Exploring disciplinary differences in semantic uniformity

A computational approach to codification

Elaheh Sadat Ahmadi

1. Introduction

The study of disciplinary variation has long centered on the notion that fields differ not only in subject matter but also in how they organize knowledge and establish consensus. Zuckerman and Merton (1972: 507) introduced the concept of codification to capture these differences, defining it as "the consolidation of empirical knowledge into succinct and interdependent theoretical formulations." Highly codified disciplines such as physics and chemistry were seen as internally coherent, equipped with precise criteria for evaluating new contributions, and characterized by rapid citation turnover, what Derek Price called "citation immediacy." In these fields, the clarity of theoretical orientation allowed younger scholars to participate meaningfully at early stages in their careers. By contrast, in less codified disciplines such as zoology and botany, expertise depended on the accumulation of descriptive knowledge and loosely connected theories, making consensus harder to achieve and contributions slower to be recognized. Cole and Zuckerman (1975) extended this framework to their own field, arguing that sociology of science itself had, by the 1970s, achieved the level of codification necessary to become a recognized specialty. Their analyses of citation patterns, peer evaluations, and publication growth suggested that codification was linked to the consolidation of a specialty's identity and intellectual maturity. Yet subsequent work, including Cole, Cole and Dietrich (1978), complicated this picture. Although they sought to measure disciplinary consensus through indicators such as faculty ratings of scholars, reviewer agreement, and immediacy indices, the results were often inconsistent, statistically weak, or sensitive to the growth of literature. Stephen Cole (1983) went further, questioning the assumption that natural sciences inherently exhibit greater consensus and theoretical development than social sciences. His analyses showed that even in fields regarded as "hard sciences," disagreements at the research frontier were substantial. He concluded that disciplinary differences could not be neatly ordered along a fixed hierarchy, and that consensus was shaped as much by institutional and social mechanisms (peer review, funding structures, gatekeeping) as by cognitive organization alone.

Critics soon highlighted a deeper methodological limitation. Although the frameworks introduced by Zuckerman and Merton and later elaborated by subsequent scholars made important contributions by linking codification to observable scholarly behaviors, they fell short by relying on assumed field rankings and intuitive classifications, often grounded in Kuhnian authority or disciplinary tradition rather than empirical indicators (Cozzens, 1985). This lack of direct measurement rendered the interpretation of empirical findings ambiguous, as codification itself remained an unmeasured abstraction. Consequently, failures to confirm expected behavioral patterns could indicate that the codification hypothesis is incorrect, that commonsense rankings of disciplines along this dimension are flawed, or that codification is not a discipline-level phenomenon at all (Braxton and Hargens, 1996).

Addressing these debates and their methodological shortcomings, Gläser et al. (2024) reframe codification as a "global epistemic property" of research fields, one that cannot be attributed to individual research processes or their interrelations but exists only at the level of the scientific community. They contrast codification with analytical properties like "resource intensity" and structural ones like "epistemic diversity" of research processes and fields, emphasizing that codification reflects the holistic structure of a field's knowledge and language. Drawing on the foundational insights of Zuckerman and Merton, Gläser and colleagues define codification through two dimensions: theoretical organization and linguistic standardization. The former refers to how clearly concepts are defined, hierarchically structured, and theoretically integrated. The latter concerns the degree to which terminology is used consistently, avoiding homonyms and synonyms, and exemplified most extremely by mathematical language. To approximate codification empirically, Gläser et al. propose methods such as content analysis of scientific texts, bibliometric reconstruction of conceptual networks, and linguistic metrics. They particularly highlight the use of Contextualized Word Embeddings (CWEs) to detect semantic ambiguity and polysemy in field-specific language, a strategy this study adopts and explains in the following section.

The aim of this chapter is to report on an approach to a linguistically grounded measure of codification by examining the contextual stability of terminology in scientific discourse. Whereas previous research has largely relied on proxy indicators to approximate the degree of codification, this study builds on recent work that emphasizes the "standardization of a field's language" (Gläser et al., 2024). It introduces an embedding-based metric to capture how consistently key terms are used within disciplinary corpora, composed of journal articles in sociology and astrophysics, two disciplines conventionally positioned at opposite ends of the consensus spectrum. In doing so, the study investigates whether established differences in disciplinary consensus are reflected in their patterns of language use, particularly in their respective terminologies, and seeks both to operationalize the linguistic component of codification and to explore the use of large language models in the scalable comparative analysis of scientific fields.

This chapter is structured as follows: Section 2 outlines the methodological approach, including corpus construction, preprocessing, and the embedding-based procedure for computing Semantic Uniformity. Section 3 reports the empirical results, comparing sociology and astrophysics in terms of the contextual stability of their terminology. Section 4 addresses potential sources of error in corpus design and implementation, and Section

5 considers the broader conceptual and methodological implications of using contextual embeddings to operationalize codification.

2. Measuring codification with contextual embeddings

The experiment reported here focuses on one dimension of codification identified by Gläser et al. (2024: 12): the "standardization of a field's language". In their framework, a highly standardized language is one in which each term has a single, clearly defined meaning, and each concept is represented by only one term, minimizing ambiguity from synonyms and homonyms:

> "*The standardisation of a field's language* is high when among the concepts used there are few homonyms and synonyms, i.e. when only one word is used to represent a phenomenon, and when this word is used to represent only this particular phenomenon."

Building on this definition, standardization can be examined empirically through semantic uniformity, that is, by assessing whether a term is used consistently to refer to the same concept. A discipline's semantic uniformity score thus reflects the degree of consistency in how terms are employed within a given corpus of scientific writing. In highly codified fields, terms are expected to maintain stable, precisely defined meanings and therefore exhibit high semantic uniformity across contexts. In less codified disciplines, however, terms are more likely to carry multiple or even conflicting meanings, leading to greater polysemy and lower semantic uniformity.

2.1 Semantic Uniformity Score

Semantic Uniformity Score (SUS) can be built using contextual embeddings generated by a pre-trained BERT (Bidirectional Encoder Representations from Transformers) model (Devlin et al., 2019). BERT is a deep contextual language model based on Transformer architecture, pretrained on large-scale text corpora and its self-attention mechanism enables it to capture long-range dependencies and assign greater importance to contextually relevant words, producing embeddings that reflect a term's precise meaning in its specific context. BERT's effectiveness in detecting semantic ambiguity and polysemy is well established: it can distinguish word senses, with instances forming distinct clusters in the embedding space (Wiedemann et al., 2019) and can even predict a word's polysemy level from instances describing the same sense (Garí Soler and Apidianaki, 2021). These capabilities make BERT particularly well-suited for assessing the consistency of terminology in this experiment.

Leveraging these properties, BERT-generated contextual embeddings can be applied to quantify semantic uniformity across term occurrences in a corpus. Each occurrence of a word is represented as a high-dimensional vector $e_{w,i}$, informed by its co-occurring words and syntactic environment. For a given word w, appearing n_w times in the corpus D, all pairwise cosine similarities between its contextual embeddings are computed. The mean of these similarities, $Sim(w) = \frac{1}{\binom{n_w}{2}} \sum_{1 \leq i < j \leq n_w} \cos(e_{w,i}, e_{w,j})$, serves as

an indicator of how stable the word's meaning is across contexts: tightly clustered embeddings indicate semantic consistency, while scattered vectors suggest contextual variability or polysemy. This type of calculation is mathematically equivalent to the *SelfSim* metric reported by Garí Soler and Apidianaki (2021), though here it is applied for a different analytical purpose by aggregating the word-level similarity scores into a discipline-level SUS. This aggregation can be performed either without weighting (treating all words equally) or with frequency-based weighting, which gives greater influence to commonly used terms. This is formalized as:

$$SUS_D^{(\alpha)} = \frac{\sum_{w \in W} [n_w^{\alpha}.sim\,(w)]}{\sum_{w \in W} n_w^{\alpha}}$$

where $\alpha = 0$ produces the unweighted SUS and $\alpha = 1$ the weighted variant. A higher SUS suggests that a field's language leans toward standardized, well-defined usage, while a lower SUS indicates that terms are deployed with more interpretive flexibility. Comparing weighted and unweighted scores can further show whether semantic consistency is concentrated in a high-frequency core vocabulary or distributed more evenly across the lexicon.

2.2 Data and method

To empirically examine disciplinary differences in semantic uniformity, the analysis focuses on two corpora, astrophysics and sociology. Using the Web of Science (WoS) categories *Astronomy & Astrophysics* and *Sociology* in combination with *Journal Citation Reports* (JCR), five high-impact journals were selected for each field based on disciplinary focus, original research content, consistent article length, and archival depth. From each journal, 100 highly cited research articles were sampled (25 from 1990, 2000, 2010, and 2020) resulting in 500 articles per discipline.

The raw PDF documents were first processed using the NOUGAT OCR model (Blecher et al., 2023), which extracts structured text directly into Markdown format while preserving mathematical notation and visual layout. This was followed by an extensive post-processing step, where abstract, reference and the appendix sections containing tables were removed, equations and variables were replaced with placeholder tokens ([EQUATION], [VARIABLE]), and errors such as misrecognized or missing content were manually corrected. In contrast to Storer (1967), who treated the number of tables as an indicator of codification, this study removed tables and figures from the corpus. While their captions were retained for their explanatory value, the tables themselves consist primarily of numerical data; since the present analysis focuses on linguistic units, they were excluded. Standard text normalization steps like stopword removal and lemmatization were deliberately avoided, as they risk removing syntactic and semantic cues essential for embedding quality (Alzahrani and Jololian 2021). Instead, efforts focused on normalizing Markdown artifacts (e.g., hyphens, asterisks, headers), and correcting formatting inconsistencies. Punctuation and capitalization were maintained, while noise reduction refined the corpus by removing extraneous fragments around placeholders (e.g., "m[VARIABLE]"), consolidating consecutive placeholders, and deleting common

scholarly abbreviations (e.g., *et al.*, *cf.*) that added little value for embedding-based analysis. Finally, the text was segmented into blocks optimized for BERT's maximum input length of 512 tokens. Segmentation proceeded in stages: shorter blocks were merged to enhance coherence, oversized ones were split at sentence boundaries to prevent truncation, and residual undersized blocks were either expanded by borrowing sentences from adjacent segments or removed if they contained too little content. These procedures ensured that the resulting segments were structurally coherent, contextually rich, and of comparable length distributions across disciplines, thereby preparing the corpus for reliable embedding extraction.

To generate contextual embeddings for semantic analysis, the corpus was first initialized from preprocessed segmented text. Thresholds were applied to exclude low-frequency terms, defined as those appearing fewer than ten times in the corpus or in fewer than five documents (to reduce noise from typos or idiosyncratic vocabulary, ensure terms are representative of the broader discourse, and improve the stability of similarity estimates). Embedding extraction was performed using the *bert-base-uncased* implementation from the Hugging Face library, with a customized tokenizer that treated the predefined mathematical placeholder tokens as indivisible units. Since these tokens were absent from the original BERT vocabulary, they were assigned initial vectors by copying the pretrained embeddings for the base terms "variable" and "equation." This provided semantically grounded starting points instead of random vectors, reducing the risk of unpredictable contextual representations during inference. During inference, BERT's self-attention mechanism used these fixed embeddings as inputs and combined them with surrounding context to produce per-occurrence contextualized representations. Segments were tokenized in this way, padded, and passed through the BERT model, with final token embeddings computed by averaging the last four hidden layers. Subword units produced by WordPiece tokenization were reassembled and averaged to produce a single embedding per word.

After tokenization and embedding generation, a final vocabulary filter was applied before saving the terms for similarity analysis. This step did not alter the embeddings already produced, any token present in the segments, including noise items, had already contributed to the contextual encoding. The filter simply determined which tokens would be carried forward into the analysis. Non-alphabetic strings (e.g., numbers, mixed codes such as "fig_3"), domain-irrelevant artifacts from document extraction (e.g., "http", "html"), and very short tokens under three characters (e.g., "c4", "in") were removed, along with untagged mathematical expressions left over from incomplete NOUGAT recognition. Although such tokens remained in the original segments and influenced surrounding context during embedding generation, excluding them at this stage ensured that similarity calculations were based only on genuine lexical items, and that the Semantic Uniformity Score of the discipline was calculated from actual words rather than formatting artifacts or non-lexical symbols.

After that, cosine similarity scores were computed to assess how consistently each word was used across different contexts. To handle the large number of pairwise comparisons without exceeding memory limits, embeddings were first L2-normalized and then divided into fixed-size blocks. Each block was compared both internally (within-block) and externally (across-block) so that every unique pair of token embeddings was

included exactly once. For each word type all occurrence-level scores were aggregated to produce a mean similarity value that reflects its overall contextual stability across the corpus and a standard deviation that captures how much its meaning varied across different uses.[1]

2.3 Results

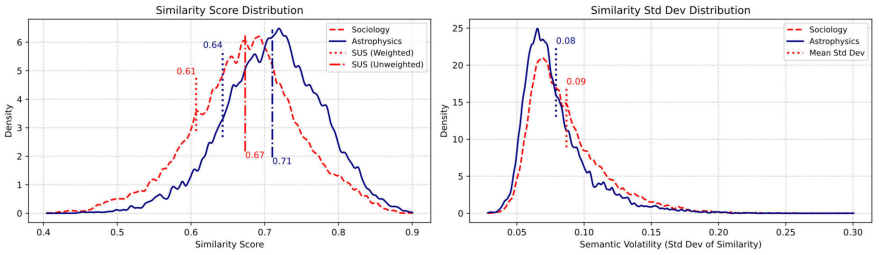
The analysis reveals notable disciplinary differences in the consistency of language use. Comparing SUS values between corpora, astrophysics exhibits higher semantic uniformity than sociology, a distinction evident in both unweighted and frequency-weighted scores. While the gap between fields is relatively modest in absolute terms (around 0.04), its consistency across both weighted and unweighted measures lends it analytical significance. In the unweighted comparison, which treats all word types equally, astrophysics reaches a mean SUS of 0.71, compared to 0.67 in sociology. When weighting by word frequency, scores decline in both fields, reflecting the broader contextual variation of more commonly used terms, but the gap remains: astrophysics registers 0.64, while sociology falls to 0.61. The results indicate that language in astrophysics is more tightly codified, with terms tending to retain clearer and more stable meanings across contexts. In contrast, sociological language appears more interpretively flexible, particularly among frequently used terms, which exhibit greater contextual variability.

The distinctions between disciplines are evident not only in the average similarity scores but also in their distributional patterns. Kernel density estimates (KDE; Figure A, left plot) show that similarity values in astrophysics are more densely concentrated around higher scores (0.7 to 0.9), whereas sociology displays a broader and flatter distribution. A similar trend appears in the standard deviation of similarity scores, which serves here as a proxy for semantic volatility, with sociology exhibiting a slightly higher value (0.09) than astrophysics (0.08). This suggests more flexible and variable word usage in sociology. In contrast, astrophysics shows tighter semantic boundaries. These differences are visually reinforced in the KDE curves (right plot), where sociology reveals a wider spread and astrophysics a sharper peak. Beyond the aggregate corpus-level comparison, the analysis also examines whether the observed difference in semantic uniformity between the two disciplines persists across different parts of speech (POS). The lower panel in Figure A presents the SUS disaggregated by syntactic category, including nouns, verbs, adjectives, proper nouns, and stopwords, serving as illustrative examples. Although absolute SUS values vary across grammatical categories, the relative difference between disciplines remains stable, with astrophysics consistently exhibiting higher semantic uniformity than sociology. This pattern suggests that the observed contrast is neither confined to specific lexical classes such as technical nouns nor masked by high-frequency functional tokens like stopwords, which are not typically expected to show uniform usage. Rather, it reflects a broader tendency toward codified and consistent language use across grammatical roles. Notably, the higher SUS values in astrophysics per-

1 The implementation underlying the analyses reported in this chapter has been archived as part of a larger research project (Ahmadi 2025) and is publicly available via Zenodo: https://doi.org/10.5281/zenodo.18753591.

sist even in categories where greater contextual variability would normally be expected, such as adjectives and stopwords.

Figure A: Semantic uniformity across disciplines

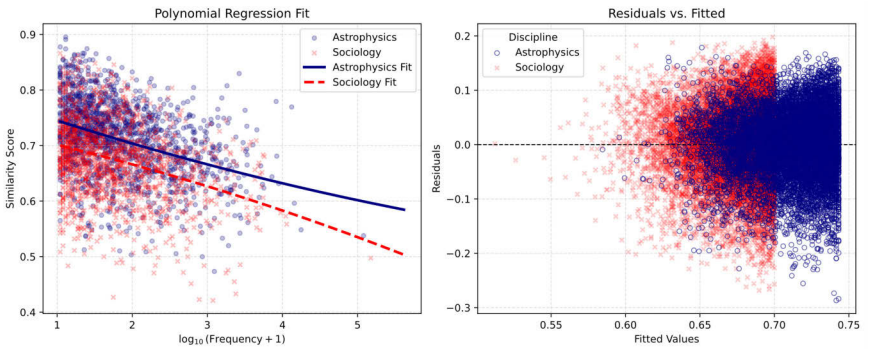


SUS by POS Group

Discipline	POS Group	Types (N)	Tokens (Σ freq)	SUS (Unweighted)	SUS (Weighted)
Sociology	Noun (NN+NNS)	5375	1278829	0.664	0.642
Sociology	Verb (all)	3358	424094	0.65	0.62
Sociology	Adjective (JJ)	2066	465541	0.679	0.641
Sociology	Proper Noun (NNP)	1925	122618	0.747	0.728
Sociology	Stopwords	108	1118273	0.56	0.535
Astrophysics	Noun (NN+NNS)	3321	1199085	0.695	0.677
Astrophysics	Verb (all)	2222	407587	0.687	0.666
Astrophysics	Adjective (JJ)	1274	425623	0.702	0.664
Astrophysics	Proper Noun (NNP)	2272	174721	0.769	0.751
Astrophysics	Stopwords	92	1143179	0.602	0.58

To examine the relationship between disciplines in greater detail with respect to frequency effects, an OLS regression model was estimated with word-level similarity as the dependent variable and log-transformed frequency, discipline, and their interaction as predictors. The results, presented in Figure B, confirm a general trend across both disciplines: semantic similarity tends to decrease as word frequency increases. However, the form and steepness of this decline differ notably.

Figure B: Model fit and residual diagnostics



In astrophysics, the relationship is largely linear, reflecting a gradual decrease in similarity among more frequent terms. In sociology, by contrast, the interaction terms reveal a significant nonlinear effect, with a sharper drop in similarity at higher frequencies. This suggests that the core vocabulary of sociology (the most frequently used terms) is

less semantically uniform than its periphery, whereas the gradient of variation in astrophysics is more restrained. Residual diagnostics, shown in the right panel of Figure B, support this interpretation: residuals in sociology are more widely dispersed, particularly at lower fitted values, while in astrophysics, they are more tightly clustered around zero, indicating a more consistent and predictable relationship between frequency and similarity.

Table A: OLS model summary with BP test

	Polynomial Interaction Model
Intercept	0.7112***
	(0.0007)
C(Discipline)[T.Sociology]	-0.0381***
	(0.0010)
log_count_c	-0.0401***
	(0.0013)
C(Discipline)[T.Sociology]:log_count_c	0.0036**
	(0.0018)
I(log_count_c ** 2)	0.0018
	(0.0012)
C(Discipline)[T.Sociology]:I(log_count_c ** 2)	-0.0039**
	(0.0017)
R-squared	0.1647
R-squared Adj.	0.1645
N	23372
R-squared	0.165
Adj. R-squared	0.165
F-statistic	935.688
Prob (F-stat)	0

Standard errors in parentheses.
* p<.1, ** p<.05, ***p<.01

Breusch-Pagan LM Stat: 285.9098
Breusch-Pagan LM p-value: 1.07e-59
Breusch-Pagan F Stat: 57.8753
Breusch-Pagan F p-value: 4.594e-60
Note: Standard Errors are heteroskedasticity robust (HC3)
log_count_c = log_count - mean(log_count)

Taken together, these results indicate a more semantically stable lexical structure in astrophysics, whereas sociology displays greater contextual flexibility in word usage. Although word frequency accounts for a meaningful share of the observed variation, it explains only approximately 16.5% of the variance in similarity scores. This implies that additional linguistic, contextual, or extra-contextual factors likely contribute to the disciplinary differences in semantic uniformity.

To evaluate the interpretive validity of the contextual embeddings as a proxy for codification, a qualitative validation was conducted using HDBSCAN. The aim was to test whether embeddings for key terms formed coherent clusters reflecting known or expected senses, thereby supporting the SUS metric derived from them. Consistent usage should produce tight clusters; polysemy should yield distinct groupings. This approach builds on Simons (2026), who showed that domain-tuned BERT models can recover multiple senses of terms like *Planck* in astrophysics. To examine both strengths and limitations, two terms from each field were analyzed: *order* and *paradigm* in sociology, *Planck* and *wave* in astrophysics.

The results present a mixed picture. Despite lacking domain-specific training, the general-purpose BERT model exhibited some promising capabilities. In sociology, the

clustering of the term *order* (Sim(w) = 0.43, N = 2392) successfully differentiated several distinct meanings, including grammatical constructions (e.g., *in order to*), sociological concepts (*in-order-to motive*), ordinal scales (*first-order*), statistical modeling terminology, and macro-social structures (*Interaction Order*). These clusters demonstrated strong semantic coherence, suggesting that the embeddings effectively captured clear polysemy, though this may have been partially driven by syntactic cues. By contrast, the term *paradigm*, which showed moderate semantic consistency (Sim(w) = 0.66, N = 102), resisted clear semantic partitioning. Most instances referred to theoretical frameworks or research orientations, ranging from established analytic traditions (e.g., *network paradigm*, *race-comparative paradigm*) to disciplinary shifts described as *paradigm shifts*. Although many uses carried clear Kuhnian connotations, the term's meaning was often shaped more by pragmatic context than by discrete conceptual boundaries. The difficulty in forming distinct clusters and the frequent classification of plausible uses as noise highlight *paradigm*'s conceptual looseness and functional adaptability. Manual analysis confirmed that the term often serves less as a fixed label and more as a flexible framing device conveying intellectual orientation across diverse empirical contexts.

The astrophysics cases offered complementary insights. For the highly polysemous term *Planck* (N = 2991; Sim(w) = 0.71), clustering successfully identified three prominent senses: institutional references (e.g., Max Planck Institutes), the ESA satellite mission (*Planck Collaboration*), and the technical *Fokker-Planck equation*. However, the clustering also revealed notable limitations. Embeddings tended to overemphasize dominant senses, particularly the satellite mission, sometimes fragmenting conceptually unified instances into separate clusters. Conversely, certain semantically distinct senses, such as *Planck units* (marked by mathematical notation and identified in prior work, e.g., Simons, 2026), were not detected. This omission was likely a byproduct of preprocessing steps, especially the substitution of equations with placeholder tokens.

The term *wave* (N = 730; Sim(w) = 0.65) further illustrates the challenges of clustering semantically broad and polysemous terms using general-purpose language models. Although three primary clusters emerged (gravitational waves, shock or blast waves, and plasma waves) with reasonable internal coherence, the boundaries between them remained weak, and over half of the occurrences were classified as noise. This outcome reflects the conceptual and discursive overlap among different wave phenomena in astrophysics, where wave types frequently co-occur and are described using similar modeling frameworks. Overall, the clustering results suggest that the contextual embeddings used in this study are capable of capturing some meaningful distinctions, but lack the granularity required to reliably separate closely related technical senses, particularly when terms are both frequent and broadly applied.

Across both disciplines, clustering analyses surfaced key methodological challenges stemming from the limited sensitivity of general-purpose BERT embeddings to subtle contextual variation (cf. Simons, 2026). In specialized scientific language, conceptually coherent usages were sometimes split due to superficial lexical differences, while distinct meanings blurred when surface cues aligned too closely. The clustering algorithm, though well-suited for irregular data, struggled at times to balance noise filtering with semantic granularity, either fragmenting dominant senses or merging subtly distinct ones. These effects were amplified by frequency disparities and the lack of a gold-stan-

dard sense inventory, highlighting both the interpretive demands and the technical indeterminacy of unsupervised clustering in scientific discourse.

3. Limitations

The corpus design inevitably introduces potential sources of error that may influence the absolute values of the SUS. These can be disentangled into several categories.

Genres: This study draws exclusively on journal articles, which, while offering structural comparability, omit other genres such as books, conference papers, grant proposals, or letters. Codification may differ across these contexts, particularly between formal publications and more exploratory forms of communication. For the observed relationship to reverse (i.e., sociology appearing more codified than astrophysics), one would have to assume that language in sociology's excluded genres is *more* semantically uniform than in its journal articles, while language in astrophysics' excluded genres is *less* uniform than in its journal articles. This seems theoretically unlikely: it would require that sociological monographs differ substantially from research articles in ways that increase codification, while astrophysics books are less semantically uniform than astrophysics articles. In such a case, including all genres, or only books, would show the opposite relationship between disciplines. Nonetheless, this assumption could be tested directly by varying the corpus composition to include additional genres.

Field delineation: Field delineation also matters for interpreting the results in light of earlier arguments about codification. Cole and Zuckerman (1975) proposed that codification becomes most visible when a research area consolidates into a specialty with shared evaluative standards and a stable vocabulary. The present study, however, relies on WoS subject categories that define disciplines at a broad level, combining heterogeneous subfields with distinct terminologies. Such breadth introduces semantic diversity that can lower SUS values and makes codification harder to detect at the disciplinary scale. A narrower delineation at the specialty level would align more closely with the conditions under which Cole and Zuckerman argued codification becomes identifiable, and would likely increase absolute SUS scores in both sociology and astrophysics. Yet even under this assumption, a reversal of the observed disciplinary ordering would require that sociological specialties be systematically more codified than those in astrophysics, for which there is no empirical basis.

Temporal coverage: The corpus includes articles from 1990, 2000, 2010, and 2020. Spanning multiple decades tends to increase semantic diversity, as vocabularies and meanings change over time. For the observed relationship to reverse under a different sampling window, for example, using only articles from 2020–2025, astrophysics would need to have become substantially less uniform after 2020, to the point that sociology's SUS would overtake it. Conversely, for the reversal to occur by extending coverage backwards, sociology would need to have been significantly more codified before 1990 than it is in later decades. Both scenarios seem improbable given what is known about the historical

development of the two fields, although this assumption requires empirical validation through time-specific corpora.

Journal quality: Restricting the sample to journals with high JCR metrics reduces the inclusion of material less likely to contribute to the body of knowledge, concepts, and methods that the research community actively uses and builds upon. For the observed relationship to reverse, sociology's lower-JCR literature would have to exhibit more semantically uniform language use than its higher-JCR literature, while astrophysics' lower-JCR literature would have to exhibit less, a questionable pattern. Although JCR rankings are not a perfect proxy for cognitive core membership, higher-ranked journals often have more standardized editorial practices, which can encourage (but do not guarantee) terminological consistency. Nonetheless, this could be tested by including mid- and low-JCR journals and comparing the results.

The overall measurement pipeline also remains sensitive to several implementation choices, each of which can influence the absolute values of the SUS.

Frequency thresholds: Removing infrequent terms is an effective way to reduce noise, yet it can also disproportionately exclude highly specialized terms that are consistently used within subfields, potentially lowering the SUS for fields with more niche vocabularies. This effect can be tested by recomputing the SUS under different cut-off points to verify whether the relative disciplinary ordering remains stable.

Segmentation: Segmentation strategy presents another source of variation which directly affects which words are included and how *context* is defined, ultimately influencing the generated embeddings. The size of the textual context fed into BERT affects the amount of co-text available to each term; shorter segments (e.g., 100 tokens) limit context and may increase variability in embeddings, whereas longer segments (e.g., 500 tokens) capture broader linguistic environments and may result in more stable representations. Robustness here can be assessed by recalculating the SUS with alternative segment lengths and comparing the results.

Placeholders: A third consideration concerns the treatment of mathematical notation in astrophysics. In the present corpus, equations and variables were replaced with placeholders to facilitate processing, yet these expressions are among the most standardized elements in the field. If considered part of the language, their omission likely underestimates astrophysics' semantic uniformity and may contribute to the relatively small observed difference between the two disciplines. This can be tested by reprocessing the astrophysics corpus with a math-aware BERT model capable of embedding equations directly, then comparing the resulting SUS with the placeholder-based approach.

Taken together, these considerations underline that no corpus design or modeling pipeline can entirely eliminate uncertainty. However, the potential biases identified here are unlikely to systematically favor one discipline over the other in ways that would invert the observed pattern. The numerical results reported here are specific to the analyzed disciplinary corpora and should thus be understood as relative measures of semantic

uniformity and the relative degree of language standardization, rather than as absolute indicators of codification. Within this scope, the relative difference seems to represent a robust feature of the two fields, though its precise magnitude may vary under alternative design choices. This study therefore can be further refined through targeted variations in corpus construction and processing.

4. Conclusion

This chapter has explored how computational models of meaning can be leveraged to operationalize the linguistic dimension of codification in scientific discourse. By applying the Semantic Uniformity Score to corpora from astrophysics and sociology, it provides an empirical lens on long-standing claims about disciplinary differences in cognitive consensus. The results suggest a visible contrast: astrophysics tends toward more semantically uniform usage of terms, whereas sociology exhibits greater contextual variability, particularly among frequently used vocabulary. While these patterns are suggestive, their interpretation must remain exploratory rather than definitive.

The attempt to capture codification through an embedding-based semantic uniformity metric necessarily entails reducing the complexity of meaning-making in language. The model employed here, pretrained on general-purpose corpora, represents meaning through statistical regularities in language use, specifically patterns of word co-occurrence and attention distributions. Ontologically, this reflects the distributional hypothesis: concepts are treated as emergent from patterns of use rather than grounded in stable, context-independent definitions. Yet this assumption is marked by theoretical ambiguity. While some perspectives hold that words possess relatively stable core senses modulated by context, others argue that meaning is wholly relational and constructed within situated usage (Ravin and Leacock, 2000). Contextual embeddings further inherit the inherent ambiguity of both natural language and the training corpora on which they are based. Consequently, high variability in embeddings does not necessarily signal genuine semantic ambiguity (polysemy); it may instead arise from topic shifts, syntactic variation, or genre differences.

In this sense, the Semantic Uniformity Score should be understood as a partial indicator of codification. It captures regularities in how terms are distributed across contexts, but it does not fully reflect the conceptual clarity or epistemic consensus such terms may embody. Complementary approaches, such as clustering contextual embeddings, inspecting the sentences from which they are drawn, or qualitative content analysis, can help validate SUS and mitigate its limitations by moving beyond static textual windows and incorporating broader interpretive and paratextual elements. Ultimately, interpreting meaning requires sensitivity to the conceptual traditions, interpretive frameworks, and historical developments that shape how terms like *structure*, *agency*, or *order* are deployed. In fields like sociology, where such terms span multiple schools of thought, disentangling meanings often demands knowledge that exceeds what statistical modeling can extract from local linguistic context alone.

Beyond the general limitations of distributional semantic models, which affect all research using contextual embeddings for semantic variation or shift detection, this study

also faced constraints specific to corpus construction and preprocessing. Such design choices can influence the absolute values of the SUS, but they are unlikely to overturn the relative disciplinary differences observed here, since reversing them would require empirically implausible conditions. Within these boundaries, the approach nevertheless illustrates how semantic features of disciplinary language can be examined systematically at scale, in ways that complement more established qualitative and bibliometric perspectives. Rather than presenting a definitive tool, it adds to the repertoire of methods available for studying how knowledge is articulated through shared and stabilized terms. Future work could refine this approach by incorporating expert annotation, testing models better suited to specialized language, and focusing more closely on specialty domains where conceptual alignment is often clearer. Expanding the analysis to multiword expressions or terminological units may also capture more stable conceptual constructs. Ultimately, integrating computational indicators like SUS with interpretive analysis can offer a way to trace how concepts are negotiated into shared meanings, situated within broader theoretical frameworks, and stabilized within scientific communities.[2]

References

Ahmadi ES (2025). Measuring the codification of scientific knowledge with large language models. Discussion Paper, Social Studies of Science and Technology, Institute of History and Philosophy of Science, Technology, and Literature, Technische Universität Berlin.

Alzahrani E and Jololian L (2021) How Different Text-preprocessing Techniques Using The BERT Model Affect The Gender Profiling of Authors. arXiv:2109.13890. arXiv.

Blecher L, Cucurull G, Scialom T, et al. (2023) Nougat: Neural Optical Understanding for Academic Documents. arXiv:2308.13418. arXiv.

Braxton J and Hargens L (1996) Variation among academic disciplines: Analytical frameworks and research. In: *Reprinted from: Higher Education: Handbook of Theory and Research*. Agathon Press. New York.

Cole JR and Zuckerman H (1975) The Emergence of a Scientific Specialty: The Self-Exampli fying Case of the Sociology of Science. In: Coser LA (ed.) *The Idea of Social Structure. Papers in Honor of Robert K. Merton*. New York, Chicago, San Francisco, Atlanta: Harcourt Brace Janovich, Inc., pp. 139–174.

Cole S (1983) The hierarchy of the sciences? *American Journal of Sociology* 89: 111–139.

Cole S, Cole JR and Dietrich L (1978) Measuring the cognitive state of scientific disciplines. In: Elkana Y, Lederberg J, Merton RK, et al. (eds) *Toward a Metric of Science*. New York: Wiley, pp. 209–251.

2 This chapter was written with support from large language models (LLMs). All model-generated text was reviewed and, where necessary, rewritten by the authors, who remain fully responsible for the final version. For details on the use of LLMs in this volume, see the statement in the volume's introduction.

Cozzens SE (1985) Comparing the Sciences: Citation Context Analysis of Papers from Neuropharmacology and the Sociology of Science. *Social Studies of Science* 15(1): 127–153.

Devlin J, Chang M-W, Lee K, et al. (2019) BERT: Pre-training of deep bidirectional Transformers for language understanding. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (eds J Burstein, T Solorio, and C Doran), Minneapolis, Minnesota, 2019, pp. 4171–4186. Association for Computational Linguistics.

Garí Soler A and Apidianaki M (2021) Let's Play Mono-Poly: BERT Can Reveal Words' Polysemy Level and Partitionability into Senses. *Transactions of the Association for Computational Linguistics* Roark B and Nenkova A (eds) 9. Cambridge, MA: MIT Press: 825–844.

Gläser J, Hoffmann M, Laudel G, et al. (2024) The empirical investigation of epistemic properties of research processes and fields. Berlin.

Ravin Y and Leacock C (2000) Polysemy: An Overview. In: *Polysemy: Theoretical and Computational Approaches*. Oxford, New York: OUP Oxford, pp. 1–29.

Simons A (2026) Meaning at the Planck scale? Contextualized word embeddings for doing history, philosophy, and sociology of science. In: Simons A, Wüthrich A, Zichert M, et al. (eds) *Understanding Science with Large Language Models? Potentials for the History, Philosophy, and Sociology of Science*. Bielefeld: transcript, part-3.

Storer NW (1967) The Hard Sciences and the Soft: some sociological observations. *Bulletin of the Medical Library Association*. 55(1): 75–84.

Wiedemann G, Biemann C, Chawla A, et al. (2019) Does BERT Make Any Sense? Interpretable Word Sense Disambiguation with Contextualized Embeddings. arXiv:1909.10430, Preprint.

Zuckerman H and Merton RK (1972) Age, aging, and age structure in science. In: Riley MW, Johnson M, and Foner A (eds) *Aging and Society, Volume 3: A Sociology of Age Stratification*. New York: Russell Sage Foundation, pp. 292–356.