International Journal of Applied Linguistics

WILEY

ORIGINAL ARTICLE

Assessing EFL Learners' Phraseological Competence Through Association Measures: The Role of Reference Corpus Selection

Tingting Wang[1] | Dandan Zhou[1] | Magali Paquot[2]

[1]School of Foreign Studies, Nanjing University, Nanjing, Jiangsu, China | [2]Fonds de la Recherche Scientifique & Centre for English Corpus Linguistics, UCLouvain, Louvain-la-Neuve, Belgium

Correspondence: Dandan Zhou (ddzhounj@nju.edu.cn)

Received: 14 December 2025 | **Revised:** 9 July 2026 | **Accepted:** 20 July 2026

Keywords: association measures | collocational assessment | phraseological competence | reference corpus

关键词: 短语能力 | 关联度指标 | 参照语料库 | 搭配评估 | 评估效度

ABSTRACT

Association measures, such as pointwise mutual information (PMI), have been widely used to explore learners' phraseological competence. Yet, the validity of PMI-based evaluations critically depends on the reference corpus employed, and the extent to which this methodological choice affects assessment outcomes has not been systematically examined. This study investigates how reference corpus selection influences PMI scores and, consequently, the evaluation of learners' phraseological competence. We calculated PMI scores for *amod*, *dobj*, and *advmod* dependency pairs from the oral performances of 98 test-takers, comparing the outcomes when using four reference corpora: the full COCA corpus (COCA-ALL) and its specialized academic (COCA-ACAD), news (COCA-NEWS), and spoken (COCA-SP) sub-corpora. Results revealed that, first, for the same set of dependency pairs, the choice of reference corpus led to statistically significant differences in their calculated PMI scores for all three dependency types ($p < 0.001$). This effect was most pronounced for *dobj* relations, where 88.9% of pairs exhibited substantial PMI differences across certain reference corpora. Second, this systematic bias translated directly into unreliable learner assessments. COCA-ALL consistently assigned significantly lower median PMI scores ($p < 0.001$), leading to unstable proficiency rankings (correlation coefficient dropping to 0.59). For *dobj* relations, this caused over half of learners (55.6%) to shift by 10 or more rank positions. These findings demonstrate that reference corpus selection is a critical factor that can systematically affect and undermine the validity of phraseological assessments.

摘要

点互信息等关联度指标已被广泛用于对学习者短语能力的考察。然而, 基于点互信息的评估在很大程度上依赖于所选用的参照语料库, 而这一方法学选择对短语能力评估结果的具体影响, 迄今尚缺乏系统检验。本研究探讨参照语料库的选择如何影响点互信息得分, 进而作用于对学习者短语能力的评估。研究从98名受试者的口语产出中提取形名搭配、动宾搭配和状中搭配三类依存搭配, 分别计算其点互信息值, 并比较了在COCA全库及其学术、新闻和口语三个子库四种参照语料库条件下的结果差异。结果表明, 第一, 对同一组依存搭配而言, 参照语料库的选择在所有三类依存类型中均引发点互信息值的统计显著差异($p < .001$), 其中动宾搭配受影响最大, 在部分参照语料库对比中有88.9%的搭配对呈现出大幅点互信息差异。第二, 这一系统性偏差直接导致学习者评估结果的不稳定。相较于其他语料库, COCA全库持续赋予显著较低的点互信息中位数($p < .001$), 致

使短语能力排名出现波动, 相关系数降至.59。在动宾关系中, 该问题尤为突出, 超过一半的学习者(55.6%)的排名变动达到或超过10个位次。上述结果表明, 参照语料库的选择是影响短语能力评估效度的关键因素, 可能对评估结果产生系统性偏差。

1 | Introduction

Phraseological use is an important part of nativelike, fluent, and idiomatic language use (Durrant and Schmitt 2009; Ellis et al. 2008; Paquot and Granger 2012). Over the past few decades, a growing body of research has adopted a frequency-based approach to phraseology, examining word combinations in learner corpora using semi-automated methods (Paquot and Granger 2012). Within this tradition, statistical association measures (AMs), such as Pointwise Mutual Information (PMI), have played an important role in explorations of collocational knowledge (see Ellis et al. 2008; Durrant and Schmitt 2009; Granger and Bestgen 2014; Paquot 2019). These measures quantify the statistical relationship between words, reflecting the exclusivity of collocates—the degree to which two words co-occur more often than expected by chance. This exclusivity is linked to the predictability of co-occurrence, facilitating recognition, acquisition, and storage (Wray 2002). By computing PMI values against a reference corpus, researchers can gain insights into the extent to which L2 learners produce native-like phraseological patterns (Ellis et al. 2008; Granger and Bestgen 2014; Paquot 2019). Empirical studies consistently show that advanced L2 learners produce combinations with higher PMI scores, reflecting more native-like phraseological patterns, whereas lower-proficiency learners tend to rely on simpler, word-by-word combinations with weaker associative strength (Bestgen and Granger 2014; Durrant and Schmitt 2009; Paquot 2018, 2019).

However, one critical methodological issue that persists is the choice of the reference corpus used to calculate these measures. The linguistic and situational characteristics of a corpus—especially in terms of genre, register, and modality—can influence PMI scores, even when the same word combinations are analyzed across corpora (Gablasova et al. 2017; Paquot and Naets 2017). Nevertheless, L2 phraseological research has often relied on broad-coverage, multi-register corpora such as the British National Corpus (BNC) or the Corpus of Contemporary American English (COCA), often without considering the degree of comparability between the reference corpus and the target learner data (i.e., Siyanova-Chanturia 2015).

This misalignment is more than a minor methodological detail: it directly impacts the validity, fairness, and interpretability of phraseological assessments (Gablasova et al. 2017; Paquot and Naets 2017). Since phraseological patterns can vary across linguistic settings (e.g., involving different social and situational characteristics), relying on a mismatched reference corpus can distort PMI-based evaluations, either inflating or underestimating learner competence (Biber and Barbieri 2007).

Given these challenges, a systematic examination of reference corpus selection is urgently needed to safeguard the validity, reliability, and pedagogical relevance of phraseological proficiency assessment in L2 research. Without such scrutiny, conclusions about learner competence risk being misleading, and the theoretical understanding of L2 phraseological proficiency and development may be built on unreliable empirical foundations.

This study investigates the impact of reference corpus selection on PMI-based assessments of L2 phraseological competence. Specifically, we compare PMI scores derived from general versus specialized corpora to measure how corpus choice influences results. Our findings aim to improve methodological accuracy in assessing L2 phraseology, thereby addressing a key limitation in current corpus-driven research.

2 | Literature Review

2.1 | Phraseological Competence as an Index of L2 Proficiency

While early assessment frameworks treated lexis and grammar separately (Lado 1961; Bachman and Palmer 1996), current research recognizes language as fundamentally formulaic (Sinclair 1991; Wray 2002). Phraseology—covering collocations, idioms, and lexical bundles—is now considered essential for nativelike fluency and idiomaticity (Granger and Paquot 2008; Paquot and Granger 2012).

A growing body of research demonstrates that phraseological use differentiates performances across proficiency levels, and traces language development. Studies comparing native and non-native production, as well as performances across proficiency levels, reveal consistent differences in the accuracy, diversity, and sophistication of multiword patterns (Laufer and Waldman 2011; Granger and Bestgen 2014; Paquot 2019). In particular, more advanced English learners typically use more collocations (Tsai 2008), exhibit a broader range of phraseological patterns (Fan 2009), and employ more sophisticated, low-frequency collocations (indexed by higher mutual information [MI] scores) (Durrant and Schmitt 2009; Granger and Bestgen 2014). Converging evidence further establishes the predictive validity of phraseological measures via correlations with L2 writing and speaking proficiency (Jiang et al. 2023; Paquot et al. 2022; Vandeweerd et al. 2023; Wang and Zhou 2025).

2.2 | Identifying Collocational Patterns Using Association Measures: The Impact of Reference Corpus Selection

Statistical association measures, particularly PMI, are widely used to assess L2 collocational knowledge (Ellis et al. 2008; Gries 2022). However, PMI scores are not intrinsic properties of word pairs but estimates strictly dependent on the probability distributions of the reference corpus (Church and Hanks 1990). Because of this sensitivity, any register mismatch between learner and reference data undermines comparability and validity (Egbert et al. 2022; Gablasova et al. 2017). Consequently, aligning the

reference corpus with the learner data is not a preference but a methodological prerequisite (Bestgen and Granger 2014; Paquot and Naets 2025).

Empirical evidence confirms that PMI scores are highly sensitive to reference data. Gablasova et al. (2017) demonstrated that scores for identical collocations (e.g., *make a decision*) varied drastically (3.67–7.91) across different sub-corpora, suggesting that estimates from a single general corpus may be unstable. Extending this to learner assessments, Paquot and Naets (2017) compared PMI profiles derived from three reference corpora (BNC, ENCOW14, L2RC). They found that corpus selection significantly altered the evaluation of learners' *amod* and *dobj* combinations. Crucially, they observed that the "reference corpus effect" varied by dependency type, implying that corpus choice does not impact all grammatical relations uniformly.

Despite this, many studies continue to rely on general corpora that include a variety of registers and/or genres as well as spoken and written modalities, such as the BNC and the COCA, as reference corpora for PMI calculation. This reliance is often due to the need for extensive data to explore phraseological patterns, coupled with the lack of availability of fully comparable datasets of sufficient size. Although such corpora provide coverage and size, they often underrepresent the specific contextual features of L2 production (Bestgen and Granger 2014; Durrant and Schmitt 2009). Furthermore, some studies compute collocational strength using reference corpora with mismatched genres to assess collocations in a different genre. This methodological issue is particularly pressing in L2 speech research, where the scarcity of large spoken corpora often compels researchers to use reference corpora dominated by written texts. For instance, Paquot et al. (2022) analyzed oral production from the Trinity Lancaster Corpus yet calculated MI scores using ENCOW14, a web corpus predominantly composed of written texts. Similarly, Vandeweerd et al. (2023) assessed oral data from the TEF exam against FRCOW16, a massive web-scraped corpus of written French. While these choices may be pragmatic in the absence of a more suitable comparator, their limitations should be made explicit. Specifically, this reliance on written reference corpora raises important concerns about the fairness and validity of current L2 phraseological assessment practices, especially given that spoken language is inherently characterized by greater immediacy, interactivity, and situational variability than written text (Biber et al. 1999; Biber 2009).

However, current research still lacks empirical evidence examining how reference corpus selection influences the computation of PMI scores. To address this gap, the present study systematically investigates the influence of reference corpus register—comparing a general standard against specialized subcorpora—on the evaluation of EFL learners' spoken phraseology. Specifically, we aim to determine how corpus choice affects both the calculation of collocational strength and the subsequent assessment of learner proficiency. The study is guided by two research questions:

RQ1: How does the register of the reference corpus influence PMI-based measures of collocational strength in EFL learners' oral production?

RQ2: How do differences in PMI scores derived from reference corpora representing distinct registers affect the assessment of phraseological competence in EFL learners?

By addressing these research questions, this study extends Gablasova et al. (2017), which documented the variation in collocational strength across different linguistic settings, by examining the role of reference corpus selection for assessing phraseological competence in L2 speaking. It also aims to empirically examine Paquot and Naets' (2017) suggestion that different reference corpora can be used to evaluate different aspects of phraseological competence.

3 | Methodology

3.1 | Learner Corpus

The dataset comprises transcripts from 98 Chinese EFL learners who participated in the Test for English Majors Band 8 Oral (TEM 8-Oral), a standardized proficiency assessment for fourth-year undergraduates in mainland China. The task required participants to deliver a three-minute monologic commentary on a social issue after four minutes of preparation.

Performances were double-rated using the official rubric, with final scores (averaged across two raters, with arbitration for large discrepancies) grouped into four bands: *excellent* ($\geq$ 80), *good* (70–79), *pass* (60–69), and *not pass* ($<$ 60).

From an initial pool of 287 transcripts, length-based outliers were removed using the Tukey IQR method. The final analytic sample (N = 98) was established through stratified random sampling to ensure balanced representation across proficiency levels and controlled text lengths. Table 1 summarizes the sample distribution. All transcripts were de-identified prior to analysis and handled in accordance with institutional ethical guidelines for secondary use of assessment data. The Institutional Review Board of Nanjing University determined that this study did not constitute human subject research.

3.2 | Phraseological Units

This study analyzes phraseological use in three grammatical relations frequently examined in English learner corpus research: adjectival modifiers (*amod*; adjective + noun), adverbial modifiers (*advmod*; adverb + verb/adjective/adverb), and verb + direct object (*dobj*). The focus on these structures is informed by previous studies that have identified their centrality in English phraseology and their significance in characterizing L2 proficiency (Laufer and Waldman 2011; Paquot 2018, 2019; Wang and Zhou 2025). Adopting these widely studied dependency relations also enables systematic comparison with related research.

Dependency arcs for the three target relations were automatically extracted from the learner transcripts using **spaCy 3.7.2** with the *en_core_web_sm* model (v3.7.1). The extraction process began with text preprocessing. Transcripts were used in their original form without applying spelling or grammar correction, removing disfluencies, or filtering utterances. As the data consist

TABLE 1 | Distribution of texts and text length by TEM-8-Oral score band.

Score band (TEM-8-Oral)	Number of texts	Text length (running-word tokens)		Task score	
		M	*SD*	*M*	*SD*
90–100	3	366.67	22.51	90.00	0.00
80–90	18	307.56	30.07	82.44	2.18
70–80	24	278.75	22.90	75.42	5.00
60–70	24	264.50	31.39	65.25	4.74
0–60	29	236.38	31.29	54.76	9.32
Total	98	283.05	67.92	65.25	15.84

TABLE 2 | Illustration of dependency relations analyzed in the study.

Dependency type	Sentence	Extracted relation
dobj (direct object)	I decided to **face** the **challenge**.	face/VERB-challenge/NOUN
Amod (adjectival modifiers)	I decided to **face** the **big challenge**.	big/ADJ-challenge/NOUN
Advmod (adverbial modifiers)	I decided to **quickly face** the challenge.	face/VERB-quickly/ADV
	The challenge was **very big**.	big/ADJ -very/ADV
	I decided to face the challenge **quite quickly**.	quickly/ADV—quite/ADV

of transcribed speech rather than learner-produced written texts, spelling errors are not expected, as the transcripts follow standardized transcription conventions. This decision was made to preserve the authenticity of learner language and to avoid introducing artificial regularities or altering non-native linguistic patterns. Filled pauses (e.g., um/uh) were retained because, as interjections (INTJ) in spaCy's POS tagging, they rarely form the target *amod*, *advmod*, or *dobj* dependency relations, thus having negligible impact on our primary data extraction. Using spaCy's full pipeline (tokenization, POS tagging, lemmatization, dependency parsing), we extracted all amod, advmod, and dobj arcs and represented each instance by its dependent and head lemmas. The complete technical details of the extraction procedure are provided in the accompanying OSF repository (https://osf.io/3q9fs/overview?view_only=64a64969b8424f4787f4dffcc5805eeb).

Table 2 provides an illustration of the dependency relations analyzed.

To assess the robustness of dependency parsing on learner speech, two trained annotators (the first author and a second coder, a PhD student in applied linguistics with expertise in phraseology) independently annotated a stratified random 10% subsample (118 utterances from 10 test takers, covering the full proficiency range). Annotators worked on the spaCy-tokenized transcripts (i.e., the same token boundaries used for automatic extraction) and, using UD-based definitions, identified all instances (dependency arcs) of *amod*, *advmod*, and *dobj* in each utterance. Inter-annotator agreement was high (Cohen's κ: *amod* = 0.98, *dobj* = 0.92, *advmod* = 0.81). The 42 disputed relation instances were adjudicated to create a consensus gold standard.

We then evaluated spaCy's output against this gold standard using precision, recall, and F1-score for each relation (Table 3). Performance was very strong for *amod* (precision = 1.00, recall = 0.98, F1 = 0.99), and good for *advmod* (precision = 0.76, recall = 1.00, F1 = 0.86) and *dobj* (precision = 0.80, recall = 0.86, F1 = 0.83). These results collectively indicate acceptable accuracy for extracting the three target relations within our dataset.

To further limit the impact of residual parsing errors, we applied a frequency threshold (≥ 5 occurrences in the reference corpus), which reduces chance co-occurrences (Evert 2005). Erroneous pairs from learner language are highly unlikely to meet this frequency threshold, as they typically reflect non-standard or idiosyncratic usage that is not attested in native speaker corpora. This combined strategy of high parsing accuracy and frequency-based filtering ensures that any parsing errors have a minimal and non-systematic impact on our PMI calculations and overall results.

3.3 | Measurement for Collocational Strength

To quantify the strength of dependency-based collocations, we computed **dependency-pair frequencies** and **PMI values** from the COCA reference corpus using the same lemma representation and dependency relations as in the learner pipeline. This ensured full methodological consistency across datasets. Each dependency pair was represented as lowercased dependent–head lemmas.

Collocational strength was quantified using PMI (Church and Hanks 1990), defined as:

$$PMI\,(dep, head) = log_2\,(P\,(dep, head)\,/\,(P\,(dep) \times P\,(head)))$$

where $P(dep,\ head) = f(\text{dep, head}) / N_{\text{pairs, r}}$, $P(dep) = f(dep) / N_{\text{pairs, r}}$, $P(head) = f(head) / {}_{\text{pairs, r}}$. Here, $f(dep, head)$ is the

TABLE 3 | Precision, recall, and *F*-scores for phraseological unit extraction.

Relation	Precision	Recall	*F1*
amod	1.00	0.98	0.99
advmod	0.76	1.00	0.86
dobj	0.80	0.86	0.83

TABLE 4 | Reference corpora overview.

Sub-corpus	Word count	Number of texts
COCA	1,001,610,938 words	485,179 texts
COCA Spoken	127,396,932 words	44,803 texts
COCA Newspaper	122,958,016 words	90,243 texts
COCA Academic	120,988,361 words	26,137 texts

frequency of a dependency pair within a given relation type r (i.e., *amod*, *advmod*, or *dobj*), and $N_{\text{pairs},r}$ is the total number of dependency-pair tokens for that relation in the reference corpus. $f(dep)$ and $f(head)$ refer to the frequencies of the dependent and head lemmas within the extracted dependency relations of the same type r. PMI was calculated separately for each dependency relation to account for distributional differences across relation types. This formulation estimates both joint and marginal probabilities within the same dependency-defined event space, ensuring internal consistency in the association measure. Only dependency pairs that occurred at least five times in the reference corpus were included to reduce the influence of chance co-occurrence (Evert 2005)[1]. All extractions and calculations were implemented in Python. Full computational details are provided in the OSF script repository (https://osf.io/3q9fs/overview?view_only=64a64969b8424f4787f4dffcc5805eeb).

3.4 | Reference Corpus

To examine how reference corpus choice affects PMI-based measures of phraseological competence, we used the full COCA (Davies 2020) and three subcorpora (spoken, newspaper, academic) as reference corpora.

The full COCA serves as a methodological baseline. As a widely used, large, and balanced corpus in phraseological research (e.g., Tavakoli and Uchihara 2020; Yoon 2016), it allows us to situate our results within the field and assess whether findings based on broad-coverage corpora hold when compared with register-specific alternatives.

The three subcorpora were selected to align with key characteristics of the learner data (university students' oral commentaries on social issues): COCA-Spoken captures the oral mode, COCA-Newspaper reflects argumentative discourse on social topics, and COCA-Academic represents formal academic conventions. Table 4 summarizes the corpora.

All reference corpora were parsed using the same pipeline as the learner data (spaCy v3.7.2; en_core_web_sm v3.7.1). We extracted dependency arcs labeled amod, advmod, and dobj, using the same preprocessing conventions (lemmatization, lowercasing, exclusion of punctuation/whitespace). Dependency-pair token frequencies were then aggregated within each corpus to compute PMI.

3.5 | Data Analysis

To address RQ1 (influence on collocational strength), we conducted a two-step analysis. First, we calculated data coverage—the proportion of learner pairs meeting the frequency threshold (≥5)—to determine how register differences affected the inclusion of phraseological data. Second, focusing on the "common set" of pairs shared across all corpora, we quantified the magnitude of score variation using pairwise absolute PMI differences $|\Delta\text{PMI}| \geq 1$. Following Gablasova et al. (2017), a threshold of $|\Delta\text{PMI}| \geq 1$ was used to identify substantial, register-induced fluctuations in collocational strength.

To address RQ2 (impact on the assessment of learner competence), we computed median PMI scores[2] for each learner across the four reference corpora and ranked them accordingly. The median was selected due to the heavy-tailed nature of PMI distributions, making it less sensitive than the mean to occasional extreme PMI values arising from rare collocations or parsing noise. Importantly, an analysis of learner-level counts of PMI-eligible pairs confirms that the data underpinning these medians are generally robust. Detailed descriptive statistics are provided in our OSF repository (https://osf.io/3q9fs/overview?view_only=64a64969b8424f4787f4dffcc5805eeb). We then evaluated the stability of these rankings using two metrics: (1) rank concordance, assessed via pairwise Spearman's correlations (ρ); and (2) rank variability, quantified by the maximum absolute rank change per learner and the proportion of learners whose rankings shifted by at least 10 positions.

Throughout this study, a significance threshold of $p < 0.05$ was adopted for all inferential statistical tests, and all reported p-values are two-tailed. Following Cohen (1988), effect sizes were interpreted as small ($r \geq 0.10$), medium ($r \geq 0.30$), and large ($r \geq 0.50$).

TABLE 5 | Coverage of dependency pairs in learner texts across reference corpora.

Dependencytype	Extracted number of dependency pair tokens	Reference corpus							
		COCA-ALL		COCA-ACAD		COCA-NEWS		COCA-SP	
		N	%	*N*	%	*N*	%	*N*	%
advmod	1836	1782	97.06%	1432	78.00%	1512	82.35%	1521	82.84%
amod	2795	2642	94.53%	2349	84.04%	2072	74.13%	2044	73.13%
dobj	1270	1195	94.09%	973	76.61%	977	76.93%	947	74.57%

Note: N and % represent the number and percentage of dependency pair tokens used for PMI calculation.

4 | Results

4.1 | Effects of Reference Corpus Selection on PMI Scores (RQ1)

4.1.1 | Coverage of Dependency Pairs in Learner Texts

Our analysis of data coverage revealed that the choice of reference corpus largely determined how many learner-produced dependency pairs could be included for PMI calculation. As shown in Table 5, coverage varied markedly across the four reference corpora.

COCA-ALL provides very high coverage for all three dependency types (*advmod*: 97.06%, *amod*: 94.53%, *dobj*: 94.09%), indicating that almost every dependency pair identified in the learner data is also attested in COCA and can therefore be included in the calculation of PMI scores. In contrast, all three COCA subcorpora exhibit lower coverage, with notable variation depending on both the reference corpus and dependency type. The most pronounced effect occurs with *amod* pairs, which exhibit the widest coverage variation. Their coverage ranges from 84.04% in COCA-ACAD down to just 73.13% in COCA-SP. The coverage for *dobj* and *advmod* dependencies is also lower in the specialized corpora but shows less variability among them, with rates consistently falling between 77% and 83%.

4.1.2 | PMI Score Variation for the Same Dependency Pairs Across Reference Corpora

Having established the set of dependency pairs common to all four corpora (N = 3668; including 1760 *amod*, 875 *dobj*, and 1033 *advmod* tokens), we next examined whether the choice of reference corpus led to statistically significant differences in their calculated PMI scores.

Table 6 provides a descriptive summary of the median PMI scores for each dependency type as calculated using the four reference corpora. The resulting median PMI scores varied across corpora and dependency relations. The *dobj* relation showed the widest fluctuations, with median PMI scores ranging from a negative association when benchmarked against COCA-ALL (Mdn = −1.51) to a positive one when using COCA-SP as the reference (Mdn = 0.44). For *amod*, the median PMI values were positive in all corpora, ranging from 1.44 (based on COCA-NEWS) to 2.07 (based on COCA-SP), with relatively smaller cross-corpus variation. In contrast, *advmod* dependencies exhibited the smallest differences across reference corpora, with median PMI values fluctuating narrowly between −0.13 and 0.21, suggesting minimal corpus-related variation in this dependency type.

Given the non-normal distribution of the PMI scores, a non-parametric Friedman test was conducted to determine whether the choice of reference corpus had a significant effect on the resulting PMI values. The results showed a significant effect for all dependency types: *amod* ($\chi^2(3) = 274.644$, $p < 0.001$), *advmod* ($\chi^2(3) = 99.885$, $p < 0.001$), and *dobj* ($\chi^2(3) = 1300.537$, $p < 0.001$). This indicates that PMI-based measurements of collocational strength are highly sensitive to the reference corpus used for calculation. Follow-up Wilcoxon signed-rank tests were performed to identify which specific pairs of reference corpora produced significantly different PMI scores (effect sizes reported as r; see Table 7 for full results). For *dobj* and *amod* relations, all six pairwise comparisons of the PMI scores—each set calculated from a different reference corpus—yielded significant differences ($p < .001$), with mostly medium to large effect sizes. For *advmod*, however, significant differences primarily emerged only when PMI scores calculated using the general COCA-ALL corpus were compared against those calculated using a specialized sub-corpus (i.e., COCA-NEWS, COCA-ACAD, or COCA-SP), again with small-to-medium effects. In contrast, there were no statistically significant differences between the PMI scores when the reference benchmarks were switched among the specialized sub-corpora themselves (e.g., scores based on COCA-NEWS vs. scores based on COCA-ACAD).

4.1.3 | Magnitude and Distribution of PMI Differences

Table 8 summarizes the magnitude of PMI score variation across the six pairings of reference corpora, presenting the proportion of dependency pairs that exhibited substantial ($|\Delta PMI| \geq 1$) versus minor ($|\Delta PMI| < 1$) score differences for each corpus combination and dependency type.

The choice of reference corpus had a marked impact on the resulting collocational strength measurements, though the magnitude of this effect varied by which pair of corpora was being compared. The most pronounced divergence in PMI scores arose when comparing values derived from COCA-ALL against those from COCA-ACAD, where 42.2% of all common pairs showed a substantial PMI difference. A similarly high proportion of substantial differences (38.6%) was observed when contrasting the PMI scores based on COCA-ACAD and COCA-SP. In contrast, the greatest stability was found when comparing PMI scores from

TABLE 6 | Descriptive statistics of median PMI by dependency type and reference corpus.

Dependency type	Metric	COCA-ALL	COCA-ACAD	COCA-NEWS	COCA-SP
advmod	Median PMI	0.07	0.21	0.21	−0.13
	IQR	3.6	3.69	3.6	3.67
	Q1	−1.34	−1.29	−1.39	−1.32
	Q3	2.26	2.39	2.21	2.34
dobj	Median PMI	−1.51	0.25	0.36	0.44
	IQR	3.63	4.58	3.88	3.83
	Q1	−3.24	−1.81	−1.40	−1.33
	Q3	0.39	2.76	2.48	2.50
amod	Median PMI	1.95	1.80	1.44	2.07
	IQR	3.46	3.12	3.39	3.68
	Q1	−0.25	−0.58	−0.26	−0.22
	Q3	3.21	2.54	3.14	3.47

TABLE 7 | Friedman tests and pairwise Wilcoxon signed-rank tests for median PMI across reference corpora.

Dependency type	χ^2 (*df*), *p*	ACAD vs. NEWS *r, p*	ACAD vs. SP *r, p*	ACAD vs. ALL *r, p*	NEWS vs. SP *r, p*	NEWS vs. ALL *r, p*	SP vs. ALL *r, p*
amod	$\chi^2(3) = 274.644$, $p < 0.001$	−0.112, < 0.001	−0.358, < 0.001	−0.200, < 0.001	−0.373, < 0.001	0.078, < 0.001	0.268, < 0.001
advmod	$\chi^2(3) = 99.885$, $p < 0.001$	−0.068, 0.028	−0.057, 0.0646	0.259, < 0.001	−0.036, 0.2508	0.337, < 0.001	0.235, < 0.001
dobj	$\chi^2(3) = 1300.537$, $p <$ 0.001	−0.136, < 0.001	−0.166, < 0.001	0.81, < 0.001	−0.137, < 0.001	0.818, < 0.001	0.166, < 0.001

Note: Pairwise comparison cells report effect size r and Holm-adjusted p-values from Wilcoxon signed-rank tests ("r, p"). Effect sizes were computed as $r = Z/\sqrt{N}$, *where Z is the standardized test statistic, and N is the number of dependency pairs scorable in both reference corpora for that comparison. df = degrees of freedom.*

COCA-NEWS and COCA-SP, where only 15.3% of pairs crossed the $|\Delta PMI| \geq 1$ threshold.

A closer analysis reveals that this instability in PMI scores was not uniform across dependency relations. *Dobj* pairs consistently showed the greatest variability, with the proportion of substantial PMI differences ranging from 21.7% (when comparing scores based on COCA-NEWS vs. COCA-SP) to as high as 88.9% (when comparing scores based on COCA-ALL vs. COCA-NEWS). *Amod* pairs also displayed considerable variation, particularly in the COCA-ACAD vs. COCA-SP comparison (42.56%). In contrast, *advmod* pairs remained relatively stable across all corpus combinations, with most proportions below 30%.

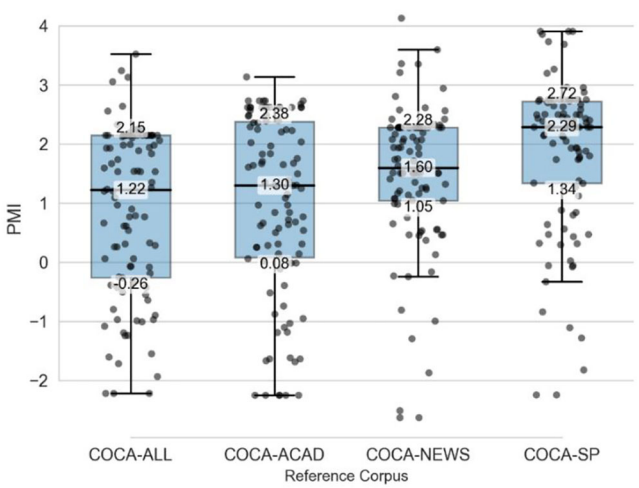

FIGURE 1 | Figure 1 Median PMI scores across reference corpora for the *amod* relation. [Color figure can be viewed at wileyonlinelibrary.com]

4.2 | Impact of Reference Corpus Selection on PMI-Based Assessment of Phraseological Competence

This section presents findings on how the choice of reference corpus systematically affects learners' overall median PMI scores and, crucially, the stability of their resulting proficiency rankings.

Figures 1–3 illustrate the distributions of learner-level median PMI scores, with each set of scores calculated using one of

TABLE 8 | Distribution of dependency pairs with substantial (| ΔPMI |≥ 1) vs. minor (| ΔPMI |< 1) absolute PMI differences across reference corpora.

Reference corpus	Dependency type	Range of |PMI difference|		Total
		Count (%) (≥ 1)	Count (%) (< 1)	
COCA-ALL and COCA-ACAD	*advmod*	261 (25.27%)	772 (74.73%)	1033 (100%)
	amod	590 (33.52%)	1170 (66.48%)	1760 (100%)
	dobj	696 (79.54%)	179 (20.46%)	875 (100%)
	Total	1547 (42.18%)	2121 (57.82%)	3668 (100%)
COCA-ALL and COCA-NEWS	*advmod*	123 (11.91%)	910 (88.09%)	1033 (100%)
	amod	141 (8.01%)	1619 (91.99%)	1760 (100%)
	dobj	778 (88.91%)	97 (11.09%)	875 (100%)
	Total	1042 (28.41%)	2626 (71.59%)	3668 (100%)
COCA-ALL and COCA-SP	*advmod*	121 (11.71%)	912 (88.29%)	1033 (100%)
	amod	130 (7.39%)	1630 (92.61%)	1760 (100%)
	dobj	776 (88.69%)	99 (11.31%)	875 (100%)
	Total	1027 (28.00%)	2641 (72.00%)	3668 (100%)
COCA-ACAD and COCA-NEWS	*advmod*	208 (20.14%)	825 (79.86%)	1033 (100%)
	amod	681 (38.69%)	1079 (61.31%)	1760 (100%)
	dobj	235 (26.86%)	640 (73.14%)	875 (100%)
	Total	1124 (30.64%)	2544 (69.36%)	3668 (100%)
COCA-ACAD and COCA-SP	*advmod*	299 (28.94%)	734 (71.06%)	1033 (100%)
	amod	749 (42.56%)	1011 (57.44%)	1760 (100%)
	dobj	368 (42.06%)	507 (57.94%)	875 (100%)
	Total	1416 (38.60%)	2252 (61.40%)	3668 (100%)
COCA-NEWS and COCA-SP	*advmod*	130 (12.58%)	903 (87.42%)	1033 (100%)
	amod	240 (13.64%)	1520 (86.36%)	1760 (100%)
	dobj	190 (21.71%)	685 (78.29%)	875 (100%)
	Total	560 (15.27%)	3108 (84.73%)	3668 (100%)

Note: ΔPMI denotes the difference in PMI for the same dependency pair computed from the two reference corpora being compared; | ΔPMI | is its absolute value.

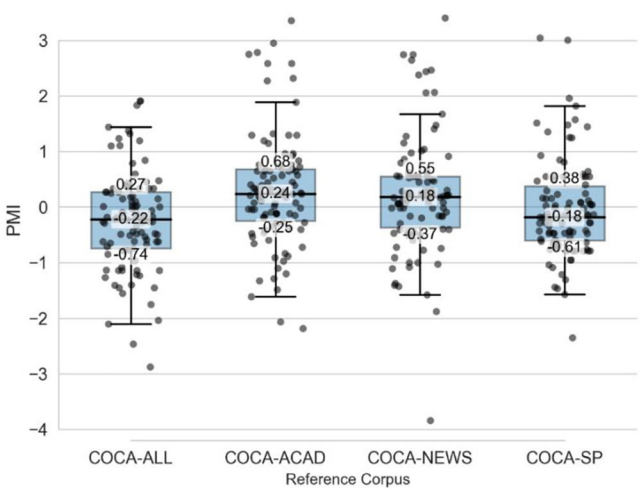

FIGURE 2 | Figure 2 Median PMI scores across reference corpora for the *advmod* relation. [Color figure can be viewed at wileyonlinelibrary.com]

the four reference corpora. A consistent pattern emerges from these distributions: PMI scores derived from the general corpus (COCA-ALL) were systematically lower than those derived from the three specialized corpora. For the *dobj* relation, for instance, the median learner score was −1.74 when benchmarked against COCA-ALL, a stark contrast to the positive median scores (≥ 0.28) obtained when using the specialized corpora as reference. This pattern of lower PMI scores when using COCA-ALL as a benchmark also held for *amod* (Mdn = 1.22 vs. ≥ 2.15) and *advmod* (Mdn = −0.22 vs. ≥ 0.18).

To confirm these visual patterns, Friedman tests were conducted, which revealed significant overall effects of corpus choice for all three dependency types: *amod* ($\chi^2(3) = 123.77$, $p < 0.001$), *advmod* ($\chi^2(3) = 174.16$, $p < 0.001$), and *dobj* ($\chi^2(3) = 67.19$, $p < 0.001$). Post-hoc Wilcoxon signed-rank tests with Holm-adjusted *p*-values (reported in Table 9) pinpointed the sources of this variation. The patterns were complex but consistently showed that the most substantial and frequent significant differences arose when scores derived from COCA-ALL were compared against those derived from any specialized sub-corpus. For *advmod*, for example,

TABLE 9 | Friedman tests and Wilcoxon signed-rank post-hoc comparisons for learners' median PMI scores across reference corpora.

Dependency	Omnibus: Friedman χ^2(df), p; Kendall's W	ACAD vs. NEWS		ACAD vs. SP		ACAD vs. ALL		NEWS vs. SP		NEWS vs. ALL		SP vs. ALL	
		r	*p*	*r*	*p*	*r*	*p*	*r*	*p*	*r*	*p*	*r*	*p*
amod	$\chi^2(3) = 123.771$, $p < 0.001$, $W = 0.421$	0.42	< 0.001	0.78	< 0.001	0.13	0.209	0.71	< 0.001	0.58	< 0.001	0.81	< 0.001
advmod	$\chi^2(3) = 174.159$, $p < 0.001$, $W = 0.592$	0.1	0.297	0.24	0.061	0.86	< 0.001	0.21	0.068	0.87	< 0.001	0.87	< 0.001
dobj	$\chi^2(3) = 67.188$, p < 0.001, $W = 0.229$	0.08	0.456	0.48	< 0.001	0.65	< 0.001	0.43	< 0.001	0.65	< 0.001	0.18	0.138

Note: Pairwise cells report effect size r and Holm-adjusted p-values from Wilcoxon signed-rank tests. Effect sizes were computed as $r = Z/\sqrt{N}$, where Z is the standardized Wilcoxon test statistic and N is the number of paired observations (learners) included in that comparison. Omnibus effect sizes are reported as Kendall's W (coefficient of concordance).

FIGURE 3 | Figure 3 Median PMI scores across reference corpora for the *dobj* relation. [Color figure can be viewed at wileyonlinelibrary.com]

comparisons against COCA-ALL produced large effect sizes (r = 0.86 to 0.87); comparisons between PMI scores derived from the different specialized corpora (e.g., COCA-ACAD-based scores vs. COCA-NEWS-based scores) were not statistically significant. A similar, though less pronounced, pattern was observed for *dobj*. For *amod*, while significant differences were widespread, the largest effect size was again found when comparing scores based on COCA-SP against those based on COCA-ALL (r = 0.81).

Results regarding the stability of learner assessments are presented in Table 10 (rank concordance) and Table 11 (rank order variation).

Spearman's correlations (Table 10) revealed that rank-order stability was highly dependent on dependency type. Rankings for *amod* (mean ρ = 0.90) and 0 *advmod* (mean ρ = 0.89) were generally stable. However, the stability for *dobj* was substantially lower (mean ρ = 0.78). Critically, the correlation dropped to just ρ = 0.59 when comparing rankings derived from COCA-ACAD against those from COCA-NEWS and COCA-SP, indicating that learner rankings can be significantly reshuffled depending on the reference corpus used.

The magnitude of this reshuffling is detailed in Table 11. For *dobj*, more than half the learners (55.6%) would see their rank change by at least 10 positions, with an average maximum shift of nearly 13 places. This level of rank instability is markedly higher than that for *amod* (35.2%) and *advmod* (42.6%).

5 | Discussion

This study set out to examine how the choice of reference corpus used for PMI calculation affects the measurement of collocational strength in EFL learners' spoken production (RQ1), and in turn, how such variation shapes the assessment of their phraseological competence (RQ2). To address these aims, PMI values were computed based on four reference benchmarks: COCA-ALL, COCA-ACAD, COCA-NEWS, and COCA-SP.

TABLE 10 | Spearman's ρ correlations for median PMI rankings across reference corpora.

Pairwise comparison	*amod*	*advmod*	*dobj*
COCA-ALL vs COCA-ACAD	0.92	0.90	0.75
COCA-ALL vs COCA-NEWS	0.92	0.93	0.90
COCA-ALL vs COCA-SP	0.93	0.92	0.90
COCA-ACAD vs COCA-NEWS	0.84	0.92	0.59
COCA-ACAD vs COCA-SP	0.88	0.80	0.59
COCA-NEWS vs COCA-SP	0.93	0.85	0.93

Note: Each cell reports the Spearman's ρ correlation between learner rankings derived from two different reference corpora.

TABLE 11 | Rank change based on median PMI scores across reference corpora by dependency.

Dependency	ALL–ACAD	ALL–NEWS	ALL–SP	ACAD–NEWS	ACAD–SP	NEWS–SP	Overall % with ΔRank ≥ 10	Overall mean maximum rank change
amod	4.33	4.33	4.22	6.00	5.50	4.15	35.2	8.53
dobj	8.27	4.75	4.67	9.73	10.12	4.31	55.6	12.67
advmod	4.69	4.31	4.44	4.70	6.76	5.85	42.6	9.41

Note: ΔRank denotes the absolute change in a learner's rank position when median PMI scores are computed using one reference corpus versus another. Columns 2–7 report the mean absolute rank difference between each pair of reference corpora. "% with ΔRank ≥ 10" indicates the proportion of learners whose rank shifted by 10 or more positions (up or down) across any corpus comparison. "Mean max rank change" is the average, across learners, of each learner's largest rank shift observed in any pairwise comparison.

5.1 | Effects of Reference Corpus Selection on PMI Scores (RQ1)

Our investigation of RQ1 yielded two key insights. First, the size and composition of the reference corpus determine the extent of learner data that serve as a basis for evaluation. Because PMI is calculated only for pairs that meet a minimum frequency threshold in the reference corpus, the content and size of that corpus directly dictate which of the learner's productions are "scorable." As our results show, a large, aggregated corpus such as COCA-ALL provided near-complete coverage (> 94%), ensuring that almost all learner productions were considered. In contrast, the three specialized corpora served as stricter filters, excluding considerable portions of learner data, with up to 27% for *amod* pairs when using COCA-SP. This difference is plausibly attributable to corpus size: COCA-ALL contains approximately 1,001,610,938 words, compared with 127,396,932 words in COCA-SP, 122,958,016 words in COCA-NEWS, and 120,988,361 words in COCA-ACAD. This finding aligns with the discussion by Paquot and Naets (2017), which underscores the importance of corpus size and scope in achieving sufficient data coverage for PMI-based evaluations.

Second, and more critically, for the same set of dependency pairs that are scorable across corpora, the analysis revealed that, when different reference corpora were used, the resulting PMI scores differed significantly. Specifically, substantial differences in PMI scores were observed for *dobj*, *amod*, and *advmod* pairs when comparing scores derived from the general COCA-ALL corpus with those from any of the specialized corpora (COCA-NEWS, COCA-ACAD, or COCA-SP). This observation is consistent with Gablasova et al. (2017), who also noted that MI scores varied significantly based on the reference corpus used, even for the same dependency pairs. Our results confirm that the statistical strength of a collocation is not fixed, but is sensitive to the distributional characteristics of the selected reference data. Additionally, our findings echo the argument made by Paquot and Naets (2017) regarding the trade-off between corpus size and comparability over coverage. While a large aggregate corpus can provide broad coverage, it may obscure varying distributions of co-occurrences across different registers and the written versus spoken divide. We note, however, that corpus size and register are partially confounded in currently available corpora (specialized registers are often represented by smaller corpora), which we treat as a methodological limitation when interpreting corpus effects.

Crucially, the *dobj* relation proved uniquely sensitive to corpus choice, with substantial shifts in over 80% of comparisons (peaking at ~89% in comparisons involving COCA-ALL with COCA-NEWS or COCA-SP). This sensitivity may plausibly be linked to the role of *dobj* pairs in encoding core propositional content (e.g., *present an argument*), which is more directly constrained by task demands. Since the assessment elicited formal, academically oriented patterns, evaluating them against reference corpora dominated by informal or narrative discourse (e.g., COCA-SP) created a structural mismatch. Consequently, valid, high register-specific combinations received artificially low PMI scores not because they were infelicitous, but because the reference corpus privileged a different set of more colloquial verb–object patterns. The observed performance drop may therefore reflect a corpus-driven register bias: PMI-based measures can penalize

TABLE 12 | Sample dependency pairs with|ΔPMI|< 1 across reference corpora.

Dependency type	Dependency pair	COCA-ALL	COCA-ACAD	COCA-NEWS	COCA-SP
dobj	broaden horizon	7.63	7.94	8.40	9.05
	alleviate poverty	6.87	6.73	7.42	7.89
	fulfill expectation	5.86	5.12	6.21	6.48
	enter workforce	5.77	5.57	5.07	6.62
	gain access	5.25	4.69	5.14	5.93
	overcome difficulty	4.10	4.71	4.53	5.49
	find niche	3.82	3.72	4.79	5.04
	acquire knowledge	3.62	4.50	3.49	4.84
	devote themselves	3.25	4.09	4.18	3.86
	provide opportunity	2.44	4.31	3.83	4.16
amod	real estate	8.22	8.34	8.65	8.11
	rapid growth	7.31	6.40	7.47	7.33
	stiff competition	6.59	6.37	7.20	7.79
	fierce competition	6.78	6.91	6.61	7.63
	sustainable development	7.12	5.86	6.07	7.25
	cheap labor	6.14	6.72	6.09	7.12
	mental health	6.60	5.70	6.58	6.91
	little bit	6.24	5.81	6.55	6.89
	migrant worker	6.92	5.55	6.09	6.65
	slow pace	5.67	6.72	5.87	6.77
advmod	weigh heavily	6.84	7.64	6.75	7.04
	perfectly reasonable	6.84	6.38	6.06	7.38
	highly competitive	6.59	5.50	6.85	7.22
	firmly believe	5.05	5.42	5.60	5.81
	relatively easy	5.20	5.59	5.36	5.72
	less demanding	5.12	5.05	4.79	5.02
	brave enough	4.43	5.94	4.39	4.92
	too much	4.47	5.18	4.88	4.23
	very difficult	4.70	4.21	5.00	4.80
	live happily	4.70	4.21	5.00	4.80

contextually appropriate choices when the reference benchmark fails to mirror the specific register of the learner output.

To illustrate how the statistically established cross-corpus PMI differences manifest linguistically, we conducted a qualitative analysis of dependency pairs, distinguishing between a "stable" set (pairs with minimal fluctuation, |ΔPMI|< 1.0 in all pairwise comparisons) and a "sensitive" set (|ΔPMI|≥ 1.0 in at least one comparison). Table 12 reports the top 10 stable dependency pairs for each type, selected by ranking them according to their mean PMI across all four corpora. This ranking ensures the selected examples are not merely stable, but are also strong and widely used collocations. Table 13 shows representative corpus-sensitive pairs for which the PMI computed from one reference corpus is higher than the corresponding PMI values from the other corpora, highlighting benchmark-specific peaks in association strength.

Our analysis reveals a clear distinction between collocations whose PMI is stable across reference corpora and those sensitive to the corpus choice. The corpus-stable pairs, shown in Table 12, largely consist of entrenched lexical units with broad applicability. For *amod* relations, items like *real estate*, *mental health*, and *fierce competition* are lexicalized compounds or widely recognized terms, resulting in consistently high PMI scores. Similarly, for *dobj* relations, pairs such as *broaden horizon*, *alleviate poverty*, and *overcome difficulty* represent core semantic concepts that transcend specific contexts. The stability of *advmod* pairs like

TABLE 13 | Sample dependency pairs with|ΔPMI|≥1 across reference corpora.

Dependency type	COCA-ALL > others	COCA-ACAD > others	COCA-NEWS > others	COCA-SP > others
dobj	/	fill gap watch movie save money spark debate earn money	give birth spark discussion give chance land job realize dream use platform	hone skill pay attention earn living achieve goal sharpen skill overcome obstacle play piano sharpen skill improve quality fulfill dream pay price
amod	social relation social security future career critical thinking	other hand big city second reason little money future career	heated discussion cultural activity different lifestyle first class stable life	upward mobility human being cutthroat competition fresh air virtuous circle doctoral student monthly salary economic growth financial institution industrial pollution
advmod	strongly advocate more balanced	let alone bring back push forward especially true very glad really want well as well perform	highly regard very peaceful also mean	much thank very significant very solid very late

firmly believe and *weigh heavily* is also consistent with their common function in expressing stance and evaluation. The robustness of these pairs suggests that a core lexicon of general-purpose collocations exists whose PMI is largely independent of the reference corpus.

In contrast, the corpus-sensitive pairs (Table°13) demonstrate that the choice of reference corpus systematically shapes which collocations are rewarded or downplayed. This variation is not random but directly reflects each corpus's lexical and functional priorities. For instance, COCA-ACAD assigns a uniquely high PMI to collocations central to academic discourse, such as *spark discussion*, and the formal discourse marker let alone. COCA-NEWS, conversely, favors pairings related to journalism and public affairs, like *spark debate* and the modifier *heated discussion*. Meanwhile, COCA-SP shows a clear preference for items related to socio-economic life and personal evaluation, such as *cutthroat competition*, *upward mobility*, *monthly salary*, and the informal phrase *little money*. The results for COCA-ALL highlight the inherent trade-off of a large, mixed corpus. Because it averages many text types, it only assigns top value to collocations common across most domains, like *social security* or *critical thinking*. In doing so, it masks the value of specialized collocations, neutralizing the distinct peaks of competence that the focused corpora reveal.

Taken together, these findings are consistent with Paquot and Naets' (2017) proposition that different reference corpora reveal distinct facets of phraseological competence. The present analysis further clarifies how this differentiation operates at an evaluative level: changing the reference corpus leads to a direct re-evaluation of the same learner production. For example, a learner's use of *spark discussion* would be highly valued when measured against COCA-ACAD, signaling proficient academic language use. However, its PMI score would be considerably lower if assessed against COCA-SP, where the pairing is less salient. Thus, the PMI score should not be interpreted as a context-independent measure of phraseological competence; rather, it rewards alignment with the target register encoded in a given reference corpus and penalizes alignment with alternative registers represented in others.

The broader implications of this systemic bias for assessing and interpreting EFL learners' phraseological competence are discussed in relation to RQ2 below.

5.2 | Impact of Reference Corpus Selection on Assessment of Phraseological Competence (RQ2)

Our analysis reveals a systematic bias when using general corpora for assessment: median PMI scores from COCA-ALL

were consistently lower than those from specialized sub-corpora. This attenuation appears to stem from the heterogeneity of general corpora, where pooling diverse registers tends to dilute statistical associations compared to the more homogeneous, register-specific profiles of specialized datasets. While COCA-ALL offers superior coverage (> 94%), our findings call into question its status as a "neutral" default. The trade-off is clear: broad coverage comes at the expense of specificity, thereby obscuring the distinct norms of the specialized registers learners aim to approximate. Consequently, using a composite baseline like COCA-ALL risks underestimating learner competence by failing to reward register-appropriate collocations, underscoring the need for register-aware benchmarking rather than a blind reliance on corpus size. Beyond this general bias, our analysis also revealed more fine-grained, specific register-driven effects from specialized corpora. Notably, COCA-Spoken produced substantially higher PMI scores for certain relations, such as the *amod* relation, compared to COCA-Academic or COCA-News. This pattern highlights the importance of register proximity between reference and learner corpora. COCA-Spoken and our learner corpus share key features: both involve planned, formal, topic-driven spoken discourse. This register alignment means that COCA-Spoken captures the type of elaborated adjectival modification (e.g., *cutthroat competition, upward mobility*) that characterizes our learners' prepared monologic speech. However, this alignment is only partial—COCA-Spoken draws from broadcast media (news interviews, talk shows), which differs from the academic commentary task performed by our learners. The resulting PMI inflation for *amod* pairs in COCA-Spoken illustrates a consequential methodological point: when the reference corpus is register-proximate to the learner corpus, PMI scores tend to be systematically higher, which may lead to an overestimation of learners' assessed competence. Conversely, register-distant corpora are liable to produce lower scores, thereby underestimating it. This sensitivity underscores the need to carefully match reference corpora to learner production tasks.

We also identified that the impact of corpus choice on the assessment of individual learners' phraseological competence varies across grammatical relations. *Dobj* dependencies exhibited the most volatility. This aligns with Paquot and Naets (2017), who noted systematic variation in MI profiles for *dobj* pairs across different reference corpora. Our study extends this observation by quantifying its concrete impact at the level of individual learners. The consequences were most pronounced in the reordering of the learner cohort: for the *dobj* relation, over half the students (55.6%) experienced a rank shift of 10 or more positions and an average maximum change of nearly 13 places (Table 10). This addresses a critical research gap by empirically demonstrating that a mismatched reference corpus, an issue already noted in the literature (e.g., Durrant and Schmitt 2009), can fundamentally alter not only scores but also the ranking of a student's perceived proficiency level. In contrast, *amod* and *advmod* dependencies demonstrated greater stability across the different reference corpora. We hypothesize that this stability is a direct reflection of the specific communicative task assigned to the learners. The TEM 8-Oral commentary task required advanced university students to deliver a prepared oral commentary on a social issue. This context naturally elicits a hybrid register that blends features from multiple domains: it is spoken in its delivery, argumentative and topical in content (similar to COCA-NEWS), and influenced by learners' formal academic training (similar to COCA-ACAD). We propose that the collocational preferences governing adjectival and adverbial modification (*amod* and *advmod*) are more portable across formal spoken and academic registers than preferences for verb–object (*dobj*) selection, which tend to be more tightly tied to specific registers and topics. In this sense, learners' recurrent use of academic-style modifiers (e.g., *significant increase*, *critically important*) can be seen as evidence of a relatively generalized formal phraseological competence that is activated in both prepared spoken commentary and written academic contexts.

These findings have important implications for the validity and fairness of automated assessment. A lack of alignment between the communicative function of a task and the reference corpus introduces construct-irrelevant variance: a learner's score becomes an artifact of the corpus's statistical properties rather than a true measure of ability. As our rank analysis shows, using different baselines (e.g., COCA-NEWS vs. COCA-SP) can fundamentally reorder student rankings, leading to inconsistent evaluations of the same underlying performance. This instability may compromise the fairness of the assessment, as high-quality, register-appropriate collocations may be penalized simply because they deviate from the norms of a poorly matched benchmark.

To mitigate these risks, we propose a principle of triangulation through multi-corpus benchmarking. Since perfect register matching is often difficult to achieve —and as our coverage analysis shows, even specialized corpora may fail to fully capture learner data—evaluators should score productions against multiple carefully selected baselines and report the resulting register sensitivity. This approach reframes corpus divergence from a source of error into an interpretative tool. For instance, convergence across corpora (as seen with *amod*) may indicate robust general competence, while divergence (e.g., a high score against COCA-SP but low against COCA-ACAD) may reflect register-specific mastery rather than a deficit. Ultimately, transparently reporting how scores fluctuate across benchmarks allows for a more nuanced, evidence-based evaluation that better captures the context-dependent nature of phraseological competence.

6 | Conclusion

This study demonstrates a central principle for PMI-based phraseological assessment: the reference corpus selection is not a methodological detail, but a primary determinant of validity. Our findings show that corpus choice generates substantial differences in PMI scores, with direct consequences for assessment outcomes. Crucially, the use of mismatched corpora or seemingly "safe" composite baselines like COCA-ALL does not simply lower precision; it actively distorts measurement. As demonstrated, register-mismatched corpora may penalize target-like features, whereas broad general corpora dilute register-specific norms, thereby obscuring the very competence the measure is intended to capture. In both cases, construct validity is compromised. Two key findings emerge. First, the trade-off between corpus size and comparability requires careful consideration. Second, different corpora reveal complementary facets of phraseological competence. Together, these insights underline the need to

critically evaluate the alignment between the reference corpus and the communicative context of learner data.

These conclusions, however, should be viewed in light of several scope-related constraints. First, the analysis is limited to three dependency relations (*amod*, *advmod*, *dobj*), which may underrepresent other relation types (e.g., subject–verb, prepositional complements) and longer multiword expressions beyond dependencies, which are closely linked to fluency and overall oral proficiency (Tavakoli and Wright 2020). Second, the reference corpora are U.S.-English dominant (COCA and its subregisters), with the "spoken" register skewed toward broadcast/semi-scripted speech, potentially mismatching the semi-planned monologues in our dataset. Third, a critical methodological limitation is the partial confounding between corpus size and register representation in currently available corpora. Specialized registers are often represented by smaller corpora, making it challenging to isolate the independent effects of register characteristics versus corpus size when interpreting our findings regarding corpus effects.

Despite these limitations, the study offers implications for assessment practice. To ensure fair and interpretable evaluations, researchers should move beyond default corpus choices and prioritize principled alignment between the reference corpus and the assessment task. This requires greater transparency in reporting the rationale for corpus selection. Future inquiries should systematically investigate how this corpus choice impacts assessment outcomes across diverse L1 backgrounds, proficiency levels, and genres. Ultimately, careful attention to reference corpus selection is essential for building valid, reliable, and context-sensitive tools for assessing phraseological competence.

Funding

This work was supported by the Program of China Scholarship Council (Grant No. 202406190117) and The International Research Foundation for English Language Education (TIRF) 2025 Doctoral Dissertation Grant.

Ethics Statement

We confirm that the manuscript has not been published or is under consideration for publication elsewhere. Further, this submission has been approved by the institution where the study was conducted. All participants have given their consent for the data to be used for research.

Conflicts of Interest

No potential conflict of interest was reported by the authors.

Data Availability Statement

The data underlying this article cannot be shared publicly to protect the privacy of individuals that participated in the study. The data that support the findings of this study is available from the corresponding author upon reasonable request.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the authors used ChatGPT (OpenAI) to refine the language, suggest alternative phrasings, and improve the overall readability of the manuscript. After using this tool, the authors carefully reviewed, revised, and approved all content and take full responsibility for the content of the published article.

Endnotes

[1] We set $f \geq 5$ because it is widely used in association-measure research to reduce chance co-occurrences and stabilize PMI estimates (Evert, 2005; Paquot, 2019). Lower cutoffs (e.g., $f \geq 3$) increase noise from very rare pairs, whereas higher cutoffs (e.g., $f \geq 10$) reduce coverage by excluding many moderately frequent combinations

[2] To ensure that learner-level median PMI scores are not driven by unequal data availability, we examined the number of PMI-eligible dependency-pair tokens per learner. Although minima are low in some cases, the interquartile ranges across corpora and dependency types are consistently moderate (typically ~9–28 tokens), indicating that most learners' medians are based on sufficient evidence. Full descriptives are available in the OSF repository (https://osf.io/3q9fs/overview?view_only=64a64969b8424f4787f4dffcc5805eeb).

References

Bachman, L., and A. Palmer. 1996. *Language Testing in Practice*. Oxford University Press.

Bestgen, Y., and S. Granger. 2014. "Quantifying the Development of Phraseological Competence in L2 English Writing: An Automated Approach." *Journal of Second Language Writing* 26: 28–41. https://doi.org/10.1016/j.jslw.2014.09.004.

Biber, D. 2009. "A Corpus-Driven Approach to Formulaic Language in English: Multi-Word Patterns in Speech and Writing." *International Journal of Corpus Linguistics* 14, no. 3: 275–311. https://doi.org/10.1075/ijcl.14.3.08bib.

Biber, D., and F. Barbieri. 2007. "Lexical Bundles in University Spoken and Written Registers." *English for Specific Purposes* 26, no. 3: 263–286. https://doi.org/10.1016/j.esp.2006.08.003.

Biber, D., S. Johansson, G. Leech, S. Conrad, E. Finegan, and R. Quirk. 1999. *Longman Grammar of Spoken and Written English*. Longman.

Church, K. W., and P. Hanks. 1990. "Word Association Norms, Mutual Information, and Lexicography." *Computational Linguistics* 16, no. 1: 22–29. https://aclanthology.org/J90-1003/.

Cohen, J. 1988. *"The Effect Size." In Statistical Power Analysis for the Behavioral Sciences*, 77–83. Routledge.

Davies, M. 2020. *The Corpus of Contemporary American English (COCA)*. Brigham Young University. https://www.english-corpora.org/coca/.

Durrant, P., and N. Schmitt. 2009. "To What Extent Do Native and Non-Native Writers Make Use of Collocations?" *IRAL—International Review of Applied Linguistics in Language Teaching* 47, no. 2: 157–177. https://doi.org/10.1515/iral.2009.007.

Egbert, J., D. Biber, and B. Gray. 2022. "Corpus Representativeness: A Conceptual and Methodological Framework." In *Designing and Evaluating Language Corpora: A Practical Framework for Corpus Representativeness*, 52–67. Cambridge University Press.

Ellis, N. C., R. Simpson-Vlach, and C. Maynard. 2008. "Formulaic Language in Native and Second Language Speakers: Psycholinguistics, Corpus Linguistics, and TESOL." *TESOL Quarterly* 42, no. 3: 375–396. https://doi.org/10.1002/j.1545-7249.2008.tb00137.x.

Evert, S. 2005. "*The Statistics of Word Co-Occurrences: Word Pairs and Collocations*." Doctoral diss., Institut für maschinelle Sprachverarbeitung, University of Stuttgart.

Fan, M. 2009. "An Exploratory Study of Collocational Use by ESL Students—A Task-Based Approach." *System* 37, no. 1: 110–123. https://doi.org/10.1016/j.system.2008.06.004.

Gablasova, D., V. Brezina, and T. McEnery. 2017. "Collocations in Corpus-Based Language Learning Research: Identifying, Comparing, and

Interpreting the Evidence." *Supplement, Language Learning* 67, no. S1: S155–S179. https://doi.org/10.1111/lang.12225.

Granger, S., and M. Paquot. 2008. "Disentangling the Phraseological Web." In S. Granger and F. Meunier, 27–49. John Benjamins Publishing Company. https://doi.org/10.1075/z.139.07gra.

Granger, S., and Y. Bestgen. 2014. "The Use of Collocations by Intermediate and Advanced L2 Writers: A Bigram-Based Study." *International Review of Applied Linguistics in Language Teaching* 52, no. 3: 229–252. https://doi.org/10.1515/iral-2014-0011.

Gries, S. Th 2022. "What Do (Some of) Our Association Measures Measure (Most)? Association?" *Journal of Second Language Studies* 5, no. 1: 1–33.

Jiang, J., P. Bi, N. Xie, and H. Liu. 2023. "Phraseological Complexity and Low- and Intermediate-Level L2 Learners' Writing Quality." *International Review of Applied Linguistics in Language Teaching* 61, no. 3: 765–790. https://doi.org/10.1515/iral-2019-0147.

Lado, R. 1961. *The Construction and Use of Foreign Language Tests*. McGraw-Hill.

Laufer, B., and T. Waldman. 2011. "Verb-Noun Collocations in Second Language Writing: A Corpus Analysis of Learners' English." *Language Learning* 61, no. 2: 647–672. https://doi.org/10.1111/j.1467-9922.2010.00621.x.

Paquot, M. 2018. "Phraseological Competence: A Missing Component in University Entrance Language Tests? Insights From a Study of EFL Learners' Use of Statistical Collocations." *Language Assessment Quarterly* 15, no. 1: 29–43. https://doi.org/10.1080/15434303.2017.1405421.

Paquot, M. 2019. "The Phraseological Dimension in Interlanguage Complexity Research." *Second Language Research* 35, no. 1: 121–145. https://doi.org/10.1177/0267658317694221.

Paquot, M., and H. Naets. 2017. "The Role of the Reference Corpus in Studies of EFL Learners' Use of Statistical Collocations." Paper presented at ICAME38, Prague, Czech Republic, May 24–28.

Paquot, M., and H. Naets. 2025. "Phraseological Sophistication as a Multidimensional Construct: Exploring the Relationship Between Association, Register Specificity and Frequency of Word Combinations." *International Journal of Learner Corpus Research* 11, no. 1: 217–244. https://doi.org/10.1075/ijlcr.23033.paq.

Paquot, M., and S. Granger. 2012. "Formulaic Language in Learner Corpora." *Annual Review of Applied Linguistics* 32: 130–149. https://doi.org/10.1017/S0267190512000098.

Paquot, M., D. Gablasova, V. Brezina, and H. Naets. 2022. "Phraseological complexity in EFL learners' spoken production across proficiency levels." In A. Leńko-Szymańska and S. Götz (Eds.), *Studies in corpus linguistics* (Vol. 104, pp. 115–136). John Benjamins. https://doi.org/10.1075/scl.104.05paq.

Sinclair, J. M. 1991. *Corpus, Concordance, Collocation*. Oxford University Press.

Siyanova-Chanturia, A. 2015. "Collocation in Beginner Learner Writing: A Longitudinal Study. System." 53: 148–160. https://doi.org/10.1016/j.system.2015.07.003.

Tavakoli, P., and T. Uchihara. 2020. "To What Extent Are Multiword Sequences Associated With Oral Fluency?" *Language Learning* 70, no. 2: 506–547. https://doi.org/10.1111/lang.12384.

Tsai, K.-J. 2015. "Profiling the Collocation Use in ELT Textbooks and Learner Writing." *Language Teaching Research* 19, no. 6: 723–740. https://doi.org/10.1177/1362168814559801.

Vandeweerd, N., A. Housen, and M. Paquot. 2023. "Proficiency at the Lexis–Grammar Interface: Comparing Oral Versus Written French Exam Tasks." *Language Testing* 40, no. 3: 658–683. https://doi.org/10.1177/02655322231153543.

Wang, T., and D. Zhou. 2025. "Predictive Effects of Phraseological Complexity on the Quality of Chinese EFL Learners' L2 Oral Production." *International Journal of Applied Linguistics* 35, no. 2: 810–823. https://doi.org/10.1111/ijal.12662.

Wray, A. 2002. *Formulaic Language and the Lexicon*. Cambridge University Press.

Yoon, H.-J. 2016. "Association Strength of Verb-Noun Combinations in Experienced NS and Less Experienced NNS Writing: Longitudinal and Cross-Sectional Findings." *Journal of Second Language Writing* 34: 42–57. https://doi.org/10.1016/j.jslw.2016.11.001.