Raising Genre Awareness through Visualizing Language Features

Abstract

This paper introduces the Feature Visualizer, an open-access AI-powered tool designed to raise genre awareness among novice academic writers through inductive learning, a process that includes approaches such as discovery learning. The tool houses an annotated corpus of scientific research articles written by computer science majors and allows learners to explore authentic texts using on-demand visualizations and multimodal explanations. By engaging with the corpus, learners identify recurring language patterns and rhetorical structures at macro-, meso-, and micro-levels, facilitating the bottom-up discovery of genre conventions. A longitudinal study with Japanese undergraduate computer science majors showed that the tool enhanced learners' awareness of academic writing conventions and genre features. Focus group interviews further confirmed the usability and pedagogical value of the Feature Visualizer. We conclude by discussing practical applications for genre-based writing instruction informed by inductive learning principles.

1. Introduction

Novice academic writers often struggle to meet the linguistic and rhetorical expectations of scientific genres, especially when composing research articles in English (Blake, 2021; Flowerdew, 2015; Swales & Feak, 2012). This challenge is particularly acute for learners in disciplines such as computer science, where familiarity with genre conventions is essential for academic success (Zobel, 2014). Writing in the style expected by a community of practice (Lave & Wenger, 1991; Hyland, 2012) is difficult without sufficient familiarity with the genre, which makes it challenging to maintain conventional features (Bhatia, 1999). While explicit instruction of genre features has its place, inductive learning approaches allow learners to follow their own discovery-driven pathways to internalize these conventions (Reppen, 2016). For students who are self-directed this is potentially more engaging than more traditional deductive learning approaches (Smart, 2014).

The Feature Visualizer was developed as an AI-powered, open-access online tool designed to raise genre awareness through interactive visualization of language features. By allowing users to engage with authentic texts, focus on genre-relevant features, and access further details and examples on demand, the Feature Visualizer aims to foster a more autonomous, inductive learning experience. Our expectation was that learners would read texts that were closest to the type of research article they were about to write. Should learners want to focus on particular linguistic or rhetorical features, they can select the feature from a menu, which automatically visualizes the target feature in the article that is loaded. This enables them to notice genre-specific

patterns that they might not otherwise have discovered. Rather than restrict learners to inductive learning, we also incorporated additional materials in the form of brief and extended textual explanations alongside multimodal explanations, which users can access. The multimodal support operates at two levels: first, through on-demand highlighting and annotation triggered when users select features to focus on; and second, through narrated video slideshow explanations that provide a more comprehensive, guided walkthrough of selected genre features.

Data-driven learning (DDL) approaches that incorporate corpora present multiple challenges to learners which increase the cognitive and logistical burden placed on learners (Boulton & Vyatkina, 2021; Jablonkai & Csomay, 2022; Johns, 2002). These challenges include selecting appropriate corpora, identifying relevant search terms, interpreting partially contextualized results from key-word-in-context searches (Anthony, 2019), and parsing authentic language (Gilquin & Granger, 2022). These tasks, while pedagogically valuable, may also overwhelm novices and distract from higher-level genre analysis. The Feature Visualizer alleviates this burden by pre-curating a focused corpus and embedding automated, feature-specific searches. This removes the need for learners to manually locate or define salient patterns; instead, they can simply choose an article and activate pre-selected visualizations of key genre features. In doing so, the tool retains the exploratory spirit of DDL while eliminating much of the associated drudgery, enabling learners to concentrate on interpretation and application.

The remainder of this article is structured to demonstrate how the Feature Visualizer both draws on and contributes to current understandings of genre pedagogy. Section 2 outlines the theoretical framework, situating the tool within traditions of inductive learning, noticing, and corpus-based pedagogy while connecting it to prior visualization projects. Section 3 introduces the design and functionality of the Feature Visualizer, illustrating how its interface and processes make genre features visible to learners. Section 4 outlines the longitudinal mixed-methods study conducted with Japanese computer science majors, providing the empirical basis for evaluating the tool. Section 5 presents the findings, combining quantitative evidence of learning gains with qualitative insights into how learners engaged with the tool. The final section offers conclusions, addresses limitations, and discusses implications for future research and pedagogy, highlighting the broader significance of visualization in supporting inductive approaches to genre-based writing instruction.

2. Theoretical Framework

AI-driven visualization tools have been developed to help learners of English focus on

specific linguistic or rhetorical features. For example, web application to identify and visualize aspects of information structure, such as information flow, end focus, and end-weight was developed to help advanced learners of English (Blake et al., 2023). Similarly, TrendScribe, an AI-powered application that generates descriptions of time-series data at different levels of language proficiency, offering learners of English graded exemplar texts based on numerical inputs (Blake et al.,2025). These projects illustrate how visualization and AI can be combined to scaffold language learning, providing a backdrop for the development of the Feature Visualizer.

The Feature Visualizer is grounded in the principles of inductive learning, where learners are guided to identify patterns and derive rules by engaging directly with input (Ellis, 2021; Prince & Felder, 2006). Unlike deductive approaches, which begin with explicit explanation, inductive methods encourage exploration, pattern recognition, and hypothesis formation (Prince & Felder, 2007). Corpus-based pedagogy (Li et al., 2025) aligns well with an inductive approach, offering learners access to authentic language data for self-guided analysis.

The integration of AI further enhances this experience by enabling real-time interaction with text and dynamic visualization of linguistic features. For example, by drawing on machine learning libraries for named entity recognition, we can identify repeated references to the same entities. This facilitates the detection of anaphoric references and helps make patterns of cohesion more visible to learners, supporting their understanding of how ideas are connected within and across paragraphs.

A key mechanism underlying inductive learning is noticing, viz. the process by which learners consciously register language features in the input. According to the noticing hypothesis (Schmidt, 1990), conscious attention to linguistic forms is a necessary condition for language acquisition. Empirical studies have supported the role of noticing in the development of second language linguistic competence (Ekanayaka & Ellis, 2020, 2021; Ishikawa & Révész, 2020). The Feature Visualizer is designed to foster this noticing process by making genre-specific language features more salient through visual cues, which are intended to not only focus attention on salient patterns, but pique the interest and curiosity of learners, encouraging them to interact with the content housed in the Feature Visualizer.

In the context of genre awareness, inductive learning through corpus exploration supports learners in identifying structural, lexical, and rhetorical norms at different levels of textual organization. Visualizing language patterns at different scales, namely: macro (e.g., move structures), meso (e.g., sentence types), and micro (e.g., verb tenses), may help learners internalize how scientific texts are constructed. Colour coding, pithy comments, toggled explanations, and multimodal input serve as a bridge between raw textual data and learner interpretation, reinforcing discovery and

noticing as central mechanisms in the learning process. Learners can rely on their own inferences in a purely inductive approach, or may choose to move along the cline towards deductive approaches by opting to access increasingly more detailed descriptions and explanations.

Although previous studies have demonstrated the feasibility of corpus-based learning and visualization tools for supporting noticing and inductive learning (e.g., Blake et al, 2023; Blake et al, 2025), relatively little is known about how AI-driven visualization of corpora affects learner awareness of genre conventions. This study addresses this gap by examining how interaction with the Feature Visualizer influences students' genre awareness, asking specifically: *How does interaction with the Feature Visualizer impact the genre awareness of students?*

3. The Feature Visualizer

The Feature Visualizer is a web-based application built to support inductive genre learning through interaction with a curated corpus of scientific research articles in computer science. This tool shows features in their full textual context, which contrasts with typical DDL explorations that rely on keyword-in-context (KWIC) searches in learner-friendly corpus tools such as CorpusMate (Crosthwaite & Baisa, 2024) and Sketch Engine for Language Learners (SKELL; Kilgarriff et al., 2015).

The Feature Visualizer employs a dual-layered analytical system combining machine learning and rule-based string matching to identify discourse features at macro-, meso-, and micro-levels. Table 1 shows the discourse features that can currently be visualized within the tool.

Table 1: *Features incorporated in the Feature Visualizer*

Level	Feature	Details
Macro	Sections	abstract, introduction, method, etc.
	Moves	related works, importance, novelty
Meso	Functions	referring to figures, tables and equations
Micro	Connections	coherence and cohesion within paragraphs
	Linking	using prepositions, conjunctions and adverbs
	Tense	present simple, past perfect, etc.
	Passive voice	passive voice verb phrases
	Modality	hedges, boosters and approximation

Users begin by selecting an article from a prepared pre-loaded corpus of 12 research articles written by students and categorized by methodological focus (empirical, applied, experimental, theoretical). Figure 1 shows the user interface for each research article, which is split into three main areas. The collapsible menu bar on the left enables users to select rhetorical and language features to visualize. The

central area houses the selected research article. Textual explanations and embedded video explanations are displayed on the right.

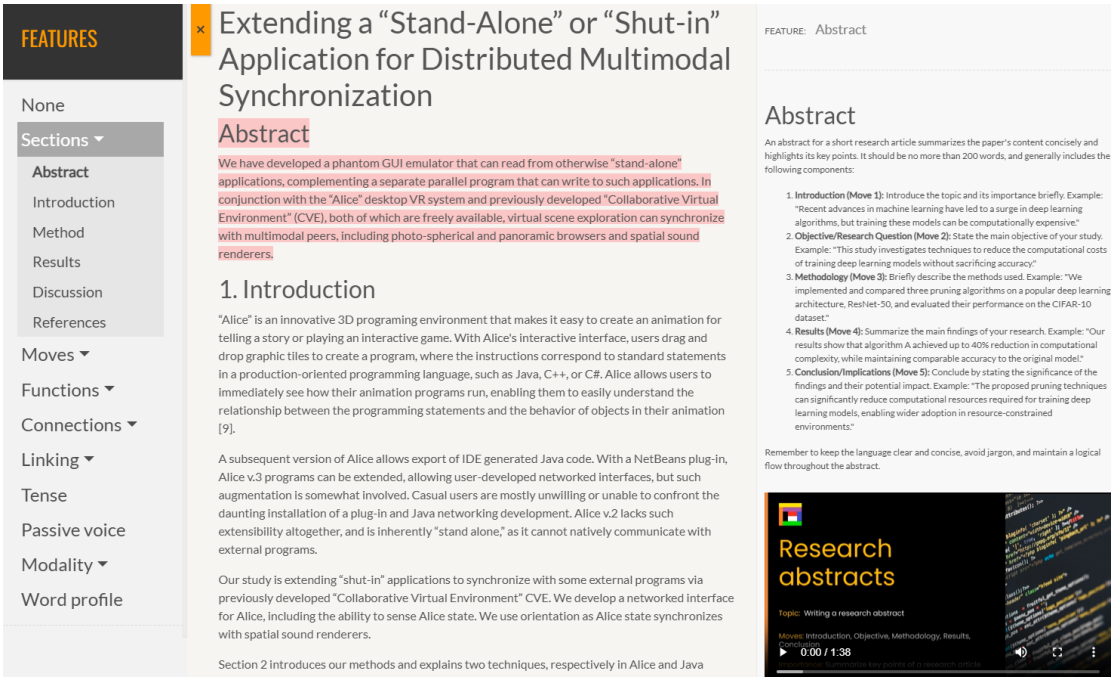

Figure 1: *Screenshot of the Feature Visualizer with the abstract feature selected.*

Once loaded, the article is presented in its original form. Toggle buttons in the menu allow users to reveal or hide visualizations. These include colour-coded indicators of structural elements such as moves and steps, highlighting of cohesive devices, verb patterns, academic phrases, and grammatical constructions characteristic of the genre. Each highlighted feature is linked to textual and multimodal explanations that can be accessed on demand. Figure 2 shows a research article with the tense feature selected. Each finite verb phrase is highlighted and the name of the tense is automatically identified and displayed before the verb. Past, present and future forms are assigned different colours.

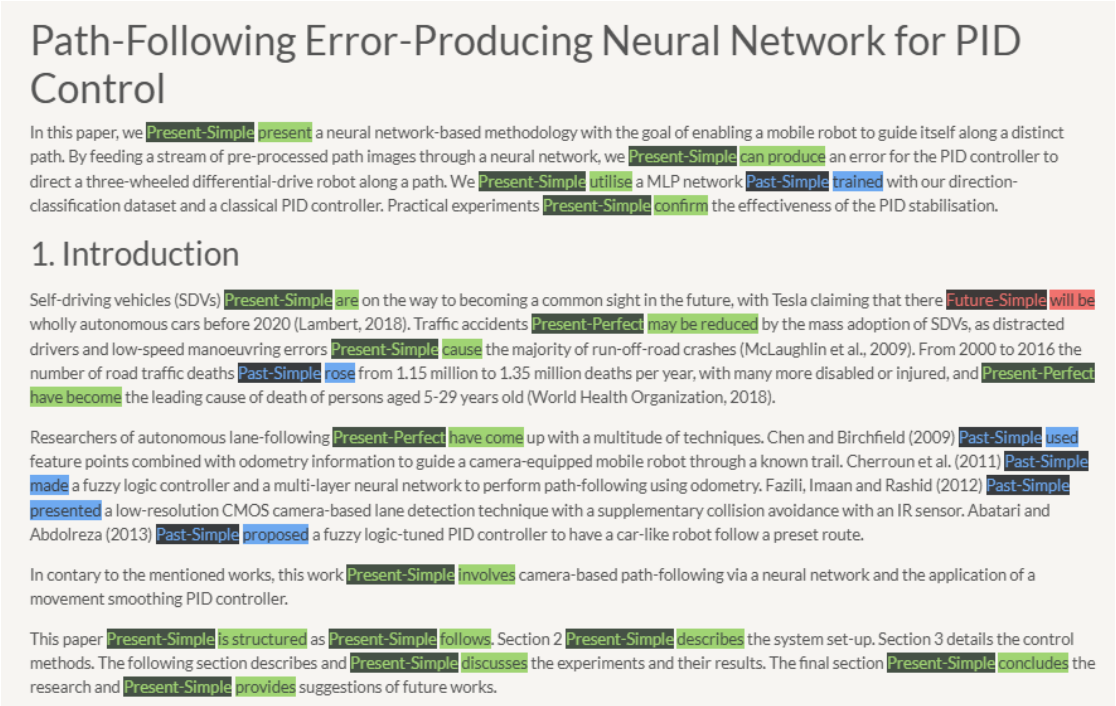

Figure 2: *Screenshot of a research article with the tense feature selected. Each identified verb form is assigned a tense label and colorized.*

The Feature Visualizer allows learners to control the depth of explanations, moving along a continuum shown in Figure 3 from implicit colorization and simple descriptions to explicit descriptions and explanations. This approach resonates with the Sociocultural Theory view of mediated learning (Poehner, 2018; Vygotsky, 1978), where learners actively negotiate the level of assistance they require.

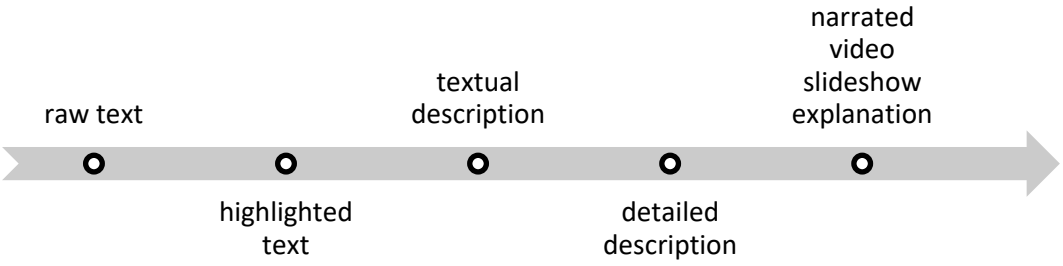

Figure 3: *Cline showing increasing information availability. Learners adopting an inductive approach are expected infer from raw and highlighted text.*

This layered interaction design encourages learners to explore patterns independently while also providing support when needed. The tool thus facilitates a self-directed, inductive pathway through the complexities of academic language using an intuitive interface, ameliorating the common challenges in data-driven corpus-based learning (Anthony, 2019).

The identification of linguistic features is achieved through a combination of rule-based parsing and AI-powered methods. The corpus was manually annotated for sections (e.g., Introduction, Method) and rhetorical moves (e.g., showing importance, showing novelty) in order to reduce the number of false positive results that often occur with probabilistic approaches.

Part-of-speech (POS) tagging is performed using the NLTK PerceptronTagger, a machine learning-based tagger (Bird et al., 2015). Many of the micro-level features, such as tense, modality, passive constructions, and linking expressions, are identified using regular expressions applied to the POS-tagged corpus. For instance, passive voice constructions are located by matching POS patterns corresponding to auxiliary + past participle sequences, while hedges, boosters and approximations within the modality function are detected through curated lists of expressions (Hyland, 1998; 2000). While rule-based string matching suffices for many syntactic and lexical features, two advanced functions that reveal discourse-level semantic relationships, namely the cohesion function and coherence function, rely on AI-driven processes to detect deeper discourse-level relations that are not amenable to pattern matching alone. Here, AI-driven refers to established natural language processing (NLP) models, such as the Stanford CoreNLP coreference resolution system (Clark & Manning, 2015) and semantic similarity metrics based on WordNet synsets (Wu & Palmer, 1994), rather than large language models (LLMs), which were only employed in generating learner-facing explanations.

The cohesion function draws on the statistical coreference resolution algorithm (Clark & Manning, 2015) from the Stanford CoreNLP toolkit (Manning et al., 2014) to automatically detect cohesive links between referring expressions within a paragraph. The coreference resolution process constructs chains of expressions (e.g., noun phrases and pronouns) that refer to the same discourse entity. Expressions that are resolved to the same entity are highlighted with the same background colour in the visualization. This allows users to see at a glance how entities are linked across sentences, revealing the density and distribution of referential cohesion within a paragraph as can be seen in Figure 4.

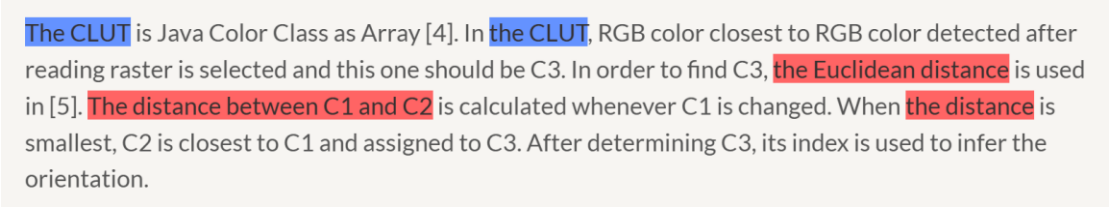

Figure 4: *Extract of a research article with the cohesion feature selected. Two different entities are identified, namely the Color Look Up Table (CLUT) and Euclidean distance.*

The coherence function applies semantic similarity measures to identify and highlight conceptually related content within texts. This function extracts all expressions tagged as nouns by the NLTK Perceptron POS-tagger and compares their meanings using synsets from WordNet (Miller, 1995), applying the Wu-Palmer similarity metric (Wu & Palmer, 1994). Pairs of nouns with a similarity score above 0.9 are treated as semantically reinforcing one another, and their frequencies are boosted accordingly.

Following this similarity-based reinforcement, the system identifies the most central noun(s) in each paragraph, i.e., those with high frequency and strong semantic connections. These central nouns are displayed prominently at the start of each paragraph in the user interface. For instance, in a paragraph that mentions *"climate," "weather," "temperature,"* and *"precipitation,"* the system may identify *"climate"* as the key unifying concept and visually highlight it to reflect its semantic centrality. This AI-assisted feature supports learners in understanding how thematic coherence is built across sentences, providing a scaffolded pathway to noticing abstract discourse-level patterns.

4. Methodology

This longitudinal mixed-method study investigated how interaction with the Feature Visualizer influenced genre awareness among Japanese undergraduate computer science majors. Participants were enrolled in a 14-week thesis writing course in which they drafted a short research article in English. The primary aim was to examine whether regular engagement with the tool supported the development of genre-related knowledge.

Throughout the course, students were allotted 10–15 minutes per session to use the Feature Visualizer. Rather than assigning specific tasks, students were encouraged to explore the tool freely to deepen their understanding of the research article genre and apply this understanding to their own writing.

Student engagement with the Feature Visualizer was monitored through classroom observation. The course tutor recorded patterns in tool use, noting variations in interaction styles. Midway through the semester, students participated in a short verbal survey to provide feedback on how they were using the tool. Additionally, semi-structured focus group interviews were conducted with six students upon course completion to explore their experiences and perspectives in greater depth. After each focus group interview, a member check was conducted to confirm the veracity of the interviewer's notes.

A pre-test was administered at the start of the course to assess students' baseline knowledge of the genre of short computer science research articles. The same

test was repeated at the end of the semester as a post-test to measure any changes in genre awareness. Table 2 lists the twelve genre-related concepts that are assessed. A copy of the test is available in Appendix 1. The test was administered on paper on both occasions but students were allowed to look up unknown vocabulary on their computer.

To evaluate statistical significance in performance changes, McNemar's test was applied to the subset of participants who completed both the pre-test and post-test. This non-parametric test is appropriate for paired nominal data and is commonly used to detect changes in dichotomous outcomes across two time points within the same participants. Unlike parametric tests such as the paired-sample t-test or effect size metrics like Cohen's d, which require continuous interval data and assume normality, McNemar's test is specifically suited to detect within-subject changes in categorical paired data, providing an appropriate measure of significance for this study.

Of the 23 enrolled students, 19 completed the pre-test, and 14 completed both pre- and post-tests, forming the sample for quantitative analysis. For each concept, responses were coded as correct or incorrect, allowing construction of 2×2 contingency tables. McNemar's test with continuity correction was used to assess whether the proportion of correct responses increased significantly after instruction.

Table 2: *Genre-related Concepts Covered in Course and Test.*

Concept	**Relevant content**
Organization concepts	Move from general to specific; first to last; most important to less important
Organization of article	Sections detailing answers to why, how, what and so what, i.e. introduction, method, results and discussion
Front and end matter	Abstract, references and appendices
Introductions	Describing novelty, significance, providing background and overview of remainder of article
Paraphrasing	Linking current section to previous section
Signposts	Helping reader understand organization using adverbs
Voice	Using passive voice to focus on processes not people
Tense	Using past tense for completed actions
Definition	Creating shared understanding with readers
Approximation	Describing exact values in a more reader-friendly way
Abstraction	Repacking processes using nominalization
IEEE citation system	Using numbers in square brackets to refer to sources

5. Findings

This section presents the findings of our study in relation to the guiding research question: *How does interaction with the Feature Visualizer impact the genre awareness of students?* To address this question, we first examine whether the tool influenced learners' recognition of academic writing concepts, drawing on pre- and post-test data to provide quantitative evidence of change. We then turn to qualitative data from classroom observations, surveys, and focus group interviews to explore how learners engaged with the tool and what strategies they adopted. By combining statistical results with insights into learner behaviours and perceptions, we provide a comprehensive account of both the extent (whether) and the manner (how) in which the Feature Visualizer shaped students' awareness of genre conventions.

As shown in Table 3, post-test scores demonstrated substantial gains in the recognition of all targeted academic writing concepts. Awareness of concepts such as typical content of introductions, the selection of appropriate grammatical tenses, and the purpose of paraphrasing one's own words showed clear improvement. The post-test results related to the content of the front and end matter, the use of passive voice particularly in the method section, and details of the IEEE citation system improved substantially. Even the less familiar concepts at the outset, such as the purpose of providing definitions and the use of abstraction, saw increases from zero to between 7 and 10 students.

McNemar's test results confirmed that these gains were statistically significant for most concepts. Notably, the increase in recognition of the content of introductions ($\chi^2 = 9.0909$, $p = .0026$) and use of paraphrasing ($\chi^2 = 6.125$, $p = .0133$) reached strong significance. With the exception of the use of approximation, which rose from 36% to 64% but did not reach statistical significance, all improvements were significant at the $p < .05$ level, indicating a strong learner-driven developmental effect resulting from independent engagement with the materials.

Table 3: *Results of Pre-test and Post-test for Participants who Completed both Tests.*

Concept	Pre-test	Post-test	χ^2	p-value	Significant?
Organization concepts	4 (28%)	12 (86%)	5.1429	0.0233*	Yes
Organization of article	5 (36%)	14 (100%)	5.1429	0.0233*	Yes
Front and end matter	6 (42%)	14 (100%)	4.1667	0.0412*	Yes
Introductions	3 (21%)	14 (100%)	9.0909	0.0026*	Yes
Paraphrasing	0 (0%)	8 (57%)	6.125	0.0133*	Yes
Signposts	3 (21%)	14 (100%)	9.0909	0.0026**	Yes
Voice	4 (29%)	12 (86%)	6.125	0.0133*	Yes
Tense	2 (14%)	12 (86%)	8.1	0.0044**	Yes
Definition	0 (0%)	10 (71%)	8.1	0.0044**	Yes
Approximation	5 (36%)	9 (64%)	2.25	0.1336	No
Abstraction	0 (0%)	7 (50%)	5.1429	0.0233*	Yes

IEEE citation system	5	(36%)	14	(100%)	6.125	0.0133*	Yes

*p <.05, **p < .01

The findings suggest that inductive learning supported by AI-enhanced, corpus-based tools can meaningfully raise learners' awareness of genre conventions. The design of the Feature Visualizer enabled them to engage directly with curated texts, fostering self-directed exploration and discovery of genre conventions. By visualizing abstract rhetorical features, the Feature Visualizer made genre norms of short research articles more accessible and actionable, enabling learners to notice and understand them with greater ease.

Based on the observation log, verbal mid-course survey and focus group interviews, most students engaged actively with the Feature Visualizer, exhibiting two common usage patterns, which can be described as intensive article-level exploration and extensive feature-level comparison.

The first pattern, *intensive article-level exploration*, involved in-depth exploration of a single research article. Students adopting this approach tended to load an article closely aligned with the methodological focus or content of their own research topic and systematically examined its structural and linguistic features. Most students began their exploration with macro features. They toggled various visualizations on and off, focusing on how genre conventions were realized within their chosen article.

The second pattern, *extensive feature-level comparison*, was characterized by focused investigation of a particular linguistic or rhetorical feature across multiple articles. Students employing this strategy selected a feature of interest, such as tense usage or move structures, and rapidly navigated through several articles to compare how the feature was realised across different contexts. This comparative analysis approach enabled them to notice patterns of variation and commonality, thereby developing a more flexible and fine-grain understanding of genre conventions.

Observation logs indicated frequent interaction with the colour-coded visualizations. Students often used the Feature Visualizer alongside other digital resources such as machine translation services (e.g., DeepL and GoogleTranslate), search engines, and LLMs (e.g., ChatGPT and Gemini), primarily accessed through their mobile devices or personal laptops rather than the university workstation.

Focus group interviews revealed that students found the tool intuitive and particularly helpful in discovering and clarifying genre expectations for research writing. All six participants reported increased confidence in evaluating and revising drafts of their own research articles.

The layered design of the tool, ranging from raw text to multimodal explanations, allowed students to control the depth of information accessed. However,

contrary to expectations, video explanations were rarely used. When asked, only three of the 16 respondents reported watching any video content, and they explained that they did so either because they wanted deeper insight into a particular feature or because the labelled highlights alone were insufficient for full understanding. For example, one student commented, "I could not understand the differences between the different types of linking words from the highlighting alone." By contrast, the majority of students indicated that they did not feel the need for additional information. One pragmatic student noted, "I understood enough to write my thesis, and I am time-pressured, so it was not necessary." Another added, "I could work out the meaning from seeing the highlighted terms in context and felt confident." (All quotations are idiomatic translations from Japanese.)

All focus group participants agreed that the tool enhanced their awareness of the genre conventions of short computer science research articles. Tutors also observed notable improvements in the clarity, coherence, and formality of student writing. The integration of visualization, corpus-based interaction, and learner autonomy was consistently highlighted as a strength of the Feature Visualizer.

The two engagement patterns (intensive article-level exploration and extensive feature-level comparison) were noticed, which reflect different learner strategies for inductive genre learning. The first pattern, intensive article-level exploration, aligns with model-based approaches to learning from exemplars. Research on exemplar-based learning demonstrates that working closely with a single, detailed case can help learners notice and facilitate the abstraction of general principles (Renkl, 2014). In the context of genre pedagogy, a single text often functions as a comprehensive model through which learners can identify rhetorical moves and organisational structures (Blake, 2001; Swales, 1990). Focusing on one exemplar also reduces the cognitive complexity of the task: according to cognitive load theory, novices benefit from limiting the range of input in order to allocate more attention to structural and linguistic features that might otherwise be overlooked (Bahari, 2023). By systematically engaging with a single article, learners employing this strategy are able to build a coherent reference model for academic writing conventions, which then may serve as a foundation for their own production.

The second pattern, extensive feature-level comparison, reflects a contrastive learning strategy, drawing on cross-textual variation to infer rules and conventions. In this approach, learners actively search for patterns of similarity across texts, which they can then use to formulate generalisations about genre norms, a form of "frequency-biased abstraction of regularities" (Ellis, 2002, p.143). In genre pedagogy, comparative work across texts highlights how rhetorical conventions can vary within and across disciplines, enabling learners to move beyond fixed models toward more

flexible understandings of discourse (Hyland, 2012). This aligns with research in contrastive rhetoric, which emphasises the importance of cross-textual comparisons for recognising discourse organisation and rhetorical choices (Connor, 2002). By engaging with multiple texts through feature-focused toggling, students following this strategy are able to abstract genre principles through systematic contrast. Both patterns highlight the affordances of the Feature Visualizer in supporting autonomous, data-driven exploration, with learners adjusting their interaction style based on their immediate learning goals.

The flexible toggle-based interface supported different learning preferences, allowing users to select the depth and mode of explanation most suited to their needs. The use of a corpus provided a firm foundation for explorations in authentic disciplinary usage, while the tool's design promoted a shift from passive rule-following to active pattern discovery. This guided autonomy, in which learners navigate independently within a structured environment, proved successful in sustaining engagement and fostering deeper learning.

In addition to using the Feature Visualizer, students frequently accessed other digital resources, including machine translation tools, search engines, and LLMs. This blended engagement aligns with connectivism (Siemens, 2004), which emphasizes the role of networked tools and distributed knowledge sources in modern learning environments. It also reflects the principles of distributed cognition (Hollan, Hutchins, & Kirsh, 2000), wherein cognitive processes are shared across individuals, tools, and environments, and situated learning theory (Lave & Wenger, 1991), which views learning as embedded within authentic contexts and social practices. The Feature Visualizer functioned not as a stand-alone instructional aid but as a central node in a broader ecology of tools that learners actively drew upon to construct genre knowledge.

6. Conclusions, limitations and implications

The Feature Visualizer is an open-access, AI-powered tool designed to raise genre awareness through inductive learning. Grounded in the principles of corpus-based pedagogy and interactive discovery, the tool enables learners to explore authentic texts, visualize key features, and access explanatory support on demand. A longitudinal study showed that learners using the tool improved their genre knowledge of discipline-specific writing and developed greater confidence in navigating the conventions of scientific genres. Learners viewed the tool as effective and user-friendly. This study reinforces the value of integrating AI and corpora in ways that promote inductive learning, learner autonomy, and genre awareness. Future

research will explore the application of the Feature Visualizer in other domains and with larger, more diverse cohorts of learners.

An additional contribution of this study lies in the identification of two distinct engagement patterns: intensive article-level exploration and extensive feature-level comparison. These patterns illustrate the different ways in which learners approach inductive genre learning using AI-driven feature visualization. The first pattern reflects a model-based approach in which a single exemplar text serves as the main reference point, while the second demonstrates a contrastive strategy that draws on similarities and differences across texts. Recognising these patterns provides insight into how learners adapt exploratory strategies to their goals and highlights the tool's flexibility in accommodating multiple learning pathways.

These results echo broader pedagogical trends advocating for the integration of AI and corpora in ways that prioritize learner agency, real-world language exposure, and interactive exploration. Projects, such as CorpusChat, illustrate how AI-powered chatbots can scaffold corpus exploration in real time (Cheung & Crosthwaite, 2025) while the online interface of English-Corpora.org which houses multiple corpora, offers the option to use LLMs to assist in the identification of patterns.

However, this study also revealed certain limitations. Despite the tool's multimodal capabilities, some features, such as video explanations, were underutilized, suggesting a need for further investigation into how learners perceive and engage with different forms of support. Another limitation concerns the corpus size and the reliance on pre-annotated features. While the tool is designed to work with raw text as much as possible, some degree of manual annotation was necessary to avoid false positives, as automated section identification using machine learning alone did not achieve sufficient accuracy for deployment. Finally, although the learning outcomes were promising, the sample size for quantitative analysis was limited to 14 students. This small cohort size constrains the generalizability of the findings, though it does offer a valuable proof of concept.

Future research should examine long-term effects, differences across proficiency levels, and usage patterns in larger and more diverse cohorts to better understand how the tool supports genre learning across contexts.

The Feature Visualizer offers practical benefits for language teachers and curriculum designers seeking to integrate corpus-based, inductive learning into writing instruction. Its application is particularly relevant in fields that rely on highly structured texts, such as STEM disciplines. The tool can be incorporated into classroom activities, independent study tasks, or revision workshops, and requires minimal training due to its intuitive design.

Teachers can use the tool to support discovery-based activities that complement explicit instruction, helping learners develop both conceptual understanding and practical application of genre conventions. Its embedded multimodal explanations appeal to diverse learning styles, sustaining engagement and supporting differentiation. By promoting active learning and data-driven inquiry, the tool helps learners take ownership of their writing development. The Feature Visualizer is available online at https://fv.rt247a.ddns.me/.

Acknowledgements

This project was funded by JSPS grant-in-aid *Kakenhi* number 19K00850.

References

Anthony, L. (2019). Tools and strategies for Data-Driven Learning (DDL) in the EAP writing classroom. In K. Hyland, & L.L.C. Wong, (Eds.), *Specialised English* (pp. 179–194). Routledge.

Bahari, A. (2023). Challenges and affordances of cognitive load management in technology-assisted language learning: A systematic review. *International Journal of Human–Computer Interaction, 39*(1), 85–100. https://doi.org/10.1080/10447318.2021.2019957

Bhatia, V. (1999). Generic integrity in document design. *Document Design, 1* (3), 151–163. https://doi.org/10.1075/dd.1.3.01bha

Bird, S., Klein, E., & Loper, E. (2015). Natural Language ToolKit (Version 3.1) [Software library]. Retrieved from http://www.nltk.org/

Blake, J. (2021). *Corpus-based study of the rhetorical organization and lexical realization of scientific research abstracts.* [Unpublished doctoral dissertation]. Aston University.

Blake, J., Pyshkin, E. and Pavlic, S. (2023). Automatic detection and visualization of information structure in English. *Proceedings of the 6th International Conference on Natural Language Processing and Information Retrieval* (pp.200–204) ACM. https://doi.org/10.1145/3582768.3582784

Blake, J., Zhao, P. and Pyshkin, E. (2025). TrendScribe: Design and Development of a Pedagogic Trend Description Generator for Learners of English. In C. Sombattheera, Weng, P. and Pang, J. (Eds.), *Multi-disciplinary Trends in Artificial Intelligence: 17th International Conference, MIWAI 2024. Proceedings, Part I. Lecture Notes in Artificial Intelligence, Vol 15431* (pp.89–101). Springer, Singapore. https://doi.org/10.1007/978-981-96-0692-4_8

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.

Cheung, L., & Crosthwaite, P. (2025). CorpusChat: integrating corpus linguistics and generative AI for academic writing development. Computer Assisted Language Learning, 1–27. https://doi.org/10.1080/09588221.2025.2506480

Clark, K., & Manning, C. D. (2015, July). Entity-centric coreference resolution with model stacking. In C. Chengqing & M. Strube (Eds.), *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (pp. 1405–1415). https://doi.org/10.3115/v1/P15-1

Connor, U. (2002). New directions in contrastive rhetoric. *TESOL quarterly, 36*(4), 493–510. https://doi.org/10.2307/3588238

Crosthwaite, P., & Baisa, V. (2024). A user-friendly corpus tool for disciplinary data-driven learning: Introducing CorpusMate. *International Journal of Corpus Linguistics, 29*(4), 595–610. https://doi.org/10.1075/ijcl.23056.cro

Ekanayaka, W. I., & Ellis, R. (2020). Does asking learners to revise add to the effect of written corrective feedback on L2 acquisition?. *System, 94*, 102341. https://doi.org/10.1016/j.system.2020.102341

Ekanayaka, W. I., & Ellis, R. (2021). Which is the most effective technique for inducing learners' attention to Written Corrective Feedback (WCF): Revision or discussion? Approximate replication of Ekanayaka and Ellis (2020). *SSRN Electronic Journal.* https://doi.org/10.2139/ssrn.3809061

Ellis, N. C. (2002). Frequency effects in language processing: A review with implications for theories of implicit and explicit language acquisition. *Studies in Second Language Acquisition, 24(*2), 143–188. https://doi.org/10.1017/S0272263102002024

Ellis, R. (2021). *Reflections on task-based language teaching.* Multilingual Matters.

Flowerdew, L. (2015). Corpus-based research and pedagogy in EAP: From lexis to genre. *Language Teaching, 48*(1), 99–116. https://doi.org/10.1017/S0261444813000037

Gilquin, G., & Granger, S. (2022). Using data-driven learning in language teaching. In A. O'Keeffe & M.J. McCarthy (Eds.), *The Routledge handbook of corpus linguistics* (pp. 430–442). Routledge. https://doi.org/10.4324/9780367076399

Hollan, J., Hutchins, E., & Kirsh, D. (2000). Distributed cognition: Toward a new foundation for human-computer interaction research. *ACM Transactions on Computer-Human Interaction, 7*(2), 174–196. https://doi.org/10.1145/353485.353487

Hyland, K. (1998). Boosting, hedging and the negotiation of academic knowledge. *Text & Talk, 18*(3), 349–382.

Hyland, K. (2012). *Disciplinary identities: Individuality and community in academic discourse*. Cambridge University Press.

Hyland, K. (2000). Hedges, boosters and lexical invisibility: Noticing modifiers in academic texts. *Language awareness*, *9*(4), 179–197. https://doi.org/10.1080/09658410008667145

Ishikawa, M., & Révész, A. (2020). L2 learning and the frequency and quality of written languaging. In W. Suzuki, & N. Storch (Eds.), *Languaging in language learning and teaching: A collection of empirical studies* (pp. 220–240). John Benjamins.

Jablonkai, R. R., & Csomay, E. (Eds.). (2022). *The Routledge handbook of corpora and English language teaching and learning.* Routledge.

Johns, T. (2002). Data-driven learning: The perpetual challenge. In B. Kettemann & G. Marko (Eds.), *Teaching and learning by doing corpus analysis. Proceedings of the Fourth International Conference on Teaching and Language Corpora* (pp. 105–117). Brill. https://doi.org/10.1163/9789004334236_010

Kilgarriff, A., Marcowitz, F., Smith, S., & Thomas, J. (2015). Corpora and language learning with the Sketch Engine and SKELL. *Revue Française de Linguistique Appliquée, (1),* 61–80.

Lave, J., & Wenger, E. (1991). *Situated Learning: Legitimate Peripheral Participation.* Cambridge University Press.

Li, D., Noordin, N., Ismail, L., & Cao, D. (2025). A systematic review of corpus-based instruction in EFL classroom. *Heliyon, 11*(2), e42016.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S., & McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. In K. Bontcheva & J. Zhu (Eds.), *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations* (pp. 55–60).

Poehner, M. E. (2018). Probing and provoking L2 development: The object of mediation in dynamic assessment and mediated development. In J. Lantolf, M.E. Poehner & M. Swain (Eds.), *The Routledge handbook of sociocultural theory and second language development* (pp. 249–265). Routledge.

Prince, M. J., & Felder, R. M. (2006). Inductive teaching and learning methods: Definitions, comparisons, and research bases. *Journal of Engineering Education, 95*(2), 123–138. https://doi.org/10.1002/j.2168-9830.2006.tb00884.x

Prince, M., & Felder, R. M. (2007). The many faces of inductive teaching and learning. *Journal of College Science Teaching, 36*(5), 14.

Renkl, A. (2014). Toward an instructionally oriented theory of example-based learning. *Cognitive Science, 38*(1), 1–37. https://doi.org/10.1111/cogs.12086

Reppen, R. (2010). Using corpora in the language classroom. In B. Tomlinson (Ed.), *Materials development in language teaching* (pp. 35–50), Cambridge University Press.

Schmidt, R. (1990). The role of consciousness in second language learning. *Applied Linguistics, 11*(2), 129–158. https://doi.org/10.1093/applin/11.2.129

Siemens, G. (2004, December 12). Elearnspace. Connectivism: A learning theory for the digital age. *Elearnspace.org.* 1–7.

Smart, J. (2014). The role of guided induction in paper-based data-driven learning. *ReCALL, 26*(2), 184–201. https://doi.org/10.1017/S0958344014000081

Swales, J. M. (1990). *Genre analysis: English is academic and research settings*. Cambridge University Press.

Swales, J. M., & Feak, C. B. (2012). *Academic writing for graduate students: Essential tasks and skills (3rd Edition)*. University of Michigan Press.

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Harvard University Press.

Zobel, J. (2014). *Writing for computer science,* (3rd Edition). Springer. https://doi.org/10.1007/978-1-4471-6639-9

Appendix 1: Pre- and post-test on genre of computer science research articles

Please write your answers to these questions. You can write in English and/or Japanese. You can use a dictionary if needed, but do not use generative AI.

1. How are research articles organized?
2. How are sections within the article organized?
3. What is given before the introduction or after the conclusion?
4. What is often included in the introduction?
5. Why do we paraphrase our own words within a research article?
6. Why do we use signposts?
7. When and why do we use active and passive voice?
8. What tenses are used, and why?
9. Why are definitions included?
10. How can you mention a value in a table (e.g., 4.512) in the results section?
11. Why are some processes described using nouns rather than verbs?
12. How do you cite a research article?