

An Integrated Corpus-Based Text-Mining Framework for Analyzing Student Discourse in Linguistic Landscape Pedagogy

Enid Lee^{1*}

*Correspondence:

e.lee@oku.ac.jp

¹ Department of British and American Language and Culture, Okinawa International University, Japan



Abstract

Linguistic landscape (LL) projects are widely used to foster sociolinguistic awareness and experiential learning, yet evaluations have relied predominantly on qualitative analyses of interviews and reflective journals. Consequently, systematic examination of student writing for broader discourse patterns remains limited. This study introduces a corpus-based text-mining framework that reconceptualizes student writing as a research corpus, applying triangulated quantitative analyses to LL project discourse. The corpus comprised discussion and conclusion sections from 95 undergraduate papers written in Japanese by English majors at a Japanese university following neighborhood LL investigations. Using KH Coder, four analytical procedures—word frequency, co-occurrence networks, multidimensional scaling, and hierarchical clustering, identified recurring lexical and conceptual patterns. Findings revealed three main tendencies: students interpreted LLs through functional, tourism-oriented perspectives; employed research-oriented discourse marked by comparison and evidence-based reasoning; and conceptually linked abstract language notions with specific varieties and communicative functions. Beyond substantive insights, the framework offers a scalable, replicable approach for corpus-level analysis that complements qualitative methods, advancing both LL pedagogy and educational text-mining methodologies.

Keywords: linguistic landscape pedagogy; corpus-based analysis; text mining; KH Coder; student writing

1. Introduction

LL research has developed from a relatively specialized area of sociolinguistics into an interdisciplinary field spanning applied linguistics, multilingualism, literacy studies, language policy, and language education. Originally defined by Landry and Bourhis (1997) as the visibility and salience of languages on public and commercial signs, the concept has subsequently expanded to encompass the social, cultural, ideological, and educational meanings embedded in written language displayed in public spaces. As a result, LLs are now widely recognized not only as objects of sociolinguistic inquiry but also as valuable pedagogical resources that enable students to investigate a variety of language- and communication-related issues through direct engagement with authentic language use.

Within language education, LL pedagogy has attracted increasing attention because it offers students opportunities to connect theoretical concepts with their everyday linguistic environments. Instead of learning about multilingualism solely through textbooks or classroom discussion, students actively investigate the languages, scripts, symbols, and semiotic resources that shape communication in public spaces. Previous studies have shown that LL projects can foster sociolinguistic awareness, critical language awareness, intercultural understanding, inquiry-based learning, and reflective thinking (Allen, 2025; Dagenais et al., 2009; Gorter et al., 2021; Li & Marshall, 2020; Rowland, 2013; Sayer, 2010; Shohamy & Waksman, 2009; Spinelli, 2024). Through photographing, documenting, and analyzing public signs, students cultivate a heightened awareness of linguistic diversity and cross-cultural communication while simultaneously engaging in observation, comparison, interpretation, and evidence-based reasoning.

Despite this growing body of research, evaluating the educational value of LL projects remains methodologically challenging. Most previous studies have relied on qualitative approaches, including interviews, reflective journals, classroom observations, and thematic analyses of selected student reflections. These approaches provide rich insights into students' experiences and meaning-making processes. However, they are inherently difficult to scale across large datasets and often require researchers to focus on relatively small numbers of participants or illustrative excerpts. While such qualitative approaches remain indispensable for understanding how students construct meaning, they provide only a partial view of the collective discourse generated across an entire cohort of learners.

Consequently, relatively little attention has been paid to how student writing itself can be analyzed systematically as a corpus. When considered collectively rather than as individual assignments, student papers provide a rich source of learning-related discourse that can reveal recurring lexical, semantic, and conceptual patterns across large numbers of learners. Corpus-level analyses make it possible to identify dominant interpretive frames, recurring thematic structures, and patterns of academic discourse that may remain difficult to detect through conventional close reading alone. Importantly, this perspective does not replace qualitative interpretation; rather, it provides an empirical foundation from which richer qualitative analyses can subsequently investigate how and why these broader discourse patterns emerge.

2. The Present Study

The present study addresses this methodological gap by treating student writing not simply as evidence for assessment but as a research corpus through which collective patterns of learning-related discourse can be systematically investigated. As opposed to examining individual papers in isolation, the study applies a triangulated text-mining framework to identify recurring lexical, semantic, and conceptual patterns across an entire corpus of student writing. The framework integrates four complementary analytical procedures implemented in KH Coder—word frequency analysis, co-occurrence network analysis, multidimensional scaling, and hierarchical cluster analysis. Each procedure contributes a distinct perspective on the corpus, collectively forming a robust and replicable analytical framework for examining student discourse at scale. While each analytical procedure offers valuable insights independently, their combined application enables triangulation across multiple levels of textual organization, thereby strengthening the reliability of the overall interpretation.

The study also contributes to ongoing methodological discussions in applied linguistics concerning the relationship between quantitative and qualitative analyses of educational discourse. Student writing has traditionally been examined through close reading of selected texts or qualitative coding of relatively small datasets. Such approaches remain essential for understanding students' interpretive processes and meaning making; however, they become increasingly challenging as datasets grow larger. Corpus-based text mining offers a complementary perspective by enabling researchers to identify systematic discourse patterns across entire collections of student writing before investigating representative cases in greater depth. Far from positioning quantitative and qualitative approaches as competing methodologies, the present study argues that they can be understood as complementary stages within a broader analytical workflow. Corpus-level analyses identify recurring patterns that subsequently provide productive starting points for qualitative interpretation, while qualitative analyses explain how and why those patterns emerge within individual students' reflections.

The study was guided by the following research questions:

RQ1. How can a triangulated corpus-based text-mining framework be used to systematically analyze student discourse generated through LL projects?

RQ2. What dominant lexical, thematic, and conceptual patterns emerge from students' discussions and conclusions following LL projects?

RQ3. What methodological and pedagogical implications does this corpus-based approach offer for LL pedagogy and educational research in applied linguistics?

The purpose of the study is twofold. First, it demonstrates a systematic and replicable methodology for analyzing student discourse through corpus-based text mining. Second, it applies this methodology to investigate how undergraduate students interpreted LLs following individual field-based research projects. This dual focus aims to enhance pedagogical frameworks for LL assignments while providing researchers with a more rigorous, scalable approach to analyzing student reflections.

3. Methodology

3.1 Data

The corpus consisted of discussion and conclusion sections extracted from 95 research papers written in Japanese by undergraduate English majors enrolled in an LL course at a university in Okinawa, Japan. As part of the course, students independently selected an LL site within their local neighborhood or surrounding towns, documented the linguistic and semiotic features they observed through photographs and field notes, reviewed relevant literature, and produced an academic research paper presenting their findings and interpretations.

Only the discussion and conclusion sections were included in the corpus. These sections were selected because they represent the stage at which students move beyond describing observations and begin interpreting findings, evaluating sociolinguistic phenomena, and reflecting on the broader implications of their analyses. Whereas earlier sections primarily report procedures and observational data, the discussion and conclusion sections contain students' attempts to synthesize evidence, construct arguments, and evaluate the communicative and sociolinguistic significance of LLs. They therefore provide the richest evidence of learning-related discourse and are the most appropriate sections for examining how students interpret LL phenomena.

The papers were written in Japanese to allow students to express complex sociolinguistic observations and critical reflections in their first language. Conducting the analyses in Japanese minimized linguistic constraints associated with second-language academic writing and enabled the corpus to more accurately represent students' conceptual understanding. English translations are provided only when presenting representative examples in the Results and Discussion sections.

Following extraction from each student paper, the discussion and conclusion sections were compiled into a single corpus, converted into plain text, and imported into KH Coder (Version 3), a software package designed for quantitative content analysis and text mining of language corpora (Higuchi, 2016, 2017). Standard preprocessing procedures, including morphological analysis, removal of function words, and normalization of lexical forms, were conducted within KH Coder prior to textual analysis.

3.2 Analytical Procedures

Avoiding reliance on a single analytical technique, the present study employed four complementary procedures that synthesize into a triangulated text-mining framework. Each procedure examines a different dimension of the corpus, enabling lexical prominence, semantic relationships, conceptual organization, and thematic structure to be investigated from multiple analytical perspectives. Interpreting patterns that recur across these complementary analyses increases confidence that the findings reflect stable characteristics of the corpus instead of being artifacts of any individual analytical technique.

Word frequency analysis was first conducted to identify the most salient lexical items appearing throughout the corpus. Frequency analysis provides an overview of the dominant topics and concepts discussed by students and establishes the lexical foundation for subsequent analyses. Co-occurrence network analysis was then employed to examine semantic relationships among frequently occurring words. Words that regularly appeared

within the same contextual unit were connected to visualize conceptual associations, thereby revealing how students linked ideas throughout their written discussions. Multidimensional scaling analysis was subsequently performed to explore broader conceptual organization within the corpus. By positioning words according to the similarity of their co-occurrence patterns, multidimensional scaling provides a visual representation of semantic proximity and highlights relationships that may not be immediately apparent through frequency counts or network diagrams alone. Finally, hierarchical cluster analysis was conducted to identify higher-order thematic groupings within the corpus. Cluster analysis groups words according to similarities in their co-occurrence patterns, providing an additional perspective on conceptual organization and allowing the consistency of thematic structures identified through the preceding analyses to be evaluated.

In combination, these four analytical procedures provide complementary evidence concerning the discourse patterns characterizing the corpus. Instead of treating each analysis independently, the present study interprets convergence across the four procedures as evidence of robust lexical, semantic, and conceptual patterns within students' learning-related discourse.

3.3 Unit of Analysis

Paragraphs were selected as the co-occurrence unit throughout the analyses because they generally represent coherent units of academic argument not isolated propositions. Compared with sentence-level analysis, paragraph-level co-occurrence captures broader semantic relationships while reducing fragmentation of conceptual patterns across individual sentences. Because students typically developed a single argument or interpretive theme within each paragraph, this unit of analysis was considered most appropriate for examining reflective academic writing produced through LL projects.

3.4 Interpretation of Results

Interpretation focused on patterns that emerged consistently across the four analytical procedures instead of isolated lexical items or individual visualizations. Recurring relationships identified through word frequency analysis, co-occurrence network analysis, multidimensional scaling, and hierarchical cluster analysis were interpreted as representing stable characteristics of the corpus. This triangulated approach strengthened the robustness of the analysis by reducing reliance on any single analytical procedure and providing complementary evidence concerning the dominant forms of discourse generated by students following their LL investigations.

4. Results

4.1 Word Frequency Analysis: Salient Lexical Patterns

The word frequency analysis (Table 1) provided an overview of the lexical characteristics of the student corpus by identifying the most frequently occurring content words in the discussion and conclusion sections. The most salient lexical items included 言語 (language, $f = 626$), 英語 (English, $f = 474$), 日本語 (Japanese, $f = 420$), 中国語 (Chinese, $f = 291$), 韓国語 (Korean, $f = 145$), 簡体字 (simplified Chinese characters, $f = 113$), 繁体字 (traditional Chinese characters, $f = 66$), 情報 (information, $f = 201$), 外国人

(foreigner, $f = 319$), 観光 (tourism, $f = 173$), 観光客 (tourist, $f = 272$), and 多言語 (multilingual, $f = 304$). Terms related to signage were also prominent, including 表記 (notation, $f = 422$), 表示 (display, $f = 395$), 案内 (guidance, $f = 221$), 看板 (sign/signboard, $f = 180$), ピクトグラム (pictogram, $f = 173$), and サイン (sign, $f = 67$). Words referring to specific contexts, such as 沖縄 (Okinawa, $f = 286$), 空港 (airport, $f = 78$), ホテル (hotel, $f = 71$), メニュー (menu, $f = 66$), and 店舗 (store, $f = 92$), were also common. In addition, evaluative and functional terms such as 使用 (use, $f = 425$), 利用 (utilize, $f = 203$), 必要 (necessary, $f = 93$), 理解 (understand, $f = 74$), and 重要 (important, $f = 73$), and research-related vocabulary such as 研究 (research, $f = 232$), 調査 (investigation/investigate, $f = 229$), 考察 (discussion, $f = 175$), 結果 (results, $f = 149$), 分析 (analysis/analyze, $f = 118$), and 比較 (comparison/compare, $f = 83$) appeared frequently.

The prominence of these lexical items indicates that students' discussions centered on multilingual communication, the informational functions of signage, and the practical needs of visitors navigating public spaces. At the same time, the frequent occurrence of research-related vocabulary suggests that students framed their observations within an academic discourse characterized by comparison, interpretation, and evidence-based reasoning as opposed to simple description. Cumulatively, the frequency analysis establishes the principal lexical domains that characterize the corpus. Students' written discussions were dominated by references to multilingual language choice, tourism, information provision, and research-oriented inquiry, providing the lexical foundation for the subsequent analyses of semantic relationships and conceptual organization.

Table 1. Top 100 Most Frequent Words in Students' Discussions and Conclusions

Rank	Word	English Translation	Freq	Rank	Word	English Translation	Freq
1	言語	language	626	51	思う	think	82
2	英語	English	474	52	図	figure	81
3	使用	use	425	53	設置	installation/install	81
4	表記	notation	422	54	大きい	large	81
5	日本語	Japanese	420	55	関係	relationship	80
6	考える	consider	395	56	内容	content	80
7	表示	display	395	57	場合	case	79
8	外国人	foreigner	319	58	対応	treatment/deal with	79
9	多言語	multilingual	304	59	空港	airport	78
10	中国語	Chinese	291	60	結論	conclusion	76
11	沖縄	Okinawa	286	61	対象	target	76
12	見る	see	282	62	理解	understand	74
13	多い	many	281	63	共通	common	73
14	観光客	tourist	272	64	重要	important	73
15	研究	research	232	65	注意	notice	73
16	調査	investigation/investigate	229	66	デザイン	design	71
17	案内	guidance	221	67	ホテル	hotel	71
18	利用	utilize	203	68	今後	future	70
19	情報	information	201	69	記載	describe/record	69
20	写真	photograph	197	70	明らか	clarify	69
21	標識	sign/mark	196	71	サイン	sign	67

22	看板	sign/signboard	180	72	メニュー	menu	66
23	言語景観	linguistic landscape	179	73	繁体字	traditional Chinese characters	66
24	考察	discussion	175	74	海外	overseas	64
25	ピクトグラム	pictogram	173	75	確認	confirmation/confirm	64
26	観光	tourism	173	76	商業	commercial	63
27	施設	establishment	170	77	役割	role	63
28	結果	results	149	78	景観	landscape	62
29	韓国語	Korean	145	79	高い	high	62
30	ブランド	brand	144	80	道路	road	62
31	今回	the present	142	81	意味	meaning	60
32	場所	place	128	82	改善	improve	60
33	分かる	know/comprehend	124	83	特に	particularly	60
34	分析	analysis/analyze	118	84	訪れる	visit	60
35	感じる	feel	113	85	目的	purpose	60
36	簡体字	simplified Chinese characters	113	86	数	number	59
37	地域	region	112	87	用いる	use	59
38	人	people	105	88	異なる	different	58
39	方言	dialect	101	89	主	main	58
40	使う	use	100	90	違い	difference	57
41	行う	conduct	99	91	可能	possible	57
42	必要	necessary	93	92	示す	show	57
43	中国	China	92	93	少ない	few	57
44	店舗	shop	92	94	公共	public	56
45	文字	script	90	95	持つ	hold	56
46	生産	manufacture	89	96	国際	international	55
47	多く	many	88	97	方法	method	55
48	外国	foreign country	87	98	優先	priority	55
49	日本	Japan	84	99	禁止	prohibition/prohibit	54
50	比較	comparison/compare	83	100	台湾	Taiwan	54

4.2 Co-occurrence Network Analysis: Semantic Relationships

The co-occurrence network analysis (Figure 1) identified several prominent clusters of semantically related terms, revealing how students organized and connected ideas throughout their written discussions.

(1) Multilingual language cluster: A central cluster linked the words 英語 (English), 日本語 (Japanese), 中国語 (Chinese), 韓国語 (Korean), 繁体字 (traditional Chinese characters), and 簡体字 (simplified Chinese characters). These terms were closely associated with 表記 (notation) and 使用 (use), indicating that students frequently discussed language choice in relation to practical communicative needs within multilingual environments.

(2) Tourism and foreign visitors cluster: A second major cluster centered on 外国人 (foreigner), 観光客 (tourist), and 沖縄 (Okinawa). These words were strongly associated with 案内 (guidance) and 表示 (display), suggesting that students commonly interpreted LL issues through the communicative needs of international visitors.

(3) Signage and information design cluster: A third cluster grouped together ピクトグラム (pictograms), 表示 (display), and 案内 (guidance). This cluster reflects students' attention to the visual and semiotic design of public signage and its effectiveness in communicating information across linguistic boundaries.

(4) Research and analytical discourse cluster: A fourth cluster contained 調査 (investigation), 研究 (research), and 結果 (results), demonstrating that students consistently employed elements of academic discourse when interpreting their observations and constructing arguments.

Far from representing isolated themes, these clusters illustrate how students connected observations of multilingual signage with broader considerations of communication, usability, and research-based interpretation. The network therefore provides the first indication that students' discourse extended beyond simple description toward analytical engagement with LL phenomena.

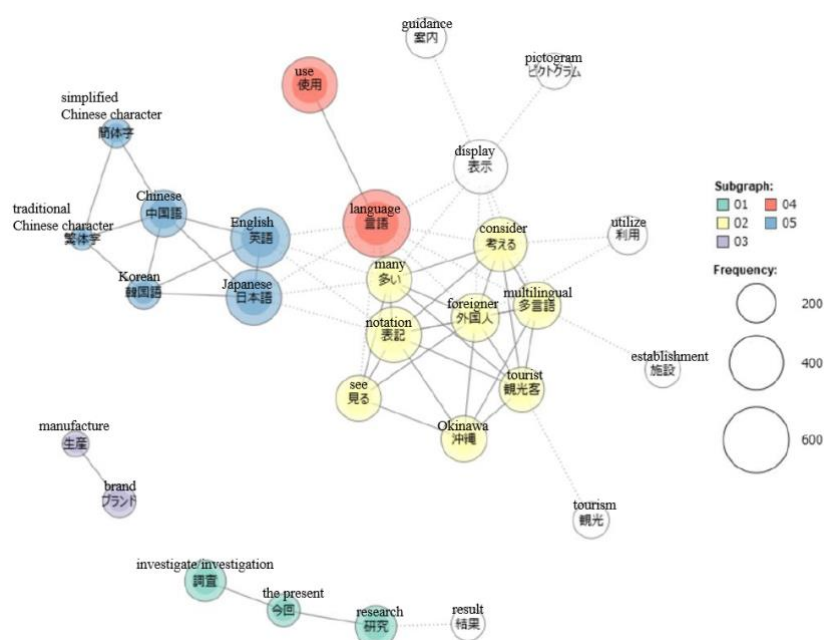


Figure 1. Co-Occurrence Network Graph of High-Frequency Words in Students' Discussions and Conclusions (with English Translations)

4.3 Multidimensional Scaling: Conceptual Organization

The multidimensional scaling analysis (Figure 2) revealed a three-part conceptual structure underlying students' discourse.

(1) Core conceptual anchor: The word 言語 (language) occupied the central position within the semantic space, functioning as the primary conceptual anchor around which other themes were organized.

(2) Language categories: A second cluster contained the specific language names 英語 (English), 日本語 (Japanese), 中国語 (Chinese), and 韓国語 (Korean), together with 繁体字 (traditional Chinese characters) and 簡体字 (simplified Chinese characters). This grouping represents the concrete linguistic categories that students encountered during their field investigations.

(3) Functional communicative context: A third cluster comprised 情報 (information), 観光客 (tourist), 外国人 (foreigner), 案内 (guidance), 看板 (sign/signboard), 利用 (utilize), and 理解 (understand). Collectively, these words represent the practical communicative functions of multilingual signage within public space.

The spatial configuration suggests a conceptual organization in which the abstract concept of language serves as a bridge between specific linguistic varieties and their practical communicative functions. Instead of pointing to separate themes, the multidimensional scaling analysis indicates that students integrated general sociolinguistic concepts with concrete observations of multilingual communication in everyday environments.

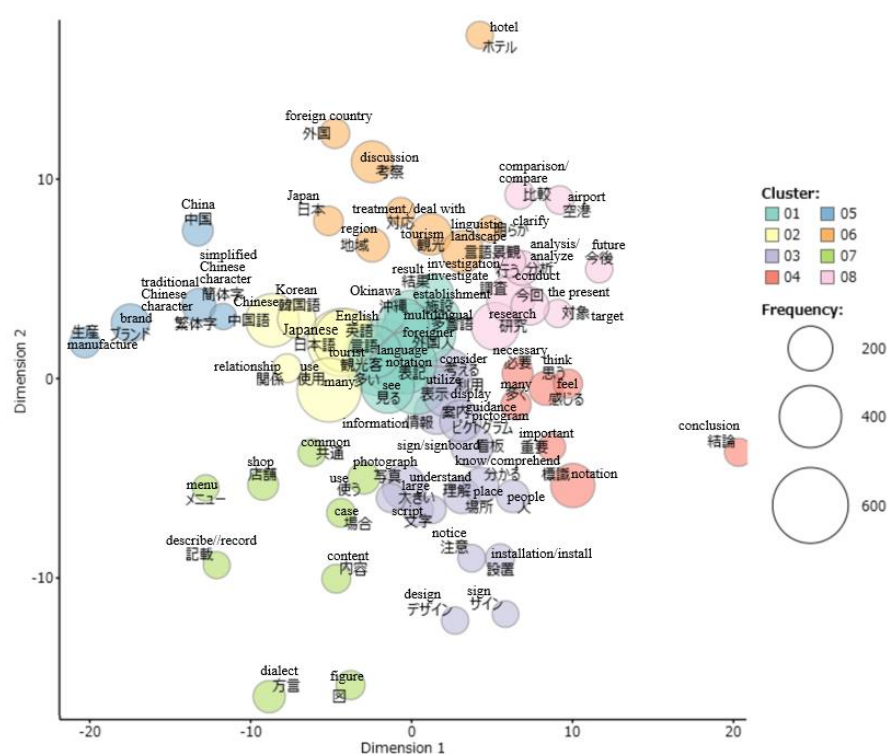


Figure 2. Multidimensional Scaling Representation of High-Frequency Words in Students' Discussions and Conclusions (with English Translations)

4.4 Hierarchical Cluster Analysis: Thematic Domains

The hierarchical cluster analysis (Figure 3) confirmed the thematic patterns identified through the preceding analyses by revealing four higher-order conceptual domains.

(1) Multilingual signage: This cluster comprised 英語 (English), 日本語 (Japanese), 中国語 (Chinese), 韓国語 (Korean), 繁体字 (traditional Chinese characters), and 簡体字 (simplified Chinese characters).

(2) Tourism and foreign visitors: This cluster grouped together 外国人 (foreigner), 観光客 (tourist), and 沖縄 (Okinawa), highlighting students' tendency to interpret multilingual signage within a tourism-oriented context.

(3) Information design and communicative usability: This cluster included 案内 (guidance), 表示 (display), 多言語 (multilingual), and ピクトグラム (pictograms), emphasizing the communicative effectiveness and accessibility of public signs.

(4) Analytical reflection: This cluster consisted of 調査 (investigation/investigate),

比較 (comparison/compare), 分析 (analysis/analyze), 結果 (results), and 研究 (research), indicating sustained engagement with academic reasoning and systematic inquiry.

The close correspondence between these thematic domains and the semantic clusters identified in the co-occurrence network analysis demonstrates strong convergence across the analytical procedures. Because they do not merely represent findings unique to any single technique, these recurring thematic structures appeared consistently across the triangulated text-mining framework, strengthening confidence that they reflect stable characteristics of the student corpus.

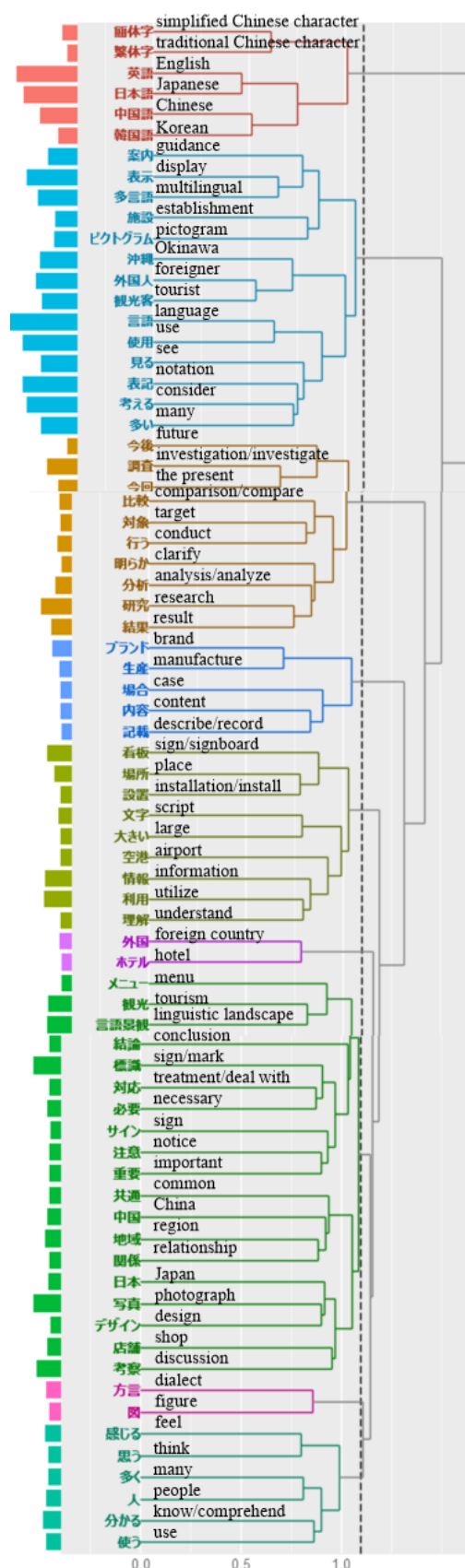


Figure 3. Hierarchical Cluster Analysis Dendrogram of Frequently Occurring Words in Students' Discussions and Conclusions (with English Translations)

5. Discussion

The triangulated text-mining framework employed in this study revealed four consistent corpus-level patterns across students' discussions and conclusions. First, students interpreted LLs primarily through their communicative and informational functions, particularly in relation to multilingual accessibility for foreign visitors. Second, tourism emerged as the dominant interpretive frame through which students understood language use in public space. Third, students consistently employed research-oriented discourse characterized by comparison, analysis, and evidence-based reasoning. Finally, students organized their discussions around conceptual links between abstract notions of language, specific linguistic varieties, and practical communicative functions. The convergence of these findings across four complementary analytical procedures demonstrates the value of corpus-based text mining for systematically identifying recurring discourse patterns within large collections of student writing.

5.1 Methodological Contributions

Beyond its substantive findings, the principal contribution of this study is methodological. Previous evaluations of LL pedagogy have relied predominantly on qualitative analyses of interviews, reflective journals, classroom observations, or selected excerpts from student writing. While these approaches provide rich insights into students' experiences and interpretive processes, they are less well suited to identifying broader discourse patterns across large cohorts.

The present study demonstrates that student writing can be reconceptualized as a research corpus through which collective patterns of learning-related discourse become visible. Without displacing qualitative inquiry, the triangulated text-mining framework provides an empirical foundation for subsequent interpretive analysis by systematically identifying recurring lexical, semantic, and conceptual structures across an entire corpus. The convergence observed across word frequency analysis, co-occurrence network analysis, multidimensional scaling, and hierarchical cluster analysis provides further corroboration that the identified themes represent stable characteristics of students' written discourse as opposed to isolated observations associated with any single analytical procedure.

This complementary relationship between quantitative and qualitative approaches is particularly important in educational research. Corpus-level analyses reveal what students collectively write about, whereas qualitative analyses are better suited to explaining how students construct those interpretations and why particular perspectives emerge. Viewed holistically, these approaches offer a more comprehensive understanding of learning-related discourse than either methodology alone.

5.2 Functional Orientation and the Informational Function of LLs

Across all four analyses, students consistently framed LLs in terms of their practical communicative functions. High-frequency lexical items such as 情報 (information), 案内 (guidance), 必要 (necessary), and 理解 (understand) indicate that students primarily perceived public signs as resources for facilitating communication. This functional orientation closely aligns with the informational function of LLs originally proposed by

Landry and Bourhis (1997), in which public signage serves to communicate information efficiently to diverse language users.

At the same time, relatively few references were made to the symbolic, ideological, or identity-related dimensions of LLs that have become central to more recent scholarship (Blommaert, 2013; Karpava, 2024; Ruddy & Ben Said, 2015; Shohamy & Gorter, 2009). Far from being a limitation, this emphasis on informational functions may represent a natural stage in students' development as novice researchers. When first encountering LL research, students appear to focus on immediately observable communicative functions before progressing toward more abstract questions concerning language, identity, ideology, and power. This interpretation is also reflected in individual students' written discussions. For example, one student emphasized the complementary relationship between textual language and visual communication:

"The analysis revealed that pictograms and textual language function in a complementary relationship." (S2-22)

Similarly, another student observed:

"Information that users need—such as prohibitions, hazards, transportation information, route maps, and station names—becomes easier to understand through multilingual signs and pictograms, contributing to a more satisfying travel experience." (S6-22)

These excerpts closely reflect the corpus-level pattern identified through the quantitative analyses, illustrating students' emphasis on accessibility, information provision, and the communicative functions of multilingual signage overshadowing its symbolic or ideological dimensions.

These findings suggest that explicit pedagogical scaffolding may help students extend their analyses beyond functional description toward more critical engagement with the ideological and symbolic dimensions of LLs. Such developmental processes are explored in greater depth in the companion qualitative study, which examines how students' interpretations evolve through reflective writing.

5.3 Tourism as an Interpretive Frame

The strong association between multilingual signage and tourism-related vocabulary indicates that students interpreted LLs predominantly through the lens of tourism and public communication. This finding reflects patterns observed in LL studies conducted in tourist destinations, where language choice is often shaped by economic considerations and the need to accommodate international visitors (Gorter, 2006; Cenoz & Gorter, 2009). Terms such as 外国人 (foreigner), 観光客 (tourist), 空港 (airport), ホテル (hotel), メニュー (menu), and 店舗 (store) suggest that students viewed multilingual signage primarily as part of the infrastructure supporting international mobility and visitor services. The finding should also be interpreted within the sociolinguistic context in which the study was conducted. The corpus was derived from students investigating LLs in Okinawa, a region characterized by a substantial tourism industry and highly visible multilingual signage in Japanese, English, Chinese, and Korean. Consequently, the prominence of tourism-related discourse likely reflects both the linguistic environments students encountered and the interpretive frameworks through which they understood those

environments. In this sense, the thematic patterns identified here should not be interpreted as universal characteristics of LL pedagogy but rather as products of a particular educational and sociolinguistic context.

Although the specific thematic distributions may differ in other locations, the analytical framework itself is readily transferable to investigations conducted in urban multilingual settings, minority language communities, heritage language contexts, or regions where issues of language policy and identity are more salient. Future comparative studies may therefore reveal how local linguistic ecologies shape the discourse students produce while simultaneously demonstrating the broader applicability of the methodological framework proposed here.

5.4 Emergence of Research-Oriented Discourse

Another notable finding is the consistent presence of research-related vocabulary, including 分析 (analysis/analyze), 比較 (comparison/compare), 結果 (results), 調查 (investigation/investigate), and 研究 (research). These lexical patterns indicate that students adopted academic discourse conventions and structured their discussions around systematic observation, comparison, interpretation, and evidence-based reasoning to the exclusion of personal opinion alone.

Previous research has suggested that LL projects foster critical thinking, inquiry-based learning, and research skills (Dagenais et al., 2009; Sayer, 2010). The present findings extend these claims by providing corpus-based evidence that students not only collected observational data but also employed the discourse practices associated with introductory academic research. This tendency is also evident in students' reflections on their own research processes. For example, one student critically evaluated the limitations of the investigation:

“Because the photographs were taken randomly throughout the shopping mall, some were not suitable for comparison. This is one aspect of the research design that should be improved.” (S28-25)

Another student similarly reflected on methodological improvement:

“This study relied heavily on objective observation, and the lack of additional data limited the findings. In future research, I would like to collect more materials and gather objective opinions through questionnaires.” (S19-25)

Moving beyond simple descriptions of LL features, these excerpts demonstrate students' engagement with methodological reasoning, critical evaluation, and evidence-based reflection. They therefore complement the corpus-level findings by illustrating how research-oriented discourse was expressed in individual student writing.

The prominence of research-oriented discourse suggests that LL projects function not only as activities promoting sociolinguistic awareness but also as authentic opportunities for students to develop foundational research competencies. Because students formulate questions, collect data, compare observations, evaluate evidence, and justify conclusions, LL assignments provide an accessible introduction to empirical inquiry within applied linguistics.

5.5 Conceptual Organization of Learning-Related Discourse

The multidimensional scaling analysis revealed a conceptual structure in which the abstract notion of language functioned as a central organizing concept linking specific linguistic varieties with practical communicative functions. Although the cross-sectional design of the present study does not permit claims regarding individual developmental trajectories, this conceptual organization is consistent with experiential approaches to language learning in which students connect theoretical concepts with authentic observations made in real-world contexts.

Importantly, the analyses presented here identify collective discourse patterns not direct evidence of learning outcomes. The corpus demonstrates how students expressed their interpretations through written discourse following LL investigations, but it cannot determine the depth of individual conceptual change or explain why particular interpretive frames emerged. These questions require complementary qualitative analyses capable of examining students' reflective meaning-making processes in greater depth.

5.6 Pedagogical and Methodological Implications

The findings have several implications for LL pedagogy. First, the predominance of functional and tourism-oriented discourse suggests that novice researchers naturally attend to the informational functions of public signage. Instructors may therefore wish to incorporate explicit discussion of symbolic meaning, language ideology, identity, and power to encourage more critical engagement with LLs.

Second, the consistent attention to multilingual communication and information design indicates that LL projects effectively promote awareness of accessibility, audience, and communicative effectiveness in multilingual public spaces. Such findings reinforce the value of LL pedagogy as a form of experiential learning that connects classroom concepts with authentic language use.

Third, the frequent use of research-oriented discourse suggests that LL assignments function as valuable introductory research experiences. Students engaged in observation, comparison, analysis, and interpretation while employing discourse practices characteristic of empirical inquiry, thereby supporting the development of foundational academic literacy alongside sociolinguistic awareness.

Beyond LL pedagogy, the present study has broader methodological implications for applied linguistics research. Student writing is routinely collected in educational settings but is typically treated only as assessment evidence. The present study demonstrates that such writing can also be conceptualized as a research corpus capable of yielding systematic insights into learning-related discourse. Because the analytical procedures employed here are transparent, scalable, and readily replicable, the framework may be adapted to analyses of reflective journals, study-abroad reports, project papers, service-learning portfolios, teacher education reflections, and other forms of educational writing commonly investigated in applied linguistics. By complementing rather than replacing qualitative inquiry, corpus-based text mining offers a practical and transferable methodology for examining large collections of student discourse while preserving opportunities for subsequent interpretive analysis.

6. Conclusion

This study introduced and demonstrated a triangulated corpus-based text-mining framework for analyzing student discourse generated through LL projects. By treating student writing as a research corpus instead of viewing it solely as assessment evidence, the study showed how four complementary analytical procedures—word frequency analysis, co-occurrence network analysis, multidimensional scaling, and hierarchical cluster analysis—can be integrated to identify recurring lexical, semantic, and conceptual patterns across a large collection of student papers. The convergence of findings across these analytical procedures illustrates the robustness of the proposed framework and demonstrates its value as a systematic approach to examining learning-related discourse in educational settings.

As previously demonstrated, the analyses revealed four consistent corpus-level patterns. First, students emphasized the functional role of multilingual signage, particularly in relation to communication with foreign visitors. Second, LLs were interpreted primarily through a tourism-oriented framework that foregrounded accessibility and information provision. Third, students consistently employed research-oriented discourse characterized by comparison, analysis, and evidence-based reasoning. Finally, students' conceptual organization linked abstract notions of language with concrete linguistic varieties and practical communicative functions. Broadly considered, these findings demonstrate that corpus-based text mining can systematically reveal collective discourse patterns that might remain difficult to identify through close reading of individual papers alone.

The principal contribution of this study is methodological. While previous research has demonstrated the educational value of LL pedagogy primarily through qualitative analyses of interviews, reflective journals, and selected student reflections, the present study illustrates how student writing itself can be reconceptualized as a research corpus. Rather than replacing qualitative inquiry, the proposed framework complements interpretive approaches by providing a transparent, scalable, and replicable means of identifying corpus-level discourse patterns across large collections of student writing. This sequential relationship between corpus-based text mining and qualitative analysis offers a productive methodological model for educational research in applied linguistics, enabling researchers first to identify recurring patterns across an entire corpus and subsequently to investigate representative cases in greater interpretive depth.

The study also contributes to LL pedagogy by providing empirical evidence that students' written discourse reflects sustained engagement with multilingual communication, language choice, and the communicative functions of public signage. The prominence of research-oriented vocabulary further suggests that LL projects provide authentic opportunities for students to develop analytical reasoning, evidence-based argumentation, and introductory research skills alongside sociolinguistic awareness. These findings reinforce the value of LL pedagogy as an experiential approach that integrates classroom learning with authentic linguistic environments while simultaneously supporting the development of academic literacy.

More broadly, the analytical framework proposed here extends beyond LL pedagogy. Student writing is routinely collected across educational contexts but is seldom examined

systematically as a research corpus. Because the framework presented in this study is transparent, transferable, and readily replicable, it has potential applications to a wide range of educational settings, including analyses of reflective journals, project reports, study-abroad reflections, service-learning portfolios, teacher education narratives, and other forms of learning-related discourse commonly investigated in applied linguistics. In this light, the project stands as both a pedagogical validation of LL fieldwork and a transferable methodological model for macro-level discourse analysis in educational settings.

Several limitations should be acknowledged when interpreting the findings. First, the corpus consisted of student papers produced within a single university course in Japan. Consequently, the discourse patterns identified in this study inevitably reflect the particular sociolinguistic characteristics of the local LL, including the high visibility of multilingual tourism signage and the prominence of Japanese, English, Chinese, and Korean in public spaces. While the analytical framework itself is transferable to other educational contexts, the thematic patterns observed in this corpus should not be assumed to generalize to LL projects conducted in different sociolinguistic environments. Second, the study analyzed only the discussion and conclusion sections of students' papers. This decision was made because these sections contain the richest evidence of interpretation, synthesis, and reflection, making them particularly appropriate for investigating learning-related discourse. Nevertheless, future studies may benefit from examining complete papers to explore how students' observations, analyses, and conclusions develop across different stages of academic writing. Third, corpus-based text mining identifies recurring lexical and conceptual patterns but does not directly capture the processes through which students construct meaning or explain why particular interpretive frames emerge. Although the triangulated analytical framework provides robust evidence of corpus-level discourse patterns, it cannot determine the depth of individual conceptual development or the reasoning underlying students' written interpretations. A companion qualitative study employing reflexive thematic analysis addresses this limitation by examining representative student papers in greater depth and exploring how students articulate their evolving understandings of LLs (Lee, 2026). The dual studies provide complementary perspectives on the educational value of LL pedagogy by combining corpus-level pattern identification with detailed qualitative interpretation.

The present study points to several directions for future research. Comparative studies conducted across different linguistic, cultural, and educational contexts would help determine the extent to which corpus-level discourse patterns vary according to local LLs and pedagogical practices. Such research would contribute to a more nuanced understanding of how contextual factors shape students' interpretations of multilingual public space. Future studies might also examine changes in student discourse over time through longitudinal corpus analyses. Comparing writing produced before and after LL projects, or across multiple stages of a curriculum, could provide valuable insights into the development of sociolinguistic awareness, critical language awareness, and academic discourse.

From a methodological perspective, further work should explore the integration of corpus-based text mining with qualitative analytical approaches in sequential mixed-methods research designs. Corpus-level analyses may first be used to identify dominant

discourse patterns across large datasets, after which qualitative analyses can investigate representative cases in greater depth. Such complementary analytical workflows have the potential to strengthen educational research by combining the breadth of corpus-based analysis with the interpretive richness of qualitative inquiry. Future research may also extend the analytical framework to other forms of educational writing in applied linguistics, including teacher reflections, peer feedback, language learning journals, digital portfolios, and study-abroad narratives. Investigating the applicability of corpus-based text mining across these diverse genres would further establish its value as a versatile methodological approach for examining learning-related discourse.

As LL pedagogy continues to expand across diverse educational contexts, researchers require methodological approaches capable of examining increasingly large and complex collections of student writing systematically and transparently. The present study demonstrates that treating student writing as a research corpus and analyzing it through a triangulated corpus-based text-mining framework provides one such approach. By serving as a complement to, instead of a replacement for, qualitative inquiry, the framework offers a scalable, replicable, and transferable methodology for identifying corpus-level discourse patterns while preserving opportunities for rich interpretive analysis. More broadly, the study illustrates how routinely collected student writing can become a valuable source of evidence for educational research, contributing not only to the continued development of LL pedagogy but also to methodological innovation in applied linguistics and language education.

Acknowledgements

This work was supported by JSPS KAKENHI Grant Number JP25K04333.

Declaration of GenAI Use in the Writing Process

The author used ChatGPT to improve language, clarity, and readability. The author reviewed and edited all AI-generated content as necessary and takes full responsibility for the content of the published article.

Conflict of Interest

The author has no conflict of interest to declare.

References

- Allen, T. J. (2025). Experiential learning in an intercultural communication class in Japan: Doing and reflecting on linguistic landscape group projects. *Language, Culture and Curriculum*, 38(2), 274 – 292. <https://doi.org/10.1080/07908318.2024.2449076>
- Blommaert, J. (2013). Ethnography, superdiversity and linguistic landscapes: Chronicles of complexity. *Multilingual Matters*. <https://doi.org/10.21832/9781783090419>
- Cenoz, J., & Gorter, D. (2009). Language economy and linguistic landscape. In E. Shohamy & D. Gorter (Eds.), *Linguistic landscape: Expanding the scenery* (pp. 55 – 69). Routledge. <https://doi.org/10.4324/9780203930960-14>
- Dagenais, D., Moore, D., Sabatier, C., Lamarre, P., & Armand, F. (2009). Linguistic landscape and language awareness. In E. Shohamy & D. Gorter (Eds.), *Linguistic landscape: Expanding the scenery* (pp. 253 – 269). Routledge. <https://doi.org/10.4324/9780203930960-26>
- Gorter, D. (Ed.). (2006). Linguistic landscape: A new approach to multilingualism. *Multilingual Matters*. <https://doi.org/10.21832/9781853599170>
- Gorter, D., Cenoz, J., & van der Worp, K. (2021). The linguistic landscape as a resource for language learning and raising language awareness. *Journal of Spanish Language Teaching*, 8(2), 161-181. <https://doi.org/10.1080/23247797.2021.2014029>
- Higuchi, K. (2016). A two-step approach to quantitative content analysis: KH Coder tutorial using Anne of Green Gables (Part I). *Ritsumeikan Social Sciences Review*, 52(3), 77 – 91. <https://doi.org/10.34382/00003706>
- Higuchi, K. (2017). A two-step approach to quantitative content analysis: KH Coder tutorial using Anne of Green Gables (Part II). *Ritsumeikan Social Sciences Review*, 53(1), 137 – 147. <https://doi.org/10.34382/00003731>
- Karpava, S. (2024). Multilingual linguistic landscape of Cyprus. *International Journal of Multilingualism*, 21(2), 823 – 861. <https://doi.org/10.1080/14790718.2022.2096890>
- Landry, R., & Bourhis, R. Y. (1997). Linguistic landscape and ethnolinguistic vitality: An empirical study. *Journal of Language and Social Psychology*, 16(1), 23 – 49. <https://doi.org/10.1177/0261927X970161002>
- Lee, E. (2026, September 19-20). *Learning to see the unseen: A thematic analysis of student learning in linguistic landscape pedagogy* [Paper presentation]. The 27th Annual International Conference of the Japanese Society for Language Sciences, Tokyo, Japan.
- Li, J., & Marshall, S. (2020). Engaging with linguistic landscaping in Vancouver's Chinatown: A pedagogical tool for teaching and learning about multilingualism. *International Journal of Bilingual Education and Bilingualism*, 23(8), 925 – 941. <https://doi.org/10.1080/13670050.2017.1422479>
- Rowland, L. (2013). The pedagogical benefits of a linguistic landscape project in Japan. *International Journal of Bilingual Education and Bilingualism*, 16(4), 494 – 505. <https://doi.org/10.1080/13670050.2012.708319>
- Ruddy, R. & Ben Said, S. (Eds.). (2015). Conflict, exclusion and dissent in the linguistic landscape. *Palgrave Macmillan*. <https://doi.org/10.1057/9781137426284>
- Sayer, P. (2010). Using the linguistic landscape as a pedagogical resource. *ELT Journal*, 64(2), 143 – 154. <https://doi.org/10.1093/elt/ccp051>
- Shohamy, E., & Gorter, D. (Eds.). (2009). *Linguistic landscape: Expanding the scenery*. Routledge. <https://doi.org/10.4324/9780203930960>
- Shohamy, E., & Waksman, S. (2009). Linguistic landscape as an ecological arena. In E. Shohamy & D. Gorter (Eds.), *Linguistic landscape: Expanding the scenery* (pp. 313 – 331). Routledge. <https://doi.org/10.4324/9780203930960-30>
- Spinelli, B. (2024). Revisioning the linguistic landscape through a multiliteracies framework: Authorizing learners' personal semiotics in a global space. In G. Guarnieri (Ed.), *The art of teaching Italian* (pp. 92-117). Georgetown University Press. <https://doi.org/10.31389/j.ctv37v9f9t.10>

Author biodata

Enid Lee is Professor in the Department of British and American Language and Culture at Okinawa International University, Japan. She holds a Ph.D. in Linguistics and an M.A. in English as a Second Language from the University of Hawaii at Manoa. She has taught in the areas of English, linguistics, intercultural communication, and multilingual education at the undergraduate and graduate levels in Japan and Hong Kong. Her research interests include additional language teaching, intercultural education, and heritage language maintenance.

E-mail: e.lee@okiui.ac.jp