



Reassessing keyword representativeness in WoS-indexed articles related to corpus linguistics: A Biblioshiny and Gries-style analysis

Myounggho Ha
(Silla University)

Ha, Myounggho. 2026. Reassessing keyword representativeness in WoS-indexed articles related to corpus linguistics: A Biblioshiny and Gries-style analysis. *Linguistic Research* 43(Special Edition): 99-119. This study reassesses keyword representativeness in Web of Science-indexed articles related to corpus linguistics by combining Biblioshiny-based bibliometric analysis with an adapted Gries-style frequency-dispersion framework. Bibliographic data were retrieved from the Web of Science Core Collection on June 22, 2026. After post-export validation and keyword-field cleaning, 4,270 English-language journal articles published between 2001 and 2025 were retained. Three keyword fields were compared: Author Keywords, Keywords Plus, and All Keywords, defined as the cleaned record-level union of the two fields. Biblioshiny's Most Frequent Words output was used to select the top 50 candidate terms in each field. These terms were then reassessed using record frequency, year span, period coverage, source span, maximum source share, and DKL-based concentration scores for years, five-year windows, and sources. The adapted Gries-style procedure is used here as a within-dataset assessment of distributional representativeness rather than as a target-versus-reference keyness analysis. The results show that the three keyword fields provide complementary profiles: Author Keywords foreground author-selected and retrieval-related terms, Keywords Plus emphasizes broader indexing-derived descriptors, and All Keywords provides the widest overall coverage. However, high frequency does not necessarily imply broad temporal or source distribution. Keyword representativeness should therefore be evaluated through both frequency and dispersion. (Silla University)

Keywords bibliometric analysis, Biblioshiny, Author Keywords, Keywords Plus, frequency-dispersion analysis, Web of Science, corpus linguistics

1. Introduction

Bibliometric analysis provides systematic methods for examining publication patterns, conceptual structures, and research trends in large bodies of scholarly literature (Župič and Čater 2015; Donthu et al. 2021). Among bibliographic metadata, keywords are particularly important because they provide concise representations of article content and are widely used in frequency analysis, co-word analysis, and thematic mapping. Tools such as bibliometrix and its web interface, Biblioshiny, have further facilitated keyword-based bibliometric analysis (Aria and Cuccurullo 2017; Moral-Muñoz et al. 2020).

Keyword fields in Web of Science, however, are not interchangeable. Author Keywords reflect terms selected by authors to describe their studies, whereas Keywords Plus is generated through an indexing process based on cited-reference information (Garfield and Sher 1993). The two fields may therefore foreground different aspects of a research domain. A merged field can increase coverage, but greater coverage does not necessarily imply greater representativeness.

A second issue concerns the interpretation of keyword frequency. A highly frequent keyword may be broadly distributed across a dataset, but it may also be concentrated in particular years, periods, or sources. Corpus-linguistic research has long emphasized that frequency should be considered together with dispersion because raw frequency alone does not reveal how widely an item is distributed (Gries 2008, 2021; Egbert and Biber 2019). This distinction is particularly relevant to bibliometric keyword analysis, where frequent terms may reflect retrieval conditions, indexing practices, or localized research concentrations.

The present study compares Author Keywords, Keywords Plus, and All Keywords in Web of Science-indexed articles related to corpus linguistics. Biblioshiny's Most Frequent Words output is first used to identify the top 50 candidate terms in each keyword field. These candidates are then reassessed using an adapted Gries-style frequency-dispersion framework. Here, Gries-style refers to an adaptation of Gries's frequency-dispersion logic rather than to his original target-versus-reference keyness procedure. The present analysis evaluates how broadly bibliographic keywords are distributed across publication years, fixed five-year windows, and sources within the retrieved dataset.

The study addresses the following three research questions.

1. How do Author Keywords, Keywords Plus, and All Keywords differ in their frequent keyword profiles?
2. To what extent are the most frequent keywords broadly distributed across years, fixed five-year windows, and sources, and which keywords exhibit temporal or source-based concentration?
3. What does the frequency-dispersion comparison reveal about keyword representativeness across the three keyword fields?

By addressing these questions, the study distinguishes frequency-based visibility from distributional representativeness and evaluates the three keyword fields under a common analytical framework.

2. Literature review

2.1 Bibliometric analysis and keyword-based mapping

Bibliometric analysis is widely used to examine publication growth, citation patterns, collaboration, source distributions, and conceptual structures. It provides reproducible quantitative methods for mapping large research domains and identifying major patterns in scientific literature (Župič and Čater 2015; Donthu et al. 2021). Science-mapping approaches further examine relationships among authors, documents, sources, and terms, allowing the intellectual and thematic structure of a field to be investigated (Cobo et al. 2011a, 2011b).

Keywords play a central role in this process because they provide concise descriptors of publication content. Co-word and keyword-frequency analyses use these descriptors to identify visible concepts and relationships within a research domain (Callon et al. 1983). However, the interpretation of keyword results depends on both the source of the keyword metadata and the analytical measure used.

2.2 Bibliometrix, Biblioshiny, and keyword fields

Bibliometrix is an open-source R package for quantitative research in bibliometrics and science mapping (Aria and Cuccurullo 2017), while Biblioshiny provides a

graphical interface for conducting many of its analyses (Moral-Muñoz et al. 2020). In the present study, Biblioshiny is used to identify frequent candidate keywords before their distributional profiles are reassessed.

Web of Science Author Keywords and Keywords Plus differ in their origins and descriptive functions. Author Keywords are supplied by authors, whereas Keywords Plus was developed as an algorithmic derivative indexing mechanism based on cited-reference information (Garfield and Sher 1993). Previous research has shown that these fields can differ in descriptive scope and thematic emphasis (Zhang et al. 2016). Recent research in linguistics likewise indicates that Author Keywords and Keywords Plus can foreground different aspects of the same literature (Lee and Jhang 2026). These differences support analyzing the fields separately rather than treating them as interchangeable.

The present study additionally examines All Keywords, defined as the cleaned record-level union of Author Keywords and Keywords Plus. This merged field increases keyword coverage, but its representativeness must still be evaluated rather than assumed from its greater inclusiveness.

2.3 Frequency, dispersion, and representativeness

Frequency is a fundamental measure in corpus and bibliometric analysis, but it does not indicate how evenly an item is distributed. A term may occur frequently because its occurrences are concentrated in a limited number of corpus parts. Dispersion therefore complements frequency by measuring distributional breadth (Gries 2008). Egbert and Biber (2019) similarly demonstrate that corpus-frequency keyword analysis may identify terms that are frequent but unevenly distributed, while Gries (2021) argues for retaining frequency and dispersion as analytically distinct dimensions.

A closely related methodological precedent is Liu (2025), who extended Gries's frequency-dispersion approach to bibliometric research by applying it to corpus-driven bigrams extracted from article abstracts and combining selected bigrams with Author Keywords for research-trend analysis. Liu's study demonstrates the value of combining frequency and dispersion in bibliometric research. The present study, however, addresses a different analytical problem: it neither extracts nor adds n-grams from article texts nor performs target-versus-reference corpus keyness. Instead, it evaluates

existing Web of Science keyword metadata through within-dataset temporal and source dispersion.

Accordingly, keyword frequency in the present study refers to the number of unique bibliographic records containing a keyword, while dispersion refers to the distribution of those records across publication years, five-year windows, and sources. Representativeness is therefore understood specifically as distributional representativeness within the retrieved dataset.

2.4 Research gap and contribution

Bibliometric keyword studies commonly characterize research fields through keyword frequency, co-occurrence, network structure, or thematic mapping (e.g., Araújo et al. 2024; Lee and Jhang 2026). Although recent work has demonstrated that corpus-linguistic frequency-dispersion techniques can contribute to bibliometric analysis (Liu 2025), less attention has been given to using distribution-sensitive measures to compare existing bibliographic keyword fields themselves.

The present study addresses this gap by evaluating Author Keywords, Keywords Plus, and their record-level union under a common within-dataset framework. Its contribution is therefore not to claim the first transfer of frequency-dispersion analysis to bibliometrics, but to apply distribution-sensitive analysis to the comparative assessment of bibliographic keyword fields. This provides a more cautious basis for interpreting keyword importance than frequency alone.

3. Methodology

3.1 Data search and collection

Bibliographic data were collected from the Web of Science Core Collection on June 22, 2026. The search was restricted to SCI-EXPANDED, SSCI, A&HCI, and ESCI. The Web of Science Categories field was used to restrict the disciplinary scope to Linguistics or Language & Linguistics, and the Topic field was used to identify corpus-related studies. The search query was: WC=(Linguistics OR "Language & Linguistics") AND TS=("corpus linguistics" OR "corpus-based" OR "corpus based" OR

"corpus driven" OR "corpus-driven" OR "learner corpus" OR "learner corpora" OR "corpus analysis") AND PY=(2001-2025) AND DT=(Article) AND LA=(English).

The initial export contained 4,611 records, and no duplicate Web of Science accession numbers (UT) were found. Post-export validation identified 98 records coded as PY = 2026, which were excluded to enforce the intended 2001–2025 time span. A further 243 records contained neither Author Keywords nor Keywords Plus. The two exclusion sets did not overlap, leaving 4,270 records in the final keyword-analysis dataset.

3.2 Data cleaning and keyword preparation

The title and abstract fields received only formatting corrections, including the removal of line breaks, tabs, control characters, and redundant spaces. More active normalization was limited to the keyword fields. Author Keywords (DE) and Keywords Plus (ID) were standardized through whitespace and delimiter normalization, lowercasing, within-record duplicate removal, and limited morphological normalization.

Before the final analyses, keyword forms affecting the reported top-50 candidate sets were cross-checked against the raw Web of Science export. Clearly distorted transformations introduced during normalization were restored to their source forms, and the resulting keyword fields were used for the final Biblioshiny analysis and all subsequent calculations.

After DE and ID had been prepared, KW_Merged, referred to as All Keywords, was generated as their record-level union. Duplicate terms within the same record were removed. Of the 4,270 final records, 3,966 contained Author Keywords, 2,966 contained Keywords Plus, and all 4,270 contained at least one term in All Keywords. These field-specific counts were used as denominators for relative-frequency and dispersion calculations.

3.3 Biblioshiny analysis and candidate keyword selection

The cleaned master dataset was imported into Biblioshiny for descriptive bibliometric analysis. Keyword analyses were conducted separately for Author Keywords, Keywords

Plus, and All Keywords, using DE, ID, and KW_Merged, respectively.

For each field, the top 50 terms from Biblioshiny's Most Frequent Words output were selected as candidate keywords. These lists were used only for candidate selection. All frequency and dispersion measures reported in the subsequent analysis were recalculated directly from the corrected master dataset. Because several high-frequency terms were also included in the retrieval strategy, their frequencies were interpreted in relation to the way the dataset was constructed.

3.4 Gries-style frequency-dispersion analysis

The Gries-style analysis was implemented as a within-dataset assessment of bibliographic keyword representativeness. Unlike Gries's original target-versus-reference keyness procedure, the present study does not test statistical overrepresentation relative to an external reference corpus. Instead, it evaluates how broadly the records containing each keyword are distributed within each keyword field.

Record frequency was defined as the number of unique records containing a keyword. Relative frequency was calculated using the corresponding field-specific denominator: 3,966 records for Author Keywords, 2,966 for Keywords Plus, and 4,270 for All Keywords. Year span measured the number of publication years in which a keyword occurred, source span measured the number of sources containing it, and maximum source share measured the proportion of its records contributed by the single most frequent source.

Period coverage was calculated using five fixed five-year windows: 2001–2005, 2006–2010, 2011–2015, 2016–2020, and 2021–2025. These windows are not interpreted as empirically distinct historical phases. Data-driven periodization may be appropriate when the objective is to identify substantive stages of thematic development (Lee et al. 2025; Liu 2025); here, equal-duration windows are used only as coarse-grained units of temporal dispersion. Differences in publication volume across the windows are incorporated into the field-specific baseline used in the concentration calculation.

For publication year, period, and source distributions, Kullback-Leibler divergence was calculated as:

$$D_{KL}(P||Q) = \sum_{i: p_i > 0} p_i \log_2 \left(\frac{p_i}{q_i} \right)$$

Here, p_i is the proportion of a keyword's records occurring in corpus part i , and q_i is the proportion of all records in the corresponding keyword field occurring in that same part.

For easier comparison, DKL was transformed into a bounded concentration score:

$$C_{DKL} = 1 - \exp(-D_{KL}(P||Q))$$

Values closer to 0 indicate distributions more similar to the relevant field-specific baseline, whereas larger values indicate greater distributional deviation from that baseline. The transformed score is therefore interpreted here as a baseline-relative indicator of concentration rather than as an absolute measure of concentration. It is used descriptively and comparatively and is not interpreted against a fixed cutoff or as an inferential significance statistic.

Finally, the Biblioshiny frequency rankings were compared with the corresponding frequency-dispersion profiles to determine whether highly ranked keywords were also broadly distributed across publication years, five-year windows, and sources. The comparison also identified frequent terms showing stronger temporal or source-based concentration and examined whether these patterns differed across Author Keywords, Keywords Plus, and All Keywords. No keyword field was treated as inherently superior; instead, the three fields were compared in terms of the complementary information they provide about keyword visibility and distributional representativeness.

4. Analysis and discussion

4.1 Overview of the keyword-analysis dataset

The final keyword-analysis dataset consisted of 4,270 English-language journal articles published between 2001 and 2025 and distributed across 394 sources. The annual number of records rose markedly over the study period, from 9 records in 2001 to

502 records in 2025. By five-year window, the dataset contained 45 records in 2001–2005, 250 in 2006–2010, 594 in 2011–2015, 1,142 in 2016–2020, and 2,239 in 2021–2025. Thus, more than half of the records were published in the most recent window.

The most common sources included major journals in corpus linguistics and applied linguistics, such as *Journal of Pragmatics*, *International Journal of Corpus Linguistics*, *Corpus Linguistics and Linguistic Theory*, *Journal of English for Academic Purposes*, and *English for Specific Purposes*. The strong growth in publication volume also means that the five temporal windows differ substantially in size. As explained in Section 3.4, these fixed five-year windows are used as analytical units for coarse-grained temporal dispersion rather than as empirically discrete historical phases. Their unequal record volumes are incorporated into the field-specific baseline distribution used in the DKL calculation. Accordingly, temporal concentration is assessed relative to the observed publication distribution rather than to an assumption of equal-sized periods.

4.2 Frequent keyword profiles across the three keyword fields

The analysis first compared the three keyword fields on the basis of frequency. Biblioshiny's Most Frequent Words output identifies the terms that are most visible in each field, but frequency is treated here as an initial descriptive measure rather than as direct evidence of representativeness. Table 1 presents the ten most frequent terms in Author Keywords, Keywords Plus, and All Keywords. The complete top-50 lists were retained as the candidate sets for the subsequent frequency-dispersion analysis.

Table 1. Top 10 most frequent terms in the three keyword fields

Rank	Author Keywords	Occ.	Keywords Plus	Occ.	All Keywords	Occ.
1	corpus linguistics	910	english	508	corpus linguistics	965
2	corpus	262	language	485	english	594
3	corpus analysis	256	corpus	316	corpus	559
4	learner corpus	136	discourse	191	language	490
5	collocation	117	acquisition	139	corpus analysis	263
6	academic writing	106	construction	113	discourse	228
7	english	98	lexical bundle	112	lexical bundle	161
8	corpus-based	85	frequency	105	collocation	154
9	spanish	71	learner	85	learner corpus	148
10	corpus-based study	68	grammar	83	frequency	144

Table 1 reveals clear differences among the three keyword fields. Author Keywords are strongly oriented toward corpus-specific and author-selected terminology. *corpus linguistics* is by far the most frequent term (910 records), followed by *corpus* (262), *corpus analysis* (256), and *learner corpus* (136). Several highly ranked terms, including *corpus linguistics*, *corpus analysis*, *learner corpus*, and *corpus-based*, overlap with the retrieval strategy and should therefore be interpreted partly as retrieval-related or anchoring terms rather than as independent thematic discoveries.

Keywords Plus presents a broader indexing-oriented profile. Its highest-frequency terms are *english* (508), *language* (485), *corpus* (316), *discourse* (191), and *acquisition* (139). Compared with Author Keywords, this field gives greater prominence to general disciplinary descriptors such as *language*, *discourse*, *acquisition*, *construction*, and *grammar*, reflecting the broader conceptual associations generated through Web of Science indexing.

All Keywords combines these two profiles. Its leading terms include *corpus linguistics* (965), *english* (594), *corpus* (559), *language* (490), and *corpus analysis* (263), while terms such as *lexical bundle*, *collocation*, and *learner corpus* also remain prominent. The merged field therefore increases keyword coverage, but the higher frequencies are partly a structural consequence of combining Author Keywords and Keywords Plus. Frequency alone cannot determine whether these terms are broadly distributed across the dataset, which motivates the frequency-dispersion analysis in

the following section.

4.3 Frequency-dispersion and concentration profiles of top keywords

The analysis then reassessed the top 50 candidate keywords in each field by combining record frequency with temporal and source dispersion. The analysis examines whether highly frequent keywords are also broadly distributed across years, five-year windows, and sources, and identifies cases in which high frequency coexists with temporal or source-based concentration. The measures are interpreted as defined in Section 3.4; they are therefore not re-derived here. Frequency-dispersion profiles of the top 10 keywords in each field are provided in Appendix Table A1.

4.3.1 Field-level frequency-dispersion profiles

Table 2 summarizes the top-50 profiles at the field level. Medians are used because each candidate set contains a small number of extremely high-frequency terms that would otherwise exert disproportionate influence on the comparison.

Table 2. Field-level summary of Gries-style frequency-dispersion profiles

Indicator	Author Keywords	Keywords Plus	All Keywords
Candidate keywords	50	50	50
Field-specific denominator	3,966	2,966	4,270
Median frequency	43.0	51.5	81.0
Median relative frequency (%)	1.08	1.74	1.90
Median year span	15.0	15.0	17.0
Median period coverage	4.0	4.0	4.0
Median source span	31.0	34.5	49.0
Median year concentration	0.300	0.277	0.197
Median period concentration	0.049	0.067	0.037
Median source concentration	0.930	0.882	0.871
Median maximum source share (%)	9.76	8.09	7.52

The three keyword fields differ in both frequency and distribution. Author

Keywords have a median frequency of 43.0 records, a median year span of 15 years, and a median source span of 31 sources. Keywords Plus has a higher median frequency (51.5) and a somewhat broader median source span (34.5). All Keywords shows the highest median frequency (81.0), year span (17 years), and source span (49 sources), as expected from the record-level union of the two fields.

The concentration measures add information not captured by these range statistics. Median year concentration is 0.300 for Author Keywords, 0.277 for Keywords Plus, and 0.197 for All Keywords. Under the present baseline-relative definition, these values indicate how strongly the keyword distributions deviate from the year distribution of their corresponding field-specific baselines. Median period concentration is 0.049 for Author Keywords, 0.067 for Keywords Plus, and 0.037 for All Keywords, while median source concentration is 0.930, 0.882, and 0.871, respectively. Because year, period, and source concentration scores are calculated over different partitions, their numerical magnitudes are interpreted within each analytical dimension rather than compared directly across dimensions. Together with the larger median year and source spans and the lower median maximum source share in All Keywords (7.52%, compared with 9.76% for Author Keywords and 8.09% for Keywords Plus), these results indicate that the merged field generally broadens coverage, although it does not eliminate distributional concentration.

4.3.2 Keyword-level frequency-dispersion and concentration profiles

Tables 3–5 provide selected keyword-level examples from the three top-50 candidate sets. In addition to frequency and range measures, the tables report maximum source share and year-, period-, and source-concentration scores. Frequency, range, maximum source share, and concentration scores are considered jointly to provide a more complete picture of keyword distribution. This combined assessment helps distinguish high-frequency terms with relatively broad temporal and source distributions from terms whose occurrences are more restricted to particular years, temporal windows, or sources. Lower concentration scores indicate distributions closer to the corresponding field-specific baseline; the scores are interpreted comparatively rather than against fixed cutoffs.

Table 3. Selected frequency-dispersion profiles: Author Keywords

Indicator	corpus linguistics	corpus	corpus analysis	learner corpus	collocation	corpus-based translation studies
Rank	1	2	3	4	5	14
Frequency	910	262	256	136	117	55
Year span	23	22	22	17	19	18
Year conc.	0.015	0.108	0.032	0.133	0.197	0.415
Period coverage	5	5	5	4	4	4
Period conc.	0.003	0.040	0.009	0.031	0.077	0.200
Source span	250	140	131	69	74	20
Max source share (%)	2.86	3.05	5.86	7.35	8.55	21.82
Source conc.	0.348	0.581	0.658	0.821	0.780	0.969

Table 3 shows substantial variation among frequent Author Keywords. *corpus linguistics*, the highest-frequency term, occurs in 910 records across 23 publication years, all five temporal windows, and 250 sources. Its year concentration (0.015), period concentration (0.003), and source concentration (0.348) are all comparatively low, indicating that its high frequency is accompanied by broad temporal and source distribution. This pattern is consistent with its role as both a major field label and a retrieval-related anchor term.

corpus and *corpus analysis* also show broad temporal coverage, but their source-concentration scores increase to 0.581 and 0.658, respectively. More specialized terms show clearer localization. Most notably, *corpus-based translation studies* occurs in 55 records across 18 years and four five-year windows but only 20 sources. Its year concentration (0.415), period concentration (0.200), source concentration (0.969), and maximum source share (21.82%) are substantially higher than those of *corpus linguistics*. This contrast demonstrates that relatively high frequency does not necessarily imply broad temporal or source representativeness.

Table 4. Selected frequency-dispersion profiles: Keywords Plus

Indicator	english	language	corpus	discourse	acquisition	lexical bundle	frequency
Rank	1	2	3	4	5	7	8
Frequency	508	485	316	191	139	112	105
Year span	23	23	20	22	19	16	17
Year conc.	0.022	0.031	0.089	0.100	0.144	0.157	0.125
Period coverage	5	5	4	5	4	4	4
Period conc.	0.007	0.009	0.027	0.017	0.035	0.041	0.032
Source span	169	168	125	88	68	52	68
Max source share (%)	5.71	4.12	6.01	6.28	5.04	16.07	4.76
Source conc.	0.361	0.363	0.555	0.694	0.829	0.783	0.782

Keywords Plus also shows meaningful differences between frequency and distribution. The two most frequent terms, *english* and *language*, occur in 508 and 485 records, respectively, and both span 23 publication years and all five temporal windows. Together with this broad temporal coverage, their relatively low year-concentration scores (0.022 and 0.031), period-concentration scores (0.007 and 0.009), and source-concentration scores (0.361 and 0.363) indicate distributional profiles that remain relatively close to the corresponding field-specific baselines.

Other frequent terms show greater baseline-relative concentration scores in particular dimensions. *acquisition*, for example, has a year-concentration score of 0.144 and a source-concentration score of 0.829. *lexical bundle* shows a year concentration of 0.157 and a period concentration of 0.041, while its maximum source share reaches 16.07%, the highest among the selected Keywords Plus terms. *frequency* likewise shows a higher year-concentration score (0.125) than *english* or *language*, despite remaining visible across 17 publication years. These contrasts indicate that the broader indexing orientation of Keywords Plus does not make all of its frequent terms equally representative across time and sources.

Table 5. Selected frequency-dispersion profiles: All Keywords

Indicator	corpus linguistics	english	corpus	language	corpus analysis	lexical bundle	frequency
Rank	1	2	3	4	5	7	10
Frequency	965	594	559	490	263	161	144
Year span	23	24	22	24	22	18	19
Year conc.	0.017	0.014	0.034	0.044	0.031	0.095	0.139
Period coverage	5	5	5	5	5	4	4
Period conc.	0.004	0.003	0.004	0.019	0.005	0.046	0.018
Source span	254	187	190	170	132	65	82
Max source share (%)	2.69	4.88	3.58	4.08	5.70	14.29	6.94
Source conc.	0.360	0.387	0.407	0.439	0.666	0.786	0.771

The merged All Keywords field generally shows the broadest coverage, but keyword-level variation remains. *corpus linguistics* occurs in 965 records across 23 years, all five temporal windows, and 254 sources, with comparatively low year (0.017) and period (0.004) concentration scores among the selected terms. *english* is similarly broadly distributed, with a year-concentration score of 0.014 and a period-concentration score of 0.003. *corpus* and *language* also show relatively broad temporal distributions.

However, merging the two keyword fields does not remove concentration effects. *lexical bundle* has a year concentration of 0.095, a period concentration of 0.046, a source concentration of 0.786, and a maximum source share of 14.29%. *frequency* has a year-concentration score of 0.139, a period-concentration score of 0.018, and a source-concentration score of 0.771. Because the year- and period-level scores are calculated over different temporal partitions, their numerical magnitudes should not be compared directly. Rather, each score indicates the degree of deviation from the relevant field-specific baseline at its own analytical scale. These results support examining year-level and period-level dispersion separately rather than treating temporal dispersion as a single undifferentiated property.

Taken together, Tables 3–5 demonstrate that frequency rank and distributional representativeness are related but distinct. Highly frequent terms such as *corpus linguistics*, *english*, and *language* tend to show broad distribution, whereas other visible terms exhibit stronger temporal or source concentration. The three keyword fields therefore provide complementary rather than interchangeable views of the dataset.

4.4 Methodological and bibliometric implications for keyword representativeness

Overall, the findings indicate that keyword representativeness should be interpreted as a multidimensional property rather than as a direct function of frequency. Biblioshiny efficiently identifies highly visible keywords, whereas the frequency-dispersion analysis shows whether that visibility is distributed broadly across the dataset. The two stages therefore serve complementary purposes: frequency indicates candidate prominence, while temporal and source dispersion qualify how representative that prominence is.

This distinction is especially important for the present dataset because several highly frequent Author Keywords also overlap with the retrieval strategy. Terms such as *corpus linguistics*, *corpus analysis*, *corpus-based*, and *learner corpus* helped define the scope of the retrieved dataset. Their high frequency is therefore partly retrieval-conditioned and should not be interpreted independently as evidence of newly discovered thematic prominence. Dispersion analysis does not remove this retrieval effect, but it provides additional evidence about whether such terms are broadly distributed within the resulting dataset.

The comparison also clarifies the different bibliometric functions of the three keyword fields. Author Keywords foreground author-selected and often more specialized labels and show the highest median source-concentration score relative to their field-specific baseline among the three candidate sets. Keywords Plus provides broader indexing-derived descriptors, although individual terms still differ substantially in their baseline-relative temporal and source concentration profiles. All Keywords maximizes coverage by combining the two fields and generally shows broader median coverage in terms of year span and source span, but merging does not make every keyword representative. The merged field therefore should not be treated as inherently superior; rather, each field captures a different layer of bibliographic representation.

Methodologically, these findings support the use of frequency and dispersion as complementary dimensions in bibliometric keyword analysis. The adapted Gries-style framework employed here does not reproduce corpus keyness analysis but transfers its frequency-dispersion logic to bibliographic metadata. Within this framework, representativeness refers specifically to distributional representativeness within the retrieved dataset: how widely a keyword is distributed relative to the temporal and

source structure of the relevant keyword field. This interpretation avoids treating frequent keywords as automatically representative and provides a more transparent basis for comparing Author Keywords, Keywords Plus, and their record-level union.

5. Conclusion

This study compared Author Keywords, Keywords Plus, and All Keywords in 4,270 Web of Science-indexed English-language journal articles related to corpus linguistics published between 2001 and 2025. Biblioshiny was first used to identify the top 50 candidate terms in each field, after which an adapted Gries-style framework was used to reassess their frequency and distribution across publication years, five-year windows, and sources.

The results show that the three keyword fields provide distinct but complementary representations of the dataset. Author Keywords emphasize author-selected and retrieval-related terminology, Keywords Plus foregrounds broader indexing-derived descriptors, and All Keywords provides the widest overall coverage. However, frequency and representativeness are not equivalent. Some highly frequent terms are broadly distributed across time and sources, whereas others show stronger temporal or source concentration. The merged field generally broadens coverage but does not eliminate such concentration effects.

Methodologically, the study demonstrates how frequency-based bibliometric keyword analysis can be strengthened through distribution-sensitive measures. The adapted Gries-style approach employed here is not a target-versus-reference keyness analysis; rather, it evaluates the distributional representativeness of existing bibliographic keywords within the retrieved dataset. Its principal contribution lies in comparing different keyword fields under the same frequency-dispersion framework.

Several limitations remain. The dataset is restricted to Web of Science, English-language journal articles, selected linguistic categories, and corpus-related retrieval terms; consequently, the findings should not be generalized to all corpus-linguistics research. High-frequency retrieval-related terms also partly reflect the construction of the dataset. Future studies could test the framework with other databases, research domains, and keyword systems. Overall, the findings support interpreting bibliometric keyword importance through both frequency and

distribution, while clearly distinguishing the functions of Author Keywords, Keywords Plus, and merged keyword fields.

References

- Araújo, Tanya, Alexandre Abreu, and Francisco Louçã. 2024. The evolution of complexity co-occurring keywords: Bibliometric analysis and network approach. *Advances in Complex Systems* 27(4–5): 2450006. <https://doi.org/10.1142/S0219525924500061>.
- Aria, Massimo and Corrado Cuccurullo. 2017. bibliometrix: An R-tool for comprehensive science mapping analysis. *Journal of Informetrics* 11(4): 959–975. <https://doi.org/10.1016/j.joi.2017.08.007>.
- Callon, Michel, Jean-Pierre Courtial, William A. Turner, and Serge Bauin. 1983. From translations to problematic networks: An introduction to co-word analysis. *Social Science Information* 22(2): 191–235. <https://doi.org/10.1177/053901883022002003>.
- Cobo, Manuel José, Antonio Germán López-Herrera, Enrique Herrera-Viedma, and Francisco Herrera. 2011a. Science mapping software tools: Review, analysis, and cooperative study among tools. *Journal of the American Society for Information Science and Technology* 62(7): 1382–1402. <https://doi.org/10.1002/asi.21525>.
- Cobo, Manuel José, Antonio Germán López-Herrera, Enrique Herrera-Viedma, and Francisco Herrera. 2011b. An approach for detecting, quantifying, and visualizing the evolution of a research field: A practical application to the fuzzy sets theory field. *Journal of Informetrics* 5(1): 146–166. <https://doi.org/10.1016/j.joi.2010.10.002>.
- Donthu, Naveen, Satish Kumar, Debmalya Mukherjee, Nitesh Pandey, and Weng Marc Lim. 2021. How to conduct a bibliometric analysis: An overview and guidelines. *Journal of Business Research* 133: 285–296. <https://doi.org/10.1016/j.jbusres.2021.04.070>.
- Egbert, Jesse and Douglas Biber. 2019. Incorporating text dispersion into keyword analyses. *Corpora* 14(1): 77–104. <https://doi.org/10.3366/cor.2019.0162>.
- Garfield, Eugene and Irving H. Sher. 1993. Keywords Plus—algorithmic derivative indexing. *Journal of the American Society for Information Science* 44(5): 298–299. [https://doi.org/10.1002/\(SICI\)1097-4571\(199306\)44:5<298::AID-ASIS>3.0.CO;2-A](https://doi.org/10.1002/(SICI)1097-4571(199306)44:5<298::AID-ASIS>3.0.CO;2-A).
- Gries, Stefan Th. 2008. Dispersions and adjusted frequencies in corpora. *International Journal of Corpus Linguistics* 13(4): 403–437. <https://doi.org/10.1075/ijcl.13.4.02gri>.
- Gries, Stefan Th. 2021. A new approach to (key) keywords analysis: Using frequency, and now also dispersion. *Research in Corpus Linguistics* 9(2): 1–33. <https://doi.org/10.32714/ricl.09.02.02>.
- Lee, Sunghwa, Myoungho Ha, Byong-Rae Ryu, and Se-Eun Jhang. 2025. A bibliometric comparison of grammatical development in spoken and sign languages: Thematic evolution and

- persistent keywords. *Linguistic Research* 42(Special Edition): 109–147. <https://doi.org/10.17250/khisli.42..202509.005>.
- Lee, Sunghwa and Se-Eun Jhang. 2026. A bibliometric analysis of sign language linguistics research: Trends and thematic mapping. *EONEOHAG* 104: 193–223.
- Liu, Cuilin. 2025. Integrating bibliometric techniques with corpus-driven bigrams: A research trend analysis of cruise tourism. PhD Dissertation. National Korea Maritime & Ocean University.
- Moral-Muñoz, José A., Enrique Herrera-Viedma, Antonio Santisteban-Espejo, and Manuel J. Cobo. 2020. Software tools for conducting bibliometric analysis in science: An up-to-date review. *Profesional de la información* 29(1): e290103. <https://doi.org/10.3145/epi.2020.ene.03>.
- Zhang, Juan, Qi Yu, Fashan Zheng, Chao Long, Zuxun Lu, and Zhiguang Duan. 2016. Comparing Keywords Plus of WoS and Author Keywords: A case study of patient adherence research. *Journal of the Association for Information Science and Technology* 67(4): 967–972. <https://doi.org/10.1002/asi.23437>.
- Župič, Ivan and Tomaž Čater. 2015. Bibliometric methods in management and organization. *Organizational Research Methods* 18(3): 429–472. <https://doi.org/10.1177/1094428114562629>.

Appendix

Table A1. Frequency-dispersion profiles of the top 10 keywords in each keyword field

Field	Rank	Keyword	Freq.	Year span	Year conc.	Period coverage	Period conc.	Source span	Max source share (%)	Source conc.
AK	1	corpus linguistics	910	23	0.015	5	0.003	250	2.86	0.348
AK	2	corpus	262	22	0.108	5	0.040	140	3.05	0.581
AK	3	corpus analysis	256	22	0.032	5	0.009	131	5.86	0.658
AK	4	learner corpus	136	17	0.133	4	0.031	69	7.35	0.821
AK	5	collocation	117	19	0.197	4	0.077	74	8.55	0.780
AK	6	academic writing	106	18	0.210	5	0.011	41	22.64	0.884
AK	7	english	98	21	0.161	5	0.073	59	13.27	0.862
AK	8	corpus-based	85	13	0.216	3	0.110	52	8.24	0.835
AK	9	spanish	71	17	0.167	4	0.045	50	8.45	0.892
AK	10	corpus-based study	68	17	0.241	5	0.042	52	4.41	0.861
KP	1	english	508	23	0.022	5	0.007	169	5.71	0.361

Field	Rank	Keyword	Freq.	Year span	Year conc.	Period coverage	Period conc.	Source span	Max source share (%)	Source conc.
KP	2	language	485	23	0.031	5	0.009	168	4.12	0.363
KP	3	corpus	316	20	0.089	4	0.027	125	6.01	0.555
KP	4	discourse	191	22	0.100	5	0.017	88	6.28	0.694
KP	5	acquisition	139	19	0.144	4	0.035	68	5.04	0.829
KP	6	construction	113	18	0.107	4	0.020	69	6.19	0.777
KP	7	lexical bundle	112	16	0.157	4	0.041	52	16.07	0.783
KP	8	frequency	105	17	0.125	4	0.032	68	4.76	0.782
KP	9	learner	85	16	0.182	4	0.037	51	7.06	0.882
KP	10	grammar	83	22	0.307	5	0.115	49	10.84	0.857
ALL	1	corpus linguistics	965	23	0.017	5	0.004	254	2.69	0.360
ALL	2	english	594	24	0.014	5	0.003	187	4.88	0.387
ALL	3	corpus	559	22	0.034	5	0.004	190	3.58	0.407
ALL	4	language	490	24	0.044	5	0.019	170	4.08	0.439
ALL	5	corpus analysis	263	22	0.031	5	0.005	132	5.70	0.666
ALL	6	discourse	228	22	0.070	5	0.007	98	5.70	0.672
ALL	7	lexical bundle	161	18	0.095	4	0.046	65	14.29	0.786
ALL	8	collocation	154	19	0.100	4	0.033	85	6.49	0.757
ALL	9	learner corpus	148	17	0.133	4	0.043	71	8.11	0.824
ALL	10	frequency	144	19	0.139	4	0.018	82	6.94	0.771

Notes. AK = Author Keywords; KP = Keywords Plus; ALL = All Keywords. Field-specific denominators are 3,966 records for AK, 2,966 for KP, and 4,270 for ALL. Year conc., period conc., and source conc. are DKL-based transformed concentration scores calculated relative to the corresponding field-specific baseline. Higher values indicate greater deviation from the relevant baseline. Period coverage is based on five fixed five-year windows: 2001–2005, 2006–2010, 2011–2015, 2016–2020, and 2021–2025.

Myoungho Ha

Assistant Professor

Office of Liberal Arts Education Center

Silla University

140 Baegyang-daero 700beon-gil, Sasang-gu,

Busan, 46958, Korea

E-mail: hadash21@silla.ac.kr

Received: 2026. 06. 30.

Revised: 2026. 08. 20.

Accepted: 2026. 09. 03.