Lexical Richness as Relational Sense : A Semantic-Similarity–Based Analysis of Textual Style across Semantic Fields

1. Introduction

Vocabulary is not only a collection of isolated word meanings; it is also a system of relations among words. In narrative writing, authors often rely on lexical relations—especially fine-grained similarity (near-synonymy) and semantic contrast (oppositional tendency)—to express nuances that a single lexical item cannot fully convey. This relational structure produces a richer sense of tone, character depiction, and atmosphere, thereby shaping readers' stylistic impressions.

This paper proposes a quantitative approach to textual style analysis grounded in semantic similarity. Using pre-trained word embeddings, I examine how lexical items cluster and distribute within four semantic fields—emotion, character evaluation, social class, and environment—and summarize each field with four statistical metrics. The results are intended to provide a vocabulary-based, data-driven profile of where a text places its semantic emphasis and how richly it expresses meaning through lexical relations.

2. Literature Review

Lexical richness has traditionally been examined through surface-level measures such as lexical diversity, density, and frequency-based indices, which have been widely applied in studies of writing quality and second-language proficiency. Empirical research shows that higher lexical diversity and sophistication are often associated with more advanced language use (Ha, 2019). However, subsequent studies have noted that many traditional richness metrics are sensitive to text length and fail to capture deeper semantic organization (Fergadiotis et al., 2015; Kojima, 2014).

Beyond surface counts, recent work emphasizes **semantic richness**, highlighting that meaning arises from networks of semantic relations rather than isolated word forms. Psycholinguistic evidence suggests that words embedded in richer semantic neighborhoods exert stronger cognitive effects during processing (Goh et al., 2016). This perspective aligns with the distributional hypothesis, which underlies vector-space models of meaning that represent words through contextual similarity (Lenci & Sahlgren, 2019). Moreover, embedding-based approaches have been shown capable of modeling fine-grained semantic relations, including similarity and contrast (Nguyen et al., 2016). Together, these studies motivate a relational, similarity-based approach to lexical richness and textual style.

3. Research Purpose

This study is motivated by the idea that texts appear more vivid and expressive when they employ richer **lexical relations**, rather than simply a larger number of words. Lexical richness is therefore understood as structured relational meaning: near-similar words allow subtle shifts in intensity or evaluation, while semantic contrasts sharpen narrative focus. In this view, textual *sense* emerges from interactions among words, not solely from the dictionary meaning of individual items.

Accordingly, this study aims to:

1. operationalize lexical richness as relational structure using semantic similarity modeled by word embeddings;
2. quantify semantic emphasis across four fields (emotion, character evaluation, social class, and environment) through four interpretable metrics capturing density, variety, dispersion, and dominance;
3. examine how different patterns of lexical relational richness correspond to textual style and influence the nature of vocabulary input available to readers or learners.

4.1 Results

4.1.1 Subjects

The analysis targets one or more narrative texts (to be specified in the final version). Each text is treated as a corpus. All tokens are preprocessed (e.g., lowercasing; punctuation handling; optional lemmatization depending on implementation) and mapped to embedding entries when available.

4.1.2 Method[1]

The analysis was implemented using a custom Python-based toolkit developed by the author and publicly released as an open-source repository. The toolkit supports semantic-field–based lexical extraction, semantic similarity computation using pre-trained word embeddings, and the calculation of statistical metrics for lexical richness and dominance.

1. Embedding model and similarity:

Pre-trained GloVe vectors (300 dimensions) are used to represent words in a semantic vector space. Specifically, I use the Stanford GloVe "Wikipedia 2014 + Gigaword 5" pretrained set (6B tokens; 400K vocabulary; uncased; 300d). (斯坦福自然语言处理小组)[2]

Each word w is represented by an embedding vector $v(w) \in \mathbb{R}^{300}$. Semantic similarity between two lexical items w_i, w_j is computed using cosine similarity:

$$\sin(w_i, w_j) = \cos\big(v(w_i), v(w_j)\big) = \frac{v(w_i) \cdot v(w_j)}{\|v(w_i)\|\|v(w_j)\|}$$

High similarity values correspond to near-synonymy or close semantic association, while low or negative values indicate semantic distance or potential contrast (oppositional tendency).

---

[1] https://github.com/plursef/lexrich

[2] GloVe model: https://nlp.stanford.edu/projects/glove

2. Semantic fields:

The study analyzes four semantic fields: Emotion, character evaluation, social class and environment. A seed lexicon is prepared for each field.

3. Lexical items to clusters (relational structure):

To model "relational richness," lexical items within each field are grouped into similarity-based clusters. Concretely, a clustering method is applied over word vectors (e.g., agglomerative clustering / community detection over a similarity graph / k-means). Each cluster is intended to represent a region of close semantic sense (near-synonyms or strongly associated descriptors). Semantic contrast can be reflected by clusters that are far apart in the embedding space or by the presence of lexemes whose similarity to the cluster centroid is low/negative.

4. Metrics:

Let the full text contain N tokens. For a given semantic field F, let T_F be the multiset of tokens in the text that belong to field F (by lexicon match and preprocessing), with total count $n_F = |T_F|$. Let V_F be the set of unique types in F, with size $|T_F|$. Let $f(w)$ be the frequency of type w within F.

If clustering is applied, suppose field F yields K_F clusters $C_1, \dots, C_{K_F}$. Let n_k be the token count in cluster C_k (sum of frequencies of types assigned to C_k), so $\Sigma_{\{k=1\}}^{\{K_F\}} n_k = n_F$. Define $p_k = \frac{n_k}{n_F}$.

(1) FieldCoveragePer10k (coverage density):

$$Coverage_{10k}(F) = \left(\frac{n_F}{N}\right) \times 10{,}000$$

This measures how densely the text draws vocabulary from field F relative to its length.

(2) FieldTypeCount (lexical variety):

$$TypeCount(F) = |V_F|$$

This captures how many distinct lexical types the field contributes, reflecting lexical diversity within that semantic domain.

(3) ClusterEntropy (dispersion of sense across clusters):

$$Entropy(F) = -\Sigma_{\{k=1\}}^{K_F} p_k \log p_k$$

Optionally, for comparability across different K_F, a normalized form can be reported:

$$Entropy_{norm(F)} = \frac{\text{Entropy(F)}}{\log K_F}$$

Higher entropy indicates that lexical usage is distributed across multiple sense-clusters rather than concentrated in a small region of semantic space.

(4) Dominance (degree of concentration / "one cluster or word rules all"):

A cluster-based dominance (recommended, consistent with entropy) is:

$$Dominance(F) = \max_{k \in \{1,\ldots,K_F\}} p_k$$

Alternatively, a type-based dominance (if clustering is omitted) can be defined as:

$$Dominance_{type(F)} = \max \_\{w \in V\_F\} f(w)/n\_F$$

In this project, dominance is interpreted as semantic focus: a high value suggests the field is realized through a narrow sense region (or a single repeated descriptor), whereas a lower value suggests richer alternation among related expressions.

4.1.3 Results

1) In Field environment:

Intel technology manuals[3]:

Metric	Value
FieldCoveragePer10k	57.5976
FieldTypeCount	4
ClusterEntropy	1.0309
Dominance	0.6678

Tess of d'Urbervilles (novel)[4]:

Metric	Value
FieldCoveragePer10k	5057.9791
FieldTypeCount	534
ClusterEntropy	0.3089
Dominance	0.9721

Psycology Papers (about emotion):

Metric	Value
FieldCoveragePer10k	3571.2876

[3] Intel 80386 Software Developer's Manual, https://pdos.csail.mit.edu/6.828/2018/readings/ia32/IA32-3A.pdf

[4] Tess of d'Urbervilles: https://www.gutenberg.org/files/110/110-h/110-h.htm#chap06

Metric	Value
FieldTypeCount	604
ClusterEntropy	0.3661
Dominance	0.9626

Economic Text Books:

Metric	Value
FieldCoveragePer10k	3539.2410
FieldTypeCount	453
ClusterEntropy	0.7806
Dominance	0.8922

2) In field emotion:

Intel technology manuals:

Metric	Value
FieldCoveragePer10k	0.0000
FieldTypeCount	0
ClusterEntropy	0.0000
Dominance	0.0000

Tess of d’Urbervilles:

Metric	Value
FieldCoveragePer10k	43.0109
FieldTypeCount	42
ClusterEntropy	3.8142
Dominance	0.2683

Psycology Papers (about emotion):

Metric	Value
FieldCoveragePer10k	47.4290
FieldTypeCount	31
ClusterEntropy	2.9816
Dominance	0.4304

Economic Text Books:

Metric	Value
FieldCoveragePer10k	0.7110
FieldTypeCount	1
ClusterEntropy	-0.0000
Dominance	1.0000

4.2 Discussion

4.2.1 Lexical richness and semantic focus in the field of environment

The results from the environment field reveal striking contrasts across text types, demonstrating how lexical richness and dominance jointly shape stylistic orientation. The Intel technology manuals exhibit extremely low lexical richness in this field. The FieldCoveragePer10k is minimal (57.60), accompanied by a very small FieldTypeCount (4). Although the ClusterEntropy appears moderately high (1.03), the Dominance value (0.67) indicates that the limited environmental vocabulary is still largely controlled by a small subset of terms. This pattern reflects a highly instrumental and functional use of environmental language, where references to physical or operational settings are sparse, standardized, and semantically constrained. Such texts aim to eliminate ambiguity rather than cultivate descriptive richness.

In contrast, Tess of the d'Urbervilles demonstrates an overwhelmingly dense use of environmental vocabulary, with FieldCoveragePer10k exceeding 5000 and a FieldTypeCount of 534. Despite this lexical abundance, the ClusterEntropy is relatively low (0.31), and Dominance is extremely high (0.97). This suggests that while the novel employs a wide range of environmental words, they are strongly organized around a centralized semantic core. From a stylistic perspective, this indicates that the environment in the novel is not merely decorative but thematically cohesive: repeated lexical choices reinforce a consistent rural and natural atmosphere, allowing subtle variations around a dominant sense rather than dispersing meaning across unrelated descriptions.

A similar yet distinct pattern emerges in psychology papers. These texts also display high coverage and a large type count in the environment field, reflecting frequent contextual references (e.g., experimental settings, situational descriptions). However, their low entropy and high dominance indicate that environmental vocabulary functions primarily as methodological scaffolding rather than as an expressive resource. The environment is invoked to support empirical clarity, not to enrich narrative texture.

By comparison, economic textbooks present a more balanced configuration. While their coverage and type count are substantial, their higher ClusterEntropy (0.78) and lower Dominance (0.89) suggest a more distributed semantic structure. Environmental terms in economic discourse often span physical, institutional, and abstract environments, resulting in a broader semantic spread. This pattern reflects the explanatory nature of economic writing, which must integrate multiple contextual layers rather than emphasize a single sensory or spatial image.

Overall, the environment field illustrates how high lexical variety does not necessarily entail high semantic dispersion. Instead, literary style emerges from how lexical choices are organized around dominant senses, rather than from sheer vocabulary quantity.

4.2.2 Emotional vocabulary and relational sense across genres

The emotion field offers an even clearer demonstration of how lexical relations shape textual style.

As expected, Intel technology manuals contain no emotional vocabulary. All metrics are zero, confirming the genre's deliberate exclusion of affective language. This absence is itself stylistically meaningful: the manuals enforce objectivity by systematically removing emotional sense from the lexical space.

In Tess of the d'Urbervilles, emotional vocabulary exhibits moderate coverage (43.01) and a sizable type count (42), but most notably, an exceptionally high ClusterEntropy (3.81) coupled with very low Dominance (0.27). This configuration indicates that emotional expression in the novel is highly diversified and relationally rich. Rather than relying on a few repeated emotional labels, the text navigates across many semantic clusters—grief, desire, shame, hope, fear—each occupying distinct regions in semantic space. This dispersed structure allows the novel to convey emotional sense through contrast, gradation, and shifting intensity, enabling readers to experience emotions as evolving and layered rather than fixed states.

Psychology papers, while also showing relatively high emotional coverage and lexical variety, differ sharply in structure. Their lower entropy (2.98) and higher dominance (0.43) suggest a more controlled emotional lexicon. Emotional terms are often technicalized and repeatedly anchored to specific constructs (e.g., anxiety, depression, affect), resulting in a narrower semantic range. Here, emotional vocabulary serves analytical precision rather than expressive depth.

Finally, economic textbooks show minimal emotional presence, with a FieldTypeCount of only one and absolute dominance (1.00). This pattern confirms that emotional language in economic discourse is typically restricted to rare evaluative signals, reinforcing the genre's emphasis on rational abstraction.

5. Conclusion

This paper proposes a semantic-similarity–based framework for analyzing lexical richness as a relational phenomenon. By combining GloVe embeddings with four interpretable metrics—coverage density, type variety, cluster entropy, and dominance—the approach characterizes how a text distributes lexical choices across semantic fields and how it concentrates or diversifies sense within each field.

The core contribution is conceptual as well as methodological: stylistic expressiveness is argued to arise not only from what words mean individually, but from how words relate to one another—through similarity-driven shading and contrast-driven emphasis—which together produce nuanced sense that exceeds any single lexical item. In future work, the framework can be extended by comparing multiple texts systematically, testing robustness under alternative embedding models, and connecting lexical-relational profiles to reader or learner outcomes.

References

1. Ha, Hye Seung. (2019). *Lexical Richness in EFL Undergraduate Students' Academic Writing*. *English Teaching*, 74(3), 3–28. ERIC

2. Fergadiotis, G., et al. (2015). *Psychometric Evaluation of Lexical Diversity Indices*. *Journal of Speech, Language, and Hearing Research*, (PMC4490052). PMC

3. Kojima, M. (2014). *Reliability of lexical richness measures based on word lists*. *Scientia Linguistica*, (S0346251X13001565).

4. Goh, W.D., Yap, M.J., et al. (2016). *Semantic Richness Effects in Spoken Word Recognition*. *PMC4923159*. PMC

5. Nguyen, K.A., Schulte im Walde, S., & Vu, N.T. (2016). *Integrating Distributional Lexical Contrast into Word Embeddings for Antonym-Synonym Distinction*. *arXiv*.

Plagiarism check result: 1.7%

Appendix A. Word Embedding Resources

This study employs pre-trained **GloVe (Global Vectors for Word Representation)** embeddings developed by the Stanford NLP Group. GloVe is a distributional semantic model that learns word representations from global word–word co-occurrence statistics.

The specific embedding set used in this study is:

- **Wikipedia 2014 + Gigaword 5**
- **6 billion tokens**

- **400,000-word vocabulary**
- **300-dimensional vectors**
- **Uncased**

The embeddings are publicly available at the official Stanford NLP website: https://nlp.stanford.edu/projects/glove/

Detailed documentation of the training corpora can be found at:

- Wikipedia dumps: http://dumps.wikimedia.org/enwiki/20140102/
- English Gigaword Fifth Edition (LDC2011T07): https://catalog.ldc.upenn.edu/LDC2011T07

Appendix B. Mathematical Definitions and Statistical Concepts

This appendix provides brief definitions of the mathematical and statistical concepts used in the analysis. These definitions are included for reference and are not discussed in detail in the main text.

B.1 Cosine Similarity

Cosine similarity is a standard measure of semantic similarity in vector-space models. It quantifies the cosine of the angle between two word vectors, producing values typically in the range $[-1,1]$, where higher values indicate greater semantic similarity.

A general introduction to cosine similarity in distributional semantics is available at: https://en.wikipedia.org/wiki/Cosine_similarity

B.2 Shannon Entropy

ClusterEntropy in this study is based on Shannon entropy, which measures the uncertainty or dispersion of a probability distribution. In the context of this analysis, entropy captures how lexical usage is distributed across semantic clusters.

A concise reference on entropy is available at: https://en.wikipedia.org/wiki/Entropy_(information_theory)

B.3 Dominance

Dominance is defined as the maximum proportion of tokens contributed by a single semantic cluster (or, alternatively, a single lexical type). It serves as an index of semantic concentration, indicating whether a semantic field is governed by a narrow lexical core or distributed across multiple sense regions.

Related discussions of dominance and concentration measures can be found in: https://en.wikipedia.org/wiki/Concentration_measure

Appendix C. Implementation and Reproducibility

All analyses reported in this paper were conducted using a custom Python-based toolkit developed for this study. The toolkit implements semantic-field extraction, embedding-based similarity computation, clustering, and metric calculation.

To support transparency and reproducibility, the complete implementation is publicly available as an open-source repository at:
https://github.com/plursef/lexrich

The repository includes source code, documentation, and example configurations used in the analysis.

Appendix D. Notes on Interpretation

It should be noted that distributional word embeddings capture **contextual similarity**, which does not always correspond directly to classical lexical relations such as antonymy. Words with opposing meanings may appear close in embedding space due to shared contextual environments. Accordingly, semantic contrast in this study is interpreted in terms of **distributional distance and cluster separation**, rather than strict lexical opposition.