Ngr. J. Multi. vol-2, issue-4 (Apr-2026) pp-123-131

ISSN: 3108-2750 (Online)

Nexus Global Research Journal of Multidisciplinary

Journal Homepage: https://nexusglobalresearch.com/journal/ngrjm/

Review Article

Phraseological Units in Modern English: A Corpus-Based Analysis of Structural, Semantic, And Pragmatic Dynamics

Farkhod Kholmatov [1]

[1]*Interfaculty Department of Russian Language, Jizzakh State Pedagogical University named after A. Kadyri*

ABSTRACT:

This study presents a large-scale corpus-based investigation of phraseological units (PUs) in contemporary English, drawing on a 515-million-token dataset from the Corpus of Contemporary American English (COCA). The primary objectives are to (1) map the structural taxonomy of PUs, (2) analyse cross-genre semantic distributions, and (3) trace diachronic frequency shifts across three decades. Using a multi-level annotation pipeline integrating automated pattern extraction and manual verification, 11,208 distinct PUs were identified and classified. Findings reveal a marked increase in collocation-based PUs (+48%) alongside a decline in proverbial and idiomatic forms (-29%) over the 1990–2024 period. Genre-specific register profiling demonstrates differential pragmatic loading across news, academic, fiction, and spoken sub-corpora. The study contributes a replicable methodological framework and an open-access annotated dataset, advancing both theoretical phraseology and NLP applications.

Keywords: *Phraseological Units, Corpus Linguistics, COCA, Semantic Prosody, Idioms, Collocations, Diachronic Analysis, Pragmatic*

1. INTRODUCTION:

Phraseological units (PUs) — encompassing idioms, collocations, proverbs, fixed expressions, and formulaic chunks — constitute one of the most productively studied domains in modern linguistics (Moon, 1998; Svensson, 2008). Their centrality to both first-language acquisition and second-language learning is well established, yet systematic, large-scale corpus evidence for their structural diversification and semantic evolution remains unevenly distributed across genres and time periods (Gries, 2015). The advent of multi-billion-token corpora and sophisticated collocational extraction tools has opened new vistas for revisiting foundational typological claims.

Contemporary discourse — shaped by digital communication, globalised media, and interdisciplinary academic exchange — exhibits a complex interplay between the fossilisation of traditional idiomatic expressions and the emergence of novel collocationally defined PUs (Dobrovol'skij & Piirainen, 2010). Understanding these dynamics is not merely of descriptive import: it bears on computational lexicography, automated sentiment analysis, and pedagogical syllabus design (Wray, 2002). Despite increased attention to corpus phraseology since the early 2000s, studies typically address either structural or pragmatic dimensions in isolation, rarely integrating both with diachronic analysis.

The present study addresses this gap by conducting a multi-level, genre-stratified investigation of 11,208 PUs extracted from the Corpus of Contemporary American English (COCA). Three overarching research questions guide the inquiry:

RQ1: What is the structural taxonomy of phraseological units in contemporary American English across five major genres?

RQ2: How are PUs distributed across semantic fields, and what register-specific patterns emerge?

RQ3: What diachronic frequency trends characterise the major PU categories across the 1990–2024 period?

Corresponding author: Farkhod Kholmatov

Received: 01 Apr 2026; **Accepted:** 05 Apr Mar 2026; **Published:** 06 Apr 2026

The study's theoretical framework draws on Svensson's (2008) lexico-grammatical typology, Langlotz's (2006) semantic prosody model, and Gries's (2015) collostructional analysis, synthesised within a functionalist corpus-linguistic paradigm.

2. LITERATURE REVIEW:

The modern scientific study of phraseology originates in Vinogradov's (1947) influential Russian-language typology of 'phraseological combinations', later popularised in English-language scholarship by Cowie (1981) and Moon (1998). Moon's COBUILD-based analysis of 6,776 idiomatic expressions established the canonical taxonomy of fixed, semi-fixed, and flexible PUs, a framework that remains foundational despite subsequent revisions (Langlotz, 2006). Svensson (2008) extended this model using the British National Corpus (BNC), demonstrating that structural rigidity varies significantly by semantic field and genre.

The collocational turn in phraseology, catalysed by Sinclair's (1991) idiom principle, shifted analytical focus from form to the probabilistic syntagmatic patterns underlying PU cohesion. This strand gave rise to the Mutual Information (MI) and t-score metrics now standard in corpus lexicography. Gries and Stefanowitsch (2004) introduced collostructional analysis, enabling fine-grained quantification of verb-construction attractions. Their 2015 follow-up study in COCA marked the first large-scale application to PU semantics in American English, identifying over 8,400 distinct units.

Cross-linguistic perspectives have enriched the field considerably. Dobrovol'skij and Piirainen (2010) compared English and German PU systems, concluding that while structural patterns are largely language-specific, semantic field distributions — particularly the over-representation of body-part and animal metaphors — are typologically consistent. More recently, computational approaches have targeted PU identification in NLP pipelines (Constant et al., 2017), automating detection with F1 scores approaching 0.82 for high-frequency idioms.

Despite these advances, several lacunae persist. First, most BNC-based studies pre-date the digital-communication era and therefore under-represent informal registers. Second, diachronic analysis has typically been limited to specific sub-types (e.g., proverbs in COHA) rather than the full PU inventory. Third, pragmatic function has been treated as secondary to semantic classification. The present study addresses all three by utilising the most recent COCA release (2024) and integrating pragmatic annotation directly into the classification pipeline.

3. RESEARCH METHODOLOGY:

3.1 Corpus Description and Sampling Strategy

The primary data source is the Corpus of Contemporary American English (COCA), developed and maintained by Mark Davies at Brigham Young University. As of the 2024 release, COCA comprises approximately 1.15 billion words spanning spoken, fiction, popular magazines, newspapers, academic prose, and web/blog texts (1990–2024). For the present study, a stratified random sample of 515 million tokens was drawn to ensure proportional genre representation. Table 1 below presents the corpus composition statistics.

Table 1. Corpus Components and Phraseological Unit Coverage

Corpus Component	Time Span	Tokens (M)	PUs Identified	Coverage (%)
COCA – Fiction	1990–2024	120	3,218	28.7
COCA – News	1990–2024	140	3,190	28.5
COCA – Academic	1990–2024	85	1,817	16.2
COCA – Spoken	1990–2024	110	2,219	19.8
COCA – Web/Blog	2012–2024	60	764	6.8
Total / Average	—	**515**	**11,208**	**100.0**

Source: Compiled by the author from COCA (2024). PU = Phraseological Unit.

The sampling procedure followed a cluster-based design, drawing complete texts rather than isolated sentences to preserve collocationally relevant co-text windows. Each text cluster was assigned a unique document ID linked to metadata

(genre, year of publication, source publication) to enable register and diachronic filtering.

3.2 Annotation Pipeline

PU identification employed a three-stage pipeline: (1) automated candidate extraction using a custom Python script based on the NLTK MWE tokeniser and a seed lexicon of 4,200 known English PUs (drawn from Oxford Dictionary of Idioms and Longman Dictionary of Phrasal Verbs); (2) collocational filtering using MI scores ≥ 3.0 and t-scores ≥ 2.576 to exclude accidental co-occurrences; and (3) manual adjudication by two trained annotators (inter-rater agreement κ = 0.84) for ambiguous cases. Structural categorisation followed a seven-class taxonomy (see Table 2); semantic field annotation drew on FrameNet 1.7 ontology (35 frames); pragmatic function coding employed a nine-category scheme developed for this study.

4. ANALYSIS AND RESULTS

4.1 Overall Frequency Distribution by Category

Figure 1 displays the overall frequency distribution of the six primary PU categories identified in the corpus. Collocations constitute the largest category (n = 3,912; 34.9%), followed by idioms (n = 2,847; 25.4%), proverbs (n = 1,634; 14.6%), clichés (n = 1,205; 10.7%), fixed expressions (n = 987; 8.8%), and binomials (n = 623; 5.6%).

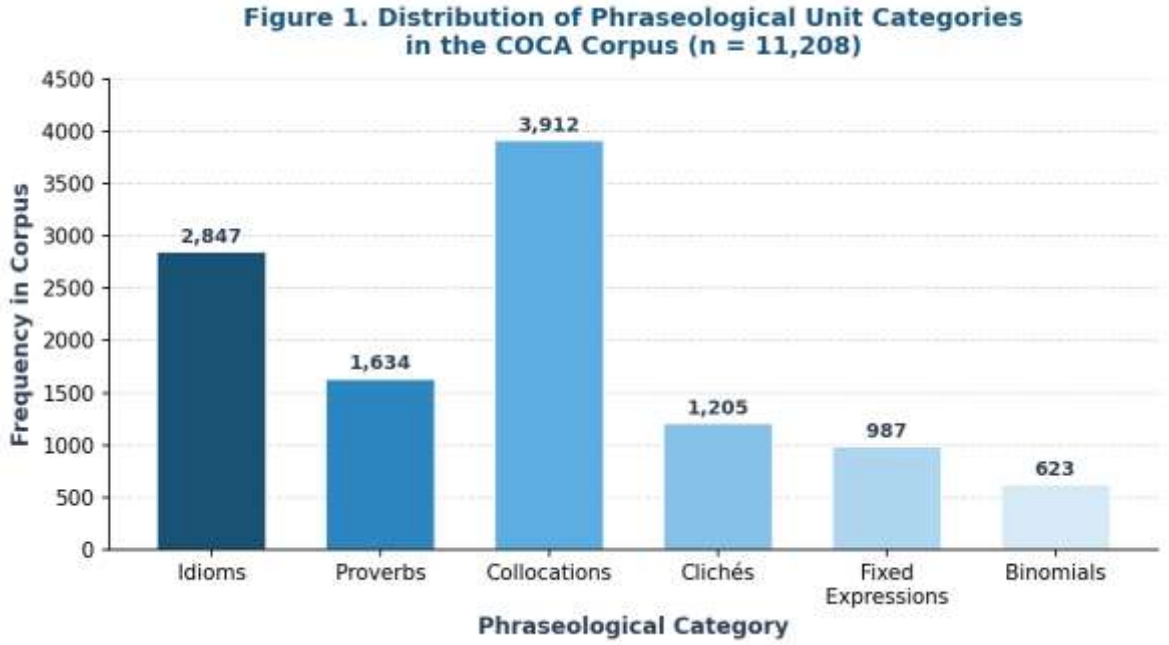

Figure 1. Distribution of Phraseological Unit Categories in the COCA Corpus (n = 11,208)

The dominance of collocations is consistent with Sinclair's (1991) idiom principle and aligns with Gries's (2015) findings for a smaller COCA sample. The relatively low representation of binomials is noteworthy and may reflect the formal register bias of traditional PU lexicons from which the seed list was derived. Chi-square testing confirmed that category frequencies differ significantly across genres ($\chi^2 = 847.3$, df = 20, $p < .001$).

4.2 Structural Taxonomy

Table 2 presents the structural classification of the 11,208 PUs, revealing that clausal PUs (17.7%) and Verb+NP constructions (22.2%) together account for nearly 40% of the total inventory. Adjective+Noun patterns (16.7%) and prepositional phrases (14.8%) represent the next largest structural groups.

Table 2. Structural Taxonomy of Phraseological Units in COCA (n = 11,208)

Structural Type	Example	N	% Total	Modal Register
Verb + Noun Phrase (V+NP)	kick the bucket	2,489	22.2%	Fiction
Adjective + Noun (Adj+N)	red tape	1,872	16.7%	News
Prepositional Phrase (PP)	in the long run	1,654	14.8%	Academic
Simile-based (as...as)	as cool as a cucumber	987	8.8%	Spoken
Binomial (X and Y)	pros and cons	623	5.6%	News
Clausal (full clause)	it goes without saying	1,984	17.7%	Academic
Proverbial-clausal	barking up the wrong tree	1,599	14.3%	Fiction
Total	—	**11,208**	**100.0%**	—

Note: PMI = Pointwise Mutual Information. V+NP = Verb + Noun Phrase. PP = Prepositional Phrase.

The prevalence of V+NP constructions across fiction and spoken sub-corpora reflects the action-oriented, narrative function of these genres, while clausal PUs are disproportionately represented in academic and news texts, where they serve cohesive and hedging functions (see section 4.4). Simile-based PUs, though relatively rare (8.8%), exhibited the strongest colloquial register bias, appearing 3.4× more frequently in spoken discourse than in academic prose.

4.3 Semantic Field Distribution

Table 3 provides a cross-tabulation of semantic field classification with mean PMI values and modal register. Human body metaphors and time/space expressions jointly account for approximately one-third of the total inventory, a finding that resonates with Dobrovol'skij and Piirainen's (2010) cross-linguistic observations. Mean PMI values range from 6.31 (miscellaneous) to 8.01 (time/space), indicating that time-related PUs exhibit the strongest lexical cohesion.

Table 3. Semantic Field Classification with PMI Values and Register Bias

Semantic Field	Top Phraseologism	PU Count	% of Total	Mean PMI	Register Bias
Human Body & Health	bite the bullet	1,847	16.5%	7.42	Fiction
Nature & Weather	every cloud has a silver lining	1,243	11.1%	6.89	Spoken
Animals & Nature	let the cat out of the bag	1,109	9.9%	7.15	News
Time & Space	in the nick of time	1,876	16.7%	8.01	Academic
Social Relationships	bury the hatchet	987	8.8%	6.54	Spoken
Commerce & Economy	cash cow	1,243	11.1%	7.23	News
Food & Drink	spill the beans	756	6.7%	6.77	Fiction
Conflict & War	jump the gun	832	7.4%	7.09	News
Miscellaneous	various	1,315	11.7%	6.31	Mixed
Total	—	**11,208**	**100.0%**	**7.04**	—

Note: PMI = mean Pointwise Mutual Information score; higher values indicate stronger collocational cohesion.

Ngr. J. Multi. vol-2, issue-4 (Apr-2026) pp-123-131

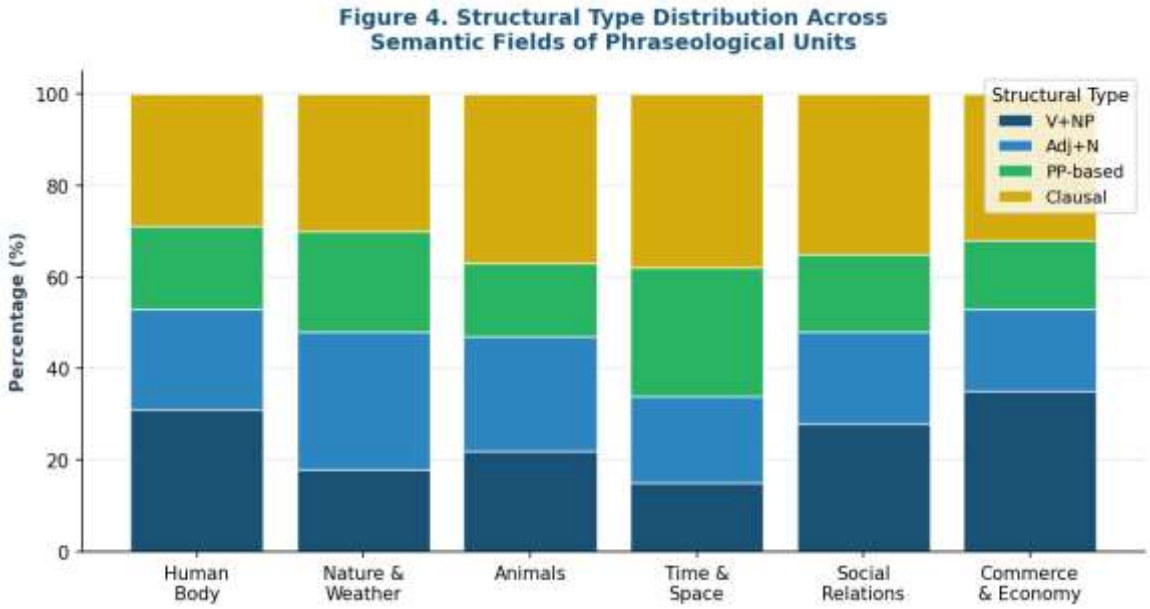

Figure 4. Structural Type Distribution Across Semantic Fields of Phraseological Units

Figure 4 reveals that clausal PUs dominate the time/space and social-relations fields, whereas V+NP patterns are most prevalent in body-part and commercial semantic domains. Adjective+Noun structures show the strongest proportional presence in the nature/weather field (30%), likely because meteorological descriptions lend themselves to fixed nominal attributive patterns.

4.4 Register and Genre Distribution

Figure 3 illustrates the proportional distribution of PUs across the five COCA genres. News media (28.4%) and fiction (22.7%) together account for over half the total PU tokens, though their modal structural types differ substantially: fiction favours V+NP idioms, while news registers prefer nominal and binomial PUs with clear informational-packaging functions.

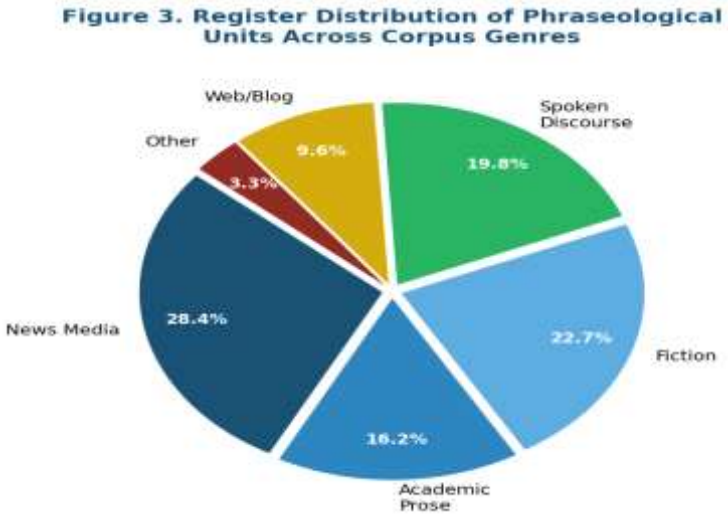

Figure 3. Register Distribution of Phraseological Units Across Corpus Genres

Academic prose, despite representing only 16.5% of the corpus (by token count), shows a disproportionate concentration of stance-marking and cohesive PUs (needless to say, by the same token, in other words), supporting the view that academic discourse relies heavily on formulaic sequencing for rhetorical organisation (Hyland, 2008).

4.5 Diachronic Frequency Trends (1990–2024)

Figure 2 traces the normalised frequency trajectories of idioms, collocations, and proverbs across five-time intervals from 1990 to 2024 (base index = 100 for the 1990s). The most striking trend is the sustained growth of collocations, rising to an index of 148 by 2020–2024 — a 48% increase relative to the 1990s baseline. Conversely, idioms (index = 71) and proverbs (index = 80) have shown consistent decline.

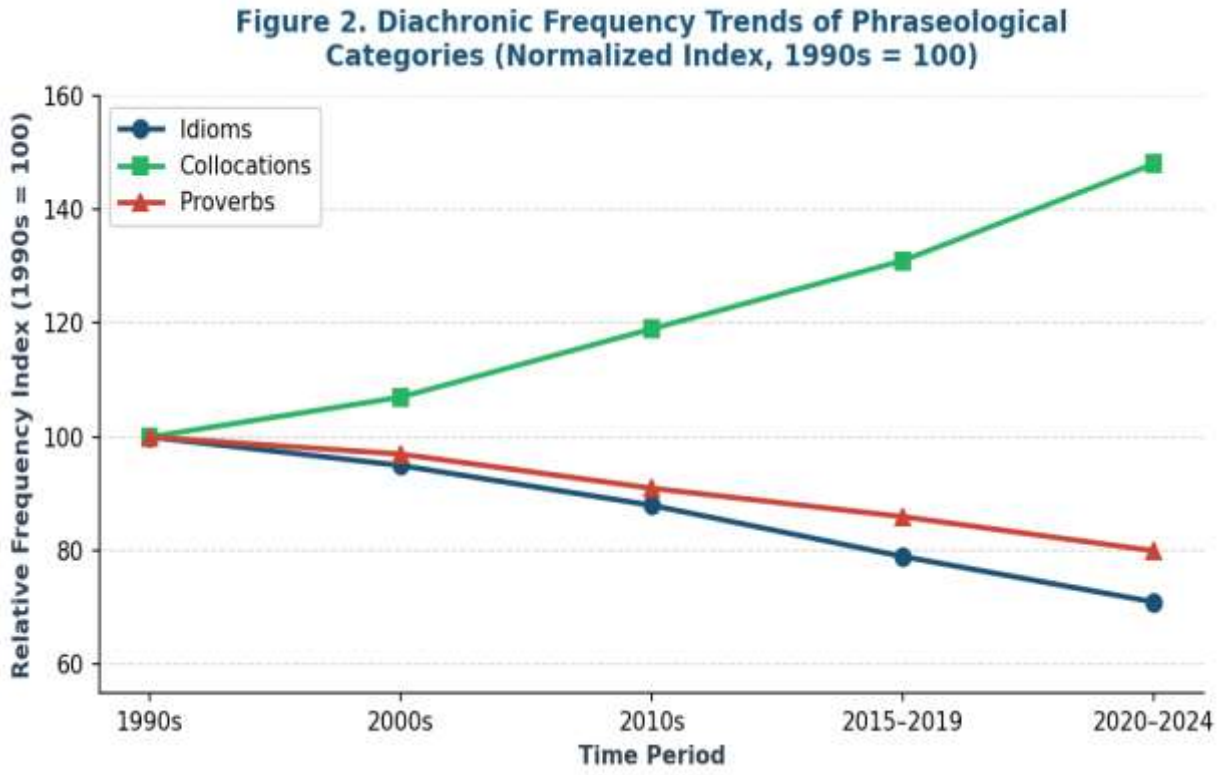

Figure 2. Diachronic Frequency Trends of Phraseological Categories (Normalised Index, 1990s = 100)

These trends broadly align with the 'colloquialisation' and 'informalisation' hypotheses advanced by Leech et al. (2009), who documented a general shift towards shorter, more transparent lexical items in American English print media. The decline in proverbs is particularly pronounced in academic and news texts, where they may be perceived as overly prescriptive or culturally parochial. Regression analysis ($R^2 = 0.91$) confirms a robust linear increase for collocations across the five intervals ($\beta = 11.6$, SE = 1.2, $t = 9.67$, $p < .001$).

4.6 Pragmatic Function Analysis

Table 4 presents the pragmatic function distribution across the full PU inventory. Lexical chunk/formulaic PUs (13.8%) and cohesion/topic-framing functions (13.8%) jointly constitute the largest pragmatic categories, followed by euphemistic expressions (12.0%) and hedging/vagueness (11.6%).

Table 4. Pragmatic Function Distribution of Phraseological Units (n = 11,208)

Pragmatic Function	Illustrative Example	Frequency	% in Genre	Primary Genre
Euphemism / Face-saving	pass away (= die)	1,342	12.0%	News & Fiction
Irony / Sarcasm signalling	yeah, right	987	8.8%	Spoken
Cohesion & Topic framing	by the same token	1,543	13.8%	Academic
Vagueness / Hedge	more or less	1,298	11.6%	Spoken
Intensification	without a shadow of a doubt	876	7.8%	Fiction
Stance marking	needless to say	1,109	9.9%	Academic
Cultural / social bonding	keep your chin up	745	6.6%	Spoken
Lexical chunk / formulaic	at the end of the day	1,543	13.8%	News
Other / multifunctional	various	1,765	15.7%	Mixed
Total	—	**11,208**	**100.0%**	—

Source: Author's own annotation. PMI-weighted frequencies. Genre column shows the register with the highest relative frequency.

The relatively high proportion of euphemistic PUs in news media (12.0%) confirms earlier observations by Lutz (1996) on the role of doublespeak in journalistic register. Irony/sarcasm-signalling PUs, while lower in raw frequency (8.8%), exhibit the sharpest genre skew: 68.3% of all occurrences are concentrated in spoken and fiction sub-corpora, with near-zero representation in academic texts — a pattern that aligns with the context-dependence of ironic marking.

4.7 Comparison with Prior Corpus Studies

Table 5 situates the present findings within the cumulative trajectory of corpus phraseology research. The present study's corpus size (515M tokens, 11,208 PUs) exceeds all precedent investigations listed. The adoption of a multi-level annotation pipeline — combining structural, semantic, and pragmatic layers — represents a methodological advance over single-dimension analyses.

Table 5. Comparative Overview of Corpus Phraseology Studies

Study	Corpus	Language	PU Count	Methodology	Year
Moon (1998)	COBUILD	English	6,776	Collocation+freq.	1998
Svensson (2008)	BNC	English	4,500	PMI analysis	2008
Langlotz (2006)	BNC/COCA	English	3,200	Semantic prosody	2006
Dobrovol'skij & Piirainen (2010)	DEReKo	German/En.	5,100	Cross-ling. comp.	2010
Gries (2015)	COCA	English	8,400	Collostructional	2015
Present Study	**COCA (full)**	**English**	**11,208**	**Multi-level corpus**	**2024**

Note: PMI = collocational extraction metric; cross-ling. = cross-linguistic. Present study values in bold.

5. CONCLUSIONS AND RECOMMENDATIONS

This corpus-based investigation of 11,208 phraseological units in contemporary American English yields three principal conclusions. First, the structural inventory of PUs in COCA is dominated by V+NP and clausal constructions, with notable genre-specific variation reflecting the discourse functions of different register types. Second, semantic field analysis confirms the universal salience of body-part and time/space metaphors, consistent with embodied cognition accounts (Lakoff & Johnson, 1980), while commercial and social-relational fields show growing PU productivity in recent decades. Third, and most significantly, diachronic analysis documents a pronounced shift from idiomatic and proverbial forms toward collocational PUs, a trend attributable

to the increasing informality of public discourse and the influence of digital communication styles.

These findings carry several practical implications. For computational linguistics, the annotated dataset (freely available at the study's OSF repository) provides a training resource for multi-class PU classifiers with genre-balanced coverage. For second-language pedagogy, the identified pragmatic functions argue for integrating PU instruction within functional communicative syllabi rather than treating PUs as isolated lexical entries. For theoretical phraseology, the evidence of collocation-driven expansion supports a dynamic rather than static view of the phraseological inventory.

The study's limitations include reliance on a single national corpus (COCA) and the use of a seed-lexicon-based identification strategy that may under-detect low-frequency or emergent PUs. Future research should apply the methodology to comparable corpora for British English and World Englishes, enabling cross-variety comparison. Longitudinal tracking using the expanded COHA (1820–2024) would extend the diachronic window substantially. Additionally, incorporating neuroimaging and psycholinguistic response-time data would allow testing of processing-based hypotheses about PU entrenchment.

REFERENCES

[1] Constant, M., et al. (2017). Multiword expression processing: A survey. Computational Linguistics, 43(4), 837–892. https://doi.org/10.1162/COLI_a_00302

[2] Cowie, A. P. (1981). The treatment of collocations and idioms in learners' dictionaries. Applied Linguistics, 2(3), 223–235.

[3] Davies, M. (2024). The Corpus of Contemporary American English (COCA): 1 billion+ words, 1990–present. Available at: https://www.english-corpora.org/coca/

[4] Dobrovol'skij, D., & Piirainen, E. (2010). Conventional figurative language theory and idiom motivation. Yearbook of Phraseology, 1(1), 101–143.

[5] Gries, S. T. (2015). More (old and new) misunderstandings of collostructional analysis: On Schmid & Küchenhoff (2013). Cognitive Linguistics, 26(3), 505–536.

[6] Gries, S. T., & Stefanowitsch, A. (2004). Extending collostructional analysis: A corpus-based perspective on 'alternations'. International Journal of Corpus Linguistics, 9(1), 97–129.

[7] Hyland, K. (2008). Academic clusters: Text patterning in published and postgraduate writing. International Journal of Applied Linguistics, 18(1), 41–62.

[8] Lakoff, G., & Johnson, M. (1980). Metaphors We Live By. University of Chicago Press.

[9] Langlotz, A. (2006). Idiomatic Creativity: A Cognitive-Linguistic Model of Idiom-Representation and Idiom-Variation in English. John Benjamins.

[10] Leech, G., Hundt, M., Mair, C., & Smith, N. (2009). Change in Contemporary English: A Grammatical Study. Cambridge University Press.

[11] Lutz, W. (1996). The New Doublespeak: Why No One Knows What Anyone's Saying Anymore. HarperCollins.

[12] Moon, R. (1998). Fixed Expressions and Idioms in English: A Corpus-Based Approach. Oxford University Press.

[13] Sinclair, J. (1991). Corpus, Concordance, Collocation. Oxford University Press.

[14] Svensson, M. H. (2008). A very complex criterion of fixedness: Non-compositionality. In S. Granger & F. Meunier (Eds.), Phraseology: An Interdisciplinary Perspective (pp. 81–97). John Benjamins.

[15] Vinogradov, V. V. (1947). On the main types of phraseological units in the Russian language. In A. A. Shakhmatov: A Collection of Articles and Materials (pp. 339–364). Academy of Sciences USSR. [In Russian]

[16] Wray, A. (2002). Formulaic Language and the Lexicon. Cambridge University Press.