driven instruction.

Through its focus on one group of twenty learners while completing an *AntConc*-based instructional course over a period of five weeks, this research moves beyond outcome-oriented research designs adopting measures which pinpoint the patterns of tool use. Supplementing screen recording data (operational data) with texts from reflection logs, offered a more detailed view of learner engagement, revealing instances of effective strategy use and pointing to caveats necessitating additional scaffolding.

The findings in this study point to the importance of introducing learners to corpus tools and teaching them strategies for corpus data interpretation in order to achieve meaningful integration of DDL within academic writing courses. In practice, this calls for designing DDL-based instruction that combines technical training in the use of corpus tools with explicit guidance for interpreting functions and using patterns appropriately within specific genres. As digital learning environments become more prevalent, the methodological approach adopted in this study — combining behavioural observation and reflection logs — can serve as guidepost for subsequent work by other teachers-researchers who may consider implementing this pedagogy and "track" their learners in order to capture the patterns of the learners' interaction with the tools and the materials and find out for themselves the impact of this strategy on learners' writing development.

Future research could build on this approach through cumulative in-depth studies across different levels or modules, allowing findings from multiple small-scale investigations to corroborate one another and gradually strengthen the evidence base. Researchers may also wish to compare different corpus tools to explore how interface design shapes learner behaviour and outcomes.

Disclosure Statement

The authors declare that there is no conflict of interest regarding the publication of this article. No financial, personal, or professional relationships have influenced the research, analysis, or conclusions presented in this work.

Notes on Contributors

Olfa Ben Amor is an assistant professor at the Faculty of Letters and Human Sciences of Sousse (University of Sousse, Tunisia). She holds an MA in English Didactics from the Higher Institute of Education and Training in Tunis and a Ph.D. in Applied Linguistics from the University of Manouba, Tunisia. Her research interests focus on corpus linguistics, the use of corpora for language learning (known as 'data-driven learning'), phraseology, ESP, EAP, and integrating technology in English classrooms. She is currently a board member of Tunisia TESOL and a member of the Laboratory for Approaches to Discourse (LAD).

BenAmor.Olfa@flsh.u-sousse.tn

Faiza Derbel is an Associate Professor of English and Applied Linguistics at the Faculty of Letters, Arts and Humanities, Manouba, Tunis. She holds a Ph.D. from the Institute of Education, London University and an MA in Applied Linguistics/TESL/CALL from Iowa State University, USA. Her research interests include ELT, Educational Technology, Computer-assisted Language Learning (CALL), and Teacher Cognition. She is currently member of the University of Manouba Scientific Council, the Scientific Council of the Laboratory for Approaches to Discourse (LAD), and a founding member of Tunisia TESOL and its Secretary General (2023-2027).

Faiza.derbal@flah.uma.tn

ORCID

Olfa Ben Amor ⓘ *https://orcid.org/0000-0001-7648-4513*

Faiza Derbel ⓘ *https://orcid.org/0000-0003-3755-6283*

References

Ben Amor, O. (2025). Exploring Phraseological Patterns in Business English Non-Finite Clauses. *Journal of the European Second Language Association*, 9(1), 168–186. DOI: https://doi.org/10.22599/jesla.135

Ben Amor, O., & Derbel, F. (2020). The identification of English non-finite structures using NooJ platform. *International Conference on Automatic Processing of Natural-Language Electronic Texts with NooJ*, 27–38. https://doi.org/10.1007/978-3-030-38833-1_3

Biber, D., Csomay, E., Jones, J. K., & Keck, C. (2004). A corpus linguistic investigation of vocabulary-based discourse units in university registers. In U. Connor, & T. A. Upton, (Eds.). *Applied Corpus Linguistics* (pp. 53–72). Brill Rodopi.

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman grammar of spoken and written English*. Longman London.

Bouabida, N. (2020). The Effect of data-driven learning & image-schema-based polysemy networks instruction on learning spatial prepositions of verticality: The case of

Moroccan EFL learners. *Jurnal Arbitrer*, 7(1), 62-70. https://doi.org/10.25077/ar.7.1.62-70.2020

Boulton, A. (2010). Data-driven learning: Taking the computer out of the equation. *Language Learning*, *60*(3), 534–572. https://doi.org/10.1111/j.1467-9922.2010.00566.x

Boulton, A. (2012). Beyond concordancing: Multiple affordances of corpora in university language degrees. *Procedia-Social and Behavioral Sciences*, *34*, 33–38. https://doi.org/10.1016/j.sbspro.2012.02.008

Boulton, A. (2017). Corpora in language teaching and learning. *Language Teaching*, *50*(4), 483–506. https://doi.org/10.1017/S0261444817000167

Boulton, A., & Vyatkina, N. (2021). Thirty years of data-driven learning: Taking stock and charting new directions over time. *Language Learning & Technology, 25(3), 66–89.* *https://doi.org/10.64152/10125/73450*

Boulton, A., & Vyatkina, N. (2024). Expanding methodological approaches in DDL. *TESOL Quarterly*, *58*(3), 1193–1204. https://doi.org/10.1002/tesq.3269

Braun, S. (2005). From pedagogically relevant corpora to authentic language learning contents. *ReCALL*, *17*(1), 47–64. https://doi.org/10.1017/S0958344005000510

Bybee, J. (2010). *Language, usage and cognition*. Cambridge University Press. https://doi.org/10.1017/CBO9780511750526

Chaalal, I. (2024). The Effectiveness of using data-driven learning activities in teaching English near-synonyms for Algerian EFL learners: A pilot study. مجلة الرسالة للدراسات والبحوث الإنسانية, 9(4), 244-268 https://asjp.cerist.dz/en/article/258753

Chambers, A. (2007). Popularising corpus consultation by language learners and teachers. In A.Cresswell, E.Hidalgo, L.Quereda, & J.Santana (Eds.). *Corpora in the foreign language classroom* (pp. 1–16). Brill. https://doi.org/10.1163/9789401203906_002

Chambers, A., & O'sullivan, I. (2004). Corpus consultation and advanced learners' writing skills in French. *ReCALL*, *16*(1), 158–172. https://doi.org/10.1017/S0958344004001211

Chapelle, C. (2003). *English language learning and technology*. John Benjamins. https://doi.org/10.1075/lllt.7

Charles, M. (2006). Phraseological patterns in reporting clauses used in citation: A corpus-based study of theses in two disciplines. *English for Specific Purposes*, *25*(3), 310–331. https://doi.org/10.1016/j.esp.2005.05.003

Charles, M. (2014). Getting the corpus habit: EAP students' long-term use of personal corpora. *English for Specific Purposes*, *35*, 30–40. https://doi.org/10.1016/j.esp.2013.11.004

Charles, M., & Hadley, G. (2022). Autonomous corpus use by graduate students: A long-term trend study (2009–2017). *Journal of English for Academic Purposes*, *56*, 101095. https://doi.org/10.1016/j.jeap.2022.101095

Cohen, A. D. (2013). Verbal report. In *C. A.* Chapelle (Ed.)*, The encyclopedia of applied linguistics* (Vol. 5). Wiley-Blackwell Oxford. https://doi.org/10.1002/9781405198431.wbeal1261

Cotos, E. (2014). Enhancing writing pedagogy with Learner Corpus Data. *ReCALL*, *26*(2), 202-224. https://doi.org/10.1017/S0958344014000019

Cotos, E., Link, S., & Huffman, S. (2017). Effects of DDL technology on genre learning. Language Learning & Technology, 21(3), 104–130. https://doi.org/10.64152/10125/44623

Cresswell, A. (2007). Getting to 'know'connectors? Evaluating data-driven learning in a writing skills course. In A. Cresswell, E. Hidalgo, L. Quereda & J. Santana (Eds.)*. Corpora in the foreign language classroom,* (pp. 267–287). Brill. https://doi.org/10.1163/9789401203906_018

Crosthwaite, P., & Cheung, L. (2019). *Learning the language of dentistry: Disciplinary corpora in the teaching of English for specific academic purposes* (Vol. 93). John Benjamins Publishing Company. https://doi.org/10.1075/scl.93

Crosthwaite, P., Wong, L. L., & Cheung, J. (2019). Characterising postgraduate students' corpus query and usage patterns for disciplinary data-driven learning. *ReCALL*, *31*(3), 255–275. https://doi.org/10.1017/S0958344019000077

Elsherbini, S. A. H., & Ali, A. D. (2017). The Effects of corpus-based activities on EFLuniversity students' grammar and vocabulary and their attitudes toward corpus. *Journal of Research in Curriculum Instruction and Educational Technology*, 3(1), 133-161. https://doi.org/10.21608/jrciet.2017.24482

Flowerdew, L. (2009). Applying corpus linguistics to pedagogy: A critical evaluation. *International Journal of Corpus Linguistics*, *14*(3), 393–417. https://doi.org/10.1075/ijcl.14.3.05flo

Flowerdew, L. (2012). Exploiting a corpus of business letters from a phraseological, functional perspective. *ReCALL: The Journal of EUROCALL*, *24*(2), 152-168. https://doi.org/10.1017/S0958344012000043

Frankenberg-Garcia, A. (2014). The use of corpus examples for language comprehension and production. *ReCALL*, *26*(2), 128–146. https://doi.org/10.1017/S0958344014000093

Gabrielatos, C. (2005). Corpora and language teaching: Just a fling or wedding bells?. *Teaching English as a Second Language—Electronic Journal*, *8*(4), 1–35. http://www.tesl-ej.org/ej32/a1.html

Gaskell, D., & Cobb, T. (2004). Can learners use concordance feedback for writing errors? *System*, *32*(3), 301–319. https://doi.org/10.1016/j.system.2004.04.001

Gilquin, G. (2022). The process corpus of English in education: Going beyond the written text. *Research in Corpus Linguistics*, *10*(1), 31–44. https://doi.org/10.32714/ricl.10.01.02

Gilquin, G., & Granger, S. (2010). How can data-driven learning be used in language teaching? (or not). In A. O'Keeffe & M. McCarthy (Eds.). *The Routledge Handbook of Corpus Linguistics*, *359-370*. https://doi.org/10.4324/9780203856949

Godwin-Jones, R. (2017). Data-informed language learning. *Language Learning & Technology*, 21(3), 9–27. https://doi.org/10.64152/10125/44629

Granger, S., Hung, J., & Petch-Tyson, S. (2002). *Computer learner corpora, second language acquisition, and foreign language teaching* (Vol. 6). John Benjamins Publishing. https://doi.org/10.1075/lllt.6

Green, C. (2017). Discourse coherence and clause combination. In C. Green (Ed.), *Patterns and Development in the English Clause System* (pp. 149–167). Springer. https://doi.org/10.1007/978-981-10-2881-6_6

Hafner, C. A., & Candlin, C. N. (2007). Corpus tools as an affordance to learning in professional legal education. *Journal of English for Academic Purposes*, *6*(4), 303–318. https://doi.org/10.1016/j.jeap.2007.09.005

Huang, X. (2022). Relationship between learning behavior and learning effect based on Big Data. In M. Atiquzzaman, N. Yen, & Z. Xu (Eds.), *2021 International Conference on Big Data Analytics for Cyber-Physical System in Smart City* (Vol. 102, pp. 371–379). Springer. https://doi.org/10.1007/978-981-16-7466-2_41

Hunston, S., & Francis, G. (2000). *Pattern grammar: A corpus-driven approach to the lexical grammar of English*. John Benjamins Publishing. https://doi.org/10.1075/scl.9.11hun

Hyland, K. (2005). *Metadiscourse: Exploring interaction in writing*. Continuum.

Johns, T. (1991). Should you be persuaded: Two samples of data-driven learning materials. In Johns, T. & King, P. (Eds.). *Classroom concordancing: English Language Research Journal* (Vol. 4, pp. 1-16.). Centre for English Language Studies, University of Birmingham.

Lai, S.-L., & Chen, H.-J. H. (2015). Dictionaries vs concordancers: Actual practice of the two different tools in EFL writing. *Computer Assisted Language Learning*, *28*(4), 341–363. https://doi.org/10.1080/09588221.2013.839567

Li, D., Noordin, N., Ismail, L., & Cao, D. (2025). A systematic review of corpus-based instruction in the EFL classroom. *Heliyon*, *11*(2). https://doi.org/10.1016/j.heliyon.2025.e42016

Liu, D., & Jiang, P. (2009). Using a Corpus-Based lexicogrammatical approach to grammar instruction in EFL and ESL contexts. *The Modern Language Journal*, *93*(1), 61–78. https://doi.org/10.1111/j.1540-4781.2009.00828.x

Lusta, A., Demirel, Ö., & Mohammadzadeh, B. (2023). Language corpus and data driven learning (DDL) in language classrooms: A systematic review. *Heliyon*. https://doi.org/10.1016/j.heliyon.2023.e22731

Ma, B. K. (1994). Learning strategies in ESP classroom concordancing: An initial investigation into data-driven learning. *Entering Texts*, 197–214.

Meunier, F., & Reppen, R. (2015). Corpus versus non-corpus-informed pedagogical materials: Grammar as the focus. In D. Biber & R. Reppen (Eds.). *The Cambridge handbook of English corpus linguistics.* (pp. 498–514). Cambridge University Press. https://doi.org/10.1007/9781139764377.028

O'Keeffe, A. (2021). Data-driven learning–a call for a broader research gaze. *Language Teaching*, *54*(2), 259–272. https://doi.org/10.1017/S0261444820000245

O'Sullivan, Í. (2007). Enhancing a process-oriented approach to literacy and language learning: The role of corpus consultation literacy. *ReCALL*, *19*(3), 269–286. https://doi.org/10.1017/S095834400700033X

Pérez-Paredes, P. (2019). A systematic review of the uses and spread of corpora and data-driven learning in CALL research during 2011–2015. *Computer Assisted Language Learning*, *35*(1–2), 36–61. https://doi.org/10.1080/09588221.2019.1667832

Pérez-Paredes, P., Sánchez-Tornel, M., Alcaraz Calero, J. M., & Jiménez, P. A. (2011). Tracking learners' actual uses of corpora: Guided vs non-guided corpus consultation. *Computer Assisted Language Learning*, *24*(3), 233–253. https://doi.org/10.1080/09588221.2010.539978

Pérez-Paredes, P., Sánchez-Tornel, M., & Calero, J. M. A. (2012). Learners' search patterns during corpus-based focus-on-form activities: A study on hands-on concordancing. *International Journal of Corpus Linguistics*, *17*(4), 482–515. https://doi.org/10.1075/ijcl.17.4.02par

Petrovitz, W. (2001). The sequencing of verbal-complement structures. *ELT Journal*, *55*(2), 172–177. https://doi.org/10.1093/elt/55.2.172

Qiu, X. (2024). Exploring the effect of corpus-based writing instruction on learner-corpus interaction in L2 revision: A study of Chinese EFL disciplinary writers. *TESOL Quarterly*, *58*(3), 1108–1137. https://doi.org/10.1002/tesq.3308

Quirk, R., Greenbaum, S., Leech, G., & Svartvik, J. (1985). *A comprehensive English grammar*. London and New York: Longman.

Schmidt, R. (1993). Awareness and second language acquisition. *Annual Review of Applied Linguistics*, *13*, 206–226. https://doi.org/10.1017/S0267190500002476

Shivaraju, P. T., Manu, G., Vinaya, M., & Savkar, M. K. (2017). Evaluating the effectiveness of pre-and post-test model of learning in a medical school. *National Journal of Physiology, Pharmacy and Pharmacology*, 7(9), 947. https://doi.org/10.5455/njppp.2017.7.0412802052017

Silberztein, M. (2020). NooJ V7.0 [software]. *Formalising Natural Languages with NooJ2020*.

Sun, Y.-C., & Wang, L.-Y. (2003). Concordancers in the EFL classroom: Cognitive approaches and collocation difficulty. *Computer Assisted Language Learning*, *16*(1), 83–94. https://doi.org/10.1076/call.16.1.83.15528

Tall, D., & Razali, M. R. (2006). Diagnosing students' difficulties in learning mathematics. *International Journal of Mathematical Education in Science and Technology*, *24*(2), 209–222. https://doi.org/10.1080/0020739930240206

Yoon, H., & *Hirvela*, A. (2004). ESL student attitudes toward corpus use in L2 writing. *Journal of Second Language Writing*, *13*(4), 257–283. https://doi.org/10.1016/j.jslw.2004.06.002

Appendices

Appendix A: The size of the corpus

Genre	Content	Number of texts	Number of words
Journal articles (JA)	The Journal of Financial Economics The Journal of International Management The Journal of Monetary Economics The Quarterly Journal of Economics	49 texts	597,387
Students' dissertations (TD)	Banking/Finance /Economics/Management	14 MAs 6 PhDs	526,338
		Total N of words = 1,123,725	

Appendix B: Performance tests

Pre-test / Post-test / Delayed Post-test

Total Score: 40 points

Section A: Pattern Recognition and Function (8 marks)

Instructions: Read the following excerpts and underline the non-finite clause constructions. Then identify their discourse function (e.g., epistemic stance, attitudinal stance).

1. It is essential to evaluate all financial risks before launching the project.

2. The software update made it easier to track online transactions.

3. Too many variables make it impossible to isolate the impact of inflation.

4. The report was clear enough to convince all stakeholders.

5. It has become increasingly important to carefully compare interest rates

6. It is reasonable to assume that foreign investment will rise under these conditions.

7. It seems necessary to revisit the initial assumptions behind the economic model.

8. It is encouraging to see that the team managed to reduce operational costs.

For each sentence:

- Underline the non-finite clause pattern.

- Briefly state its function (epistemic stance or attitudinal stance).

Section B: Phraseology Comparison (8 marks)

Instructions: Think about the phraseological patterns used to express stance. Then answer the questions below.

1. Write an example of two adjective + to-infinitive patterns that express epistemic stance (e.g., "it is likely to..."). Can you find any patterns that use stance adjectives with comparative structures (e.g., "more likely to...", "less likely to...")?

2. Write two adjective + to-infinitive patterns that express attitudinal stance (e.g., "it is essential to...").

3. Give one example of a "make it (adj) to-infinitive" pattern and explain what kind of stance it expresses.

4. What are some common modifiers found in the patterns "too (adj) to-infinitive" and "(adj) enough to-infinitive" (e.g., too complex to..., efficient enough to...)?

Section C: Rewriting Task (8 marks)

Instructions: Rewrite the sentences below using the non-finite clause patterns studied (do not change the meaning). Use a variety of structures.

1. Evaluating economic trends is important for investment planning.

2. The complexity of the data makes analysis difficult.

3. The forecast is not accurate enough to be used for decision-making.

4. The number of applicants is increasing; we cannot interview them all.

Section D: Opaque Phraseological Patterns with Non-Finite Clauses (8 marks)

Instruction:
Read the sentences below that contain less transparent, idiomatic phraseological patterns involving non-finite clauses. Then, answer the questions that follow. These expressions often serve evaluative, strategic, or cautionary functions in academic writing.

Examples:

1. *The manager urged the team to keep an eye on reducing unnecessary expenditures.*

2. *This initiative offers a window to rethinking traditional business models.*

3. *The administration must make room for integrating student feedback in the decision-making process.*

4. *The committee faced a call to reconsider the proposed framework.*

Tasks:

1. Identify the non-finite clause in each sentence and underline it.

2. Explain in your own words what the figurative expression means in each sentence (e.g., *keep an eye on*, *make room for*).

3. For each sentence, describe the **discourse function** of the expression:
   (e.g., Does it serve to recommend a strategy? Express caution? Emphasize urgency?)

4. In your opinion, which of these expressions might be more difficult for novice writers to use correctly? Explain why.

Section E: Production Task (8 marks)

Instructions: Write a short paragraph (approximately 70 words) commenting on the following table:

Table 1 Government taxation 2010

	Australia	Brazil	China	France	Germany	India	Japan	Russia	UK	USA
Total tax as % GDP	29.5	32.3	16.4	44.7	40.4	18.9	28.2	33.2	37.7	28.0

Source OECD

Your paragraph should:

- Include at least **three non-finite clause constructions**, such as:
  - *To increase market visibility...* (goal-oriented)
  - *Indicating a shift in government taxation...* (discourse organizer)
  - *It is important to note...* (stance marker)

- Demonstrate a range of **semantic functions** (e.g., stance, goal, result, condition).

- Aim for academic tone and clarity of commentary.

Appendix C: DDL Activities (A sample of a Worksheet)

Worksheet 1: Stance Expressions with adjectives and To-Infinitive Structures (Epistemic & Attitudinal)

Presentation:

Writers use non-finite clauses to convey their stance, either by expressing certainty/uncertainty (epistemic stance) or attitude/opinion (attitudinal stance). This worksheet explores patterns including *It is (adj) to-infinitive*, *make it (adj) to-infinitive*, *too (adj) to-infinitive*; and *(adj) enough to-infinitive.* Expert academic writers often embed such expressions in phraseological patterns that differ in complexity and frequency from those used by novice writers. This worksheet will guide you through querying the corpus to identify differences in frequency, phraseological variation and semantic functions.

1. *Expressing epistemic stance (certainty, probability, possibility)*
   Example 1: *it is **likely** to be viewed as particularly problematic*
   Example 2: *make it **possible** to identify both parameters*
2. *Expressing attitudinal stance (obligation/directive, desire, ability, difficulty, intention…)*
   Example 3: *it is **necessary** to know the measures variables*
   Example 4: *It is **important** to determine the forecasting models parameters*
   Example 5: *it is **too difficult** to measure the fundamental value assets*
   Example 6: *it is still **strong enough** to dominate the accommodating and anchoring effects of imperfect credibility*

Practice: Queries and Questions:

Go to the shortcut *AntConc* on your desktop, double click, AntConc will be launched automatically. Upload the texts from the file entitled JA corpus texts and Dissertations texts stored on your desktop.

1. Search for the phraseological patterns: It is/was * to */ it is/was * to *

   1.1 Record the top 5 adjectives used in each subcorpus (JA and TD).

Adjective	Frequency in JA	Frequency in TD
likely		
important		
necessary		
difficult		
possible		

1.2 Which adjectives are more dominant in expert writing? Which verbs most frequently follow the infinitive? To what extent are these verb patterns transparent or idiomatic?

1.3 Are certain combinations more fixed in JA than in TD (e.g., it is important to consider vs. it is important to note)?

1.4 Which adjectives in the pattern express epistemic stance (e.g., likely, possible as in Examples 1& 2) and which express attitudinal stance (e.g., important, necessary, difficult as in Examples 3, 4, and 5)?
   • Group the adjectives you find under these two categories.
   • Are there adjectives used exclusively in one corpus (JA or TD)?
   • Which stance type (epistemic or attitudinal) is more frequently marked in each corpus?

2. Search for the phraseological pattern: too * to *

   • Identify the adjectives used in this construction.
   • Which adjectives express **attitudinal stance** (e.g., too important, too dangerous) vs. **epistemic stance** (e.g., too uncertain)?
   • Are these constructions used to express **impossibility**, **unwillingness**, or **evaluation**?
   • Compare frequencies in JA and TD corpora. Present the counts in the following table:

Adjective	JA Frequency	TD Frequency	Stance Type
too ? to	?	?	
too ? to	?	?	

3. Search for the phraseological pattern: * enough to *

   • List adjectives found before *enough to-infinitive* (e.g., *strong enough to, clear enough to*).
   • Classify the adjectives by **stance type**: epistemic or attitudinal
   • Are these patterns used to express **ability**, **confidence**, or **appropriateness**?

4. **Compare the use of "too (adj) to" vs. "(adj) enough to" in both corpora.**

   • Which of the two patterns is more prevalent in JA vs. TD?
   • Do novice writers overuse one of the patterns?
   • Are expert writers more precise in their use of stance?

Check for slot variability

   • Do these patterns occur mostly with certain verbs (e.g., *too weak to compete, strong enough to survive*)?
   • Are some adjective + verb pairings fixed or formulaic in either corpus?

Appendix D: Learner Reflection Log

Student Name: _____   **Class #:** _____

Date: _____   **Study Week:** _____

---

1. Describe what you discovered about the target structures and their uses in today's session:

_____

_____

_____

2. Describe the steps you took when using *AntConc*.

- Which functions did you use (e.g., Concordance, File View, Show Tags)?

- Did you try new ways to search (e.g., wildcards, longer search strings)?

- How did you check whether your answers were correct?

Write your response here:

_____

_____

_____

3. Which patterns did you work with most today? (Write examples):

_____

How did you recognise their function in context?

_____

_____

4. What steps or strategies helped you most to complete the activity?

_____

What was different compared to how you usually learn grammar or writing?

_____

_____

5. What was most challenging for you today?

_____

How did you deal with these challenges, or how could the activity be improved?

_____

_____

Notes

Any additional comments about today's lesson:

_____

_____