Beyond Single Words: MWE Identification in Bioinformatics Research Articles and Dispersion Profiling Across IMRaD

Jurgi Giraud and Andrew Gargett
School of Languages and Applied Linguistics
The Open University
Milton Keynes, UK
jurgi.giraud@open.ac.uk andrew.gargett@open.ac.uk

Abstract

Multiword Expressions (MWEs) are pervasive in scientific writing, and in specialized domains they include both multiword terminology (e.g., noun compounds) and recurrent academic phrasing. This study profiles MWEs in a large corpus of bioinformatics research articles segmented by IMRaD sections. Building on recent multi-method approaches to scientific MWE identification, we extract MWEs using complementary automated strategies (semantic matching, dependency parsing, controlled vocabularies, and academic formula lists) and compare the resulting inventories by size, form, and IMRaD section distribution. We further quantify cross-document dispersion using document frequency and Gries' DP to distinguish widely reused expressions from items concentrated in a small subset of articles. Results show that bioinformatics MWEs are predominantly short and nominal, but that extraction methods differ in the extent to which they recover discourse and reporting phraseology. Dispersion is strongly long-tailed across sections with most MWEs being document-specific, while a smaller recurrent core aligns with section function and is enriched for conventional templates and standardized multiword terms. Overall, the findings argue for combining complementary identification methods with dispersion profiling to characterize domain "multiwordness" in a principled and section-sensitive way.

1 Introduction and Background

Multiword Expressions (MWEs) refer to a broad class of linguistic forms that span conventional word boundaries, consisting of two or more words that function as a single unit with semantic, syntactic, and/or lexical properties (Constant et al., 2017; Sag et al., 2002). They encompass a heterogeneous set of items including idioms, collocations, phrasal verbs, fixed or semi-fixed phrases, lexicalized compounds, and institutionalized expressions, which appear across languages with varying degrees of compositionality and predictability (Villavicencio et al., 2005; Constant et al., 2017; Masini, 2019). MWEs are not just a feature of general language: they are central to specialized discourse and scientific writing, with prior work applying MWE extraction in scientific corpora (Kim et al., 2018; Premasiri et al., 2023; Alves et al., 2024; Bagdasarov and Teich, 2024; Alves et al., 2025; Florescu and Ohniwa, 2025). Scientific research articles, in particular, make extensive use of dense nominal style and increasingly complex noun phrases to pack information efficiently (Biber and Gray, 2016; Degaetano-Ortlieb and Teich, 2018; Bagdasarov and Teich, 2024). MWEs often correspond to key domain concepts (e.g., *gene expression profile*) or conventional academic phrases (e.g., *in this study*). This tendency is especially salient in bioinformatics, an interdisciplinary field that sits at the crossroads of biology, biomedical science, and computer science (Nakaya, 2021) with a rapidly evolving terminology.

Research articles also exhibit systematic variation across sections. The IMRaD convention (Introduction–Methods–Results–Discussion), now widely adopted in the biomedical sciences, reflects distinct communicative purposes and is associated with measurable differences in rhetorical and lexico-grammatical choices (Sollaci and Pereira, 2004; Wu, 2011). A section-aware perspective is therefore valuable for locating where domain-specific terminology concentrates (e.g., procedural labels in Methods) and where formulaic discourse markers cluster (e.g., result-reporting patterns) (Sollaci and Pereira, 2004; Wu, 2011; Hyland, 2012).

Beyond identifying MWEs, it is also important to determine whether they constitute broadly shared phraseological resources or remain localized to specific papers. Dispersion profiling provides this functional perspective by quantifying how evenly an expression is distributed across doc-

Proceedings of the 22nd Workshop on Multiword Expressions (MWE 2026), pages 86–95

uments within a corpus or subcorpus (Gries, 2021). In corpus linguistics, dispersion measures complement frequency by distinguishing items that are frequent because they recur widely from items that are frequent but concentrated in a small subset of texts (Gries, 2021). This distinction is particularly relevant for scientific MWEs as some are productive and topic- or dataset-contingent (e.g., novel noun compounds), while others behave like reusable templates (e.g., reporting or framing formulas), and these differences are expected to vary by IMRaD section.

To support section-aware analyses of bioinformatics phraseology, we compiled BIOMONO_EN, a large English corpus of open-access bioinformatics research articles. Using complementary MWE identification strategies (lexicon-based semantic tagging, dependency-based extraction, ontology matching, and academic formula lists), we characterize bioinformatics MWEs across IMRaD sections and profile their dispersion using document frequency and Gries' DP (Gries, 2021). We show that (i) nominal multiword terminology dominates, with noun compounds accounting for the majority of dependency-derived MWEs; (ii) formulaic academic expressions are widely attested and contribute to the most evenly dispersed MWEs; and (iii) dispersion is strongly long-tailed, with most MWEs occurring in single documents while a smaller recurrent core is shaped by section function and enriched for multi-source overlaps and standardized terminology.

2 Methods and Materials

2.1 Corpus Compilation and IMRaD Subcorpora

To study bioinformatics MWEs in natural text, we compiled a large in-domain corpus of English research articles, named BIOMONO_EN. We leveraged the ALLOFPLOS[1] collection, a repository of ∼200k open-access articles from PLOS journals (Seiver et al., 2018). We filtered this collection by subject area metadata to retrieve articles classified under "bioinformatics". This yielded 4,707 full-text articles from journals such as *PLOS One* and *PLOS Computational Biology*, totaling approximately 24,234,000 words.

Each article was partitioned into its main sections according to the IMRaD structure (following each article's XML section tags). We extracted six section-based subcorpora: Abstracts, Introductions, Methods (including Materials), Results, Discussions, and Conclusions. This stratification enables analysis of MWEs in different communicative contexts.

Table 1 summarizes the size of each subcorpus in number of words.

Section	Word count
Abstracts	1,047,099
Introductions	3,289,196
Methods	5,770,217
Results	7,685,586
Discussions	5,941,263
Conclusions	500,764

Table 1: BIOMONO_EN corpus statistics: total words per IMRaD section.

2.2 MWE Identification Techniques

Following multi-method approaches to scientific MWE extraction (Alves et al., 2024), we used complementary automated strategies designed to capture different facets of multiwordness: (i) lexicon-based semantic matching (USAS), (ii) dependency-linked constructions (UD), (iii) controlled-vocabulary terminology (MeSH), and (iv) list-based academic formulas (AFL/ARTES). We treat these outputs as partially overlapping views rather than interchangeable inventories.

Lexicon-based semantic matching (USAS). We used the UCREL Semantic Analysis System (USAS), which supports multiword matching via lexical resources and disambiguation (Piao et al., 2003; Rayson et al., 2004). Tagging and extraction were carried out using `PyMUSAS`[2]

Dependency-based extraction (UD). We applied UD dependency parsing and extracted MWEs as sequences connected by relations commonly associated with multiword constructions, following Alves et al. (2024): `compound`, `compound:prt`, `fixed`, `flat`, and `flat:foreign`. Parsing was performed with `Stanza` (de Marneffe et al., 2021; Qi et al., 2020).

Ontology term matching (MeSH). To isolate standardized domain terminology, we performed string matching against Medical Subject Headings (MeSH)[3]. This provides high-precision matches to

[1] https://plos.org/text-and-data-mining/

[2] https://ucrel.github.io/pymusas/

[3] https://www.nlm.nih.gov/mesh/meshhome.html

controlled-vocabulary terms but does not capture novel terms absent from the ontology.

Academic formula lists (AFL/ARTES). To capture conventional academic and scientific phraseology, we matched expressions from the Academic Formulas List (AFL) (Simpson-Vlach and Ellis, 2010) and the ARTES scientific phraseology database (Kübler and Pecman, 2011). List matching estimates coverage of known formulas and complements open-ended extraction approaches. From AFL, we took the "core" list of 207 expressions (frequent in both spoken and written academia) and the "written" list of 200 expressions (specific to academic writing). From ARTES, we extracted 830 English expressions from the scientific dictionary and 420 from the cross-disciplinary dictionary.

2.3 Dispersion Analysis

To complement frequency-based profiling, we analyzed how evenly MWEs are distributed across documents within each IMRaD subcorpus. For each section, we treated each article section instance (e.g., one abstract, one introduction) as a separate document and computed two dispersion indicators for every attested MWE type. First, we calculated document frequency (DF), i.e., the number of documents in which an MWE occurs at least once, reported both as a count and as a percentage of documents (DF%). Second, we computed Gries' DP (Gries, 2021), a dispersion coefficient that quantifies the deviation of an item's observed distribution across corpus parts from an equal-share baseline. DP approaches 0 when an MWE is distributed relatively evenly across documents and approaches 1 when it is concentrated in a small subset of documents. We report DP alongside occurrences and DF/DF% to distinguish MWEs that are frequent because they recur broadly from those that are frequent but locally concentrated. Dispersion statistics were computed separately per section and stratified by MWE source (UD, USAS, MeSH, formula lists, and their overlaps) to characterize how extraction strategies differ in the degree to which they capture section-general phraseological templates versus document-specific constructions.

3 Results

3.1 MWE Extraction Results

Table 2 summarizes total and unique MWEs/entities extracted by USAS and UD, along with MeSH matches, across sections. Because sections differ in size (Table 1), raw totals partially reflect section length. To control for this, Table 3 reports rates per million words, revealing differences in extraction density that are not visible from raw counts alone. Notably, sections that are largest in raw totals (e.g., Results and Discussions) are not necessarily the highest in per-word MWE yield, underscoring the importance of normalization for section-wise comparison.

USAS method

The USAS method identified substantial inventories of MWEs across sections (Table 2). In raw terms, the largest sections contain the most MWEs. However, the contrast between raw and normalized counts is particularly informative. Although Results has the largest raw USAS total (Table 2), Methods has the highest USAS density once normalized (75,767 total USAS MWEs per million words in Methods vs. 57,119 in Results; Table 3). Similarly, Abstracts show the highest rate of unique USAS MWEs per million words (41,766), indicating comparatively high type diversity per unit of text despite being much smaller in raw size.

Figure 1a displays MWE lengths across BIOMONO_EN sections. The majority of USAS-extracted MWEs are two-word MWEs, with 59.94% for abstracts, 54.34% for introductions, 60.56% for methods, 61.15% for results, 56.31% for discussions, and 58.04% for conclusions. Three-word MWEs also represent a large proportion of extracted MWEs above 30% for each section, notably 37.77% for introductions, 35.66% for discussions, and 33.07% for conclusions.

We also quantified the prevalence of nominal MWEs in the USAS-derived inventories (Table 4). Nominal MWEs constitute a majority of unique USAS MWEs in all sections, but the proportion varies with Abstracts (86.39%), Methods (84.01%), and Discussions (84.08%) showing high nominal shares, whereas Results is notably lower (58.14%). This variability suggests that lexicon-recognized MWEs in Results include a larger proportion of non-nominal phraseology.

UD method

The dependency-based UD method produced a larger inventory of MWEs than the USAS method in all sections (Table 2). This is consistent with the productivity of compound formation and the broad coverage of dependency relations used to encode multiword constructions. Figure 1b shows that, as

Section	USAS MWEs		UD MWEs		MeSH entities	
	Total	Unique	Total	Unique	Total	Unique
Abstracts	67,448	43,733	134,817	77,006	86,669	5,226
Introductions	197,897	107,321	370,386	172,427	264,580	8,310
Methods	437,194	185,414	879,530	350,688	346,715	6,655
Results	438,990	180,405	1,310,774	372,933	421,019	7,239
Discussions	397,563	191,049	1,353,212	285,419	374,720	8,239
Conclusions	29,810	18,797	53,760	31,450	31,433	2,617

Table 2: Raw counts of total (instances) and unique MWEs/entities by section.

Section	USAS (per million words)		UD (per million words)		MeSH (per million words)	
	Total	Unique	Total	Unique	Total	Unique
Abstracts	64,414	41,766	128,753	73,542	82,771	4,991
Introductions	60,166	32,628	112,607	52,422	80,439	2,526
Methods	75,767	32,133	152,426	60,776	60,087	1,153
Results	57,119	23,473	170,550	48,524	54,780	942
Discussions	66,916	32,156	227,765	48,040	63,071	1,387
Conclusions	59,529	37,537	107,356	62,804	62,770	5,226

Table 3: Length-normalized counts (per million words) of total (instances) and unique MWEs/entities by section, computed from Tables 1 and 2.

Section		Nominal USAS MWEs
Abstracts	n=	37,782
	%	86.39%
Introductions	n=	72,494
	%	67.55%
Methods	n=	155,768
	%	84.01%
Results	n=	104,895
	%	58.14%
Discussions	n=	160,630
	%	84.08%
Conclusions	n=	14,859
	%	79.05%

Table 4: Number and percentage of unique nominal MWEs as extracted by the USAS method.

with USAS, two-word sequences dominate the UD-derived inventories, reflecting the prominence of binary compounds and short fixed constructions.

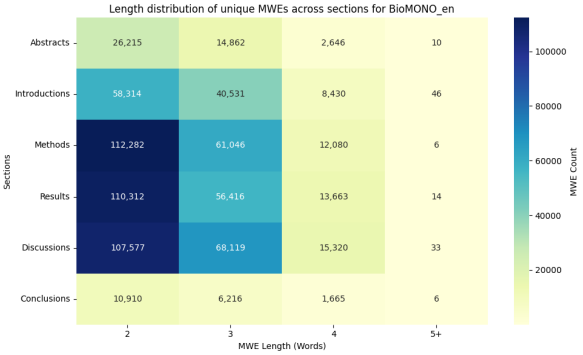
Normalization also changes how section differences are interpreted for UD MWEs. In raw terms, Discussions and Results dominate because they are long sections (Table 2). However, per-million rates show that Discussions is the densest site of UD MWEs (227,765 per million words), exceeding Results (170,550) and Methods (152,426) (Table 3).

To examine what kinds of dependency-linked constructions dominate, Table 5 reports the distribution of unique UD MWEs by relation category. More than 90% of unique MWEs extracted across BIOMONO_EN sections belong to the `compound` category. `flat` is also the second most prominent category, with for instance 8.07% of extracted MWEs from the Methods section belonging to that category, and 8.04% for the Results section. Methods and Results are the only two sections containing `flat:foreign` MWEs, although in very small number.

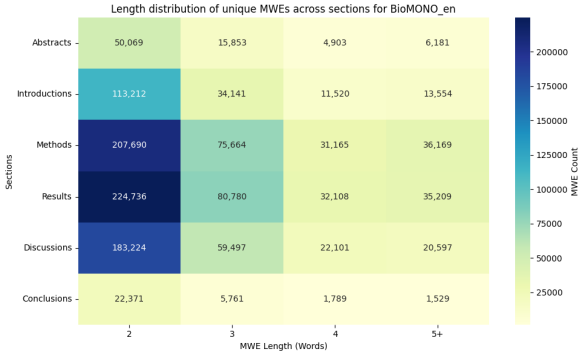
MeSH method

MeSH matching yields fewer unique items than the open-ended UD and USAS inventories (Table 2), as expected for a controlled vocabulary. However, the matches are domain-relevant by construction and provide a high-precision view of standardized terminology. For MeSH, normalized rates highlight a different profile from raw totals. Abstracts and Introductions show the highest MeSH token density (82,771 and 80,439 per million words, respectively), even though they do not contain the most raw MeSH matches (Table 3). Conversely, Methods and Results exhibit much lower unique MeSH rates per million words (1,153 and 942), suggesting heavier repetition of a narrower standardized vocabulary within those sections, whereas Conclusions show a comparatively high unique MeSH density given their short length (Table 3).

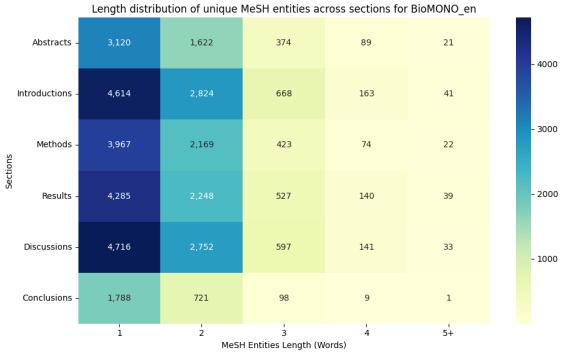
MeSH entities also vary in length from single-word entities to MWEs up to five-word long and more, as seen in Figure 1c. Single-word entities are the majority in every section: 59.70% of unique MeSH entities in Abstracts are single-word terms, 55.52% in Introductions, 59.61% in Methods, 59.19% in Results, 57.24% in Discussions, and 68.32% in Conclusions. The remaining entities are MWEs, among which two-word terms are most prevalent, and longer terms occur up to five words and beyond. This pattern indicates that controlled-



(a) Distributions of unique USAS MWEs found across BIOMONO_EN sections by length (in words).



(b) Distributions of unique UD MWEs found across BIOMONO_EN sections by length (in words).



(c) Distributions of unique MeSH entities found across BIOMONO_EN sections by length (in words).

Figure 1: Length distribution of MWEs across extraction techniques.

vocabulary terminology in bioinformatics is partly multiword and that a substantial share of standardized concepts may be missed by analyses limited to single-word types.

Academic formulaic expressions method

Table 6 presents the number and percentage of unique MWEs found in each section of the BIOMONO_EN corpus across MWE lists. The AFL lists (core and written) show near-complete coverage across all sections, with high percentages overall (82.5–100%), reaching 100% in several sections. In contrast, the ARTES lists (scientific and cross) show lower coverage, with percentages ranging between approximately 29% and 52%. Notably, the Results and Discussions sections consistently contain the highest proportion of MWEs across all lists, particularly for the ARTES lists.

3.2 Dispersion results

Dispersion is dominated by a pronounced long tail, as seen in Figure 2. The median MWE occurs once and appears in exactly one document in every section (median DF% $\approx$ 0.02–0.06%, depending on section size). Consequently, most MWEs are maximally clustered, with median DP values close to 1 throughout ($\approx$ 0.9990–0.9998). The proportion of MWEs attested in a single document is very high across the board (83.7% in Abstracts, 80.4% in Introductions, 75.4% in Methods, 81.6% in Results, 82.9% in Discussions, and 87.0% in Conclusions), rising to $\geq$89.5% in all sections when considering MWEs occurring in at most two documents. This pattern indicates that the MWE inventory is overwhelmingly driven by low-frequency, document-specific units, with only a small minority recurring across texts.

Type inventories are dominated by UD-only (53.1–62.3%) and USAS-only (25.9–32.8%) MWEs, with a stable UD+USAS overlap (10.9–13.0%). MeSH is rare (0.61–1.99%) and Formulas rarer (0.17–0.63%; Conclusions: 1.50%). Source behavior separates the long tail from the core: UD-only/USAS-only MWEs are the most document-specific (singletons: 81.2–88.5% / 77.1–90.3%), whereas overlap MWEs recur more broadly (singletons: 60.5% in Methods; 71.2% in Results). MeSH units are fewer but less singleton-heavy (53.3–67.2%), and MeSH+UD+USAS shows the

Section		compound	compound:prt	fixed	flat	flat:foreign
Abstracts	n=	73,807	132	102	2,965	-
	%	(95.85%)	(0.17%)	(0.13%)	(3.85%)	
Introductions	n=	162,188	441	327	9,471	-
	%	(94.06%)	(0.26%)	(0.19%)	(5.49%)	
Methods	n=	321,370	650	367	28,291	10
	%	(91.64%)	(0.19%)	(0.10%)	(8.07%)	(0.003%)
Results	n=	341,773	679	397	29,983	1
	%	(91.67%)	(0.18%)	(0.11%)	(8.04%)	(<0.001%)
Discussions	n=	265,856	605	434	18,524	-
	%	(93.15%)	(0.21%)	(0.15%)	(6.49%)	
Conclusions	n=	30,014	137	87	1,212	-
	%	(95.43%)	(0.44%)	(0.28%)	(3.85%)	

Table 5: Frequency and percentage of UD relation categories across the different sections of BIOMONO_EN based on unique UD MWEs.

		Abstracts	Introductions	Methods	Results	Discussions	Conclusions
ARTES	n=	250	343	277	342	397	240
scientific	%	30.86%	42.35%	34.20%	42.22%	49.01%	29.63%
ARTES	n=	123	203	159	183	216	125
cross	%	29.71%	49.03%	38.41%	44.20%	52.17%	30.20%
AFL	n=	189	205	200	204	205	189
core	%	91.30%	99.03%	96.62%	98.56%	99.03%	91.30%
AFL	n=	165	199	195	200	200	189
written	%	82.50%	99.50%	97.50%	100%	100%	94.50%

Table 6: Number of unique MWEs found in BIOMONO_EN sections across MWE lists. Percentages are calculated against the total number of unique MWEs in each list.

strongest terminological stability (singletons ∼20–33% in Abstracts–Discussions; 47% in Conclusions). Finally, Formulas behave most "core-like" (singletons: 8.9–17.8%) and are over-represented among the most evenly dispersed MWEs (share in the 500 lowest-DP MWEs: 28.0% Abstracts, 40.0% Introductions, 13.2% Methods, 24.4% Results, 37.2% Discussions, 46.4% Conclusions).

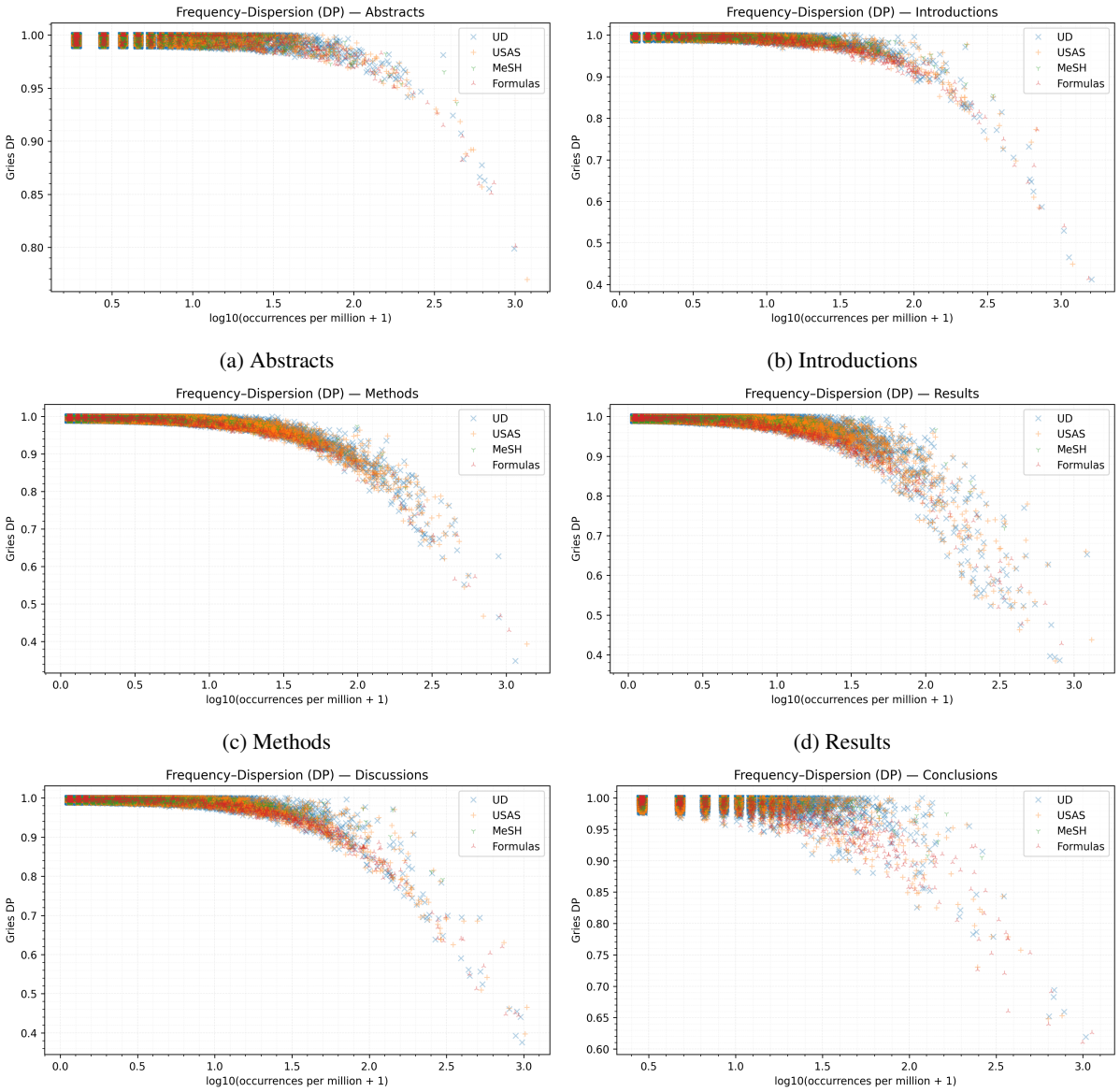
Overall, dispersion reflects a two-layered structure with a large extractor-driven long tail (UD/USAS-only) and a smaller, section-shaped recurrent core enriched for Formulas, overlap MWEs, and MeSH-overlap terminology.

4 Discussion

Across methods, our results highlight the centrality of multiwordness in bioinformatics discourse, but they also show that "MWE dominance" depends on the extraction method. The UD-based inventories are overwhelmingly compound-dominated (Table 5), confirming that noun-compound formation is a primary structural resource for expressing domain concepts succinctly. This is consistent with long-standing accounts of scientific prose as a "compressed code" that favors dense noun phrase packaging over more clausal, elaborated alternatives (Biber and Gray, 2016). This pattern also closely aligns with recent computational evidence that (i) compounds constitute the dominant UD MWE class in scientific writing overall and exhibit a clear increase over time (cf. compounds at 80.2% of UD MWEs; Alves et al., 2024), and (ii) biomedical abstracts are likewise strongly compound-heavy in UD-based inventories, with Bagdasarov and Teich (2024) reporting compound shares above 90%. For Natural Language Processing (NLP), this underscores that a large portion of domain "terminology" is not a closed list but a productive constructional space.

The lexicon-based USAS inventories also contain a large proportion of nominal MWEs in most sections (Table 4), but the proportion is notably lower in Results. This divergence is informative because lexicon-based approaches recover a broader mix of discourse and reporting phraseology, and Results is precisely the section where comparison and evidential framing are most prominent. Practically, this indicates that MWE resources for domain NLP should be section-aware, as the phraseological targets relevant to information extraction or summarization are not uniform across IMRaD.

MeSH matching provides a complementary view of standardized terminology. While most unique MeSH entities are single-word terms, a substantial fraction are multiword (Figure 1c), demonstrat-



(a) Abstracts

(b) Introductions

(c) Methods

(d) Results

(e) Discussions

(f) Conclusions

Figure 2: Frequency–dispersion profiles (occurrences vs. Gries' DP) of MWEs in BIOMONO_EN by IMRaD section. Each point corresponds to one MWE type; DP values closer to 0 indicate more even dispersion across documents, while values closer to 1 indicate stronger clustering. Points are colored by MWE source (UD, USAS, MeSH, and formulas).

ing that controlled-vocabulary terminology is not reducible to single-word naming. The elevated unique MeSH rate in Abstracts and Conclusions (Table 3) further suggests that summary sections foreground canonical entity naming, even when the running text is shorter. For downstream applications (e.g., normalization, retrieval, and MT terminology control), these multiword entities constitute high-value targets that are easy to miss under unigram-centric preprocessing.

Finally, list-based formula coverage indicates that bioinformatics writing draws broadly on general academic formulas (AFL) and on a sizeable subset of scientific phraseology documented in ARTES (Table 6). This signals substantial transferability of general scientific phraseology resources to bioinformatics, while dispersion results clarify where these templates function as shared scaffolding versus local phrasing.

Dispersion profiling adds a functional perspective on these inventories. Across IMRaD, MWE types exhibit a strongly long-tailed distribution: most occur in a single document, while a much smaller set forms a recurrent "core" whose composition shifts by section. This pattern is expected when dispersion is assessed over many documents and is captured by Gries' DP, which explicitly distinguishes frequency from distributional evenness (Gries, 2021). Importantly, the long tail should not be interpreted as a lack of phraseological structure: many UD MWEs are *productively constructed* (e.g., novel or dataset-contingent compounds) rather than retrieved as fixed strings, increasing type counts while limiting cross-document recurrence (Biber and Gray, 2016). For MWE research, this reinforces the need to separate productive constructions from reusable templates. For NLP, it suggests that robust domain handling requires both (i) mechanisms for generalizing over productive compounds and (ii) explicit modeling of recurrent templates that shape section-level discourse.

Source-specific dispersion further clarifies what constitutes the recurrent backbone. UD-only and USAS-only MWEs contribute most of the document-specific tail, whereas overlap MWEs (UD+USAS) recur more broadly, suggesting that multi-method confirmation captures sequences that are simultaneously structurally cohesive and functionally salient. The strongest "core-like" behavior is observed for list-derived formulas, which are rare as types yet disproportionately represented among the most evenly dispersed MWEs. This aligns with corpus work showing that recurrent MWEs and formulaic sequences function as register-specific building blocks in academic discourse (Biber et al., 2004; Hyland, 2008; Wray, 2002). MeSH matching complements this picture by isolating a compact set of standardized multiword terms that recur across documents when they are also recoverable by general extraction, consistent with the stabilizing role of controlled vocabularies in scientific naming.

5 Conclusion

Using complementary MWE identification strategies and dispersion profiling, this study maps bioinformatics "multiwordness" across IMRaD sections. MWEs are predominantly short and nominal, reflecting compound-heavy phrasal compression in scientific prose (Biber and Gray, 2016; Alves et al., 2024), while other extraction methods recover additional reporting and procedural templates. Dispersion is strongly long-tailed: most MWEs are document-specific, but a smaller recurrent core aligns with section function and is enriched for conventional templates and standardized multiword terminology. For NLP and MWE research, the main implication is that domain phraseology is best operationalized as a two-layer system (productive constructions plus reusable templates) and that combining multi-method identification with dispersion analysis provides a principled way to prioritize MWEs for domain-adapted preprocessing and downstream applications.

6 Limitations and Future Work

This study relies on automated MWE identification and therefore inherits method-specific biases, such as parsing sensitivity, tokenization, disambiguation, and vocabulary coverage.

A priority next step is to build a small, section-stratified manually verified subset to quantify boundary errors, false positives/negatives, and overlap reliability, enabling precision-oriented reporting in addition to coverage. Beyond validation, two extensions are particularly relevant: (i) multilingual replication to test whether the same section-conditioned multiword patterns hold under different morphosyntactic systems, and (ii) downstream evaluation to assess whether MWE-aware resources improve domain tasks such as terminology normalization, information extraction, retrieval, or domain-specific Machine Translation.

Acknowledgments

This work was supported by the Open-Oxford-Cambridge Arts and Humanities Research Council (AHRC) Doctoral Training Partnership (OOC-DTP), project reference 2739531.

Data and Code Availability

Data and code are available at https://github.com/jurgigi/BioMONO

References

Diego Alves, Stefan Fischer, Stefania Degaetano-Ortlieb, and Elke Teich. 2024. Multi-word Expressions in English Scientific Writing. In *Proceedings of the 8th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2024)*, pages 67–76, St. Julians, Malta. Association for Computational Linguistics.

Diego Alves, Stefan Fischer, and Elke Teich. 2025. Syntagmatic Productivity of MWEs in Scientific English. In *Proceedings of the 21st Workshop on Multiword Expressions (MWE 2025)*, pages 1–6, Albuquerque, New Mexico, U.S.A. Association for Computational Linguistics.

Sergei Bagdasarov and Elke Teich. 2024. Multi-word expressions in biomedical abstracts and their plain English adaptations. In *Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities*, pages 483–488, Miami, USA. Association for Computational Linguistics.

Douglas Biber, Susan Conrad, and Viviana Cortes. 2004. If you look at . . . : Lexical Bundles in University Teaching and Textbooks. *Applied Linguistics*, 25(3):371–405.

Douglas Biber and Bethany Gray. 2016. *Grammatical Complexity in Academic English: Linguistic Change in Writing*. Studies in English Language. Cambridge University Press, Cambridge.

Mathieu Constant, Gülşen Eryiğit, Johanna Monti, Lonneke van der Plas, Carlos Ramisch, Michael Rosner, and Amalia Todirascu. 2017. Multiword Expression Processing: A Survey. *Computational Linguistics*, 43(4):837–892.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308. Place: Cambridge, MA Publisher: MIT Press.

Stefania Degaetano-Ortlieb and Elke Teich. 2018. Using relative entropy for detection and analysis of periods of diachronic linguistic change. In *Proceedings of the Second Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 22–33, Santa Fe, New Mexico. Association for Computational Linguistics.

Cosmin Mihail Florescu and Ryosuke L. Ohniwa. 2025. On the creation of a corpus-derived medical multiword term list. *Information*, 16(2):118.

Stefan Th. Gries. 2021. Analyzing dispersion. In Magali Paquot and Stefan Th. Gries, editors, *A Practical Handbook of Corpus Linguistics*, pages 99–118. Springer, Cham.

Ken Hyland. 2008. As can be seen: Lexical bundles and disciplinary variation. *English for Specific Purposes*, 27(1):4–21.

Ken Hyland. 2012. Bundles in Academic Discourse. *Annual Review of Applied Linguistics*, 32:150–169.

Sun Kim, Lana Yeganova, Donald C. Comeau, W. John Wilbur, and Zhiyong Lu. 2018. Pubmed phrases, an open set of coherent phrases for searching biomedical literature. *Scientific Data*, 5:180104.

Natalie Kübler and Mojca Pecman. 2011. ARTES: an online lexical database for research and teaching in specialized translation and communication. In *ESSLLI 2011, International Workshop on Lexical Resources (WoLeR)*, Ljubljana, Slovenia.

Francesca Masini. 2019. Multi-Word Expressions and Morphology. In *Oxford Research Encyclopedia of Linguistics*.

Helder I. Nakaya. 2021. *Bioinformatics*. Exon Publications, Australia.

Scott S. L. Piao, Paul Rayson, Dawn Archer, Andrew Wilson, and Tony McEnery. 2003. Extracting Multiword Expressions with A Semantic Tagger. In *Proceedings of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, pages 49–56, Sapporo, Japan. Association for Computational Linguistics.

Damith Premasiri, Amal Haddad Haddad, Tharindu Ranasinghe, and Ruslan Mitkov. 2023. Deep learning methods for identification of multiword flower and plant names. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 879–887, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Paul Rayson, Dawn Archer, and Scott Piao. 2004. The UCREL semantic analysis system. In *Proceedings of the Beyond Named Entity Recognition Workshop*, Lisbon, Portugal.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword Expressions: A pain in the neck for NLP. In *Computational Linguistics and Intelligent Text Processing*, pages 1–15, Berlin, Heidelberg. Springer Berlin Heidelberg.

Elizabeth Seiver, M Pacer, and Sebastian Bassi. 2018. Text and data mining scientific articles with allofplos. In *Proceedings of the 17th Python in Science Conference*, pages 61 – 64.

Rita Simpson-Vlach and Nick C. Ellis. 2010. An Academic Formulas List: New Methods in Phraseology Research. *Applied Linguistics*, 31(4):487–512.

Luciana B. Sollaci and Mauricio G. Pereira. 2004. The introduction, methods, results, and discussion (IMRAD) structure: a fifty-year survey. *Journal of the Medical Library Association*, 92(3):364–371.

Aline Villavicencio, Francis Bond, Anna Korhonen, and Diana McCarthy. 2005. Editorial: Introduction to the special issue on multiword expressions: Having a crack at a hard nut. *Comput. Speech Lang.*, 19(4):365–377.

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

Jianguo Wu. 2011. Improving the writing of research papers: IMRAD and beyond. *Landscape Ecology*, 26(10):1345–1349.