

LEXICAL IDENTITY AND DISCIPLINARY DIFFERENTIATION IN ENGLISH PROGRAM CURRICULA: A CORPUS-BASED KEYNESS ANALYSIS ACROSS 38 THAI RAJABHAT UNIVERSITIES*

SAPOLACHET PRACHUMCHAI**

(Received: June 19, 2026; Revised: July 10, 2026; Accepted: July 20, 2026)

ABSTRACT

Curriculum design in higher education has traditionally relied on expert judgment rather than systematic linguistic analysis, leaving the lexical coherence and disciplinary identity of course descriptions largely unexamined. This study addresses that gap through a macro level corpus analysis of English course descriptions drawn from 38 Rajabhat universities across Thailand, spanning three professional tracks, namely General English (GE), Business English (BE), and English Education (ED). A corpus of 177,065 words compiled from 4,353 course description files was analyzed using keyness statistics, specifically Log-Likelihood, Log Ratio, and Juilland's D, with the British National Corpus and the Leipzig General Written English corpus serving as reference corpora. The findings show that 95 percent of the top 20 keywords differed significantly across curriculum types ($p < .001$). The BE curriculum was grounded in commercial discourse, the ED curriculum in pedagogical and institutional language, and the GE curriculum in humanities and translation related vocabulary, reflecting distinct English for Specific Purposes (ESP) orientations across the three tracks. A subsequent MANOVA confirmed a multivariate difference with a large effect size for the word cluster comprising "*principles*," "*educational*," and "*management*" (Pillai's Trace = 0.175, $p < .001$). Taken together, these results characterize the lexical differentiation underlying English curriculum discourse and offer a corpus-based framework for curriculum evaluation that can be applied in comparable English as a Foreign Language (EFL) context.

Keywords: Keyness Analysis, Disciplinary Discourse, Corpus Based Curriculum Analysis, Curriculum Evaluation

* Research Article, Faculty of Humanities and Social Sciences, Phetchabun Rajabhat University

* * Assistant Professor, English Program, Faculty of Humanities and Social Sciences, Phetchabun Rajabhat University, Email: sapolachet.pra@pcru.ac.th

Introduction

Course descriptions occupy an ambiguous position in academic institutions. Filed away after accreditation approvals, they are rarely revisited, yet their language performs consequential institutional work. The vocabulary of a course description encodes disciplinary assumptions: what a field is, what graduates should become, and which forms of knowledge a program treats as central or peripheral. In English language education, where the vocabulary of curriculum documents directly overlaps with what is being taught, this matters considerably. A course description that frames language learning as rule mastery implies a pedagogy different from one organized around professional communication or literary analysis, even when both documents claim identical learning outcomes.

However, curriculum development in higher education continues to depend heavily on institutional precedent and faculty consensus rather than on systematic examination of what program documentation actually says (Charttrakul and Damnet, 2021). Corpus linguistics is well positioned to interrupt this pattern: by subjecting large text collections to statistical analysis, corpus methods surface lexical regularities invisible to those working within them (Biber et al., 1998). Applied to curriculum documents, the approach has been used to identify gaps in academic vocabulary coverage (Coxhead, 2000), characterize disciplinary writing across fields (Hyland, 2008), and examine alignment between what curricula claim to foreground and what their language actually emphasizes (Curry, 2024).

Central to this approach is the concept of keyness. A keyword, in corpus linguistic terms, is not simply a frequent word; it is one whose frequency in a target corpus departs from its frequency in a reference corpus, a departure that signals discursive priorities and thematic commitments (Scott, 1997; Baker, 2004). Scott and Tribble (2006) describe keyness as capturing the *aboutness* of a corpus: the terms a discourse cannot do without.

Recent methodological refinements have strengthened the empirical basis for identifying such terms. Log-Likelihood statistics (Dunning, 1993; Rayson & Garside, 2000) provide significance testing without normality assumptions. Effect size metrics, including Log Ratio (Hardie, 2014) and percentage difference (%DIFF; Gabrielatos and Marchi, 2012), prevent inflated statistical claims from masking negligible practical differences, while dispersion measures such as Juilland's *D* (Juilland et al., 1970) confirm genuine corpus-wide recurrence rather than local clustering, though its behavior in very large or unevenly sized corpora warrants caution (Biber et al., 2016).

Nonetheless, keyness analysis has rarely been applied to English program curricula at scale. Existing corpus-based studies in language education tend to focus on single institutions, specific text types, or particular learner populations (Narkprom, 2021; Suriyapee and Pongpairoj, 2022; Metang and Narathakoon, 2025). While these studies establish that keyness analysis can surface discipline-specific vocabulary, their single-site or small-corpus designs leave open whether the lexical patterns they identify generalize beyond the institutions studied, a limitation this study's national-scale sampling is designed to address. Curry (2024) has demonstrated the value of comparative corpus analysis for regional syllabus documents, but such work has remained limited in scope. Missing is a study that examines lexical differentiation across an entire university network, comparing program types with genuinely distinct professional mandates, with sufficient statistical rigor to move from description to confirmation.

This gap carries particular weight in Thailand, where sustained policy investment in English education has not translated into improved outcomes: the country ranked 116th out of 123 countries on the 2025 EF English Proficiency Index (EF Education First, 2025), with undergraduate learners registering the steepest recent decline, a pattern corroborated by recent test-result analyses of Thai undergraduates (Sriwichai et al., 2026). Whether this reflects failures of curriculum design is a question corpus analysis cannot answer alone, but understanding how English curricula are linguistically constructed across the national university system is a necessary precondition for any evidence-based response.

The stakes of this question extend beyond curriculum documentation. When course-description vocabulary diverges from the vocabulary of the professional contexts that graduates will enter, students may complete a program without acquiring the register-specific competencies employers expect, a mismatch that surfaces only after graduation. Identifying such gaps in curriculum language is therefore a first step toward addressing them before they become labor-market problems.

This study examines that question through a macro-level keyness analysis of English course descriptions from all 38 Thai Rajabhat Universities, regional public institutions that collectively produce the majority of English teachers and English-educated professionals outside Thailand's major cities (Jantanukul, 2024). These universities offer three distinct program types: “business” English (BE), English Education (ED), and General English (GE). A corpus of 177,065 words compiled from 4,353 course description files was analyzed against two reference corpora using a validated platform (Pearson $r = 0.998$ against AntConc 4.2.4). Inferential validation through Kruskal-Wallis

tests and MANOVA supports findings that move beyond exploratory description toward statistical confirmation of inter-group differences.

Objectives

1. To compile and characterize a corpus of English course descriptions from all 38 Thai Rajabhat Universities, covering the three program types (General English, “business” English, and English Education).
2. To identify and compare the keyness profiles (Log-Likelihood, Log Ratio, and Juilland's *D*) of each program type against the British National Corpus and the Leipzig General Written English corpus.
3. To statistically test whether the keyword frequency distributions differ significantly across the three program types using Kruskal-Wallis tests and MANOVA.
4. To characterize the disciplinary identity and lexical differentiation of each program type based on concordance and contextual patterns, and to propose a corpus-based framework for curriculum evaluation in EFL higher education contexts.

Conceptual Framework

Figure 1 operationalizes keyness as a route from raw course-description text to an evidence-based account of disciplinary identity. Course descriptions, the three program types (GE, BE, ED), and the two reference corpora form the input layer. Keyness statistics (Log-Likelihood, Log Ratio, and Juilland's *D*), concordance analysis, Kruskal-Wallis tests, and MANOVA constitute the analytical stage, converting raw frequency differences into keyness profiles and statistically confirmed group differences. The outcome layer specifies what this analytical stage is designed to yield: a keyness profile for each program type, statistical confirmation of whether these profiles differ across groups, and a synthesis of both into a corpus-based framework for curriculum evaluation, corresponding to Objectives 2 through 4.

Theoretically, this framework synthesizes two traditions. Corpus linguistics supplies the empirical mechanism, keyness, for detecting lexical patterning at scale (Baker, 2004; Scott & Tribble, 2006). Disciplinary discourse theory (Hyland, 2008) supplies the interpretive lens, treating recurrent vocabulary not as incidental topic choice but as evidence of how a field constructs its object of study and its intended practitioners. Keyness analysis operationalizes disciplinary discourse theory: statistically confirmed lexical divergence becomes the empirical signature of the differing professional identities BE, ED, and GE curricula construct.

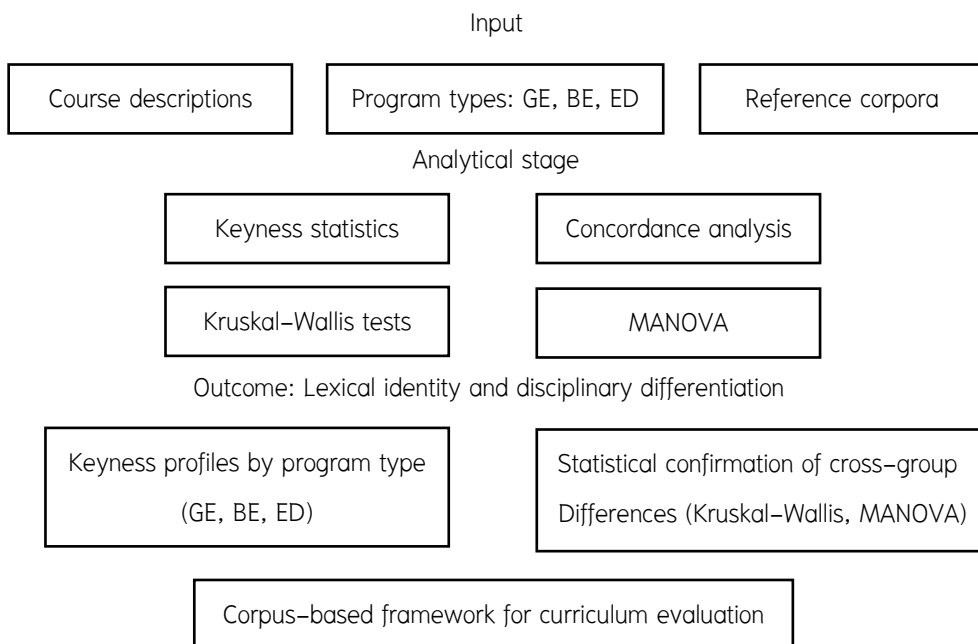


Figure 1 Conceptual framework for the corpus-based keyness analysis of English program curricula

Methodology

1. Corpus Design and Composition

The corpus was compiled from course description files representing all 38 Thai Rajabhat Universities, covering English programs active between 2019 and 2026, a window bounded at the start by Thailand's most recent national curriculum-revision cycle for Rajabhat English programs and extended to 2026 to capture the most currently accredited course descriptions available in CHECO at the time of data collection. Source documents were obtained from Thailand's national curriculum database, the Curriculum Higher Education Comparison Online system (CHECO), which requires all accredited programs to submit documentation verified against national qualification standards. This institutional sampling frame meant that all relevant programs could be included without the selection biases associated with convenience sampling.

Three program types served as the primary analytical categories: “business” English (BE), English Education (ED), and General English (GE). Each carries a distinct professional mandate within Rajabhat curricula, making them the natural basis for cross-program comparison. Table 1 summarizes corpus composition.

Table 1 Corpus composition by program type

Program Type	Files	Universities	Tokens
General English (GE)	1,277	26	39,349
“business” English (BE)	1,267	28	44,102
English Education (ED)	1,809	37	93,614
Total	4,353	38	177,065

The size disparity across program types reflects genuine differences in curriculum scope rather than sampling decisions. Because all keyness metrics used here are normalized against corpus size, the imbalance does not compromise the validity of cross-group comparisons. Three file categories were excluded: course descriptions in Thai or mixed Thai-English (to ensure linguistic consistency), files containing fewer than five words, and general education courses shared by all Rajabhat students irrespective of major, whose language reflects institutional policy rather than program-specific disciplinary orientations. Human verification of a ten-percent sample confirmed tagging accuracy before analysis proceeded.

2. Reference Corpora

Two reference corpora were used. The British National Corpus (BNC; 100 million words) represents formal written British English across a wide range of registers and was compiled in the 1990s (BNC Consortium, 2007), providing a conservative benchmark against established written English norms. The Leipzig General Written English corpus (approximately 2.3 billion words), drawn from web sources and online encyclopedias (Goldhahn et al., 2012), represents contemporary written English across a far broader and more heterogeneous range of sources. Both corpora were selected for their relevance to Thai EFL curriculum documents specifically: the BNC offers a stable, register-controlled benchmark against the formal written English that accreditation documents are expected to emulate, while Leipzig's web-scale, contemporary composition approximates the broader pool of English usage from which Thai curriculum writers, many trained internationally or drawing on online source materials, are likely to draw vocabulary. Neither corpus was compiled with EFL curriculum documents in mind, which is precisely what makes departure from both benchmarks informative.

The dual-reference design provides analytical resolution unavailable through single-reference comparisons. A word that is key against the BNC but not against Leipzig may reflect register differences rather than curriculum-specific lexis; a word that is key against both makes a stronger claim to genuine disciplinary specificity.

3. Keyness Analysis

Keyness statistics were calculated using Log-Likelihood (G^2), following Dunning (1993) and Rayson and Garside (2000). The inclusion threshold was set at $LL \geq 10.83$ ($p < .001$), a more conservative criterion than the commonly used $p < .01$ threshold, appropriate given the number of simultaneous comparisons. Because keyness comparisons were run repeatedly across program types and reference corpora, the Bonferroni correction was applied to control the family-wise error rate across these multiple comparisons.

Statistical significance alone does not establish practical meaning. In a corpus of 177,065 words compared against references of 100 million and 2.3 billion words, modest frequency differences can produce highly significant LL values. Log Ratio (Hardie, 2014) was therefore used as the primary effect size metric, with a minimum threshold of 1.0 (indicating at least twice the relative frequency in the target corpus). Percentage difference (%DIFF; Gabrielatos and Marchi, 2012) was calculated alongside as a complementary expression of effect size. Lexical dispersion was assessed using Juilland's D coefficient, with a minimum threshold of 0.70 applied to exclude words whose frequency derived from concentration in a small number of files rather than genuine corpus-wide recurrence. Only words meeting all four criteria simultaneously (LL threshold, Log Ratio ≥ 1.0 , $D \geq 0.70$, and appropriate relative frequency) were retained for interpretation.

4. Inferential Validation

Keyness analysis is exploratory by nature. Two inferential procedures were used to test whether identified differences were statistically reliable across the full corpus. Kruskal-Wallis tests were applied to the top-20 keywords from each corpus comparison, examining whether word frequency distributions differed significantly across the three program groups. This non-parametric procedure was chosen after Shapiro-Wilk tests confirmed that word frequency distributions did not meet the normality assumption required for ANOVA. Effect sizes were calculated using eta-squared where significant differences were found.

MANOVA was applied to sequential clusters of three keywords, testing whether groups of keywords jointly discriminated between curriculum types at the multivariate level. Box's M test indicated violation of covariance homogeneity ($p < .001$), so Pillai's Trace, rather than Wilks' Lambda, was used as the primary test statistic. Pillai's Trace was chosen because it remains robust even when the assumption of equal covariance across groups is violated, an advantage documented by Olson (1976). Results for both statistics are reported for transparency.

All corpus calculations were performed on a purpose-built platform validated against AntConc 4.2.4 (Anthony, 2024) using a random 100-file sample. Pearson r between the two platforms for Log-Likelihood values was 0.998, mean absolute error for LL was 2.34%, and root mean square error for normalized frequency was 0.21. All three values met the acceptance thresholds set before validation began ($r \geq 0.95$, $MAE \leq 5\%$, $RMSE \leq 0.30$), confirming that the purpose-built platform reproduces AntConc's output closely enough to support the study's conclusions.

Results

1. Shared Lexical Core

Before examining inter-program differences, it is useful to establish what the three program types share. Table 2 presents the fifteen most frequent content words across the full corpus, together with Juilland's D values.

Table 2 High-frequency content words across the full corpus (Top 15 by frequency, $D > 0.95$)

Word	Frequency	Files	D	F/10K
language	1,590	1,072	0.967	89.8
writing	1,384	886	0.963	78.2
business	1,151	618	0.954	65.0
management	1,111	691	0.958	62.8
communication	1,077	766	0.960	60.8
reading	1,059	685	0.955	59.8
principles	872	827	0.968	49.3
learners	865	510	0.947	48.9
educational	837	488	0.946	47.3
various	811	753	0.966	45.8
professional	800	456	0.949	45.2
activities	796	660	0.960	44.9
speaking	738	575	0.956	41.7
information	723	543	0.950	40.8
develop	675	552	0.956	38.1

The uniformly high D values confirm that these words are not artifacts of any single institution's documentation style, each appearing across hundreds of files distributed evenly across program types and universities. Within this shared vocabulary, writing ($F/10K = 78.2$) ranks substantially above speaking (41.7), a distribution that sits in tension with the communicative aims most contemporary English programs articulate. The co-presence of “*management*,” “*principles*,”

and “*professional*” alongside core skills vocabulary reflects the dual orientation of Rajabhat English curricula toward both language learning and professional formation, a pattern that becomes considerably more legible when program types are separated.

2. Keyword Profiles by Program Type

Table 3 presents the top ten keywords for each program type, measured against the combined corpus as reference, and filtered by $LL \geq 10.83$, $\text{Log Ratio} \geq 1.0$, and $D \geq 0.70$.

Table 3 Top 10 keywords by program type (target vs. combined corpus; $LL > 10.83$, $LR > 1.0$, $D > 0.70$)

Rank	BE Keywords	LL	LR	ED Keywords	LL	LR	GE Keywords	LL	LR
1	business	1,491.51	4.12	educational	997.07	5.63	literary	173.89	4.71
2	marketing	383.38	5.84	learners	716.09	3.22	translation	166.24	3.94
3	expressions	296.57	3.41	teacher	710.02	2.48	works	113.52	2.87
4	vocabulary	255.74	2.93	learner	533.20	5.91	selected	103.87	3.12
5	service	207.74	4.22	management	464.08	1.84	American	84.99	4.53
6	tourism	160.71	3.88	design	429.19	2.13	examine	70.15	3.28
7	Chinese	137.42	5.45	activities	354.44	1.62	texts	69.21	2.44
8	hotel	127.61	3.94	classroom	341.59	3.58	formative	67.67	3.81
9	services	127.54	3.76	teachers	289.33	2.19	written	58.27	2.63
10	international	112.78	2.67	individual	272.14	2.74	sentence	58.19	2.11

Three lexical territories emerge with reasonable clarity. BE curricula occupy a domain of commercial activity and service-sector operation, while ED curricula are organized around learner welfare and the professional responsibilities of teaching. GE curricula, by contrast, attend to texts, traditions, and the analytical labor of literary and linguistic study. The Log Ratio values are particularly informative: “*marketing*” ($LR = 5.84$) and “*learner*” ($LR = 5.91$) occur in their respective home corpora at rates between 29 and 60 times their expected general-corpus frequency, figures that mark these as constitutive terms of each program's discourse rather than incidental vocabulary.

Concordance analysis was conducted for the three highest-ranking keywords in each program type. In BE course descriptions, “*business*” functions overwhelmingly as a pre-modifier, appearing in combinations such as “*business correspondence*,” “*business documents*,” and “*business contexts*,” and framing language learning as preparation for genre-specific workplace tasks. “*Marketing*” appears in lists of commercial domains alongside “*accounting*,” “*finance*,” and “*public relations*,” positioning language as a professional medium rather than an object of study. “*Expressions*,” paired consistently with “*vocabulary*” and “*terminology*,” points toward a

pedagogical model organized around the accumulation of usable phrases for defined workplace situations.

ED descriptions construct a different kind of student. “*Educational*” appears almost always in compound form, as in “*educational psychology*,” “*educational quality assurance*,” and “*educational leadership*,” assembling the vocabulary of institutional systems rather than classroom practice. “*Learners*” appears consistently in the direct object position of verbs of professional action: to “*develop learners*,” to “*manage learning for 21st-century learners*.” The learner here is a recipient rather than an agent. “*Teacher*” appears systematically alongside “*ethics*,” “*morality*,” and “*professional spirit*,” embedding the professional role in an explicitly moral and institutional frame.

GE descriptions occupy a third discursive space. “*Literary*” collocates with “*forms*,” “*appreciation*,” “*texts*,” and “*methods*,” positioning language as something to be analyzed from critical distance rather than deployed for professional ends. “*Translation*” pairs with “*strategies*,” “*principles*,” and “*practice*,” but its left-context anchoring in “*literary works*” marks it as humanistic labor rather than professional service. The student constructed by GE descriptions is a reader and analyst: someone who approaches English through its written traditions.

Collectively, these concordance patterns suggest that BE, ED, and GE curricula are not merely discussing different topics. They are positioning students differently in relation to English: as professional users, as emerging teachers, and as critical readers. The consistency of these patterns across 38 institutions indicates structural features of curriculum discourse rather than idiosyncratic choices by individual program committees.

3. Statistical Confirmation

3.1 Kruskal-Wallis Analysis

Kruskal-Wallis tests were conducted on the top-20 keywords from the BNC comparison. Table 4 presents the results for ten representative keywords spanning the range of observed effect sizes.

Table 4 Kruskal-Wallis test results for selected keywords across GE, BE, and ED groups

Keyword	H statistic	p-value	Result	Mean F/10K BE	Mean F/10K ED	Mean F/10K GE
business	841.64	< .001	Significant	336.67	3.54	91.22
educational	712.68	< .001	Significant	1.05	121.14	1.23
learners	580.19	< .001	Significant	10.58	117.87	3.55
management	551.98	< .001	Significant	47.05	166.01	12.65
vocabulary	295.23	< .001	Significant	117.65	14.89	73.79
translation	67.23	< .001	Significant	56.46	45.03	98.26
writing	95.30	< .001	Significant	180.79	94.50	173.95
speaking	69.21	< .001	Significant	105.79	65.04	80.06
literary	88.14	< .001	Significant	2.31	1.18	38.72
contexts	1.48	.477	Not significant	54.99	45.33	65.01

Nineteen of the twenty keywords tested returned significant results after the Bonferroni correction (adjusted $\alpha = 0.0025$). The single exception, “*contexts*,” distributed evenly across all three program types, a finding that is itself interpretively useful. “*Contexts*” functions as a disciplinary placeholder, broad enough to appear wherever language learning is being described and specific enough to belong to no single professional territory. The mean F/10K values quantify the group differences in concrete terms: “*business*” appears in BE curricula at a rate more than 94 times higher than in ED, while “*educational*” shows the reverse pattern, appearing in ED at more than 115 times its BE rate. “*Translation*” is the only keyword in the top twenty whose highest mean frequency belongs to GE, at nearly twice the rate of its nearest competitor.

3.2 MANOVA

MANOVA was conducted on three sequential keywords drawn from the top-20 list. Table 5 presents the results for all six keyword clusters examined.

Table 5 MANOVA results for keyword clusters, GE vs. BE vs. ED

Cluster	Keywords	Pillai's Trace	F	p	Effect size
1	language, learners, writing	0.125	92.19	< .001	Medium
2	learners, writing, communication	0.126	93.29	< .001	Medium
3	writing, communication, reading	0.025	17.22	< .001	Small
4	communication, reading, principles	0.020	14.17	< .001	Small
5	reading, principles, educational	0.125	92.09	< .001	Medium
6	principles, educational, management	0.175	132.62	< .001	Large

All six clusters produced significant results, with effect sizes ranging from small to large, a pattern that itself merits interpretation. Clusters dominated by widely shared vocabulary (“*writing*,” “*reading*,” “*communication*”) produced smaller multivariate effects because these words, while individually differentiated, do not jointly mark any single program type with sufficient force to generate strong group separation. The large effect in Cluster 6 reflects a qualitatively different phenomenon: “*principles*,” “*educational*,” and “*management*” do not merely differ in frequency across program types; they configure together differently in each curriculum, appearing in ED with a density and combination that distinguishes it from BE and GE more powerfully than any word alone. This suggests that disciplinary identity in these curricula is carried not only by which words appear, but by which words appear together and at what intensity. Institutionally, this configuration suggests that ED programs have converged on a distinct discourse of professional formation, one in which pedagogical principles, educational institutions, and managerial oversight are not separate topics but a single, tightly bundled way of talking about what a teacher is being prepared to do. For curriculum committees, this bundling is itself a marker of program maturity: BE and GE curricula show no comparably tight three-word convergence, suggesting their professional identities remain comparatively more diffusely encoded.

Discussion

1. Disciplinary Identity as Lexical Commitment

The central finding of this study is not simply that BE, ED, and GE curricula use different words. The finding is that these differences are systematic, statistically robust, and consistent enough across 38 institutions to constitute something more than local variation. When “*business*” appears in BE course descriptions at a rate far exceeding its ED frequency (Table 4), the word is not being chosen by faculty making independent decisions. It is being reproduced by a curriculum discourse that has settled into stable conventions about what business English programs are for and what vocabulary belongs in them.

This is what Scott and Tribble (2006) describe as aboutness: not the topics a text mentions, but the lexical commitments it makes persistently enough to define its orientation. By that criterion, each program type in this study carries a defensible and legible disciplinary identity: BE is committed to the functional use of language in commercial contexts, ED to the professional formation of teachers within institutional systems, and GE to the analytical engagement with literary and translational traditions. These are not merely thematic differences but different theories of what

English is and what it is for, encoded in routine vocabulary that curriculum designers may not consciously choose but consistently reproduce.

The syntactic patterns identified in concordance analysis amplify this point, consistent with Hyland's (2008) observation that disciplinary identity in academic texts is carried as much by grammatical patterning and collocation as by topical content. In BE, "*business*" operates as a pre-modifier that frames English as an instrument for professional performance; in ED, *learners* occupies the direct-object position of professional action verbs, constructing teacher candidates as managers of others' development; and in GE, "*literary*" works recurs as the object of analytical verbs such as interpret, analyze, and criticize, positioning the student as someone who approaches language through its written traditions rather than its professional functions.

These lexical commitments carry direct implications for course design. If BE course descriptions are saturated with commercial vocabulary but assessment and learning outcomes remain generic language-skill benchmarks, the documented curriculum promises a professional orientation the assessed curriculum may not deliver. Aligning stated learning outcomes with the disciplinary vocabulary a program's own documentation already uses is a low-cost first step toward closing that gap.

2. The Significance of Shared Vocabulary

The even distribution of "*contexts*" across all three program types points to a different kind of lexical category: a shared pedagogical vocabulary that curriculum designers draw on regardless of program type, broad enough to describe the situatedness of language without signaling any particular professional orientation. Words of this kind form the common ground of curriculum discourse, while words like "*business*", "*educational*," and "*literary*" serve as diagnostic markers of disciplinary specificity. The gap between these two registers is where curriculum coherence, or its absence, is most likely to be found in practice.

3. Scope and Limitations

This study analyzes the language of curriculum documentation rather than the reality of curriculum delivery. A course description organized around 21st-century learning vocabulary does not guarantee instruction that reflects those values. Charttrakul and Damnet (2021) found precisely this kind of gap in their survey of CEFR implementation across Rajabhat institutions, where teacher understanding of the framework did not translate reliably into classroom practice. The present findings describe how these institutions construct their programs in language. A full account of whether that construction corresponds to instructional reality would require methods such as

classroom observation, student outcome analysis, and graduate surveys, which corpus linguistics alone cannot provide.

This study also establishes that BE, ED, and GE curricula are linguistically differentiated, but it does not assess whether that differentiation is appropriate or well matched to professional labor market demands. “*marketing*” appearing in BE curricula alongside *airline services* and *public relations* reveals something about how these programs conceptualize business English. It does not indicate whether that conceptualization serves graduates entering a labor market increasingly shaped by digital commerce, platform-based communication, and cross-cultural remote collaboration. Triangulating corpus findings with employer needs data and graduate employment outcomes remains a direction for future research.

4. Methodological Transferability

The analytical framework applied here is not specific to Thailand or to Rajabhat universities. Regional public universities with differentiated English program mandates exist across Southeast Asia, South Asia, and much of the Global South, and they face comparable challenges: limited resources for curriculum review, dependence on expert judgment over empirical evidence, and difficulty maintaining coherence across geographically dispersed institutions. The dual-reference-corpus design, comparing a specialized target corpus against both a formal national corpus and a broader contemporary reference, offers a transferable approach to characterizing curriculum discourse in any context where English programs serve distinct professional ends.

Curry and McEnergy's (2025) analysis of syllabus documents from regional universities in China and India identified similar patterns of lexical inconsistency between curriculum documentation and professional language demands, suggesting that the issues observed in the present study are not unique to the Thai context. Whether the lexical profiles identified here, namely BE's commercial orientation, ED's pedagogical institutionalism, and GE's humanistic textuality, are replicated across comparable university systems remains an empirical question for future research.

This study's findings converge on a replicable three-step framework for corpus-based curriculum evaluation: (1) compile course-description corpora by program type and validate them against an established analysis platform; (2) generate keyness profiles against both a formal and a contemporary reference corpus, retaining only words meeting combined significance, effect-size, and dispersion thresholds; and (3) confirm profile differences statistically via non-parametric group tests and multivariate analysis before drawing disciplinary conclusions. Curriculum committees can apply this same three-step sequence to their own program documentation to audit whether

intended professional orientations are, in fact, what the language of their course descriptions communicates.

Recommendations

For Practice

Rather than prescriptively adopting the specific keywords identified here, curriculum committees across the Rajabhat network should treat the lexical profiles reported in this study as diagnostic benchmarks. Educational stakeholders can conduct localized keyness analyses of their own course descriptions and compare the results against these benchmarks to audit whether their documentation accurately reflects its intended professional orientation (BE: “business,” “marketing,” “tourism,” “hotel”; ED: “educational,” “learners,” “teacher,” “design”; GE: “literary,” “translation,” “works,” “American”) or contains unintended lexical gaps.

For Future Research

Future research should conduct a systematic comparative validation of the corpus analysis platform against established tools such as LancsBox X, Sketch Engine, and spaCy, examining accuracy, processing speed, tokenization and part-of-speech tagging, and metadata handling, to establish full criterion validity.

References

- Anthony, L. (2024). *AntConc* (Version 4.2.4) [Computer software]. Waseda University. <https://www.laurenceanthony.net/software/antconc>
- Baker, P. (2004). Querying keywords: Questions of difference, frequency, and sense in keyword analysis. *Journal of English Linguistics*, 32(4), 346–359.
- Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus linguistics: Investigating language structure and use*. Cambridge University Press.
- Biber, D., Reppen, R., Schnur, E., & Ghanem, R. (2016). On the (non)utility of Juilland's D to measure lexical dispersion in large corpora. *International Journal of Corpus Linguistics*, 21(4), 439–464. <https://doi.org/10.1075/ijcl.21.4.01bib>
- BNC Consortium. (2007). *The British National Corpus* (Version 3, BNC XML edition). Oxford University Computing Services. <http://www.natcorp.ox.ac.uk/>
- Charttrakul, K., & Damnet, A. (2021). Role of the CEFR and English teaching in Thailand: A case study of Rajabhat universities. *Advances in Language and Literary Studies*, 12(2), 82–89.
- Coxhead, A. (2000). A new academic word list. *TESOL Quarterly*, 34(2), 213–238.

- Curry, N., & McEnery, T. (2025). Corpus linguistics for language teaching and learning: A research agenda. *Language Teaching*, 58(2), 232–251. <https://doi.org/10.1017/S0261444824000430>
- Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1), 61–74.
- EF Education First. (2025). *EF English Proficiency Index 2025*. <https://www.ef.com/wwen/epi/>
- Gabrielatos, C., & Marchi, A. (2012, September). *Keyness: Appropriate metrics and practical issues* [Paper presentation]. CADS International Conference, University of Bologna, Bologna, Italy.
- Goldhahn, D., Eckart, T., & Quasthoff, U. (2012). Building large monolingual dictionaries at the Leipzig Corpora Collection: From 100 to 200 languages. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)* (pp. 759–765). European Language Resources Association.
- Hardie, A. (2014). *Log ratio: An informal introduction* [Unpublished working paper]. ESRC Centre for Corpus Approaches to Social Science, Lancaster University. <http://cass.lancs.ac.uk/log-ratio-an-informal-introduction/>
- Hyland, K. (2008). *Academic discourse: English in a global context*. Continuum.
- Jantanukul, W. (2024). Rajabhat University concept for local development: Strategies, implementation, and community impact. *Journal of Education and Learning Reviews*, 1(6), 47–60.
- Juilland, A., Brodin, D., & Davidovitch, D. (1970). *Frequency dictionary of French words*. Mouton.
- Metang, P., & Narathakoon, A. (2025). A corpus-based study of lexical bundles of Keywords Found in Online News Articles. *THAITESOL Journal*, 38(1), 1–21.
- Narkprom, N. (2021). *A comparative corpus study on lexical bundles and keywords* [Master's thesis, Thammasat University]. TU Digital Collections.
- Olson, C. L. (1976). On choosing a test statistic in multivariate analysis of variance. *Psychological Bulletin*, 83(4), 579–586.
- Rayson, P., & Garside, R. (2000). Comparing corpora using frequency profiling. In *Proceedings of the ACL Workshop on Comparing Corpora* (pp. 1–6). Association for Computational Linguistics.
- Scott, M. (1997). PC analysis of key words – and key key words. *System*, 25(2), 233–245.
- Scott, M., & Tribble, C. (2006). *Textual patterns: Key words and corpus analysis in language education*. John Benjamins.
- Sriwichai, C., Duangfai, C., Paicharoen, N., & Yuensak, S. (2026). *Insights into English proficiency: The analysis of test results among undergraduates in Thailand*. *Indonesian Journal of Applied Linguistics*, 15(3), 630–644. <https://doi.org/10.17509/hv44am08>



Suriyapee, P., & Pongpairoj, N. (2022). Corpus to enhance verbal complements among low English proficiency Thai learners. *PASAA: Journal of Language Teaching and Learning in Thailand*, 63, 205–231. <https://doi.org/10.58837/CHULA.PASAA.63.1.8>