Ento-Linguistics: Language, Ambiguity, and Scientific Communication in Entomology

Complex Systems, Measurement, and a Six-Domain Framework

Daniel Ari Friedman
Active Inference Institute, APOIDEAS
daniel@activeinference.institute
ORCID: 0000-0001-6232-9096

Tucker Cahill Chambers
APOIDEAS
ORCID: 0009-0008-3793-7872

2026-10-10 – Version 1.3.1

DOI: 10.5281/zenodo.19574117

Contents

1 Abstract **3**

2 Introduction **5**
2.1 Terminology as a Research Question . . . 5
2.2 The Challenge of Terminological Reform . . . 5
2.3 From Labels to Biological Explanation . . . 5
2.4 Six Analytical Domains . . . 5
2.5 Research Approach . . . 6

3 Methods **7**
3.1 Data Acquisition . . . 7
3.2 Statistical Analysis . . . 8
3.3 Network Representation Sensitivity . . . 10

4 Results: Corpus Analysis and Terminology Networks **12**
4.1 Terminology Extraction Across Domains . . . 12
4.2 Terminology Network Structure . . . 12
4.3 Framing Analysis . . . 15

5 Results: Domain-Specific Findings **19**
5.1 Unit of Individuality . . . 19
5.2 Power & Labor . . . 19
5.3 Behavior & Identity . . . 19
5.4 Sex & Reproduction . . . 19
5.5 Kin & Relatedness . . . 19
5.6 Economics . . . 19
5.7 Historical Interpretation . . . 20

6 Discussion **25**
6.1 Language and Scientific Practice . . . 25
6.2 Network Structure and Its Conditional Reference . . . 25
6.3 Biological Precision and Terminological Continuity . . . 25

6.4 Active Inference as a Theoretical Perspective . . . 26
6.5 A Complex Systems Reading of Terminology . . . 26
6.6 Practical Use of CACE . . . 26
6.7 Cross-Domain Communication . . . 27
6.8 Measurement Validation and Terminology Intervention . . . 27
6.9 Limitations . . . 27
6.10 Complexity as an Explicit Modeling Question . . . 28

7 Conclusion **29**
7.1 Future Directions . . . 29

8 Related Work **30**
8.1 Scientific Language and Categories . . . 30
8.2 Terminology in Social-Insect Research . . . 30
8.3 Computational Mapping and Source-Layer Analysis . . . 30
8.4 Validation and Experimental Evidence . . . 30
8.5 Active Inference and Colony Modeling . . . 31
8.6 Collective Behavior and Complex Systems . . . 31
8.7 Positioning This Work . . . 31
8.8 Complexity, Representations, and Dynamics . . . 31

9 Acknowledgments **32**
9.1 Institutional Support . . . 32
9.2 Collaborations . . . 32
9.3 Data and Software . . . 32

10 Supplemental Methods: Text Processing and Term Extraction **33**
10.1 Source Layers and Custody . . . 33
10.2 Normalization and Token Streams . . . 33
10.3 Candidate Extraction and Domain Assignment . . . 33
10.4 Document Co-occurrence and Concept Categories . . . 34
10.5 Reproducible Inputs and Failure Behavior . . . 34

11 Supplemental Methods: Statistical and Scoring Infrastructure **35**
11.1 Sentence-context Entropy . . . 35
11.2 Exploratory Comparisons and Intervals . . . 35
11.3 Shared CACE Sample and Scores . . . 35
11.4 Framing and Discourse Proxies . . . 36
11.5 Historical OCR and Layer Coverage . . . 36
11.6 Execution and Artifact Verification . . . 36
11.7 Fixed-Margin Network Robustness Extension . . . 37

12 Supplemental Results **38**
12.1 Exploratory Pairwise Domain Comparisons . . . 38
12.2 CACE Scoring for Key Terms . . . 38
12.3 Semantic Entropy Distribution . . . 38
12.4 Interval Reporting . . . 38
12.5 Full-Text Parallel Layer . . . 39

12.6 Discourse and Rhetorical Layer . . . 40
12.7 Statistical and Source-Layer Figures . . . 41

13 Supplemental Analysis: Theoretical Extensions **44**
13.1 Individuality and Conditional Independence . . . 44
13.2 A Proposed Framing Score . . . 44
13.3 Ambiguity Annotation . . . 44
13.4 Cross-Domain Mapping . . . 44
13.5 Temporal Networks . . . 45
13.6 Multiscale Interpretation . . . 45

14 Supplemental Analysis: Case Studies and Validation **46**
14.1 Validation Agenda . . . 46
14.2 Historical Terminology Analysis . . . 46
14.3 Complete Document Coverage and Bounded Entropy . . . 46
14.4 A Reproducible Case-Study Protocol . . . 46

15 Symbols and Notation Glossary **48**
15.1 Mathematical Notation . . . 48
15.2 Theoretical Terms . . . 49
15.3 Pipeline Modules . . . 49

16 References **51**

1 Abstract

Terms such as *queen*, *worker*, and *colony* connect biological descriptions to familiar social concepts. This study introduces a six-domain Ento-Linguistic framework and an open-source descriptive text-analysis pipeline. The headline layer contains 7599 source-identified abstracts from 7609 stored strings; 10 unreconciled strings are retained but excluded. This headline layer contains 998,645 processed tokens and 11,700 candidate terms, of which 1330 receive rule-based domain assignments. The pipeline separates observed document-level term co-occurrence from a map of 6 predefined concept categories with 15 vocabulary-overlap relationships. Among assigned terms, 11.9% receive multiple labels; this measures classification overlap rather than semantic drift. Complementary analyses use 6997 PMC full-text records, 2430 historical OCR documents, and a separate arXiv layer. TF-IDF clustering and Shannon entropy summarize sentence-context distributions, while lexical patterns identify candidate framing contexts. A separately receipted fixed-margin comparison finds fewer edges and lower clustering in the selected-term graph than in its finite-chain randomized reference, indicating concentration of co-occurrence relative to the preserved incidence margins. Clarity, Appropriateness, Consistency, and Evolvability (CACE) are proposed as heuristic evaluation dimensions. Neither cluster entropy nor marker occurrence establishes distortion of biological understanding, and CACE scores have not been validated against independent human judgments. A deterministic vocabulary–threshold diagnostic makes the dependence of density and clustering on network representation explicit. Complexity motivates questions about units, coupling, feedback and scale; it is not inferred from a dense lexical projection. Provenance gaps, mixed-topic retrieval, OCR errors, overlapping groups, and explicitly bounded analyses limit interpretation. The contribution is a reproducible descriptive workflow and a framework for subsequent annotated, hypothesis-driven research, rather than a causal test of language shaping scientific thought. Code and data lineage: https://github.com/docxology/ento_linguistics.

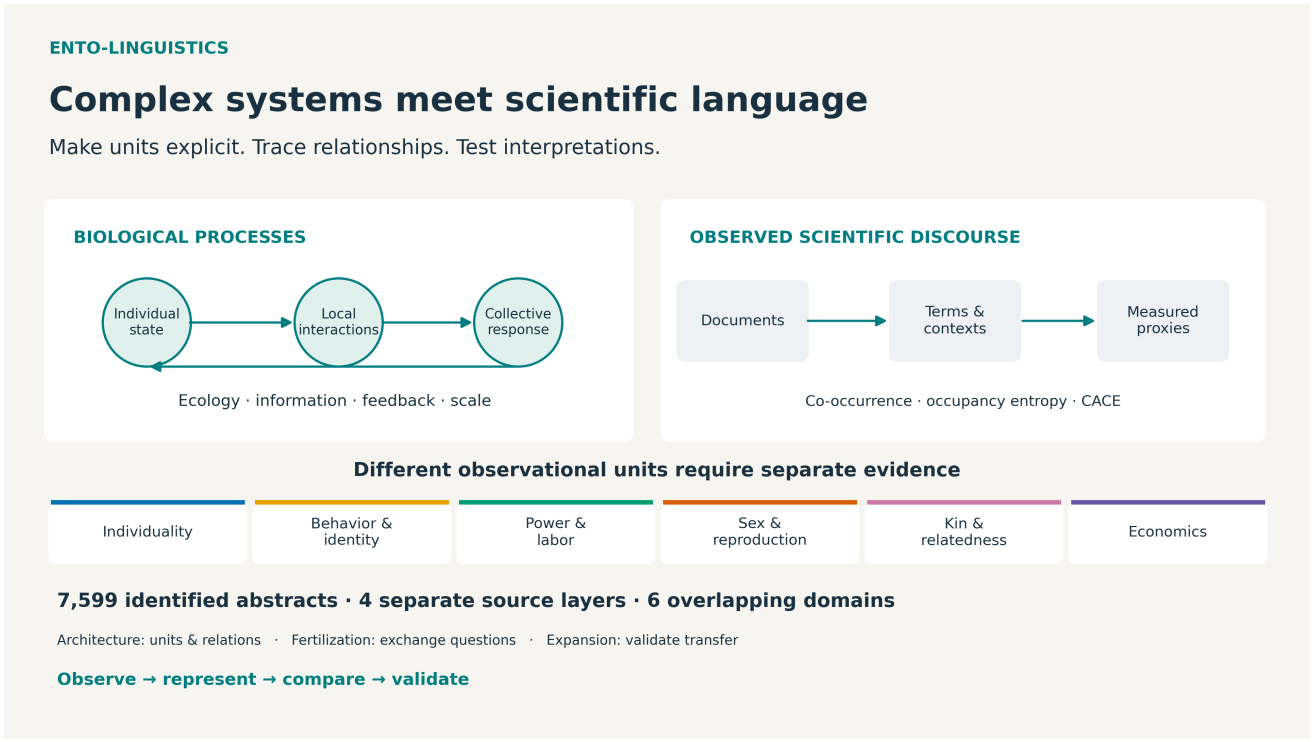


Figure 1. Graphical abstract. Biological processes, scientific discourse and computational representations have distinct observational units. The six overlapping domains organize questions; observed textual proxies and proposed validation designs retain separate evidential roles. Arrows in the biological lane are schematic, not a fitted model or newly measured interaction network.

2 Introduction

2.1 Terminology as a Research Question

Scientific terminology connects observations, categories, and explanatory models. In social-insect research, familiar labels such as queen, worker, caste, kin, and colony can serve technical functions while retaining ordinary-language connotations. The relevant question is when those connotations help communication, when explicit definitions are needed, and how either possibility can be studied without inferring beliefs from word counts alone.

Philosophical and linguistic work supplies motivation for examining terminology within scientific practice (Lakoff and Johnson, 1980, Latour, 1987, Longino, 1990). Entomological discussions of categories and loaded metaphors provide field-specific context (Gordon, 1992, Herbers, 2006, 2007). These perspectives motivate empirical questions; they do not make every conventional term misleading or demonstrate that a replacement improves scientific understanding.

2.2 The Challenge of Terminological Reform

Established terms support literature discovery and can carry precise operational meanings. Proposed alternatives therefore require context-specific comparison rather than automatic replacement. A task-based description may be useful for a behavioral observation without replacing a developmentally defined caste category. Molecular and developmental research makes that distinction especially important (Qiu et al., 2022, Sumner et al., 2018).

CACE—Clarity, Appropriateness, Consistency, and Evolvability—is introduced as an explicit set of evaluation questions. Its numerical scoring rules are one inspectable implementation, rather than an independently validated measure of understanding. The biological mechanism, intended referent, reader population, and definition supplied by an author remain central to evaluating terminology.

2.3 From Labels to Biological Explanation

A useful terminology analysis separates the expression, its referent, and the mechanism invoked to explain that referent. *Worker* can identify a developmental caste (typically non-reproductive females) or describe an individual performing a task; neither usage alone specifies how task allocation is regulated. Definitions of caste have themselves been examined as a conceptual problem (Villet, 1992), while experimental work shows that individual experience can contribute to persistent division of labor (Ravary et al., 2007). These examples locate the research question in the relationship between definitions and evidence, rather than in a presumption that social vocabulary is necessarily inaccurate.

The six domains provide questions to ask of a passage: what entity is described, at which biological scale, using which observable criteria, and with what explanatory commitment? Frequency and co-occurrence help locate passages for this inquiry. Assessing whether a label obscures a mechanism requires examining those passages and evaluating readers' inferences. This distinction makes the framework useful for comparative reading without treating the taxonomy as a discovered ontology.

2.4 Six Analytical Domains

The framework defines six overlapping domains:

1. **Unit of Individuality:** terms identifying ants, nestmates, colonies, collectives, and superorganisms; how an author specifies the unit being observed or modeled.
2. **Behavior and Identity:** distinctions between a task, an observed behavior, a persistent propensity, and a developmental category.

3. **Power and Labor:** queen, worker, caste, and related labels; whether descriptions imply control mechanisms or simply name biological roles.
4. **Sex and Reproduction:** reproductive and developmental categories; how definitions accommodate variation across taxa and mating systems.
5. **Kin and Relatedness:** pedigrees, demographic assumptions, and relatedness terminology. For example, expected sister relatedness of $r = 0.75$ requires an outbred haplodiploid pedigree with one mother and one haploid father; it is not a universal property of colonies.
6. **Economics:** resource, allocation, investment, and related expressions; distinctions between measured energetic expenditure, fitness consequences, and metaphorical usage.

These are predefined organizational categories. A term receiving several labels does not establish that its meaning changed over time. Each domain also requires qualitative annotation before lexical patterns can be interpreted as ambiguity or framing.

2.5 Research Approach

The headline computation analyzes 7599 source-identified abstracts, yielding 998,645 processed tokens and 11,700 candidates, of which 1330 receive domain assignments. Complementary PMC, BHL, and arXiv layers retain separate source identities. The workflow distinguishes corpus frequencies, observed document co-occurrence, vocabulary overlap, context-cluster entropy, and heuristic scores.

Active Inference and multiscale modeling provide a theoretical perspective (Friedman et al., 2021, Friston, 2010), but the present pipeline does not fit a generative model of scientific language or measure variational free energy. A Markov blanket specifies conditional-independence relationships in a model; it is not a lexical security filter or a biological boundary established by terminology alone.

The contribution is a descriptive workflow, a proposed taxonomy, and explicit questions for subsequent validation. Historical OCR frequencies provide dated source-layer observations rather than proof of conceptual origins or causal reform. Methods and limitations state source custody gaps, convenience-sample retrieval, statistical dependence, and bounded analyses so that interpretation remains tied to the evidence actually produced.

Complex systems supplies an organizing question: which units, relations, feedbacks and timescales does a scientific passage specify? Network representations and dynamics address different problems (Newman, 2003), while different complexity measures require different targets and assumptions (Ladyman et al., 2013). This study therefore keeps observed biological mechanisms, lexical descriptions and computational proxies distinct. The v1.3.0 representation diagnostic makes vocabulary and edge-threshold dependence inspectable; separate validation is required before transferring those measures to meaning, communication or biological organization.

3 Methods

The workflow separates acquisition, descriptive analysis, and theoretical interpretation. Supplemental Methods 10 and 11 describe the implementation. Section 13 develops conceptual proposals rather than fitted generative models.

3.1 Data Acquisition

3.1.1 Search Strategy and Source Selection

The archived input is the ordered abstract list in *data/corpus/abstracts.json*. The headline analysis selects strings whose digest maps to an identified PubMed record. Unreconciled strings remain archived but are excluded, without altering their bytes. The acquisition script delegates to *src/pipeline/corpus_build.py*; its base queries and the named growth queries in *src/data/literature_mining.py* are the executable query definitions. Digest-indexed provenance records PMIDs, DOIs, titles, years, journals, and query names where available. This corpus accumulated across retrieval runs; a single illustrative query does not reconstruct its history.

Three complementary layers remain separate: PMC title, abstract, and body text in *data/fulltexts/* shards; BHL mirror OCR in *data/bhl/* shards; and arXiv title/abstract records in *data/corpus/arxiv_records.json*. The arXiv harvester uses category-restricted queries over q-bio.PE and nlin.AO. Neither full texts nor preprints are merged into the headline abstract list. Query and retrieval metadata are retained in source sidecars.

3.1.2 Corpus Composition and Cleaning

Headline metric	Value
Stored abstract strings	7609
Source-identified analyzed abstracts	7599
Unreconciled strings excluded	10
Processed tokens	998,645
Unique token types	47,216
Candidate terms	11,700
Domain-assigned terms	1330

These are archived-input and selected-corpus counts, not estimates of all entomological literature. Broad searches retrieve adjacent biological topics and algorithmic uses of ant terminology; a keyword match does not establish relevance to ant biology. Historical volumes can contain substantial non-entomological material. Of 7073 stored PMC records, 76 whose titles mark them as correction, erratum or retraction notices or as retracted items remain archived but are excluded from analysis. This removes the standardized repeated notice text. A title rule is not a complete article-type or retraction-status screen, and relevance screening is not complete. The custody audit in *output/reports/corpus_audit.json* records missing digest provenance, repeated text, identifier mismatches, and unused sidecar entries. Gaps are retained and disclosed; undocumented source identities are not inferred. Exact text-digest matches to retrieved PubMed abstracts can recover metadata without rewriting text. This establishes source identity rather than ant-topic relevance. Digest-keyed sidecars retain only one metadata entry when several records contain identical text. Relevance annotation and complete source reconciliation are needed before treating a layer as a curated field-wide sample.

TextProcessor normalizes text, tokenizes with NLTK, filters punctuation and stop words, and optionally lemmatizes with WordNet. Extraction disables lemmatization when counting surface-form terms. NLTK datasets

are installed separately from Python dependencies; selected English tokenizer, stopword and WordNet contents are bound into analysis signatures and receipts. Article-level licenses require individual checking; indexing in PubMed/PMC does not establish reuse permission.

3.1.3 Domain Coverage Verification

The six vocabularies are defined by TerminologyExtractor.DOMAIN_SEEDS: Unit of Individuality, Behavior and Identity, Power and Labor, Sex and Reproduction, Kin and Relatedness, and Economics. Coverage means that extraction assigned terms to these labels; it does not measure retrieval recall or validate the taxonomy against independent annotation.

3.2 Statistical Analysis

3.2.1 Term Extraction and Classification

Extraction counts candidate tokens, retains configured seed words subject to length/frequency and preprocessing rules, applies remaining lexical-pattern filters, and assigns domains through direct seed matching, compound-word overlap, and fallback patterns. Three-token windows provide up to thirty deduplicated short contexts for heuristic scoring. They do not expand assignments by co-occurrence. Candidate filtering also admits scientific substrings and compound tokens, so candidates can include unrelated words and OCR artifacts. Candidate inclusion does not itself assign a domain or establish biological relevance. An n-gram utility is available but the headline extractor does not invoke it to extract multi-word phrases. The BHL literal-frequency pass separately matches multi-word seed phrases as consecutive tokens.

Minimum frequencies differ: one occurrence for headline abstracts, twenty for PMC, and two for arXiv and the BHL stack. These choices affect vocabulary size and prevent interpreting raw term counts as directly comparable measures across layers.

3.2.2 Semantic Entropy

Terms with at least five usable sentence contexts undergo TF-IDF vectorization and seeded KMeans clustering (random state 42, ten initializations). Shannon entropy measures occupancy:

$$H(t) = -\sum_{i=1}^{k} p_i \log_2 p_i. \tag{3.1}$$

The requested cluster count is $k = \max(2, \min(5, n-1, \max(2, \lfloor\sqrt{n}\rfloor)))$; occupied clusters may be fewer. Normalized entropy divides by $\log_2 k_{\text{occupied}}$, with zero for a single occupied cluster. Only successfully computed estimates enter domain means; insufficient-context and computation failures are recorded as exclusions. Domain figures and statistical descriptives share this sentence-context calculation. The $H > 2$ bits flag is exploratory, without independent calibration. Clusters are computational partitions, not annotated word senses; high entropy does not by itself demonstrate ambiguity or temporal semantic drift.

3.2.3 Measurement Targets and Validation

Each computational quantity has a defined observational unit. Frequencies count processed occurrences; network edges count shared documents; domain overlap counts classifier memberships; entropy summarizes sentence-context cluster occupancy. Framing markers and CACE scores apply configured rules to selected contexts and terms. None directly measures an author's intention or a reader's understanding. Problem-specific validation is a

central requirement of automated text analysis (Grimmer and Stewart, 2013); reproducible calculation and valid interpretation address different questions.

A future evaluation should freeze annotation rules before inspecting score rankings and retain a held-out set of source contexts. Annotation should distinguish biological referent, task versus developmental category, technical definition, figurative use, and uncertainty. Sampling should include frequent and rare terms, assigned and unassigned candidates, and contexts without marker matches. These strata allow false positives and false negatives to be examined, rather than assessing only examples selected by the pipeline. Agreement, adjudication rules, and performance by source layer should be reported. This is a proposed validation design, not an executed component of the current analysis.

3.2.4 Domain-Level Statistical Tests

Welch tests, Benjamini–Hochberg adjustments, small-sample corrected standardized differences, and one-way ANOVA summarize valid per-term entropies. Groups with fewer than two usable values are omitted. These are exploratory comparisons. Terms can belong to multiple groups and derive contexts from shared documents, violating the simple independence interpretation. Multiplicity correction does not repair dependence, retrieval bias, or unequal context counts. Threshold crossings do not establish calibrated population significance or causal effects. Confirmatory inference requires independent annotation and a document-level sampling model accounting for shared terms and overlapping domains.

3.2.5 Conceptual Network Analysis

The conceptual map assigns terms to six predefined categories. Its 15 links summarize vocabulary overlap:

$$w_{AB} = \frac{|A \cap B|}{\min(|A|, |B|)}. \tag{3.2}$$

The terminology graph counts actual documents containing each pair among the hundred most frequent domain-assigned terms, breaking frequency ties lexically. Whole-word matching is case-insensitive; repeated mentions within a document count once. Shared labels never create an observed edge. Document co-occurrence does not establish a semantic or causal relationship. The separate overlap heatmap summarizes classifier assignments. A fixed-margin Curveball randomization of the binary document–term incidence matrix supplies a conditional reference for the projected edge count and mean local clustering (Supplemental Section 11.7).

3.2.6 Rhetorical and Discourse Analysis

Regex and keyword analyzers identify candidate rhetorical, argumentative, and framing expressions. Framing is evaluated in three-token windows around extracted terms. Domain framing proportions measure occurrence contexts containing an anthropomorphic marker; multi-domain occurrences contribute to each domain but once to the overall tally. These are lexical proxies without gold-standard annotation validation. Discourse counts are accompanied by analyzed-document counts and sampling fractions. PMC discourse uses a deterministic approximately one-fifth sample of eligible texts rather than a full-text census.

3.2.7 CACE Evaluation

Four heuristic dimensions define CACE:

$$\text{Clarity}(t) = \min(1, \max(0, 1 - H(t)/3.32)), \tag{3.3}$$

$$\text{Appropriateness}(t) = \max(0, 1 - [0.4\,\mathbf{1}_{t\in\mathcal{A}} + 0.1|\text{overlap}(t, \mathcal{A})| + 0.05\max(|D(t)| - 1, 0)]), \tag{3.4}$$

$$\text{Consistency}(t) = \frac{2}{n(n-1)} \sum_{i<j} \frac{\mathbf{x}_i \cdot \mathbf{x}_j}{\|\mathbf{x}_i\|\|\mathbf{x}_j\|}, \tag{3.5}$$

$$\text{Evolvability}(t) = \tfrac{1}{2}\min(1, |D(t)|/3) + \tfrac{1}{2}\min(1, |S_t|/3). \tag{3.6}$$

The implemented Clarity normalization constant is 3.32, approximately $\log_2 10$; it is a design choice rather than the five-cluster ceiling. Here $\mathcal{A}$ is the configured anthropomorphic vocabulary, $\mathbf{x}_i$ the TF-IDF vector of the i-th of n short extraction contexts, $D(t)$ the domain assignments, and S_t scale-marker categories found in contexts. The aggregate is their arithmetic mean. Domain CACE aggregates and their figure use the same up-to-fifty-term sample, ordered by decreasing frequency and lexical tie-break, and at most ten short extraction contexts per term. Computed entropy enters Clarity where available. Default zero entropy for unavailable estimates and Consistency of 0.5 for fewer than two contexts are conventions, not evidence of clarity or consistency. Weights and thresholds are design choices; no completed coefficient sensitivity study, inter-rater reliability measurement, or independent predictive validation is claimed.

3.2.8 Historical Analysis and Coverage

All 2430 stored BHL documents enter literal seed-term frequencies per 10,000 OCR tokens by era. Literal spellings, hyphens, and OCR errors affect rates. The computational stack streams extraction and framing over all stored documents by default. Coverage metadata record available and analyzed documents and characters. Entropy is restricted to twenty frequent candidates per era, drawing contexts from all era documents. An optional positive development character budget selects complete documents by deterministic dyadic traversal across shard order; its fingerprint and coverage metadata distinguish it from a complete run. Neither the full stored corpus nor a bounded development selection is presumed representative of historical entomology.

3.2.9 Validation and Reproducibility

Tests exercise numerical examples, real files, corpus slices, and local HTTP servers. Negative controls reject malformed text, stale content, missing receipts, and failed renderer commands. They establish software behavior, not scientific validity of the proxy measures. Python dependencies are resolved by uv.lock; NLTK resources are separate prerequisites. Content fingerprints bind analyses to ordered records, project source, and locked dependencies. A completed manifest binds input and output files; rendering rejects changed artifacts and nonzero toolchain exits.

3.3 Network Representation Sensitivity

Using the verified document–term incidence and stable frequency ranking, we examine nested vocabularies of 25, 50, 75 and 100 terms at inclusive shared-document thresholds of 1, 2, 5, 10 and 20. This declared representation grid complements the fixed-margin reference; it is neither a tuned model nor a random sample. Projection and dynamics address distinct questions (Newman, 2003).

Every selected term remains a node, including isolates. An edge exists when $w(u, v) \geq \tau$, and density is $\rho_{n,\tau} = 2|E_{n,\tau}|/[n(n-1)]$. Mean local unweighted clustering is zero for degree-zero/one nodes. Component count, largest-component size, isolates and documents containing any selected term are exported. Changing vocabulary size changes both possible pairs and selected incidences; density across sizes need not be monotone.

The separate receipt binds inputs, implementation, lock and all outputs. Validation reconstructs the baseline and replays the report. No additional null model, bootstrap, uncertainty interval or community optimization is executed. The statistics diagnose representation dependence, not semantic validity or biological complexity.

4 Results: Corpus Analysis and Terminology Networks

4.1 Terminology Extraction Across Domains

Analysis of the source-identified headline layer yields 11,700 candidate terms from 7599 abstracts and 998,645 processed tokens. Of these candidates, 1330 receive domain assignments. These counts exclude 10 unreconciled archived strings; broad retrieval still limits relevance and representativeness.

Domain	Terms	Frequency	Bridging terms
Unit of Individuality	530	28,194	24
Behavior & Identity	262	17,119	91
Power & Labor	295	12,533	139
Sex & Reproduction	245	9334	62
Kin & Relatedness	84	4040	4
Economics	79	4508	3

Table 1. Rule-based domain assignments and corpus frequencies. A term can receive several labels, so domain counts and frequencies are not mutually exclusive. Bridging means multiple labels, not an observed transfer of meaning.

Among assigned terms, 11.9% have multiple domain labels. This is a property of the classifier and corpus. Temporal semantic drift would require explicit time-indexed meaning comparison, while lexical, contextual, and scale ambiguity require independent sense annotation or validated proxies. Those measurements are not established by label overlap.

The processed vocabulary has type-token ratio 0.0473. Its most frequent WordNet lemmas are ant (20,494), specie (10,620), and colony (7436). Lemmatization reduces inflected forms and can truncate words such as *species* to *specie*; extracted terms retain surface forms (Supplemental Section 10). Frequency identifies recurring lexical material; it does not establish its conceptual importance or the intentions of authors.

4.2 Terminology Network Structure

The observed terminology graph uses the hundred most frequent domain-assigned terms. Its edge weight counts documents containing both terms:

$$w(u, v) = \sum_{d=1}^{N} \mathbf{1}[u \in d]\mathbf{1}[v \in d]. \tag{4.1}$$

Whole-word matches are case-insensitive, and repeated mentions within a document do not add weight. No edge is inferred from shared labels or extraction order. The graph has mean local (unweighted) clustering coefficient 0.8955; this statistic does not measure conceptual coherence, communication quality, or resistance to reform.

4.2.1 Comparison with a Fixed-Margin Reference

Density and clustering depend on how many selected terms each abstract contains and how many abstracts contain each term. A Curveball randomization of the binary document–term incidence matrix preserves both margins (Supplemental Section 11.7). Over 3 chains and 600 retained draws for the 7599 source-identified abstracts and 100 terms, the randomized reference averages 4533.6 edges (central 95% of draws 4501–4568) and mean local clustering 0.934 (0.930–0.939). The observed graph has 4235 edges and clustering 0.895, below both envelopes (Figure 3). A sensitivity protocol with longer burn-in, wider spacing and different seeds (300 draws) gives reference

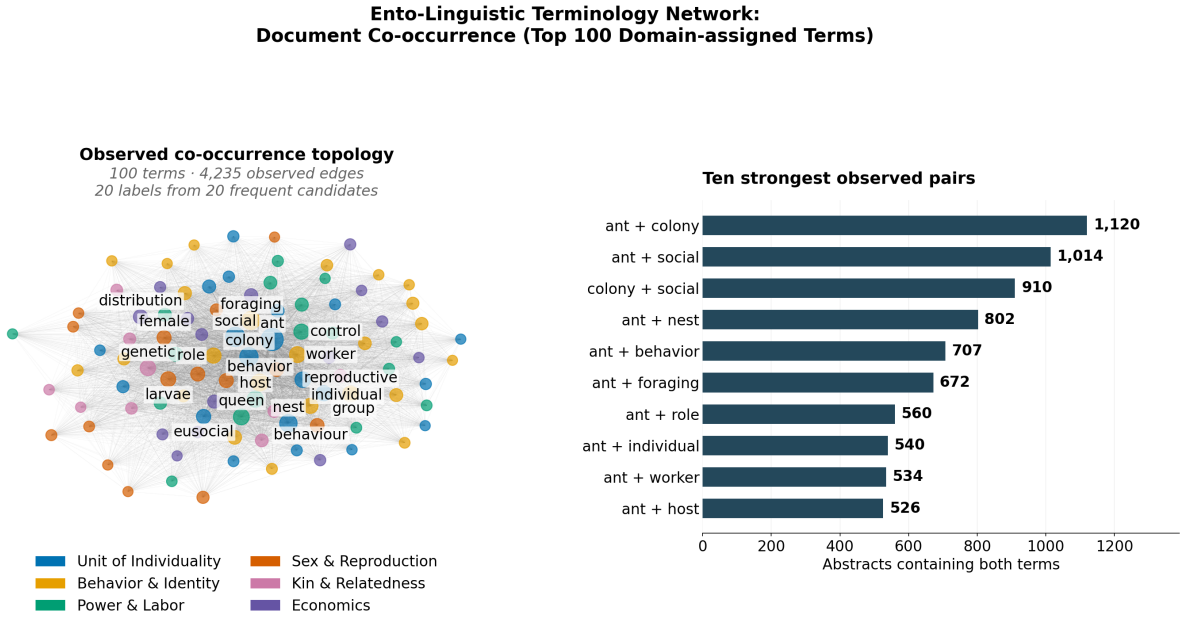


Figure 2. Observed document co-occurrence among the hundred highest-frequency domain-assigned terms in the 7599-abstract layer. Nodes represent terms, node area uses a square-root frequency scale, color identifies the primary domain, and edge width scales shared-document counts to a bounded display range (Eq. 4.1). Isolated nodes are omitted from the display and up to twenty frequent terms are considered for collision-filtered labels. The right panel ranks the ten strongest observed pairs by shared-document count, with lexical tie-breaking. All observed edges remain in the left graph, whose unweighted layout reduces visual concentration around frequent pairs. Layout and dense regions have no causal or hierarchical interpretation.

means of 4534.5 edges and 0.934 clustering. Between-chain $\hat{R}$ is 1.006 for edges and 1.003 for clustering; this diagnostic does not prove mixing.

Selected terms therefore co-occur in fewer distinct pairs, with less closed triadic structure, than their margins alone would produce: co-occurrence is concentrated among particular pairs. The envelopes describe the sampled conditional reference rather than uncertainty in the observed values, and the comparison does not identify a linguistic cause.

4.2.2 Vocabulary and Edge-Threshold Diagnostic

Across the declared 20 vocabulary–threshold combinations on the same 7599 identified abstracts, the 100-term graph at a one-document threshold has density 0.8556. Requiring at least twenty shared documents retains 1211 edges and gives density 0.2446. Figure 4 displays the full grid, including mean local clustering with isolates retained. These are deterministic descriptions of selected representations, not uncertainty intervals or evidence that one threshold recovers true semantic relations. Clustering need not decrease when edges are removed: changing neighborhoods also changes local triangle denominators. The curves separate these topological summaries rather than interpreting either as biological complexity.

Domain-assignment overlap is a different quantity, displayed separately in Figure 5.

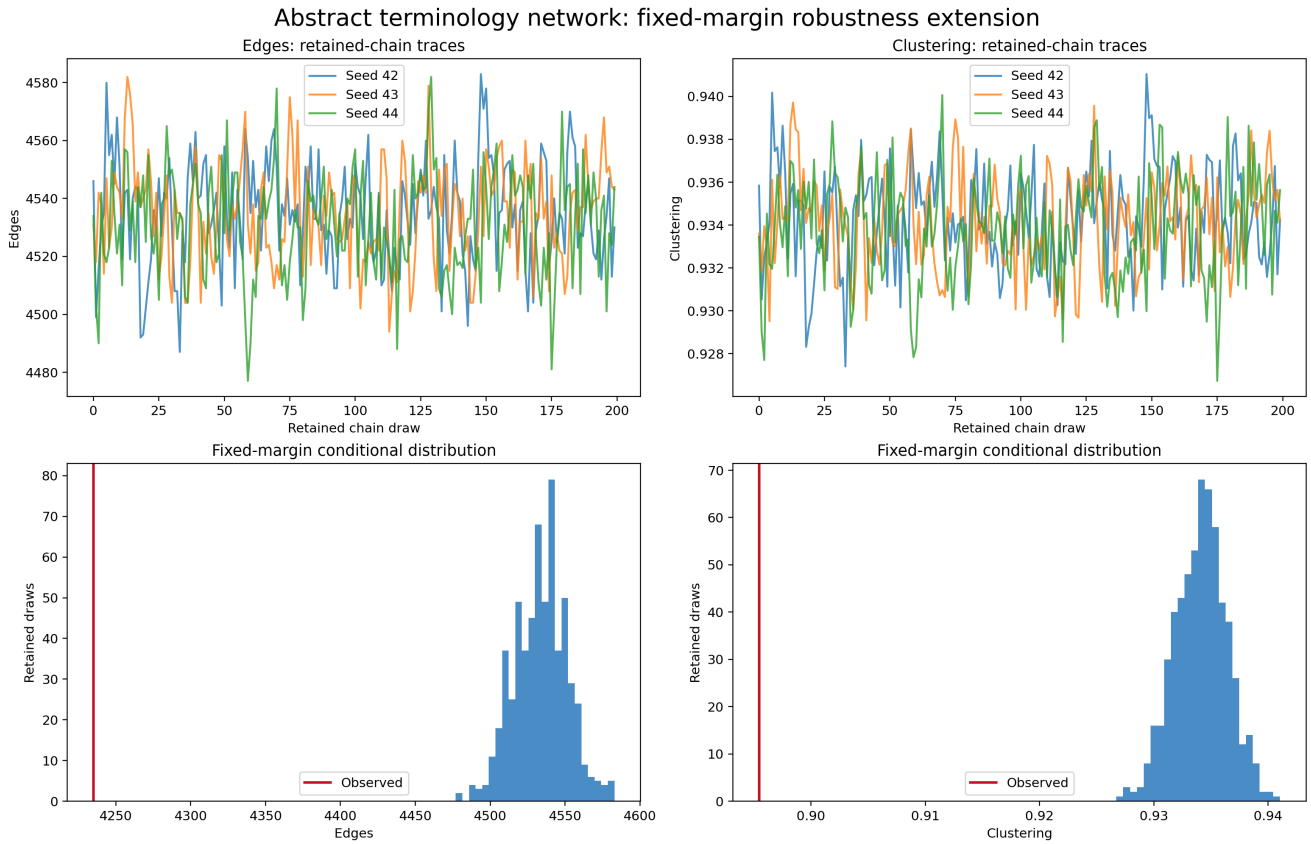


Figure 3. Fixed-margin comparison for the abstract terminology network. Top: retained-chain traces of projected edge count and mean local clustering for 3 seeds. Bottom: distributions of the 600 retained draws, with the observed value marked. Every draw preserves each abstract's selected-term count and each term's document frequency. The reference is conditional on the frequency-ranked vocabulary and the convenience corpus.

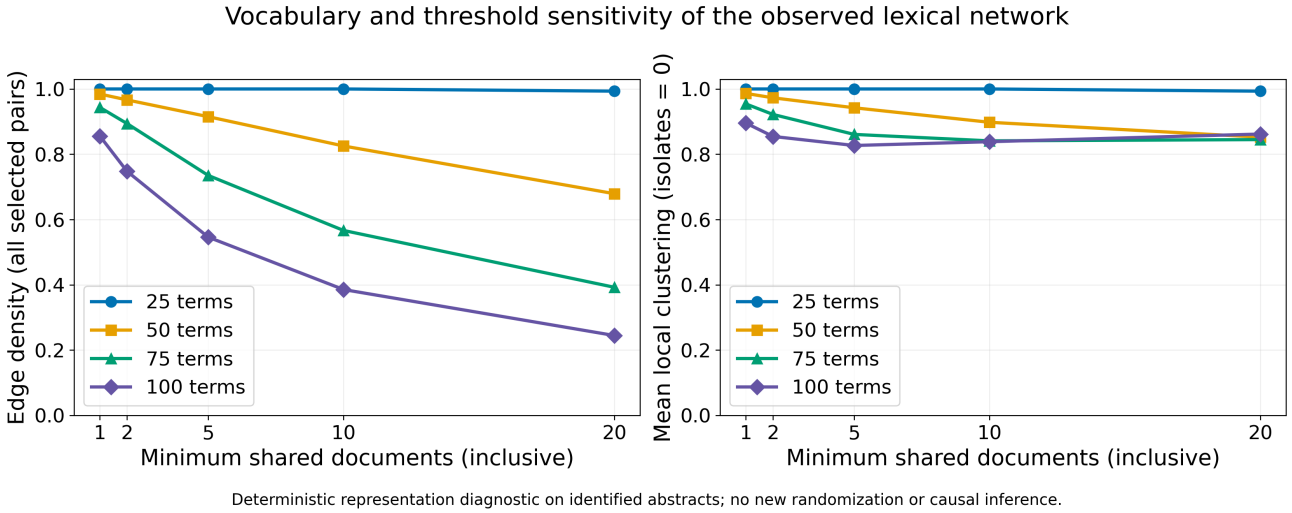


Figure 4. Vocabulary and minimum-shared-document sensitivity. Curves use nested frequency-ranked vocabularies and inclusive thresholds. The denominator includes all selected pairs, and clustering includes isolates as zero. Each point is recomputed from the same identified abstract layer; no new randomization or biological-network inference is represented.

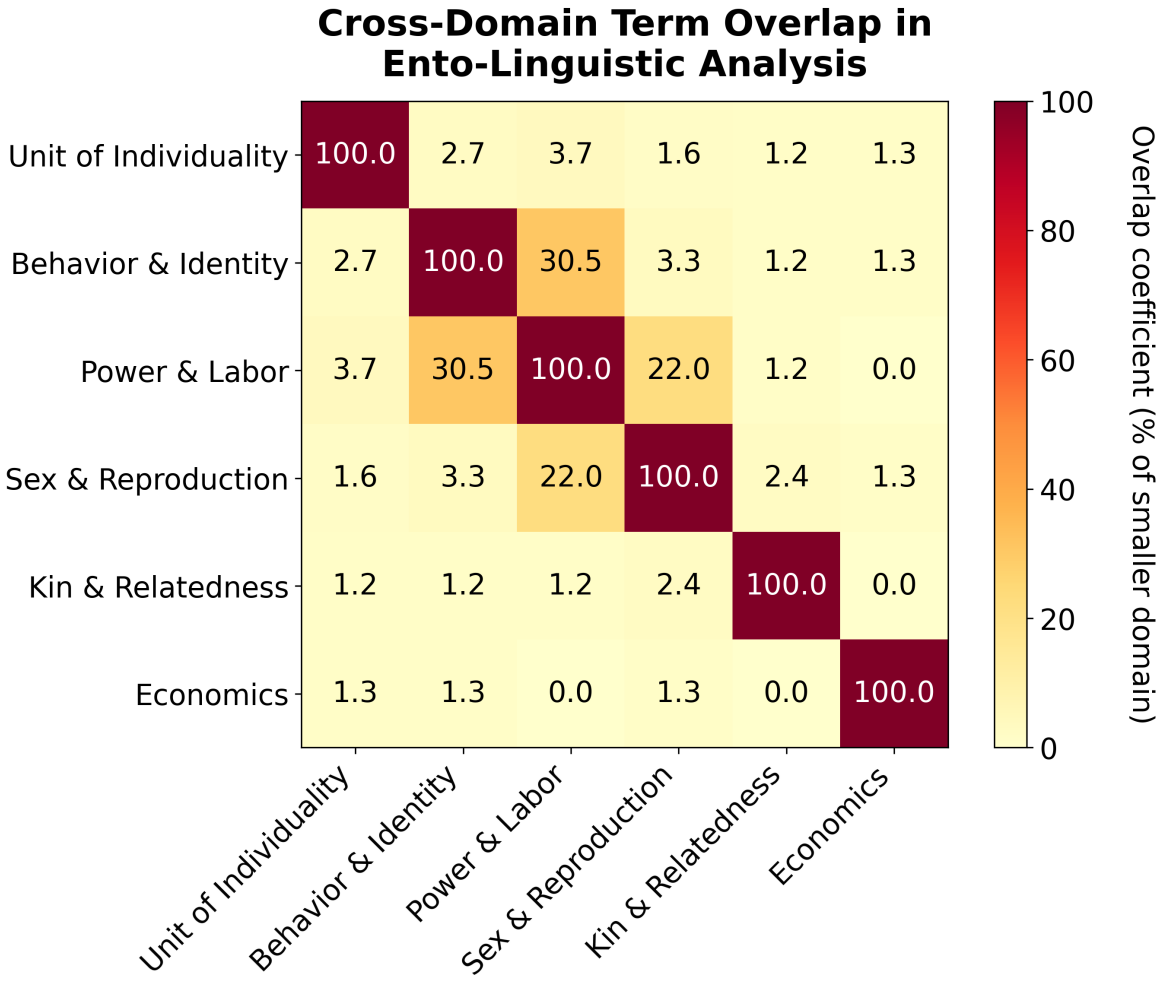


Figure 5. Szymkiewicz–Simpson overlap coefficients between domain-assigned vocabularies (Eq. 3.2). Each cell shows shared terms as a percentage of the smaller vocabulary; the diagonal is 100 by definition. Values reflect the current lexical classifier; observed zeros do not establish conceptual isolation.

4.3 Framing Analysis

Lexical markers identify contexts for further qualitative examination. They do not distinguish metaphor from technical usage, demonstrate author bias, or establish a language-induced change in biological models. The same distinction applies to the domain-specific interpretations in Section 5.

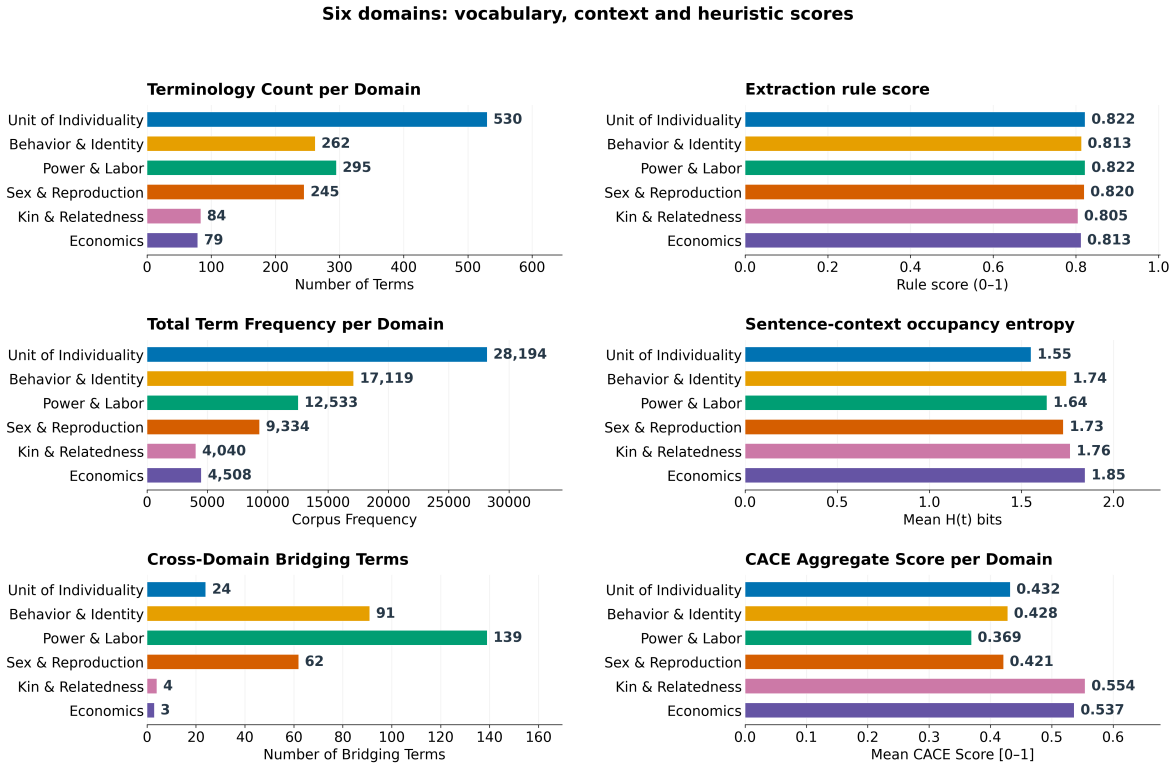


Figure 6. Six descriptive panels show distinct term counts, mean extraction confidence, total frequency, mean successfully computed sentence-context entropy, bridging counts, and heuristic CACE means over up to fifty selected terms per domain. Extraction confidence is a configured score rather than calibrated classification accuracy. Entropy and CACE sample definitions are specified in Methods; missing entropy does not establish zero ambiguity.

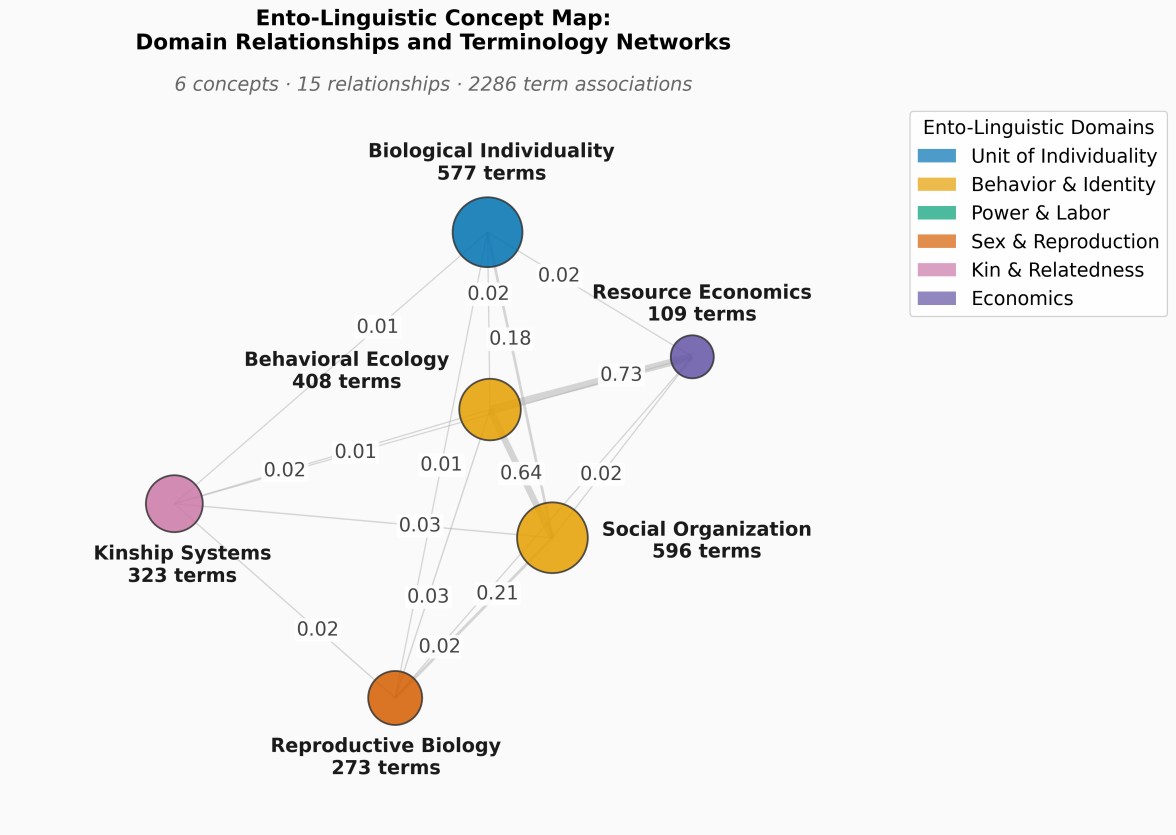
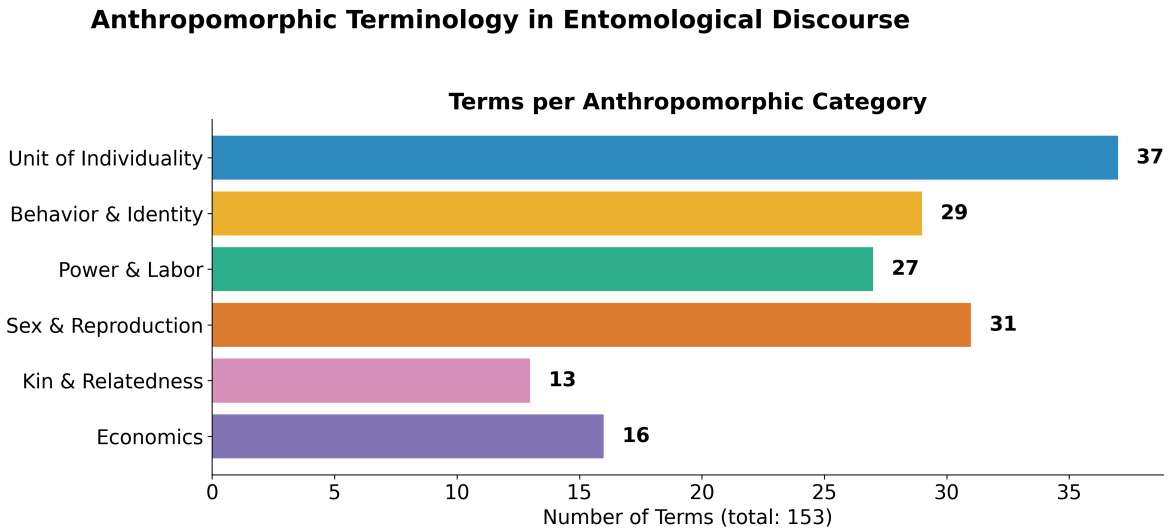


Figure 7. Six predefined concept categories connected by vocabulary overlap. Node size summarizes associated terms, which can appear in several categories; the subtitle totals term associations rather than unique terms. Edge weights are classifier-defined overlaps, not observed causal connections or a discovered biological ontology.



Sample Terms by Category

Category	Example Terms
Unit of Individuality	nest-sites-we, pre-social, nest-based, ant-frequented, ant-following, ant-symbiont ... (+31)
Behavior & Identity	food-recruitment, free-foraging, host-specialization, reproductive-to, trail-mediated, cone-foraging ... (+23)
Power & Labor	host-parasitoids, host-specialization, queen-signal, authority, host-plants, queen-derived ... (+21)
Sex & Reproduction	reproduction-immunocompetence, male-phase, queen-signal, reproduction-survival, sex-differentiation, queen-derived ... (+25)
Kin & Relatedness	non-kin, kin-recognition, sibling, selfish, kin, conflict ... (+7)
Economics	trade-off, trade-offs, trade, risk-return, cost-benefit, investment ... (+10)

Figure 8. Observed framing-marked terminology by canonical domain. Top: distinct extracted terms with at least one occurrence context matching an anthropomorphic pattern. Bottom: up to six terms per domain, selected by decreasing matched-context proportion, context count, and lexical ordering. Counts are neither curated vocabulary sizes nor occurrence frequencies. Occurrence-context framing proportions are exported separately.

5 Results: Domain-Specific Findings

The six domains organize descriptive outputs and questions for qualitative interpretation. Their assignments follow the configured vocabulary and lexical rules. Domain-specific statements below do not infer cognitive bias, biological mechanisms, or author intent from term counts.

5.1 Unit of Individuality

This vocabulary includes references to individuals, colonies, and collective organization. Its mean computed sentence-context entropy is 1.55 bits. Frequency and word-formation panels describe extracted surface forms, while the scale panel counts keyword matches in term names. These counts do not estimate biological boundaries. Figures 9, 10, and 11 support inspection of the extraction.

5.2 Power & Labor

Power and Labor contains 295 assigned terms and 139 terms with multiple labels. Its occurrence-context anthropomorphic-marker proportion is 1.2%, rather than a percentage of authors or publications using misleading language. Discussions of loaded terminology motivate contextual examination (Herbers, 2006, 2007). Molecular work on caste (Sumner et al., 2018) and developmental canalization (Qiu et al., 2022) also make it important to distinguish developmental phenotypes from temporary task categories.

Figures 13 and 14 show observed term frequency and context-cluster entropy. Figure 12 ranks predefined concept categories by summed vocabulary overlap; it does not show a biological hierarchy or term-level betweenness.

5.3 Behavior & Identity

The mean computed sentence-context entropy is 1.74 bits. Task labels provide useful candidates for context annotation. Evidence of behavioral flexibility in ants (Gordon, 2010, Ravary et al., 2007) motivates asking when a label denotes an observation, a persistent propensity, or a morphological category. The corpus statistics alone do not establish that categorical labels obscure that flexibility.

5.4 Sex & Reproduction

This domain groups reproductive and developmental terminology. Its mean computed entropy is 1.73 bits. The presence of paired labels does not by itself demonstrate a conceptual opposition. Reproductive systems and caste development vary across taxa, and the epigenetic review (Oldroyd and Yagound, 2021) provides background rather than validation of a lexical classifier.

5.5 Kin & Relatedness

The mean computed entropy is 1.76 bits. Relatedness terms require biological and demographic context: numerical coefficients depend on pedigrees, mating systems, and population structure. The present outputs do not estimate those quantities.

5.6 Economics

The classifier assigns 79 terms to Economics, with 3 receiving multiple labels. Its mean computed entropy is 1.85 bits. Low overlap can follow directly from lexicon design; a zero overlap does not establish a closed conceptual subsystem. Terms such as allocation (219 occurrences), investment (272), resource (579), and resources (713)

warrant examination of their operational definitions. Statistical or ecological compounds matching a seed word can be classification errors rather than evidence of economic framing.

5.7 Historical Interpretation

Historical readings of caste and superorganism terminology provide context (Boomsma and Gawne, 2018, Gordon, 1992, Wheeler, 1911). The BHL era analysis in Section 14.2 reports literal OCR frequencies and explicitly sampled stack outputs. Changes in those rates do not independently establish changes in conceptual commitments, the origin of a concept, or progress toward more accurate biological explanation.

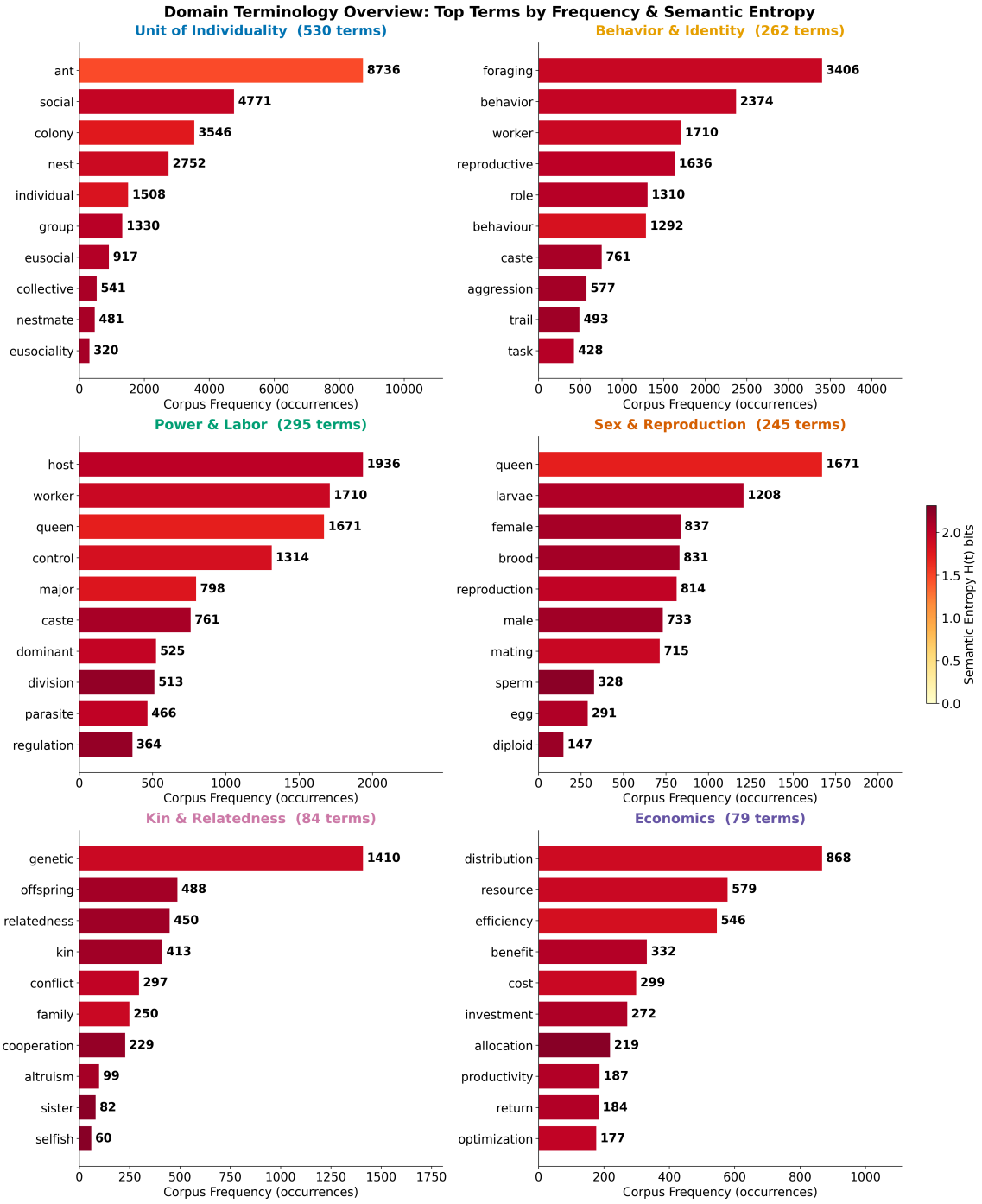


Figure 9. Ten highest-frequency extracted terms per domain. Bar length is corpus frequency and color is the attached sentence-context entropy estimate. Zero-valued defaults for insufficient contexts must not be interpreted as evidence of unambiguous meaning. These are descriptive extraction outputs from the source-identified abstract layer.

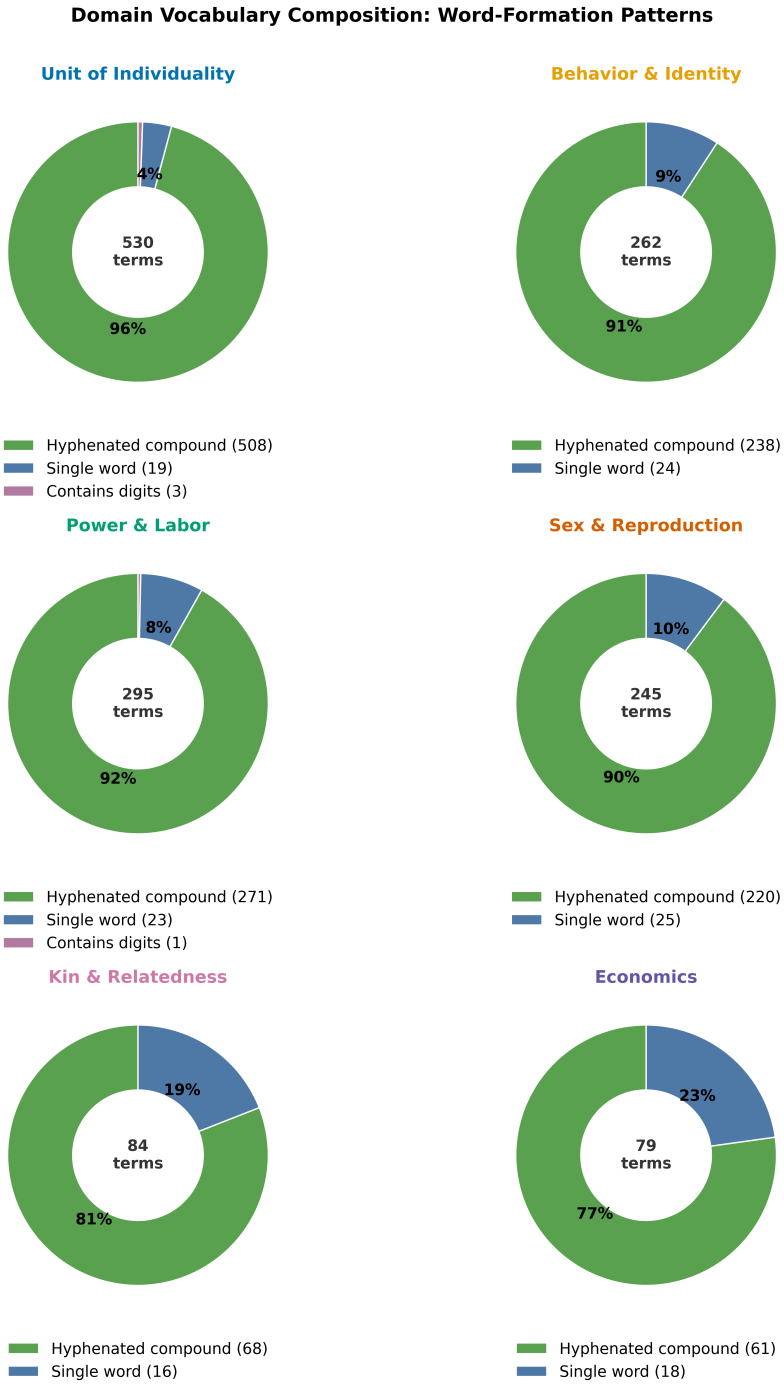


Figure 10. Surface word-formation composition of assigned vocabularies. Each term is assigned to exactly one surface category (single word, multi-word phrase, hyphenated or underscore compound, or containing digits), using the same classifier as Figure 11. The headline extractor does not independently add n-grams, so availability of a category in the plotting utility does not establish its extraction.

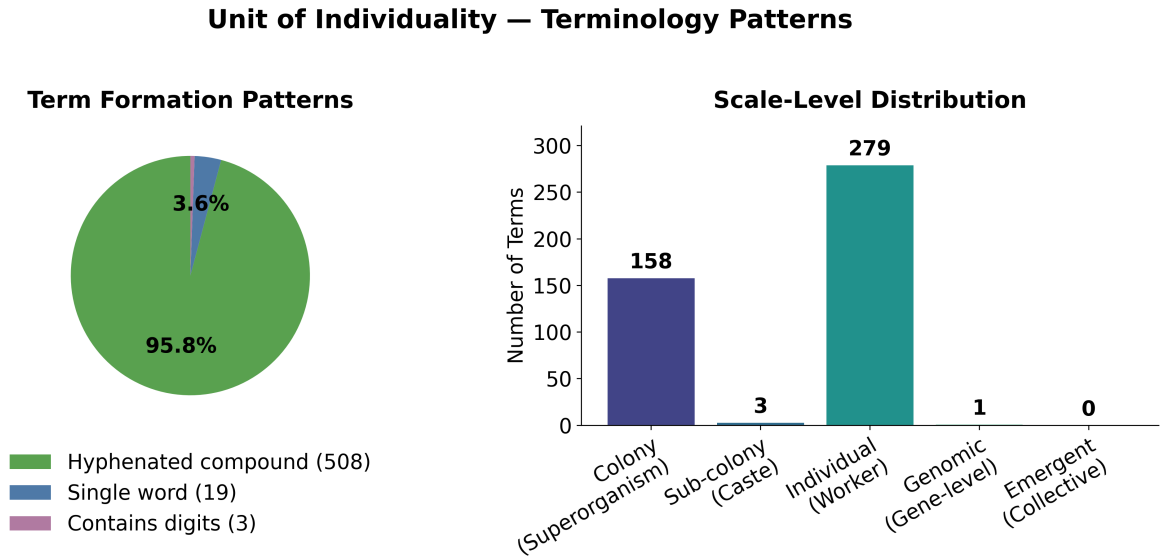


Figure 11. Unit of Individuality term-formation counts (left) and counts of term names matching scale keyword groups (right). Scale groups can overlap; true zeros are retained. Neither panel estimates actual biological scale boundaries.

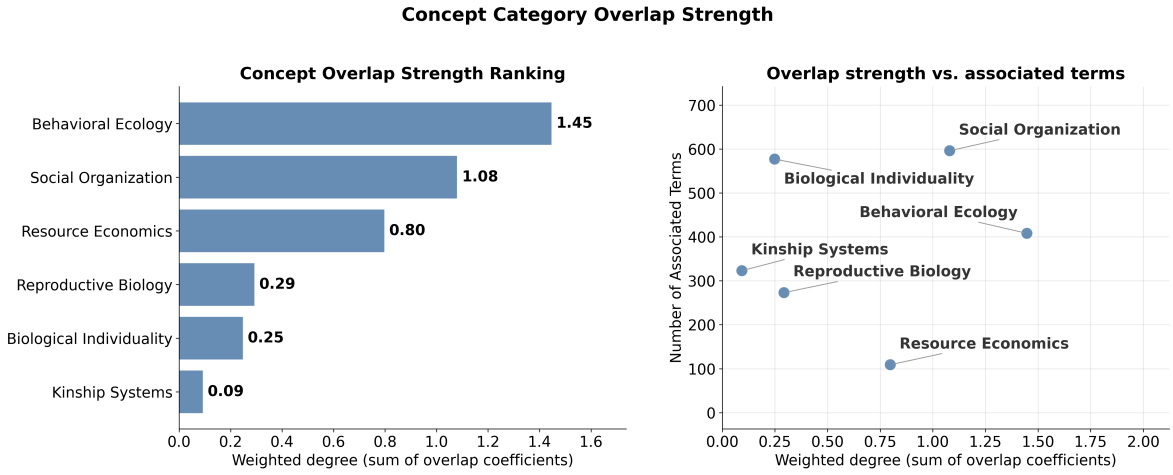


Figure 12. Overlap strength of the six predefined concept categories: each category's weighted degree, the sum of its overlap coefficients with the other categories (Eq. 3.2), ranked (left) and plotted against associated-term counts (right). Every category links to every other, so unweighted degree would not distinguish them. The display does not represent a biological command hierarchy.

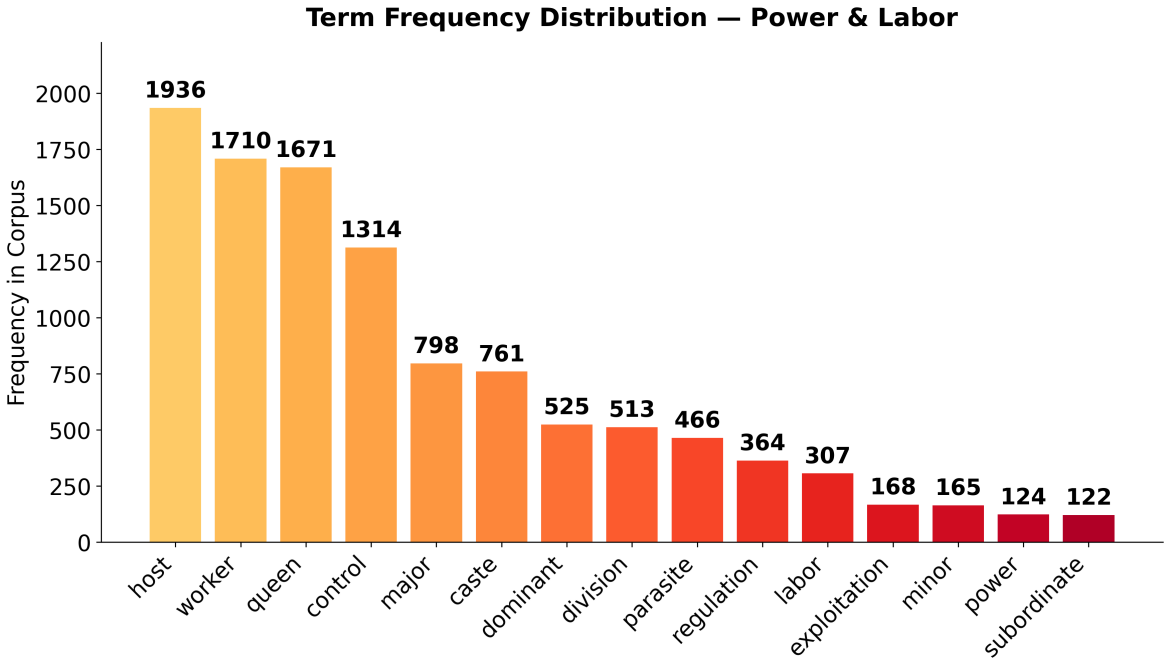


Figure 13. The fifteen highest-frequency assigned Power and Labor terms in the source-identified abstract layer. Frequency is annotated; color tracks rank. These observations do not quantify hierarchical control or bias.

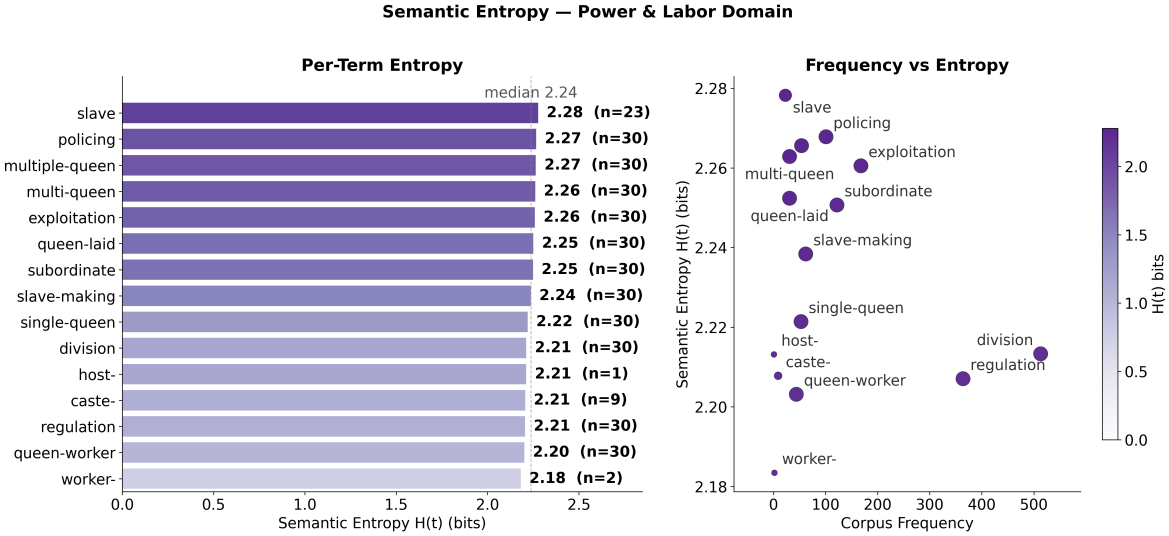


Figure 14. Power and Labor terms ranked by attached context-cluster entropy, with corpus frequency and short extraction-context counts. Sentence-context entropy and the plotted extraction-window counts use different context definitions. Cluster entropy is not an independently validated ambiguity measure.

6 Discussion

6.1 Language and Scientific Practice

The descriptive outputs organize terminology into a framework for examining the relationship between language and biological explanation. They are compatible with asking whether scientific metaphors influence inquiry (Lakoff and Johnson, 1980, Latour, 1987, Longino, 1990), but do not test that causal hypothesis. Keyword frequency, sentence-cluster entropy, and co-occurrence identify lexical patterns rather than researchers' beliefs, decisions, or errors.

Terms such as *queen*, *worker*, and *caste* deserve contextual scrutiny because biological roles and ordinary-language connotations can differ. Existing discussions of terminology reform provide a substantive motivation (Herbers, 2006, 2007). The present corpus analysis neither establishes that terminology delayed particular discoveries nor measures the adoption of alternatives across the field. Such claims would require dated source analysis and evidence about research decisions.

The observed graph has clustering coefficient 0.8955, while 11.9% of assigned terms receive multiple labels. Neither quantity demonstrates self-reinforcing conceptual bias. Co-occurrence can arise because papers discuss several biological processes together; label overlap also follows from the predefined lexicons. A visualization's arrangement must not be read as a human-style command hierarchy.

6.2 Network Structure and Its Conditional Reference

A dense co-occurrence graph needs a reference that accounts for vocabulary selection and document-level opportunity. The companion fixed-margin extension preserves selected-term counts within documents and the document frequency of each term. In the executed finite-chain comparisons, the observed graph has fewer edges and lower clustering than the randomized reference. The same direction appears under the longer-burn, wider-spacing sensitivity protocol. The substantive interpretation is concentration of co-occurrence relative to those margins, rather than exceptional density attributable to terminology alone. Topic, genre and other sources of document organization remain possible explanations. A graph's visual density is therefore an observation to explain, not evidence of conceptual bias by itself.

The extension also makes vocabulary sensitivity explicit. Changing the frequency-ranked subset changes the projection and the amount of connectivity it can display. Interpretation should state the selected vocabulary, the incidence margins and the comparison model. The comparison is separately receipted, so its inputs and outputs can be inspected independently of the corpus analysis.

6.3 Biological Precision and Terminological Continuity

A terminology decision has several possible costs: obscuring a biological distinction, inviting an unsupported analogy, or making relevant literature harder to retrieve. These costs need not move together. For example, replacing a developmental category with a task description may lose information even when the replacement sounds less anthropomorphic. The Introduction's distinction between a definitional question (what counts as a caste) and a mechanistic one (how experience shapes division of labor) is the distinction a useful glossary should preserve.

The practical unit of revision is therefore a defined use in a passage. Authors can specify the organism or collective being measured, the criteria for category membership, and the mechanism supported by their observations. Linking an alternative expression to established indexing vocabulary can retain discoverability. CACE can structure that review, but its aggregate should accompany the definition and evidential rationale rather than decide which

expression is biologically correct.

6.4 Active Inference as a Theoretical Perspective

Active Inference offers a vocabulary for discussing generative models, inference, and action (Clark, 2013, Friston, 2010). In the Active Inferants study, a simulated ant-foraging model reproduces selected colony phenomena in a laboratory-inspired setting (Friedman et al., 2021). This provides an example of mechanistic modeling without a centralized controller; it does not experimentally compare terminology choices or establish the empirical adequacy of every biological assumption.

Interpreting terminology as a prior is an analogy here; as the Introduction states, no Active Inference model is fitted to the corpus. Whether an ant or colony admits a useful blanket description depends on specified state variables, dynamics, and conditional-independence assumptions (Friston, 2013, Kirchhoff et al., 2018); it cannot be settled by changing a noun.

The Environment-Centric Active Inference and related multiscale proposals in Section 13 should be read as theoretical extensions. They require explicit models and empirical tests before supporting biological or linguistic conclusions.

6.5 A Complex Systems Reading of Terminology

The framework can organize questions about scale, connectivity and feedback without treating a lexical graph as a biological system. For each passage, specify the units (individuals, tasks, interactions or colonies), the relevant state variables, the observation timescale and the coupling being measured. The paper's three networks must remain distinct: observed term co-occurrence, configured category overlap and the biological interactions discussed by the underlying literature. A dense term graph does not supply evidence that the ants have an equally dense communication graph.

This distinction supports productive comparison rather than automatic analogy. A mechanism-based account of task allocation can connect the Power and Labor vocabulary to questions about local information and regulation, while biological individuality asks whether the reported outcome is measured at individual or colony scale. The next empirical step is to annotate these mechanistic commitments in source passages and test agreement on held-out documents. Only then can the framework assess whether a terminology choice helps readers recover the actual units, coupling and conditions of a study. Expanding to other taxa, genres or languages requires a new measurement-validation assessment; reproducible software alone does not establish transportability.

6.6 Practical Use of CACE

CACE makes evaluation criteria inspectable: clarity of operational definitions, suitability of metaphors, consistency of use, and adaptability to new findings. Its numerical implementation is a heuristic instantiation. Penalizing membership in a predefined anthropomorphic vocabulary partly encodes the desired ranking; a favorable score is not independent validation of a replacement term.

The representative-term table includes terms absent from the extraction. Their scores are derived from text features and fallback conventions, as recorded in the Extracted column. In particular, a zero default entropy for an unattested term must not be interpreted as demonstrated specificity. Comparisons between *slave* and *host worker* illustrate these rules, rather than measuring a change in readers' comprehension or research outcomes.

The computed aggregate values are 0.30 and 0.47, respectively, with Appropriateness 0.50 and 0.40. These values remain useful for inspecting the scoring implementation. Independent assessment should compare definitions and

context-specific biological accuracy, collect blinded judgments, measure agreement, and test sensitivity to the chosen weights before treating the scores as a prescriptive standard.

6.7 Cross-Domain Communication

Domain overlap can help select terms for shared glossaries and explicit operational definitions. The six-domain framework is one proposed partition; alternative lexicons and annotation schemes may produce different assignments. Maintaining links to established terminology can support discoverability while clarifying which mechanism or observational category is meant. Benefits to communication remain hypotheses for reader and author studies.

6.8 Measurement Validation and Terminology Intervention

Two study designs separate measurement quality from communication effects. The first is the annotation design specified in Section 3.2.3: it tests whether occupancy entropy and lexical indicators track reader-labeled meanings and discourse functions after accounting for term frequency, document length and topic, with agreement and uncertainty reported at the level of the annotated observation.

The second borrows the manipulation logic of metaphor experiments (Section 8). A randomized reader study can hold biological evidence constant while varying an established label, a proposed alternative and an explicit operational definition. Prespecified outcomes include inference accuracy, confidence calibration, recall and literature retrieval. This design tests a terminology effect directly. Comparing its outcomes with CACE rankings then evaluates the scoring proposal using evidence outside the scheme. Researchers and students can be included as prespecified groups rather than assumed to interpret a term identically.

6.9 Limitations

1. **Source custody and relevance.** Some archived abstract strings lack digest-indexed source metadata and are excluded from headline analysis; broad queries include adjacent biology and computational uses. Historical volumes contain mixed topics; PMC notices are excluded by a title rule rather than a complete article-type screen. Source reconciliation and relevance annotation remain incomplete.
2. **Sampling and accessibility.** Search and availability filters select a convenience corpus. Layers are not matched document samples and differ in extraction threshold, length, genre, and publication history.
3. **OCR and language.** Literal historical matching is sensitive to OCR errors, language, hyphenation, and spelling. Rare or absent matches do not date a concept's origin.
4. **Proxy validity.** Computational clusters are not annotated senses; marker patterns are not validated discourse labels. The pipeline does not measure author intent, cognitive distortion, or causal framing.
5. **Statistical dependence.** Domain groups overlap and share document-derived contexts. Reported test outputs are exploratory; no calibrated confirmatory inference is claimed.
6. **Explicitly bounded components.** PMC discourse uses an approximately one-fifth deterministic text sample; domain CACE uses at most fifty terms; BHL entropy uses twenty frequent terms per era with full-corpus contexts. Default BHL extraction, framing, and literal-frequency calculations include all stored documents; an optional development character cap produces a separately fingerprinted subset.
7. **Unmeasured theory.** No fitted generative model, empirical Markov-blanket estimation, term-reform intervention, or independent human validation is reported.

These boundaries support a focused next study: a frozen, fully reconciled corpus with independent annotations and prespecified document-level analysis, followed by a controlled evaluation of operational terminology.

6.10 Complexity as an Explicit Modeling Question

Complexity connects the paper's questions about organization, information and scale while requiring a defined target system and measurement. Accounts of complex systems distinguish candidate features and mathematical measures rather than supplying a universal interpretation of density or entropy (Ladyman et al., 2013). A complete graph has maximal density and a short description; dense word co-occurrence alone does not establish emergence, adaptation or collective cognition. One-mode projection itself can create local clustering (Newman, 2003). The new representation diagnostic consequently reports how graph statistics depend on vocabulary and edge thresholds, without converting them into biological properties.

Architecture asks which units and relations a passage specifies. Simon's near-decomposability account concerns relatively weak coupling between subsystems and different short- and long-run behavior (Simon, 1962); overlapping lexical domains do not demonstrate those dynamical conditions. Fertilization exchanges questions and models across collective behavior, network science and terminology research while retaining their observational units. Expansion asks whether measurement transfers to new taxa, genres, languages and readers after held-out validation. These three themes support a complex-systems research program; the executed corpus analyses do not by themselves complete it.

7 Conclusion

This work provides a six-domain framework and a reproducible descriptive pipeline for examining terminology in scientific text. Across 7599 source-identified abstracts, the pipeline processes 998,645 tokens and extracts 11,700 candidate terms, with 1330 receiving domain assignments. Observed document co-occurrence, predefined conceptual-category overlap, sentence-context entropy, and heuristic framing and CACE scores are reported as distinct quantities.

The 11.9% of assigned terms with multiple labels measures classification overlap, not semantic drift. The 6 concept categories are predefined rather than discovered. Complementary source layers extend the descriptive scope without establishing that language causes bias, that a terminology reform improves scientific modeling, or that numerical CACE rankings are independently validated.

The fixed-margin comparison sharpens the network contribution: visual density must be interpreted relative to document and term incidence, and the executed reference indicates more concentrated co-occurrence than those margins alone would predict. It supplies a conditional descriptive comparison without identifying a linguistic cause. The broader contribution is a method for keeping biological definitions, lexical observations, and proposed communication effects distinguishable while investigating their relationships.

7.1 Future Directions

The immediate research priorities are source reconciliation, document-level relevance and license review, and independent annotation of senses and framing. A frozen annotated corpus would support tests of extraction accuracy, agreement between annotators, context-count sensitivity, and statistical models accounting for shared documents and overlapping domain labels.

Longitudinal analysis should distinguish changes in source composition and OCR quality from changes in terminology. Multilingual comparisons require language-specific normalization and definitions. Reader or author experiments could then evaluate whether operational definitions and alternative terms improve comprehension or mechanistic explanation.

CACE offers explicit evaluation questions for such studies. Its current scores provide an inspectable starting point, with design choices and missing-data conventions disclosed. Active Inference and multiscale interpretations remain theoretical proposals until connected to specified models and empirical measurements. The repository contributes tools and auditable descriptive results for that work.

The complexity theme connects precise units, cross-disciplinary exchange and validated transfer. Architecture specifies what is represented; Fertilization exchanges questions while preserving evidence boundaries; Expansion tests measurement in new settings. The vocabulary–threshold diagnostic and graphical abstract make that program explicit. Reproducible representation is an executed contribution, whereas annotated meaning, reader effects and biological dynamics remain separate validation targets.

8 Related Work

8.1 Scientific Language and Categories

Philosophy of science and discourse research provide context for studying terminology as part of research practice (Fairclough, 1992, Kuhn, 1996, Latour, 1987, Longino, 1990, Wodak and Meyer, 2009). Conceptual and deliberate-metaphor frameworks supply complementary questions about ordinary connotations and communicative intention (Lakoff and Johnson, 1980, Steen, 2017). This study draws motivation from those traditions, while keeping computational lexical proxies separate from claims about thought, intention, or causal influence.

Work on scientific classification and cultural histories offers additional context (Berlin, 1992, Hacking, 1999, Sleigh, 2007). The present six-domain taxonomy is a proposed analytic organization; it is not independently established as exhaustive or uniquely appropriate.

8.2 Terminology in Social-Insect Research

Debates over caste categories and descriptions of ant behavior motivate explicit definitions (Boomsma and Gawne, 2018, Gordon, 1992). Research on collective behavior provides biological context for distinguishing task allocation from an assumed centralized controller (Gordon, 2010, 2019, 2023). The descriptive corpus pipeline does not itself test those mechanisms.

Discussions of racially loaded language in social-insect research provide a substantive case for examining terminology choices (Herbers, 2006, 2007). The ESA Better Common Names Project is a related institutional initiative (Entomological Society of America, 2024), and the term *colony* itself has been examined for its settler-colonial connotations (Vis, 2026). Neither the current frequency analysis nor CACE scoring measures the adoption or effects of these reforms.

Molecular accounts of caste and epigenetic mechanisms (Oldroyd and Yagound, 2021, Sumner et al., 2018) and developmental canalization research (Qiu et al., 2022) reinforce the need to distinguish developmental processes from temporary behavioral labels. Canalization must not be interpreted as proof that every caste category is labile. The historical superorganism literature (Wheeler, 1911) and later conceptual analysis (Boomsma and Gawne, 2018) also caution against dating a concept's origin from an absent OCR match.

8.3 Computational Mapping and Source-Layer Analysis

Computational literature mapping supplies methods for examining connections among terms and publications (Chen, 2006). Here, document co-occurrence is kept separate from vocabulary-overlap relationships between predefined concepts. TF-IDF/KMeans occupancy entropy is used as a descriptive measure of context distributions, without claiming independent sense annotation or validated linguistic ambiguity.

Source-layer comparisons are also descriptive. Abstracts, full texts, historical volumes and preprints differ in access, genre, length, topic and extraction threshold. An observed difference between layers is not automatically a robustness result or a historical change in meaning.

8.4 Validation and Experimental Evidence

Grimmer and Stewart (2013) emphasize, for automated content analysis, validation tailored to the substantive question. Its relevance here is methodological: word counts and computational categories require an explicit connection to the construct they are intended to measure. This study's receipts establish computational custody, not construct validity; Section 3.2.3 specifies the annotation design that would address the latter.

Experimental metaphor research supplies a complementary way to investigate reader responses. Thibodeau and Boroditsky (2011) studied how contrasting crime metaphors affected reasoning. That evidence concerns a different topic and participant setting. It supports using controlled language manipulations to formulate testable questions, while leaving the effects of technical ant terminology on specialists and students unresolved. Connecting these traditions requires both measurement validation and a direct communication experiment.

8.5 Active Inference and Colony Modeling

The Free Energy Principle and Active Inference supply theoretical vocabulary for generative modeling (Clark, 2013, Friston, 2010, 2013, Kirchhoff et al., 2018). The Active Inferants framework supplies a simulated ant-foraging example (Friedman et al., 2021); model behavior alone cannot establish how terminology choices affect scientific reasoning.

Theoretical perspectives on eusociality and biological individuality (Boomsma and Gawne, 2018, Nowak et al., 2010) motivate careful definition of units and mechanisms. The Environment-Centric Active Inference extensions in the supplement remain proposed constructions requiring explicit models and empirical testing.

8.6 Collective Behavior and Complex Systems

Collective-behavior research supplies a mechanistic counterpart to the terminology framework. Couzin (2009) relates group-level response to social interaction, individual state and environmental modification. Feinerman and Korman (2017) distinguishes collective responses supported by individual information from outcomes that require interactions across a scale gap. These accounts motivate asking which entity holds information and which process combines it; a colony-level description alone does not identify the computation.

Ecological conditions also matter to how interaction processes regulate activity (Gordon, 2014). In an experimental harvester-ant example, combined food and forager chemical cues on mimics increased outgoing foraging activity (Greene et al., 2013). Such experiments link a defined intervention to an observed response. They illustrate the level of biological evidence needed for a mechanistic claim, rather than validating any particular lexical label. The corpus network instead records pairs of expressions in documents: its vertices, edges and timescale differ from those of a measured ant-interaction network. Similar graph vocabulary should prompt an explicit comparison of observational units before a mechanism is transferred between fields.

8.7 Positioning This Work

This repository contributes a six-domain descriptive workflow with auditable source layers, inspectable computational definitions, regenerable figures and receipt-bound values. CACE is a proposed evaluation framework, not a validated measure or intervention. Its independent validation would require annotated meanings, blinded judgments, measured agreement, and tests of sensitivity and communication outcomes. The distinction between implemented measurement, theoretical motivation, and unperformed validation is part of the contribution.

8.8 Complexity, Representations, and Dynamics

Simon (Simon, 1962) motivates explicit units, coupling and timescales in hierarchical systems. Newman (Newman, 2003) distinguishes network structure, models and dynamical processes, including clustering induced by affiliation projection. Ladyman and colleagues (Ladyman et al., 2013) compare candidate features and different measures of complexity. Together these sources support precise modeling questions rather than equating a lexical graph with the biological system it describes. The present contribution is an auditable discourse representation and its sensitivity analysis; testing biological dynamics or readers' inferences requires separate observations and designs.

9 Acknowledgments

We gratefully acknowledge the contributions of individuals and institutions that made this research possible.

9.1 Institutional Support

This work was conducted at the Active Inference Institute and APOIDEAS. We thank the Institute for providing the research environment and collaborative infrastructure that supported the development of the Ento-Linguistic framework.

9.2 Collaborations

We thank colleagues and collaborators for valuable discussions and feedback throughout the development of this work, particularly regarding the six-domain framework and the design of the descriptive pipeline.

9.3 Data and Software

This research builds upon open-source software tools and publicly available datasets. We acknowledge:

- Python scientific computing stack (NumPy, SciPy, Matplotlib, NetworkX)
- Natural Language Toolkit (NLTK) for text processing and scikit-learn for TF-IDF vectorization and clustering
- LaTeX and Pandoc for document preparation
- Published entomological literature informing the domain terminology seeds

---

All errors and omissions remain the sole responsibility of the authors.

10 Supplemental Methods: Text Processing and Term Extraction

This section specifies the input and extraction stages used by the study. The numerical definitions and bounded components are detailed in Section 11. Developer APIs remain documented in source; support utilities are not additional research measurements.

10.1 Source Layers and Custody

The archived abstract input contains 7609 strings. The headline analysis includes 7599 strings whose SHA-256 digest maps to an identified PubMed record and excludes 10 unreconciled strings. Exact digest matches to retrieved PubMed abstracts can recover metadata without changing archived text. Identification establishes a source association, not relevance to ant biology, a complete retrieval history, or independent text annotation.

PMC records retain title, abstract, and body text in their own shard directory. BHL records retain historical OCR and era assignments. arXiv title/abstract records form a separate preprint layer. Those documents are not concatenated into the headline abstract input. The custody audit inspects every stored record and reports missing metadata, duplicate text, identifier disagreement, unused sidecar entries, and invalid text.

Digest-indexed sidecars can retain only one source identity when several records share identical text. These collisions are reported rather than resolved by inventing metadata. Broad searches include adjacent biology, computational terminology and mixed-topic historical volumes. The stored PMC layer includes correction and retraction notices whose repeated boilerplate produces shared body text. Records whose titles begin with a correction, erratum or retraction prefix, including retracted full articles, are retained in the shards but excluded from full-text analysis. Individual relevance, article-type screening and license review remain incomplete.

10.2 Normalization and Token Streams

TextProcessor applies Unicode normalization, lowercase conversion, word tokenization, punctuation filtering, and stop-word removal. Hyphenated and underscored tokens can remain intact. The stop vocabulary combines NLTK English stop words with configured scientific meta-language words. WordNet lemmatization is available and is used for corpus vocabulary summaries; terminology extraction disables lemmatization for its occurrence counts. Consequently a summary's most-common lemma and an extracted surface-form frequency need not have the same spelling.

NLTK sentence tokenization supplies the contexts used for semantic entropy. These sentence contexts differ from the short, processed-token windows retained on extracted Term objects. English-language resources are applied to a heterogeneous stored corpus; this is not a validated multilingual processing pipeline.

NLTK resources are separate downloaded prerequisites. The analysis receipt binds the selected English stopword file, English punkt_tab files and WordNet dictionary bytes (the complete archive when archive-backed). The same hashes enter cache signatures; changing a selected resource requires reanalysis. These hashes identify the installed inputs without vendoring them or guaranteeing automatic restoration. Open Multilingual WordNet is outside this receipt because English lemmatization does not consume it. The source implementation and uv.lock specify Python package dependencies.

10.3 Candidate Extraction and Domain Assignment

TerminologyExtractor counts tokens, applies the layer-specific minimum frequency listed in Methods, and evaluates candidate filters.

Candidate filters retain configured seed words of three to fifty characters before evaluating generic filters. Other candidates enter through scientific patterns, compound separators, or configured scientific substrings. They reject pure numbers but can admit unrelated substring matches and OCR artifacts. Candidate inclusion is therefore broader than biological terminology. For example, skin, making and queensland can remain candidates without receiving a domain label.

The six DOMAIN_SEEDS vocabularies define direct assignments. Compound tokens can inherit labels from seed-word overlap; fallback lexical patterns use word boundaries. An extracted term can receive several labels. Label overlap follows from these rules and does not demonstrate temporal semantic drift or independently annotated meanings.

Candidate iteration is sorted for deterministic output. The extractor stores surface text, lemma, frequency, domain labels, confidence, and deduplicated short contexts.

Extraction confidence combines configured frequency, context and classification features. It is not calibrated against human correctness labels. The pipeline does not report precision, recall, multilingual accuracy, or inter-rater reliability.

10.4 Document Co-occurrence and Concept Categories

The observed terminology graph uses the hundred highest-frequency domain-assigned terms, with lexical tie-breaking. Case-insensitive whole-word searches identify presence in each document. A pair's edge weight counts documents containing both terms, regardless of repeated mentions. Shared labels do not create observed edges. The saved graph statistics describe this selected vocabulary rather than the entire field.

The separate concept map uses six predefined categories and vocabulary-association rules. Terms can belong to several categories, so a total of term associations is not a count of unique terms. Category links summarize vocabulary overlap. Node positions, direct-link centrality, colors, and label selection are display conventions and do not establish biological hierarchy or causal dependence.

The overlap heatmap uses the Szymkiewicz–Simpson coefficient in Equation 3.2. Its diagonal is one by construction. A zero observed overlap does not establish conceptual separation.

10.5 Reproducible Inputs and Failure Behavior

DataLoader rejects malformed or empty corpus input, including non-string and blank records. Required pipeline stages propagate failures. Process workers merge results in input order; ENTO_ANALYSIS_WORKERS=1 selects serial execution. Resource-related pool unavailability is logged before a serial retry, while other worker failures propagate.

Caches bind ordered record contents, project Python source, locked dependencies, and explicit development limits. Changed body text invalidates an artifact even if its metadata, record count and file modification time remain unchanged. The completed receipt binds the input/output inventory and image hashes. A receipt establishes which bytes were used, not the scientific validity of the taxonomy or proxies.

11 Supplemental Methods: Statistical and Scoring Infrastructure

The canonical implementations are semantic_entropy, DomainAnalyzer, statistics_pipeline, fulltext_pipeline, arxiv_analysis, and bhl_analysis. Their generated artifacts preserve exclusions, coverage, and exploratory numerical outputs.

11.1 Sentence-context Entropy

Each source text is sentence-tokenized once for a term set. An inverted index over alphanumeric parts prunes impossible matches; a case-insensitive whole-word regular expression makes the final decision. This preserves matching sentence order. Contexts must contain more than three words, and at least five usable contexts are required for computation.

TF-IDF uses English stop words, minimum document frequency one, and at most one thousand features. KMeans uses random state 42 and ten initializations. The requested cluster count follows the bounded square-root rule in Methods, remaining below the number of contexts. Occupied clusters may be fewer than requested.

Shannon entropy is computed from occupied-cluster probabilities in bits (Equation 3.1). The normalized value divides by the occupied-cluster ceiling; a single occupied cluster has zero normalized entropy. The exploratory high-entropy flag uses a threshold above two bits. Computational partitions are not independently annotated word senses.

Results record ok, insufficient_contexts, or error status. Only ok values enter domain means and statistical groups. Excluded estimates are counted separately; their placeholder zero is not included as a measured value. A term assigned to several domains contributes to each domain group. Entropy-table sample sizes and weighted totals count valid domain memberships, not unique terms or documents.

11.2 Exploratory Comparisons and Intervals

The statistical artifact reports per-domain valid-entropy means, standard deviations where at least two values exist, and exclusions. Welch tests compare groups with at least two values. Benjamini–Hochberg adjustments apply to the emitted comparison family at a nominal threshold of 0.05. Standardized differences use a small-sample bias correction; they should not be described as uncorrected Cohen's d.

One-way ANOVA and its effect-size summary use the valid groups. Figure intervals use nominal Student-t quantiles and are omitted for groups with only one valid estimate. These numerical calculations have software tests and independent numerical comparators, but their simple population interpretation is not justified by this corpus design.

Terms can share documents, contexts and domain labels. Multiplicity adjustment does not repair that dependence, convenience retrieval, unequal context counts, or differing extraction thresholds. Threshold flags are exploratory. Confirmatory inference requires a prespecified document-level model and independently annotated evaluation data.

11.3 Shared CACE Sample and Scores

Domain figures and statistical exports use the same scoring function and sample: up to fifty terms per domain, ordered by decreasing frequency with lexical tie-breaking. Each term supplies at most ten of its stored short extraction contexts. The representative-term table is a separate named-term evaluation whose artifact records whether a term was extracted.

Clarity uses attached sentence-context entropy when available. Appropriateness uses configured vocabulary and domain-overlap penalties. Consistency uses TF-IDF context-vector cosine similarities. Evolvability uses assigned domains and scale-marker categories. Their equally weighted mean is the aggregate; Equations 3.3–3.6 specify the proposed rules.

Unavailable entropy still defaults to zero inside the heuristic, and fewer than two contexts yield a Consistency convention of 0.5. These conventions do not establish clarity or consistency. Missing figure input is labeled unavailable instead of substituting confidence as a CACE score. No completed coefficient-sensitivity study, human-rating validation, intervention benefit, or calibrated scientific-accuracy prediction is claimed.

11.4 Framing and Discourse Proxies

Occurrence-context framing uses three-token windows around domain-assigned terms. The feature extractor has four anthropomorphic, four hierarchical, and four economic regex patterns. Marker matches are lexical events, not validated author intent or judgments that a biological description is misleading.

Domain framing proportions count occurrence contexts matching an anthropomorphic pattern. Multi-domain occurrences contribute to each domain but once to the overall tally. Distinct framing-marked term counts in the figure are a different summary from those occurrence proportions.

Discourse, rhetorical, and persuasive analyzers use configured patterns and keywords. Their artifacts record eligible/analyzed texts and exclusions. PMC discourse uses a deterministic stride sample when corpus size exceeds the configured text/character targets, with an approximately one-fifth sample in this snapshot. Raw counts depend on text length and sample size; the cross-layer figure is not a normalized prevalence comparison or an annotated discourse study.

11.5 Historical OCR and Layer Coverage

BHL literal frequencies normalize NFC text, lowercase it, join hyphenated line breaks, and tokenize Latin-letter words while retaining internal hyphens. Seed phrases match exact consecutive tokens; counts are non-overlapping per phrase. Rates divide by era token counts and multiply by ten thousand. Alternative spellings, OCR errors and language affect those rates. An observed zero does not date a concept's origin.

The BHL computational stack uses cleaned OCR with the canonical extraction and classification rules. Default extraction and framing process all stored era documents using per-document counters, avoiding a corpus-wide token-position map. Stored and cleaned text are still held in memory; memory use is not constant in corpus size. Entropy evaluates twenty frequent candidates per era with sentence contexts drawn from all era documents. Unassigned candidates can enter the overall sampled entropy mean, while domain means only use assigned valid candidates.

Coverage records available and analyzed documents and characters. An explicit positive `BHL_STACK_CHARACTER_BUDGET` selects a deterministic whole-document development subset. `FULLTEXT_ANALYSIS_LIMIT` similarly bounds a development PMC run. Both limits enter fingerprints; bounded caches cannot satisfy an unbounded default run.

11.6 Execution and Artifact Verification

The standalone sequence installs locked dependencies and NLTK resources, runs tests, generates figures, checks custody, and renders the manuscript. Testing and generation run sequentially because some integration tests consume generated exports.

The analysis signature also hashes the selected English NLTK tokenizer, stopwords and WordNet contents. The

receipt records these resource hashes without machine-specific installation paths. Missing resources fail, and changed resources invalidate caches. These content hashes do not vendor or automatically restore external resource downloads.

Reported values are injected from receipted artifacts at render time. The generator checks finite/nonempty JSON, decodable PNGs, registry hashes, figure inventory and manuscript figure availability before writing its receipt. Rendering validates the receipt, rejects unresolved placeholders, and requires successful Pandoc, XeLaTeX and BibTeX execution. Undefined citations/references and missing glyphs fail the final build.

Software tests use actual corpus slices, numerical examples, real files, local HTTP and real subprocesses. Negative controls establish that malformed input, stale artifacts and failing commands are rejected. Those checks establish software behavior and provenance, not source relevance, individual reuse rights, human-validation outcomes or causal scientific conclusions.

Completed BHL eras are saved atomically in local recovery checkpoints. Each checkpoint binds the ordered era records, implementation/dependency/resource signature and any development bound, plus a digest of the completed result. A matching checkpoint can resume a disrupted run; changed inputs or bounds require recomputation, and corrupted matching results fail. The final four-layer receipt is still written only after all required stages complete.

11.7 Fixed-Margin Network Robustness Extension

A separately receipted companion workflow reconstructs the abstract terminology network from the frozen identified records and the same frequency-ranked vocabulary. It randomizes the binary document–term incidence matrix with Curveball trades (Carstens et al., 2018, Strona et al., 2014). A trade retains common terms in both selected documents and uniformly repartitions their exclusive terms while preserving both row sizes. Consequently, each term retains its document frequency and each document retains its number of selected terms. Self-loop and other no-op transitions remain part of the chain. The total weight of the projected graph equals the sum of the number of term pairs in each document; this quantity is an invariant of the conditioning margins.

Every retained draw checks the two margins and the weight invariant. The workflow reports projected edge count, mean local unweighted clustering, retained-chain traces, between-chain diagnostics and empirical conditional distributions. It also evaluates observed projections under smaller vocabulary subsets of the same ranking. A second protocol uses longer burn-in and spacing with different seeds. These checks assess structural interpretation and sampling sensitivity. Finite correlated draws and between-chain agreement do not prove convergence to the stationary distribution. Conditional null envelopes describe the sampled reference and are distinct from population confidence intervals.

The companion implementation is in research/network_robustness/. Its receipt binds the published-core input receipt, exact source inputs, extension implementation, dependency lock and generated outputs. It does not relabel cached core results as newly recomputed analyses. The executed report and trace array are provided as companion artifacts.

A validation guard additionally decodes retained numerical traces and the companion PNG, verifies full protocol dimensions, finite values and graph bounds, reconstructs the observed graph from source-identified abstracts, and recomputes the JSON summary statistics from all retained draws. The companion report's result table must agree with those summaries, and the values reported in the Results section are injected from the same summaries. Negative controls include self-hashed corrupt files and inconsistent summaries. These internal-consistency checks do not establish chain mixing, validate lexical proxies, or certify the visual interpretation of a decoded figure.

12 Supplemental Results

12.1 Exploratory Pairwise Domain Comparisons

Table 2 presents pairwise comparisons of per-term semantic entropy between all Ento-Linguistic domains using Welch's two-sample t-tests. Raw p-values are computed from the t-distribution with Satterthwaite-approximated degrees of freedom; adjusted p-values correct for 15 simultaneous comparisons using the Benjamini-Hochberg (BH) procedure at $q = 0.05$. The effect size d is a small-sample bias-corrected (Hedges-type) standardized difference. Conventional benchmarks of about 0.2, 0.5 and 0.8 (Cohen, 1988) are rules of thumb that are not calibrated for these overlapping, dependent groups. Domain descriptives entering these tests are per-term valid-entropy means with exclusions counted.

Domain A	**Domain B**	t	p **(raw)**	p **(BH)**	**Std. diff.** d	**Sig. (BH)**
Behavior & Identity	Economics	-0.9801	0.3312	0.4516	-0.2166	no
Behavior & Identity	Kin & Relatedness	-0.1601	0.8736	0.8736	-0.0402	no
Behavior & Identity	Power & Labor	1.2093	0.2289	0.4516	0.2101	no
Behavior & Identity	Sex & Reproduction	0.1864	0.8525	0.8736	0.0337	no
Behavior & Identity	Unit of Individuality	2.3743	0.0192	0.1442	0.3793	no
Economics	Kin & Relatedness	0.6153	0.5417	0.6772	0.1781	no
Economics	Power & Labor	2.0805	0.0423	0.1586	0.4216	no
Economics	Sex & Reproduction	1.1366	0.2603	0.4516	0.2350	no
Economics	Unit of Individuality	3.1260	0.0031	0.0464	0.5875	yes
Kin & Relatedness	Power & Labor	1.0421	0.3045	0.4516	0.2457	no
Kin & Relatedness	Sex & Reproduction	0.2982	0.7672	0.8736	0.0704	no
Kin & Relatedness	Unit of Individuality	1.8320	0.0767	0.2300	0.4127	no
Power & Labor	Sex & Reproduction	-0.9976	0.3203	0.4516	-0.1689	no
Power & Labor	Unit of Individuality	1.1416	0.2553	0.4516	0.1687	no
Sex & Reproduction	Unit of Individuality	2.1273	0.0353	0.1586	0.3348	no

Table 2. Exploratory pairwise Welch tests on valid per-term sentence-context entropy, with Benjamini–Hochberg adjusted p-values over 15 comparisons and standardized effect sizes. Threshold flags use $q = 0.05$; domain groups share terms and contexts, so the flags are exploratory (Supplemental Section 11). The omnibus ANOVA yields $F(5, 350) = 2.4761$, p-value 0.0319, and $\eta^2 = 0.0342$.

12.2 CACE Scoring for Key Terms

Table 3 presents full CACE evaluations for a representative set of entomological terms, comparing anthropomorphic labels with proposed functional alternatives.

12.3 Semantic Entropy Distribution

Table 4 summarizes the distribution of semantic entropy across domains.

12.4 Interval Reporting

The statistics artifact reports per-term valid-entropy descriptives (means and standard deviations, with exclusions counted) and the inferential results in Table 2; it does not compute domain-level confidence intervals, so per-domain ambiguity-score and context-variability intervals are not tabulated here. Exploratory numerical differences are summarized by the Welch t-tests and the omnibus ANOVA reported in Table 2, with per-domain entropy descriptives in Table 4 and the summary in Figure 15.

Term	C	A	Cs	E	Mean	Extracted
queen	0.48	0.45	0.04	0.33	0.33	yes
primary reproductive	1.00	1.00	0.50	0.00	0.62	no
worker	0.42	0.45	0.05	0.33	0.31	yes
non-reproductive helper	1.00	1.00	0.50	0.00	0.62	no
slave	0.31	0.50	0.05	0.33	0.30	yes
host worker	1.00	0.40	0.50	0.00	0.47	no
caste	0.36	1.00	0.06	0.33	0.44	yes
task group	1.00	1.00	0.50	0.00	0.62	no
soldier	0.35	0.50	0.10	0.17	0.28	yes
major worker	1.00	0.50	0.50	0.00	0.50	no
colony	0.47	1.00	0.03	0.33	0.46	yes
haplodiploidy	0.31	1.00	0.05	0.17	0.38	yes
trophallaxis	0.33	1.00	0.06	0.17	0.39	yes

Table 3. Heuristic representative-term CACE scores: C is Clarity, A Appropriateness, Cs Consistency, and E Evolvability; Mean weights them equally. Extracted indicates membership in the extracted vocabulary, not phrase presence anywhere in the source text. Unextracted alternatives use fallback conventions. Defaults and vocabulary penalties are not independently measured clarity, biological accuracy, or evidence of terminology-reform benefit.

Domain	Mean H (bits)	High-entropy terms (%)	N
Economics	1.85	46.2	26
Power & Labor	1.64	31.6	76
Behavior & Identity	1.74	44.6	56
Sex & Reproduction	1.73	43.8	64
Unit of Individuality	1.55	26.8	112
Kin & Relatedness	1.76	45.5	22
Overall	1.67	36.2	**356**

Table 4. Distribution of sentence-context entropy $H(t)$ across Ento-Linguistic domains (Eq. 3.1; cluster-count rule in Methods). High-entropy terms exceed $H > 2.0$ bits; the statistic summarizes occupancy of computational clusters, not annotated senses. Normalized entropy divides by $\log_2 k_{\text{occupied}}$ and is zero when one cluster is occupied. N counts terms with usable estimates; Overall sums valid domain memberships, so a multi-domain term can count more than once, and its mean and high-entropy percentage use that same denominator.

12.5 Full-Text Parallel Layer

A complementary descriptive analysis was run over 6997 open access full texts harvested from PubMed Central (PMC), after excluding 76 correction, erratum and retraction notices and retracted items from 7073 stored records. The abstract corpus remains the headline corpus of this study; the full-text layer is reported here as a separate convenience sample with a higher extraction threshold. The analysis machinery is shared: the same terminology extraction, domain assignment, semantic-entropy, and CACE scoring implementations are applied to full texts (Figure 16).

The layer comprises 44,476,595 tokens of running text, with a per-document median of 5651 tokens. Table 5 reports per-domain term counts and mean semantic entropy over the full texts; 15 pairwise Welch t-tests (Benjamini-Hochberg corrected, as in Table 2) accompany the omnibus one-way ANOVA on per-term semantic entropy, $F = 7.7409$, p-value <0.0001.

Anthropomorphic framing over the same full texts is scored as the proportion of domain-term occurrence contexts containing an anthropomorphic framing marker, using the same patterns as the abstract layer: 0.0107 overall, with

Economics at 0.0359 against 0.0074 for Unit of Individuality and 0.0078 for Behavior & Identity.

Domain	**Terms extracted**	**Mean H (bits)**
Behavior & Identity	96	1.9700
Economics	35	1.7970
Kin & Relatedness	24	1.9878
Power & Labor	118	2.0102
Sex & Reproduction	86	2.0381
Unit of Individuality	241	2.0274

Table 5. Per-domain terminology in the PMC full-text parallel layer: extracted-term counts and mean semantic entropy $H(t)$, computed over valid estimates with the same pipeline as the abstract layer; valid-estimate counts appear in Figure 16. Term extraction uses a higher minimum token frequency than the abstract layer because full texts are substantially longer.

12.6 Discourse and Rhetorical Layer

A shared discourse stage counts rule-based markers of discourse patterns, rhetorical strategies, argumentative structures, and persuasive techniques in both layers. Texts shorter than 200 characters are excluded. The abstract pass therefore covers 7583 texts (16 excluded as too short; 100.0% of the eligible texts). The full-text pass covers 1400 of 6997 texts, a deterministic 20.0% sample of eligible texts drawn to keep the pass computationally bounded. Figure 18 compares the two layers; panels spanning orders of magnitude use a symlog frequency axis.

Discourse patterns occur 1278 times as hierarchical framing in the abstract layer against 1098 occurrences in the full-text layer, with 1045 versus 1088 economic metaphors, 32 versus 52 anthropomorphic framings, and 27 versus 22 scale-ambiguous constructions.

Rhetorical-strategy marker counts in the two layers are: 1578 anecdotal markers in the abstract layer versus 9446 in full texts, 246 versus 30,330 authority markers, 456 versus 1944 analogies, and 1148 versus 4204 generalizations.

The argumentative-structure pass identifies 1803 argumentative structures in the abstract layer versus 1292 in the full-text sample. Metaphorical-language markers occur 3830 times in the abstract layer against 21,140 occurrences in full texts. All frequencies are raw marker counts. The layers differ in size and sampling, so these counts are not prevalence comparisons.

12.7 Statistical and Source-Layer Figures

Figures 15–19 summarize the headline statistics, the PMC full-text layer, the cross-layer comparison, the discourse markers and the separate arXiv preprint layer.

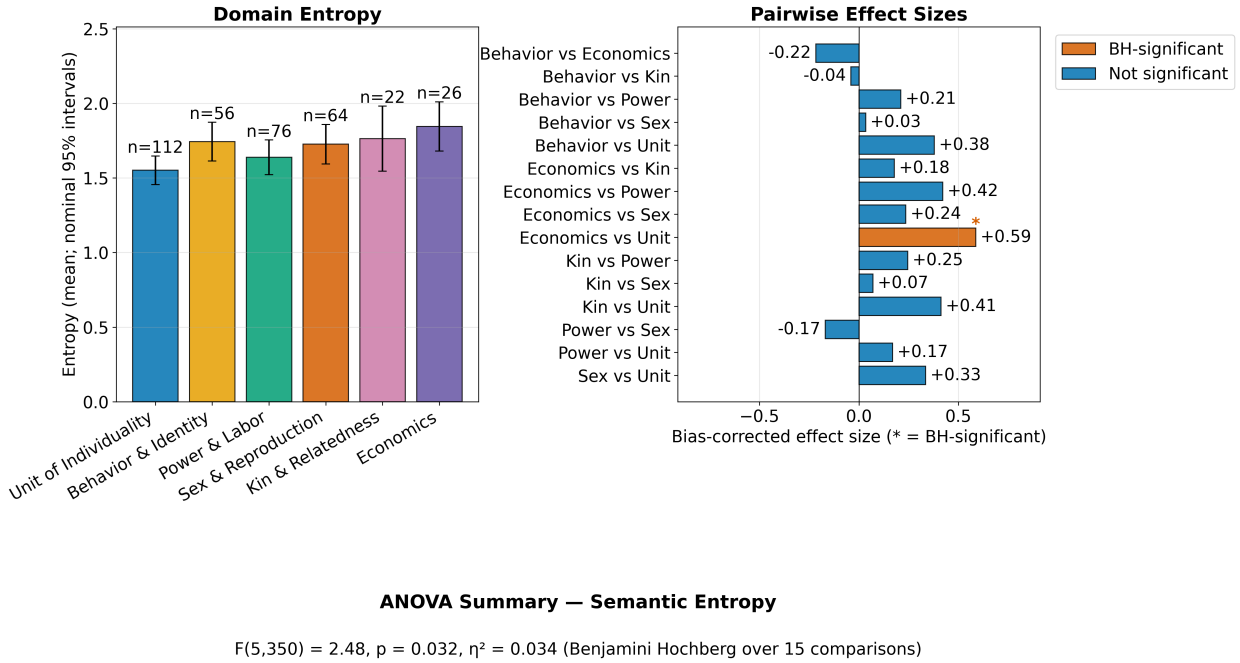


Figure 15. Headline valid-entropy descriptives with nominal intervals, bias-corrected standardized differences, and exploratory ANOVA. Bar annotations report valid term counts; overlapping domain memberships and shared document contexts limit population inference.

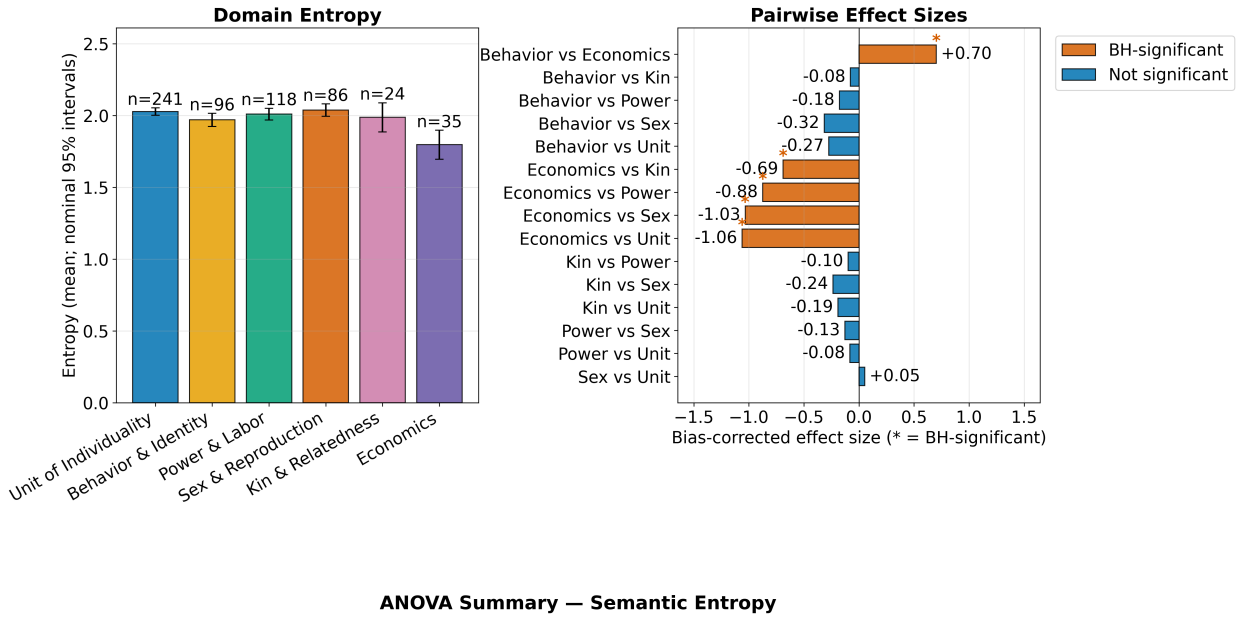


Figure 16. Separate PMC full-text statistics over all stored records by default. Extraction thresholds and context distributions differ from the abstract layer. Nominal intervals and multiplicity-adjusted threshold flags remain exploratory.

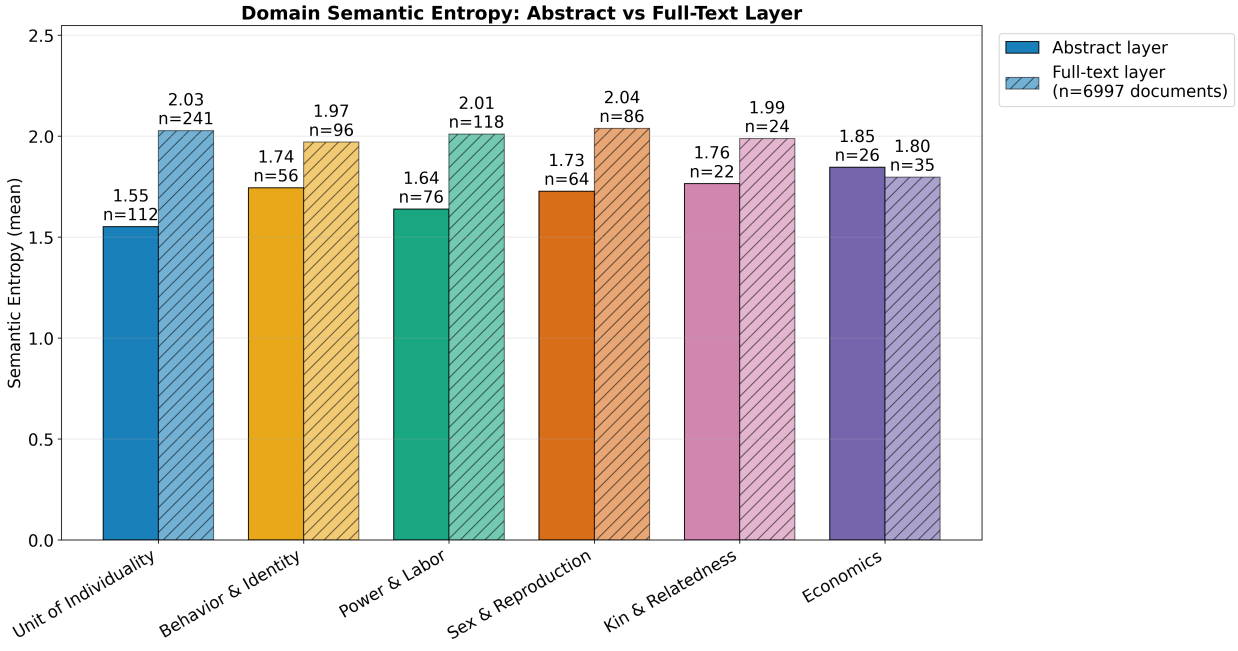


Figure 17. Mean valid context-cluster entropy in the abstract and PMC layers, with valid-term counts. This compares distinct convenience samples and extraction thresholds; it is not a matched robustness experiment or evidence of causal language effects.

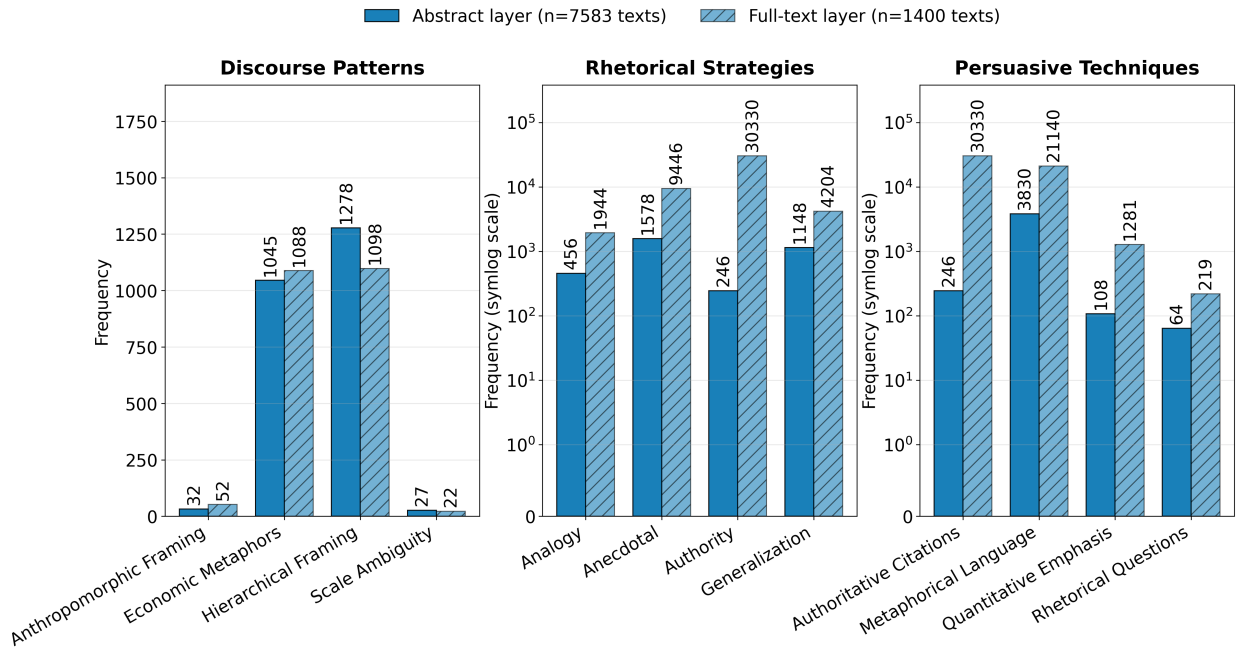
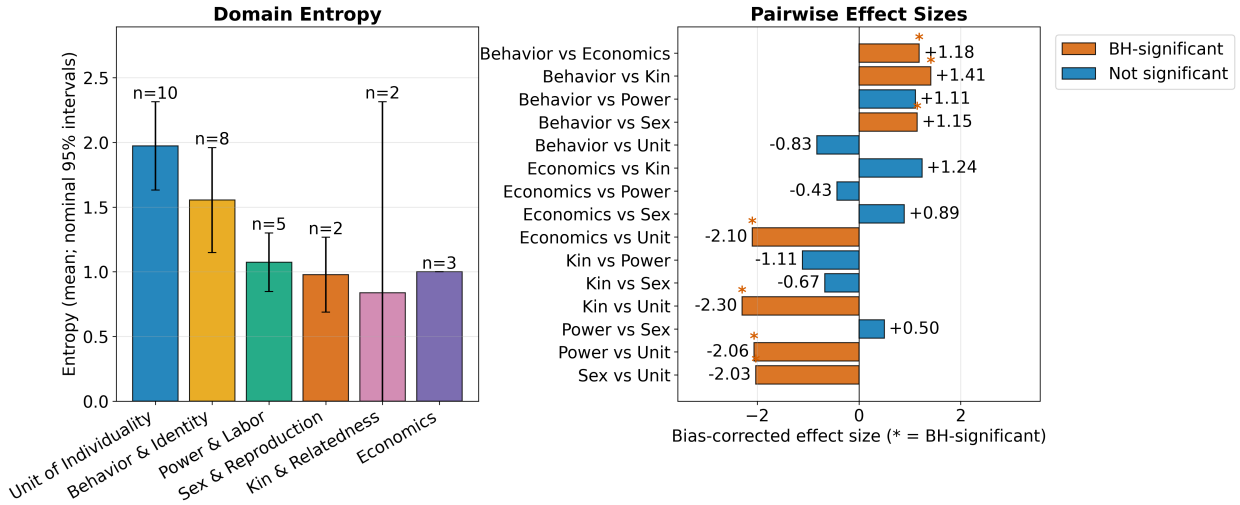


Figure 18. Lexical discourse, rhetorical and persuasive-pattern counts, with analyzed-text counts shown in the legend. PMC uses an explicitly bounded text sample. Raw frequencies depend on sample size and text length and must not be read as normalized prevalence or human-validated author intent.



ANOVA Summary — Semantic Entropy

$F(5,24) = 6.60$, $p < 0.001$, $\eta^2 = 0.579$ (Benjamini Hochberg over 15 comparisons)

Figure 19. Separate arXiv preprint-layer descriptives and exploratory comparisons. Some groups have only one valid entropy estimate, for which intervals are omitted; very small groups and nearly zero within-group variance can yield large standardized differences without establishing generalizable effects.

13 Supplemental Analysis: Theoretical Extensions

These are proposed constructions, not additional experimental results. The repository does not fit a generative model of scientific language, measure variational free energy, estimate empirical Markov blankets, or test a causal terminology intervention.

13.1 Individuality and Conditional Independence

A Markov blanket specifies a conditional-independence relation between internal variables μ and external variables η, conditional on variables B (Friston, 2013, Kirchhoff et al., 2018):

$$P(\mu \mid \eta, B) = P(\mu \mid B). \tag{13.1}$$

This relation is an assumption or a property to establish for a specified probability model, on the support where the conditional distributions are defined. A biological application requires named variables, dynamics, observations, and tests of the proposed conditional independence. A sensory/active decomposition requires further modeling assumptions.

An ant's sensory and motor variables, or colony-level interaction variables, could enter candidate models at different scales. Neither the cuticle, a nest entrance, nor a pheromone field automatically establishes a blanket. Likewise, calling a colony a superorganism is not equivalent to proving this factorization. The Active Inferants simulation provides related modeling context (Friedman et al., 2021), without validating a blanket inferred from terminology in this study.

13.2 A Proposed Framing Score

A future annotated study could define a bounded score:

$$F_{\text{frame}}(D, T) = \sum_{t \in T} w_t f_t(D) c_t, \quad w_t \geq 0, \quad \sum_{t \in T} w_t = 1. \tag{13.2}$$

Here $f_t(D)$ and c_t would be explicitly defined annotation-based quantities in $[0, 1]$, giving a score in $[0, 1]$ for nonempty T. Empty vocabularies would yield an undefined score, not an inferred absence of framing. The weights require a prespecified rationale and sensitivity analysis. This proposed quantity is distinct from variational free energy and is not an output of the current lexical-pattern pipeline.

13.3 Ambiguity Annotation

Lexical, contextual, biological-scale, and temporal ambiguity are candidate annotation categories. The implemented entropy measure summarizes context-cluster occupancy; it does not independently assign or validate these four categories. An annotation protocol should specify meaning distinctions, uncertainty, annotator training, and agreement measurement. Changes in context counts or source composition need separate controls.

13.4 Cross-Domain Mapping

Given an independently specified similarity score $s(t, D_i, D_j)$, an annotated cross-domain summary could be defined as:

$$M_{ij} = \begin{cases} |T_i \cap T_j|^{-1} \displaystyle\sum_{t \in T_i \cap T_j} s(t, D_i, D_j), & |T_i \cap T_j| > 0, \\ \text{undefined}, & |T_i \cap T_j| = 0. \end{cases} \tag{13.3}$$

This is a proposed mean over shared terms. It is not the implemented domain overlap coefficient, and an empty intersection does not establish that two biological concepts are unrelated. Similarity and domain membership require explicit definitions before interpretation.

13.5 Temporal Networks

A time-indexed analysis should first specify a common vertex registry, matching vocabulary, and sampling rules. Let $A(t)$ be weighted adjacency matrices over that registry, with zero entries for absent edges. Then an algebraically defined change is:

$$\Delta A(t) = A(t+1) - A(t). \tag{13.4}$$

Changing matrix entries can reflect document availability, vocabulary selection, source genre, or retrieval as well as changes in use. Demonstrating a conceptual shift requires controls for these alternatives and dated meaning annotation. The current BHL literal-frequency analysis does not implement this longitudinal network model.

13.6 Multiscale Interpretation

Term, domain, and cross-domain summaries can organize different observational scales. Claims about biological organization or scientific thought require additional measurements at the corresponding scale. Environment-Centric Active Inference and related interpretations remain hypotheses requiring specified models and empirical tests, rather than conclusions established by the present corpus.

14 Supplemental Analysis: Case Studies and Validation

14.1 Validation Agenda

Expert classification review, interdisciplinary assessment, historical interpretation, and cross-cultural comparison are proposed validation activities. This repository does not report completed human-annotation agreement, inter-rater reliability, multilingual validation, or a prespecified subsampling and coefficient-sensitivity study. Software tests and descriptive source-layer comparisons do not substitute for those measurements.

14.2 Historical Terminology Analysis

The historical layer contains 2430 stored BHL mirror OCR documents dated into three era buckets: 996 in 1850–1899, 1237 in 1900–1949, and 197 in 1950–1970. Retrieval selects volumes containing search expressions, not a manually curated collection of ant-only works. Document titles, dates, source identifiers, queries, and retrieval information are recorded in the source sidecar.

Term	1850–1899	1900–1949	1950–1970
caste	0.0735	0.1140	0.2253
worker	0.6394	0.7935	1.4169
queen	0.8861	1.0770	1.0064
colony	0.9283	1.1689	1.1047
superorganism	0.0003	0.0007	0.0051

Table 6. Complete-corpus literal OCR matches per 10,000 era tokens. Differences are descriptive of the selected volumes. OCR quality, topic, language, and spelling differences limit historical interpretation.

Rare matches for superorganism cannot establish that the concept originated after 1970. Alternative spellings such as super-organism are counted differently, and earlier theoretical discussion can exist outside the sampled sources. Interpretation should be anchored in dated sources such as Wheeler (1911), rather than treating an OCR absence as historical proof.

14.3 Complete Document Coverage and Bounded Entropy

The BHL extraction/entropy/framing stack uses 996, 1237, and 197 complete documents, respectively. Default extraction and framing use all era documents. Available and analyzed characters are recorded in the artifact's coverage metadata. Entropy considers twenty frequent extracted candidates per era using contexts across all documents; it is a candidate-term sample rather than an exhaustive entropy census. Candidate filtering can retain unrelated substring matches such as skin, making, or queensland; unassigned candidates enter the overall sampled entropy mean but not domain means. These are computational filter outputs, not a curated historical vocabulary. An optional development character budget produces a disclosed, separately fingerprinted document subset. Neither default storage nor that subset establishes historical representativeness.

Developmental evidence for canalized caste differentiation (Qiu et al., 2022) concerns biological mechanisms and does not establish a historical linguistic trend. Evaluating that connection would require a dated, annotated corpus with source-composition controls.

14.4 A Reproducible Case-Study Protocol

A subsequent study can select a specific term, reconcile each text to its source, annotate meaning and biological referent in dated contexts, and preregister comparisons between research traditions or eras. Controls should

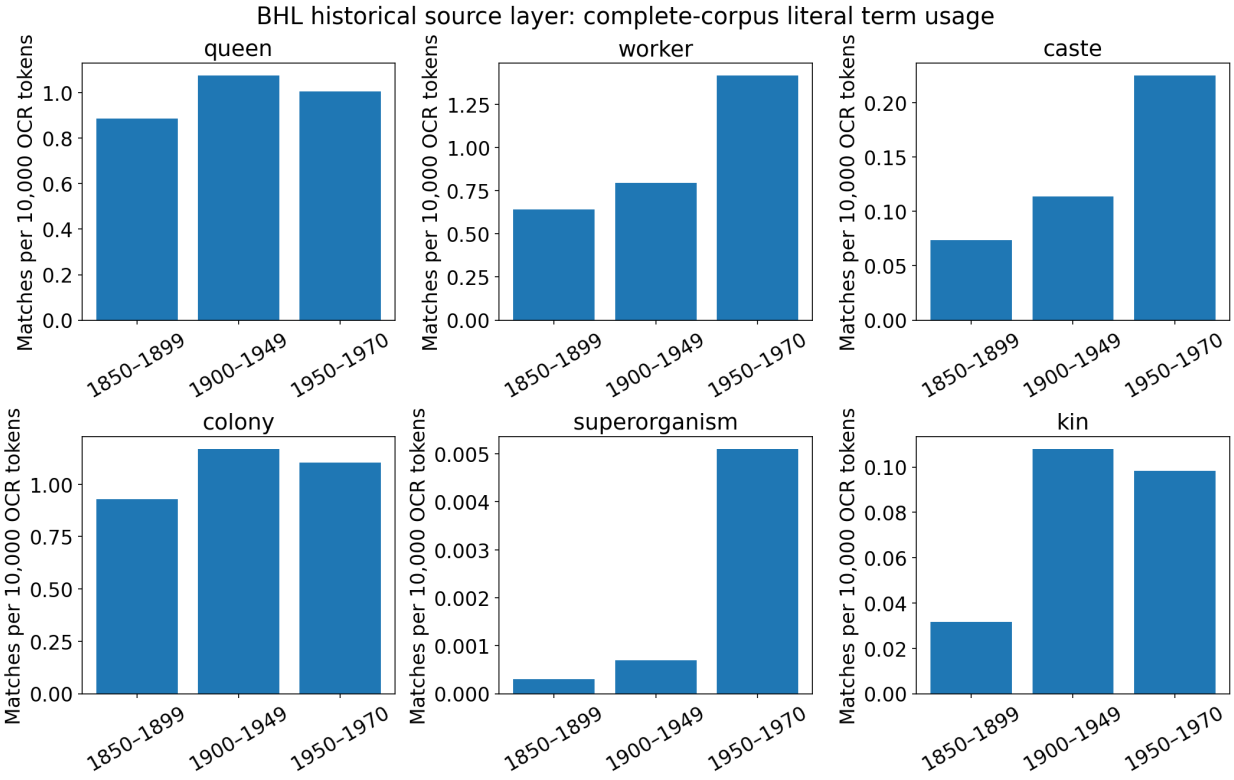


Figure 20. Full-corpus historical literal frequencies for six seed terms. Each panel uses its own rate axis. Neither zeros nor changing rates establish conceptual origin, prevalence in all entomological literature, or causal influence on research.

distinguish genuine changes in use from retrieval, OCR, spelling, and genre changes. CACE judgments should be gathered independently of the automatic scoring vocabulary, with agreement and uncertainty reported.

15 Symbols and Notation Glossary

This glossary defines the mathematical notation and domain-specific terminology used throughout the manuscript.

15.1 Mathematical Notation

Symbol	Description	First Use
$G = (V, E)$	Terminology network (graph with vertices and edges)	Eq. 4.1
$D(t)$	Set of Ento-Linguistic domains term t is assigned to	Eq. 3.4
$w(u, v)$	Observed number of documents containing both terms u and v	Eq. 4.1
$H(t)$	Semantic entropy of term t in bits (Shannon entropy over usage-context clusters)	Eq. 3.1
H^*	Configured threshold (2.0 bits); not an independently calibrated sense threshold	Eq. 3.1
$H_{\max}$	Entropy ceiling for occupied clusters, $\log_2 k_{\text{occupied}}$; zero for a single occupied cluster	Eq. 3.1
$\hat{H}(t)$	Normalized entropy $H(t)/H_{\max}$ when $H_{\max} > 0$; zero for one occupied cluster	Eq. 3.1
p_i	Empirical proportion of contexts assigned to semantic cluster i	Eq. 3.1
k	Requested number of computational clusters (k-means, $k = \max(2,\ \min(k_{\max},\ n{-}1,\ \max(2, \lfloor\sqrt{n}\rfloor)))$; $k_{\max} = 5$, $n = \lvert C_t\rvert$; $k < n$)	Eq. 3.1
C_t	Set of valid usage contexts of term t (sentences with more than three words)	Eq. 3.1
S_t	Set of biological scale levels expressed in term t's contexts	Eq. 3.6
w_{AB}	Overlap coefficient (Szymkiewicz–Simpson) between concept sets A and B	Eq. 3.2
Clarity(t)	CACE Clarity score: $\min(1, \max(0, 1 - H(t)/3.32))$	Eq. 3.3
Appropriateness(t)	CACE Appropriateness score (penalizes anthropomorphic terms)	Eq. 3.4
Consistency(t)	CACE Consistency score: mean pairwise cosine similarity of context vectors	Eq. 3.5
Evolvability(t)	CACE Evolvability score: mean of heuristic domain-breadth and scale-marker components	Eq. 3.6
$\mathcal{A}$	Set of anthropomorphic terms (queen, king, slave, worker, soldier, nurse, …)	Eq. 3.4

Symbol	Description	First Use
$F_{\text{frame}}(D, T)$	Proposed bounded annotated framing score for domain D and term set T	Supplemental Eq. 13.2
M_{ij}	Cross-domain mapping strength between domains D_i and D_j	Supplemental Eq. 13.3
$\Delta A(t)$	Change in adjacency matrices on a common vertex registry	Supplemental Eq. 13.4
B	Markov Blanket boundary of a system	Supplemental Eq. 13.1
μ	Internal states (conditionally independent of external given blanket)	Supplemental Eq. 13.1
η	External states	Supplemental Eq. 13.1

15.2 Theoretical Terms

Term	Definition	Context
Active Inference	A corollary of the Free Energy Principle stating that agents act to fulfill the predictions of their generative models.	Sec. 2
CACE	Clarity, Appropriateness, Consistency, Evolvability — four-dimensional meta-standard for evaluating scientific terminology.	Sec. 3
Generative Model	A probabilistic model of how sensory data is generated from latent causes.	Sec. 6
Markov Blanket	Variables rendering specified internal and external variables conditionally independent in a probability model.	Sec. 13
Semantic Entropy	Shannon entropy $H(t)$ over the cluster distribution of a term's usage contexts; describes computational cluster occupancy without independently validated senses.	Sec. 3
Superorganism	A colony-level organismic concept; its biological interpretation is not established by a term's frequency or by an assumed blanket.	Sec. 2; Sec. 4
Variational Free Energy	A variational bound on negative log model evidence under a specified probabilistic model; not measured by this corpus pipeline.	Sec. 6

15.3 Pipeline Modules

Module	File	Function
Text Processing	`text_analysis.py`	Tokenization, normalization, feature extraction
Term Extraction	`term_extraction.py`	Domain-aware terminology identification

Module	File	Function
Semantic Entropy	`semantic_entropy.py`	Per-term $H(t)$ computation via TF-IDF + k-means
CACE Scoring	`cace_scoring.py`	Four-dimensional terminology evaluation
Domain Analysis	`domain_analysis.py`	Per-domain framing and ambiguity analysis
Conceptual Mapping	`conceptual_mapping.py`	Cross-domain concept graph construction
Rhetorical Analysis	`rhetorical_analysis.py`	Framing detection and argumentative scoring
Discourse Analysis	`discourse_analysis.py`	Discourse pattern classification
Statistics	`statistics.py`	Statistical validation utilities
Visualization	`concept_visualization.py`	Network and domain-specific figure generation

16 References

Brent Berlin. *Ethnobiological Classification: Principles of Categorization of Plants and Animals in Traditional Societies*. Princeton University Press, Princeton, NJ, 1992.

Jacobus J. Boomsma and Richard Gawne. Superorganismality and caste differentiation as points of no return: how the major evolutionary transitions were lost in translation. *Biological Reviews*, 93(1):28–54, 2018. doi:10.1111/brv.12330.

Corrie Jacobien Carstens, Annabell Berger, and Giovanni Strona. A unifying framework for fast randomization of ecological networks with fixed (node) degrees. *MethodsX*, 5:773–780, 2018. doi:10.1016/j.mex.2018.06.018.

Chaomei Chen. CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. *Journal of the American Society for Information Science and Technology*, 57(3):359–377, 2006. doi:10.1002/asi.20317.

Andy Clark. Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, 36(3):181–204, 2013.

Jacob Cohen. *Statistical Power Analysis for the Behavioral Sciences*. Lawrence Erlbaum Associates, Hillsdale, NJ, 2nd edition, 1988. ISBN 978-0805802832.

Iain D. Couzin. Collective cognition in animal groups. *Trends in Cognitive Sciences*, 13(1):36–43, 2009. doi:10.1016/j.tics.2008.10.002.

Entomological Society of America. Better common names: Addressing racist, colonial, and otherwise problematic common names. ESA Common Names, 2024. URL https://entsoc.org/better-common-names.

Norman Fairclough. *Discourse and Social Change*. Polity Press, Cambridge, UK, 1992.

Ofer Feinerman and Amos Korman. Individual versus collective cognition in social insects. *Journal of Experimental Biology*, 220(1):73–82, 2017. doi:10.1242/jeb.143891.

Daniel Ari Friedman, Alec Tschantz, Maxwell J. D. Ramstead, Karl Friston, and Axel Constant. Active Inferants: An Active Inference framework for ant colony behavior. *Frontiers in Behavioral Neuroscience*, 15:647732, 2021. doi:10.3389/fnbeh.2021.647732.

Karl Friston. The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience*, 11(2):127–138, 2010.

Karl Friston. Life as we know it. *Journal of the Royal Society Interface*, 10(86):20130475, 2013.

Deborah M. Gordon. Wittgenstein and ant-watching. *Biology and Philosophy*, 7(1):13–25, 1992. doi:10.1007/BF00130161.

Deborah M. Gordon. *Ant Encounters: Interaction Networks and Colony Behavior*. Princeton University Press, Princeton, NJ, 2010. ISBN 978-0691138794.

Deborah M. Gordon. The ecology of collective behavior. *PLOS Biology*, 12(3):e1001805, 2014. doi:10.1371/journal.pbio.1001805.

Deborah M. Gordon. The ecology of collective behavior in ants. *Annual Review of Entomology*, 64:35–52, 2019. doi:10.1146/annurev-ento-011118-111923.

Deborah M. Gordon. *The Ecology of Collective Behavior*. Princeton University Press, Princeton, NJ, 2023. ISBN 978-0691232157.

Michael J. Greene, Noa Pinter-Wollman, and Deborah M. Gordon. Interactions with combined chemical cues inform harvester ant foragers' decisions to leave the nest in search of food. *PLOS ONE*, 8(1):e52219, 2013. doi:10.1371/journal.pone.0052219.

Justin Grimmer and Brandon M. Stewart. Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, 21(3):267–297, 2013. doi:10.1093/pan/mps028.

Ian Hacking. *The Social Construction of What?* Harvard University Press, Cambridge, MA, 1999. ISBN 978-0674004122.

Joan M. Herbers. The loaded language of science. *The Chronicle of Higher Education*, 52(41):B13, 2006.

Joan M. Herbers. Watch your language! Racially loaded metaphors in scientific research. *BioScience*, 57(2): 104–105, 2007. doi:10.1641/B570203.

Michael Kirchhoff, Thomas Parr, Ensor Palacios, Karl Friston, and Julian Kiverstein. The Markov blankets of life: autonomy, active inference and the free energy principle. *Journal of the Royal Society Interface*, 15(138): 20170792, 2018.

Thomas S. Kuhn. *The Structure of Scientific Revolutions*. University of Chicago Press, Chicago, IL, 1996.

James Ladyman, James Lambert, and Karoline Wiesner. What is a complex system? *European Journal for Philosophy of Science*, 3(1):33–67, 2013. doi:10.1007/s13194-012-0056-8.

George Lakoff and Mark Johnson. *Metaphors We Live By*. University of Chicago Press, 1980.

Bruno Latour. *Science in Action: How to Follow Scientists and Engineers through Society*. Harvard University Press, Cambridge, MA, 1987.

Helen E. Longino. *Science as Social Knowledge: Values and Objectivity in Scientific Inquiry*. Princeton University Press, Princeton, NJ, 1990.

M. E. J. Newman. The structure and function of complex networks. *SIAM Review*, 45(2):167–256, 2003. doi:10.1137/S003614450342480.

Martin A. Nowak, Corina E. Tarnita, and Edward O. Wilson. The evolution of eusociality. *Nature*, 466(7310): 1057–1062, 2010. doi:10.1038/nature09205.

Benjamin P. Oldroyd and Boris Yagound. The role of epigenetics, particularly DNA methylation, in the evolution of caste in insect societies. *Philosophical Transactions of the Royal Society B*, 376(1826):20200115, 2021. doi:10.1098/rstb.2020.0115.

Bitao Qiu, Xueqin Dai, et al. Canalized gene expression during development mediates caste differentiation in ants. *Nature Ecology & Evolution*, 6(11):1753–1765, 2022. doi:10.1038/s41559-022-01884-y.

Fabien Ravary, Emmanuel Lecoutey, Gwenaël Kaminski, Nicolas Châline, and Pierre Jaisson. Individual experience alone can generate lasting division of labor in ants. *Current Biology*, 17(15):1308–1312, 2007. doi:10.1016/j.cub.2007.06.047.

Herbert A. Simon. The architecture of complexity. *Proceedings of the American Philosophical Society*, 106(6): 467–482, 1962. URL https://www.jstor.org/stable/985254.

Charlotte Sleigh. *Six Legs Better: A Cultural History of Myrmecology*. Johns Hopkins University Press, Baltimore, MD, 2007. ISBN 9780801892141.

Gerard J. Steen. Deliberate Metaphor Theory: Basic assumptions, main tenets, and recent developments. *Intercultural Pragmatics*, 14(1):1–24, 2017. doi:10.1515/ip-2017-0001.

Giovanni Strona, Domenico Nappo, Francesco Boccacci, Simone Fattorini, and Jesus San-Miguel-Ayanz. A fast and unbiased procedure to randomize ecological binary matrices with fixed row and column totals. *Nature Communications*, 5:4114, 2014. doi:10.1038/ncomms5114.

Seirian Sumner, Emily Bell, and Daisy Taylor. A molecular concept of caste in insect societies. *Current Opinion in Insect Science*, 25:42–50, 2018. doi:10.1016/j.cois.2017.11.010.

Paul H. Thibodeau and Lera Boroditsky. Metaphors we think with: The role of metaphor in reasoning. *PLoS ONE*, 6(2):e16782, 2011. doi:10.1371/journal.pone.0016782.

Martin H. Villet. Definitions of "caste" in social insects. *Ethology Ecology & Evolution*, 4(3):213–224, 1992. doi:10.1080/08927014.1992.9523134.

Janice Vis. Whose colony?: Rethinking terminology and insect relations. *Environmental Humanities*, 18(1):78–95,

2026. doi:10.1215/22011919-12211216.

William Morton Wheeler. The ant-colony as an organism. *Journal of Morphology*, 22(2):307–325, 1911. doi:10.1002/jmor.1050220206.

Ruth Wodak and Michael Meyer, editors. *Methods of Critical Discourse Analysis*. SAGE Publications, London, 2nd edition, 2009.