A Corpus-Based Profiling of Regional English Variants in Global Media: Insights from Olympic Journalism

Felix Mao
Rye Country Day School
fymao27@gmail.com

Abstract

This paper investigates the distinctive linguistic characteristics of regional English-language journalistic editorial styles through a quantitative analysis of global media coverage. The study applies advanced classification techniques, integrating GPT-based embeddings with Support Vector Machines, to a novel corpus, the Olympic Journalism English Media Corpus. Comprising news articles related to Olympic Games covered by prominent news outlets in the United States, China, Spain, and Mexico between 2020 and 2023, this corpus enables a fine-grained analysis of 164 linguistic features across lexical, syntactic, readability, and sentiment dimensions. The findings reveal strong and interpretable distinctions in features such as verb ratio, nominality, and readability. This study not only demonstrates the enhanced classification capabilities of the model (optimized F1 score = 97.2), but also yields deeper, data-driven stylistic analysis and insights into each English journalistic style. This work provides a potential template that can be expanded to other World Englishes research.

Keywords: Corpus Linguistics, Regionally anchored English media styles, GPT-based embeddings, Support Vector Machines

1. Introduction

English functions as a global lingua franca across journalism, academia, diplomacy, and commerce. While English is often treated as a unified language, its usage varies systematically across national, institutional, and cultural contexts. Research in World Englishes has long documented lexical, grammatical, and discourse-level differences among Englishes used in different regions (Kachru, 1990; McArthur, 2001; Pennycook, 2009). However, much of this research has relied on qualitative analysis or relatively limited quantitative metrics, leaving room for more large-scale, reproducible computational approaches.

This study investigates how English-language journalism produced by media outlets in different national contexts exhibits systematic stylistic differences. Rather than claiming to model stable sociolinguistic "varieties" in the traditional sense, we examine regionally anchored English media styles—English-language journalistic production associated with outlets based in distinct national and cultural environments. These texts reflect institutional editorial norms, audience expectations, and discourse conventions that may shape measurable linguistic patterns.

To provide a controlled comparison, we construct the Olympic Journalism English Media Corpus, consisting of English-language news articles covering the Olympic Games (2020–2023) from major outlets based in the United States, China, Spain, Costa Rica, and Mexico. The Olympics provide a particularly suitable domain: journalists across regions report on a shared global event, reducing topical divergence while allowing stylistic differences to surface.

We apply a quantitative framework integrating (i) 164 linguistically motivated features spanning lexical, syntactic, readability, sentiment, and discourse dimensions, and (ii) GPT-based sentence embeddings combined with Support Vector Machine (SVM) classifiers. This dual approach enables both high-accuracy classification and feature-level interpretability.

Importantly, our goal is not to define or essentialize national English varieties. Instead, we aim to demonstrate how computational methods can identify consistent stylistic tendencies in English-language journalism across regionally distinct institutional contexts. By combining interpretable feature analysis with embedding-based modeling, this study contributes a reproducible framework for examining variation in global English media production

2. Literature Review

2.1. World Englishes and Quantitative Approaches

Research on World Englishes has documented structural and discourse differences across national contexts (Kachru, 1990; McArthur, 2001). Traditional studies often rely on qualitative description or targeted feature comparison. More recent work has adopted corpus-based approaches, enabling large-scale quantitative analysis of morphosyntactic and lexical variation (Szmrecsanyi and Kortmann, 2009).

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 6340–6348
11–16 May 2026.

Major corpora such as the **International Corpus of English (ICE)**, the **Corpus of Global Web-Based English (GloWbE)**, and the **NOW Corpus (News on the Web)** have provided foundational resources for studying regional and national differences in English. These corpora offer broad genre coverage and standardized cross-national sampling frameworks. They enable comparison of grammatical constructions, lexical preferences, and discourse markers across L1 and L2 English contexts.

Unlike these general-purpose corpora, the present study adopts a domain-controlled design: English-language Olympic journalism. By restricting the topic domain while varying the regional editorial context, we reduce confounding genre variation and allow stylistic distinctions to be examined within a shared event framework. This complements existing corpora by focusing on intra-genre stylistic differentiation rather than broad cross-register comparison.

2.2. Stylometry, Feature-Based Modeling, and Embeddings

Corpus-based stylometry traditionally relies on frequency-based indicators such as type-token ratio, entropy, readability indices, and part-of-speech distributions (Pan et al., 2015; Hänlein, 1999; Zhou et al., 2023). While effective in capturing surface-level patterns, such measures may not fully reflect higher-order semantic or discourse variation.

Recent advances in machine learning and large language model embeddings enable high-dimensional representations of text that capture contextual and semantic nuance (Tao et al., 2024). However, embedding-based models are often criticized for limited interpretability.

This study combines both approaches: interpretable linguistic features and embedding-based representations. By comparing performance across feature-only, embedding-only, and hybrid models, we aim to assess not only classification accuracy but also the extent to which stylistically meaningful features contribute to distinguishing regionally anchored English media styles.

2.3. Media Discourse and Global Events

Media reporting of global mega-events such as the Olympics provides a natural testbed for variation research. Journalists from different national contexts report on the same competitions, athletes, and results, yet their language reflects distinct editorial traditions and audience expectations (Marlina and Xu, 2018; Ritchie et al., 2010).

Prior research on Olympic discourse has focused largely on qualitative framing analysis or cross-language comparisons. Few studies have applied reproducible computational methods to compare English-language reporting across multiple national contexts within a single event domain. This study addresses this gap.

3. Methodology

3.1. Corpus Construction

To investigate regionally anchored English media styles, we compiled English-language news articles covering the Olympic Games between 2020 and 2023, including:

- Tokyo 2020 (held 2021)
- Beijing 2022 Winter Olympics
- Pre- and post-event coverage during the Olympic cycle

Articles were collected from the following English-language outlets:

- **The New York Times** (United States)
- **China Daily** (China)
- **Sur in English** (Spain)
- **The Tico Times** (Costa Rica)
- **Mexico News Daily** (Mexico)

Outlets were selected based on:

1. English-language publication
2. Substantial Olympic coverage
3. National anchoring in distinct cultural contexts

To capture a diverse range of regional Englishes, we compiled the corpus from widely-read English-language news outlets related to Olympic Games in the United States, China, Spain, and Mexico between 2020–2023. 5 outlets were selected primarily based on (i) high readership in the respective country or region, (ii) English-language based publication, and (iii) comprehensive coverage of Olympic events, consist of The New York Times (American English), China Daily (Chinese English), Sur in English and Tico Times (Spain/Costa Rica - Spanish English), and Mexico News Daily (Mexico - Spanish English).

For analytical feasibility, outlets from Spain, Costa Rica, and Mexico were grouped under a broader Spanish-dominant context category. We emphasize that this grouping does not imply a unified linguistic system; rather, it reflects English-language journalism produced in Spanish-speaking national environments.

The resulting corpus (Table 1) contains approximately 1,810 articles (~ 1.0 million tokens). Articles

Media	News #	Token #
American English		
The New York Times	780	568,478
Chinese English		
China Daily	704	305,437
Spanish English		
Sur in English	153	63,253
TicoTimes	59	39,068
Mexican News Daily	114	51,198

Table 1: Statistics of the corpus in 3 English variants.

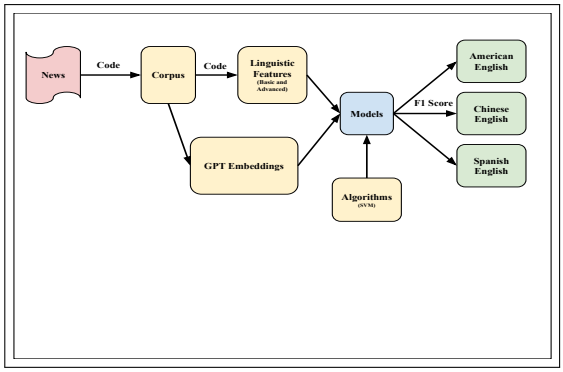

Figure 1: Model Setup and Flow

were filtered using Olympic-related keywords and publication tags. Due to limited metadata, individual journalist L1 background could not be identified; therefore, findings reflect institutional publication style rather than individual speaker competence.

All texts were preprocessed using Stanford CoreNLP (Song and Chambers), including:

- UTF-8 normalization
- Lowercasing and punctuation standardization
- Tokenization
- POS tagging
- Dependency parsing

Metadata and preprocessing scripts will be released upon publication to ensure replicability.

3.2. Linguistic Feature Extraction

We extracted 164 linguistically motivated features, organized into two tiers:

Basic Features (67)
Frequency-based indicators including

- Part-of-speech distributions (e.g., verbs, nouns, adjectives)
- Infinitive markers (POS tag: TO)
- Particles (POS tag: RP in Stanford CoreNLP)
- Function/content word proportions
- Type-token ratio (TTR)

Advanced Features (97) Higher-level stylistic metrics including:

- **Lexical Richness** (entropy, h-point, TTR variants)
- **Activity and Descriptivity Indices**, following (Heylighen and Dewaele, 1999), operationalized as ratios of verbs and adjectives relative to lexical items
- **Nominality**, defined as the proportion of nominal constructions and noun usage
- **Readability** (Flesch–Kincaid, SMOG, LIX)
- **Sentiment Metrics** (emotion frequency, polarity ratios)
- **Dependency Syntax Measures**, including mean dependency distance and dependency diversity

Feature definitions and extraction scripts are documented in the accompanying repository. The 164 features were selected through a theory-driven design rather than post hoc feature mining. Categories were determined based on prior stylometric and World Englishes research identifying lexical density, nominalization, readability, and syntactic complexity as key dimensions of register variation. No feature selection was performed based on test-set performance; instead, model-based filtering was applied only after full extraction to identify statistically meaningful features for interpretation.

Importantly, feature selection was theory-informed: categories were chosen based on prior stylometric and World Englishes research linking lexical density, nominalization, and readability to register variation.

3.3. Classification Modeling

We implemented four classification setups:

1. SVM with basic features
2. SVM with advanced features
3. SVM with GPT embeddings (text-embedding-3-large, 4096 dimensions)
4. SVM with combined embeddings + advanced features

Data were split using an 80/20 stratified train-test division.

While embedding-based models may capture latent topical signals (e.g., country-specific sports emphasis), to partially mitigate this concern, we

note that the feature-only SVM model — which relies exclusively on explicit linguistic features and not semantic embeddings — achieves a substantial macro F1 score (74.9). This suggests that stylistic signals independent of topical content contribute meaningfully to classification performance. Nonetheless, future work could incorporate topic-controlled subsampling or topic-modeling adjustments to further disentangle stylistic and content-driven variation.

Our primary objective is not merely to maximize classification accuracy, but to evaluate the extent to which explicit linguistic features contribute to distinguishing regionally anchored English media styles.

3.4. Statistical Analysis

In the final step, we aim to identify the features that are most discriminative of each English media style, and employed a two-step analytical process to filter both statistically significant and linguistically interpretable features:

- Model-based filtering: We only selected various features with high importance based on the SVM model, thus excluding those with weights or importance values <0.005.
- Statistical validation: we utilized ANOVA tests (with Bonferroni correction) to examine the statistical significance of the feature differences across 3 English media styles ($p < .05$).

By narrowing down to most significant features, the above approach contributes to improved model performance while allowing us to drive more meaningful linguistic interpretation as part of our analysis.

4. Experiments and Discussion

4.1. Experimental Setup

To ensure a robust evaluation of our methodology, we employed a standard 80/20 stratified split to create training, test, and verification sets, preserving the proportional representation of each regional English media style in both sets. Model performance was evaluated primarily using the F1-score (macro-averaged), supported by precision and recall, to provide a balanced assessment of classification accuracy across the three media style classes.

We have 4 models:

- Feature-based model with only basic linguistic features
- Feature-based model with only advanced linguistic features
- Embedding-enhanced model with only GPT embeddings
- Embedding-enhanced model with GPT embeddings and advanced features

4.2. Model Performance

We evaluated the corpus as a benchmark for English variant classification using the 4 models. Table 2 summarizes the performance measured by F1-scores for each approach.

The results show a clear trend: Feature-based models with only basic linguistic features achieved reasonable performance, confirming that fundamental metrics like verb and particle ratios contain robust signals of regional media styles. However, the model performance improved significantly when combined with advanced features, a strong evidence that those richer stylometric dimensions, such as readability indices, lexical richness, and syntactic diversity, contribute to the characteristic differences among English media styles.

In addition, we observed that the enhanced model with GPT embeddings alone delivered strong outcome, demonstrating the capacity of large language models to discern nuanced, context-sensitive stylistic distinctions between Englishes. Furthermore, with the combination of both GPT embeddings and advanced linguistic features, the model achieved the optimum result (F1 = 97.2). Although the margin over embeddings alone (96.9) is not huge, we would note the principal advantage is its interpretability. This model not only delivered best classification outcome, but also provides rich materials to analyze the specific features behind the classification, which we will elaborate further in the next section. Note the extremely low F1 score for Spanish English in the basic-feature model likely reflects corpus imbalance (Spanish subcorpus contains significantly fewer tokens). The classifier thus optimized decision boundaries primarily between American and Chinese classes.

This demonstrates that the corpus not only supports high-accuracy classification but also opens a bridge that connects black-box embeddings with those linguistic metrics that support convincing interpretations.

4.3. Feature-Level Analysis

The high classification accuracy confirms that regional English media styles in journalism are systematically distinct. Through feature importance analysis, we filtered the 33 most statistically significant linguistic features, and reviewed the values for American English, Chinese speaking English, and Spanish speaking English media styles in the Basic Linguistic Model and Advanced Linguistic

	Model + Linguistic Features	American	Chinese	Spanish	All
ML Based	Basic Features	74.6	72.7	0	66.9
	Advanced Features	80.4	81.2	34.8	74.9
LLM Based	GPT Embedding	96.3	97.5	96.1	96.6
	GPT Embedding + Advanced Features	97.2	98.2	95.2	97.2

Table 2: F1 Scores for each model combination

Model models. Tables 3 and 4 - The values shown in these tables represent normalized mean feature frequencies (per token basis) across each variant - highlight a subset of features that were most statistically significant, across 3 basic feature categories (verbs, infinitive markers, and particles) and 5 advanced feature categories (Lexical Richness, Activity & Descriptivity & Nominality, Readability, Sentiment Analysis, and Dependency Syntax).

Basic Features. From Table 3, we can observe notable, and in some cases dramatic, differences in certain basic language features across the three English media styles. We'd like to highlight 3 features including verb ratio, infinitive marker values, and the particle ratios:

- Verb ratio: Chinese (0.028) and Spanish speaking media English (0.026) exhibit a higher verb ratio, this may reflect either greater clause density or shorter clause segmentation (further qualitative validation would be required to confirm stylistic interpretation) than American English (0.022) (F(2,1807) = 17.156, p = .000).

- Infinitive marker: American (0.007) and Spanish (0.006) English use infinitives more frequently than Chinese speaking media English (0.002) (F(2,1807) = 18.152, p = .000), of which we can infer that in American and Spanish speaking media English, the use of infinitives is likely preferred to show purpose, intent, or result.

- Particle ratio: Chinese speaking English showed a significantly higher particle ratio (0.024) (F(2,1807) =50.536, p = .000), highlighting the important grammatical role of particles in this style.

Advanced Linguistic Features. From table 3, the advanced features provided a deeper, multidimensional view of stylistic differences.

- Lexical Richness: The Word Entropy values differ across the media styles with Spanish speaking English media value (0.462) is notably smaller than American (0.562) and Chinese speaking (0.568) English (F(2,1807) = 48.815, p = .000). This implies that American and Chinese speaking English have more diverse word choices. The Author's View (operationalized as the relative frequency of first-person references and evaluative stance markers, capturing the degree of explicit authorial presence) also varies across the three media styles, with the lowest being the Spanish speaking media style (0.131) (F(2,1807) = 6.421,p = .002). This implies that American English and Chinese speaking English emphasize authorial perspective slightly more. In h-point measure, American English achieves the highest value (3.305) (F(2,1807) = 72.613, p = .000), indicating that it has a more diverse vocabulary, avoiding reliance on common, high-frequency words. (Van Hout and Vermeer, 2007)

- Activity, Descriptivity, Nominality: Both Chinese speaking (0.068) and Spanish speaking (0.066) English media outlets have a higher activity and descriptivity than American (0.056) (F(2,1807) = 16.094, p = .000), an indication of a more action-oriented and vivid style compared to a more informational style assumed by American English. In addition, higher nominality value implies that Spanish speaking English media (0.062) F(2,1807) = 22.449, p = .000) relied more on noun usage, indicating a more formal style (Heylighen and Dewaele, 1999).

- Text Readability: The Flesch Kincaid Grade Scale (FKGS) measure shows large differences across the three variations, indicating that American English (6.812) (F(2,1807) = 32.769, p = .000) texts appear much easier to read, likely attributable to its focus on accessibility and simplicity. Lower American English SMOG value (3.080) (F(2,1807) = 91.495, p = .000) suggests American English assume a more straightforward style, also in line with its easier reading experience. Significant difference in LIX value and a much lower American English value (16.915) (F(2,1807) = 41.602, p = .000) translates to American English usage of shorter words and sentences to ensure easier comprehension (Štajner and et al., 2012).

- Sentiment Analysis: The American English has a lower emotion-freq value (0.036) than the Chinese and Spanish speaking English

Feature Category	Feature	American	Chinese	Spanish
Parts of Speech (uses Stanford CoreNLP, incl. NNS, NNP, NNPS, VB, VBD, VBN, CC etc)	Nouns	0.042	0.044	0.045
	Pronouns	0.024	0.023	0.023
	Verbs	0.022	0.028	0.026
	Adjectives	0.014	0.017	0.015
	Adverbs	0.016	0.014	0.014
	Determiners	0.015	0.018	0.016
	Numerals	0.014	0.015	0.016
	Prep./Conj.	0.015	0.007	0.016
	Possessive Markers	0.007	0.007	0.008
	Infinitive Markers	0.007	0.002	0.006
	Particles	0.003	0.024	0.003
Content Word Proportion (measures the proportion of one category of content words to the number of total words.)		0.023	0.026	0.025
Function Word Proportion (measures the proportion of one category of function words to the number of total words.)		0.02	0.023	0.021

Table 3: Basic Language Features and Model Output

Feature Category	Feature	American	Chinese	Spanish
Lexical Richness (refers to the quality of the vocabulary, includes TTR, CTTR, RTTR, LogTTR, h-point, alpha, word entropy, author's intent, etc)	Word Entropy	0.562	0.568	0.462
	Author's View	0.146	0.144	0.131
	HLP	0.08	0.099	0.094
	H-point	3.305	2.069	2.123
	TTR	0.097	0.073	0.085
Activity, Descriptivity, Nominality	Activity	0.056	0.068	0.066
	Descriptivity	0.056	0.068	0.066
	Nominality	0.055	0.058	0.062
Readability (measures how easy or hard to read, per complexity, syntax, and presentation)	Flesch Kincaid Grade Scale	6.812	12.603	13.947
	SMOG	3.08	4.786	4.829
	LIX	16.195	29.35	32.409
Sentiment Analysis (analysis of emotional words in a passage. e.g., neutral ratio, emotion freq, etc.)	Emotion_freq	0.036	0.044	0.041
	Positive_ratio	0.051	0.057	0.059
	Neutral_ratio	0.052	0.063	0.061
Dependency Syntax (aims to identify the relationship between words. Syntactical features include average dependency distance, dependency diversity, etc.)	Dependency Diversity	0.077	0.076	0.071
	det	0.014	0.017	0.015
	compound	0.025	0.032	0.03
	nsubj	0.016	0.018	0.017
	amod	0.013	0.016	0.015
	obj	0.009	0.01	0.01

Table 4: Advanced Language Features and Model Output

media (0.044, 0.041) ($F(2,1807) = 4.613$, $p = .000$), indicating that it assumes a less emotional style, which is reasonable given with its lower activity and descriptivity values. When examining the positive_ratio feature, we observe that the American English value (0.051) is lower than both Chinese and Spanish speaking English medias (0.057 and 0.059) ($F(2,1807) = 6.070$, $p = .002$). This implies that American English uses a more emotionless and informative style, whereas Chinese and Spanish speaking English media prefer more lively and positively emotional writing style (Taboada, 2016).

- Dependency Syntax: The mean Dependency Distance values differ numerically across variants (American: 0.908; Chinese: 0.297; Spanish: 0.333). However, this difference did not reach statistical significance ($F(2,1807) =$

2.132, p = .119). Therefore, while the pattern may suggest a tendency toward longer syntactic dependencies in American English media within this corpus, this observation should be treated as exploratory rather than conclusive. Slightly lower Dependency Diversity of Spanish speaking English media (0.071) (F(2,1807) = 45.802, p = .000) means that American and Chinese speaking English use a wider variety of grammatical relations, while Spanish speaking English media maintain a tighter range of such relations (Gibson, 2000).

4.4. Findings

By integrating the statistically validated feature differences, we can tentatively outline corpus-specific stylistic profiles for each regionally anchored English media style within Olympic journalism. These profiles reflect institutional tendencies observable in the sampled outlets and should not be interpreted as fixed or essential properties of national English varieties.

American English Media Style Within this corpus, American English journalism is characterized by comparatively lower activity and descriptivity indices (0.056), alongside lower nominality values. These patterns suggest a relatively informational and event-focused presentation style. Statistically significant readability measures further distinguish this variant: American English exhibits substantially lower Flesch–Kincaid, SMOG, and LIX scores, indicating shorter sentences and more accessible lexical choices.

Lexical richness measures (e.g., higher word entropy and h-point values) suggest broader vocabulary dispersion compared to the other variants. Sentiment-related measures show lower emotional frequency and positive ratios, consistent with a comparatively neutral reporting tone.

Although American English displays numerically higher mean dependency distance, this difference was not statistically significant; therefore, no firm conclusions about syntactic complexity can be drawn from that metric alone. However, higher dependency diversity does indicate a broader range of grammatical relations relative to Spanish-speaking English media.

Overall, within the Olympic journalism domain sampled here, American English media style appears comparatively accessible, informational, and lexically diverse, with limited overt emotional marking.

Chinese English Media Style

Chinese English media within this corpus demonstrates statistically higher activity and verb ratios, alongside elevated readability scores (Flesch–Kincaid, SMOG, LIX), indicating longer sentences and increased structural density. Higher word entropy suggests lexical variation, although a lower h-point value indicates greater reliance on high-frequency lexical items relative to American English.

Nominality measures are moderately elevated, pointing toward a more formal register. The significantly higher particle ratio (as defined by POS tag RP in Stanford CoreNLP) represents a distinct grammatical feature of this variant in the dataset.

Sentiment measures reveal slightly higher emotion frequency and positive ratio compared to American English, though these differences remain modest in magnitude.

Importantly, these findings reflect institutional journalistic practices in the sampled outlets during Olympic coverage rather than generalized properties of "Chinese English" as a sociolinguistic variety.

Spanish-Speaking English Media Style

The Spanish-speaking English media grouping (Spain, Costa Rica, Mexico) shows statistically elevated nominality values, suggesting comparatively greater reliance on noun-based constructions. Readability measures indicate the highest complexity among the three variants, with significantly higher Flesch–Kincaid, SMOG, and LIX scores.

Lexical richness indicators (e.g., lower word entropy and h-point values) suggest more concentrated lexical usage within this dataset. Activity and descriptivity measures are higher than in American English, indicating relatively more action-oriented reporting within the sampled texts.

Sentiment analysis shows higher emotion frequency and positive ratios relative to American English, pointing toward greater affective marking. Dependency diversity is slightly lower than in the other variants, indicating a narrower range of grammatical relations, though this should not be interpreted as reduced syntactic sophistication.

As noted earlier, this grouping reflects English-language journalism produced in Spanish-dominant national contexts and does not imply the existence of a unified "Spanish English" linguistic system.

Cross-Variant Interpretation

Across the three regionally anchored media styles, statistically significant differences emerge in lexical richness, nominality, activity, readability, sentiment, and selected dependency measures. These distinctions demonstrate that even within a shared global event domain, institutional editorial environments shape measurable stylistic tendencies.

However, these findings must be interpreted within the constraints of the corpus design. They represent domain-specific patterns in Olympic journalism between 2020–2023 and may not generalize to other genres or time periods. Additionally, while feature-only classification performance sug-

gests stylistic differentiation independent of topical content, residual topic effects cannot be entirely excluded.

Taken together, the results illustrate how computational stylometry can identify consistent, interpretable stylistic tendencies across regionally anchored English media contexts without presupposing essentialized national language varieties.

5. Conclusion

This study has demonstrated a quantitative framework for uncovering the distinctive stylistic fingerprints of regional English variants in global media. By integrating a comprehensive set of linguistic features with machine learning, we moved beyond high-accuracy classification (F1-score: 97.2) to deliver interpretable, data-driven profiles of American, Chinese, and Spanish speaking English media as used in Olympic journalism.

Our analysis revealed that these variants are characterized by statistically significant differences across fundamental dimensions of style. American English presents an informational and neutral profile, while Chinese speaking English is formal and analytical, and Spanish speaking English is emotional and expressive. These findings, enabled by our Olympic Journalism English Variants Corpus, provide a replicable model for quantifying stylistic variation.

We believe our findings have implications that extend beyond academic inquiry. For World Englishes research, our approach offers a bridge to connect qualitative description to quantitative comparison. In real world application, these insights can be used to facilitate content localization strategies, helping global media outlets tailor readability, tone, and style, therefore to enhance audience engagement. Furthermore, the ability to pinpoint specific features of a linguistic style could lead to more style-aware and personalized large language models (Flek, 2020), which could fine-tune their generative output to match a user's regional English media or stylistic preferences.

6. Limitations and Future Work

There are several avenues for potential future research, given the scope of this study:

The findings are based on the domain of Olympic journalism, which may employ specific conventions not representative of other genres like academic writing or everyday communication. Future work can apply this methodology to a wider array of textual sources, such as political reporting or financial news, to build a more comprehensive picture of English variation.

Although the Olympic domain reduces broad topical divergence, variation in national sports focus (e.g., differential coverage of specific athletes or events) may still introduce subtle content-based signals. While the strong performance of feature-only models indicates that stylistic markers play a substantial role, future work could employ topic modeling or balanced sport-category sampling to more strictly isolate stylistic variation from content effects.

In our research, each variant is represented by a limited number of outlets. Although these are influential sources, they may not capture the diversity of English use within their respective regions. Expanding the corpus to include a wider range of publications, including local newspapers, would strengthen representativeness of the variant profiles.

Our study currently only covers three regional media styles. A clear next step is to expand this analysis to other major World Englishes, such as Indian, other European, Australian, or African English, to enable broader comparative studies and validate the universality of the identified stylometric features.

Our results are based on a specific set of embeddings (GPT) and classifiers (SVM). Future research could benchmark a wider array of language models and neural architectures to further examine optimization result and explore the sensitivity of the linguistic insights to different computational tools.

7. Bibliographical References

Lucie Flek. 2020. Returning the n to nlp: Towards contextually personalized classification models. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 7828–7838.

Edward Gibson. 2000. The dependency locality theory: A distance-based theory of linguistic complexity. *Image, language, brain*.

Heike Hänlein. 1999. Studies in authorship recognition: a corpus based approach. *Journal of Computational Linguistics*, 13(3):221–235.

Francis Heylighen and Jean-Marc Dewaele. 1999. Formality of language: definition, measurement and behavioral determinants. *Interner Bericht, Center "Leo Apostel", Vrije Universiteit Brüssel 4.1*.

Braj B Kachru. 1990. World englishes and applied linguistics. *World Englishes*, 9(1):3–20.

Roby Marlina and Zhichang Xu. 2018. English as a lingua franca. *The TESOL encyclopedia of English language teaching*, pages 1–13.

Tom McArthur. 2001. World english and world englishes: Trends, tensions, varieties, and standards. *Language teaching*, 34(1):1–20.

Xiaxing Pan, Hui Qiu, and Haitao Liu. 2015. Golden section in chinese contemporary poetry. *Glottometrics*, 32:55–62.

Alastair Pennycook. 2009. English and globalization. In *The Routledge companion to English language studies*, pages 125–133. Routledge.

Brent W Ritchie, Richard Shipway, and P Monica Chien. 2010. The role of the media in influencing residents' support for the 2012 olympic games. *International Journal of Event and Festival Management*, 1(3):202–219.

Min Song and Tamy Chambers. Text mining with the stanford corenlp. *Measuring scholarly impact: Methods and practice*, pages 215–234.

Benedikt Szmrecsanyi and Bernd Kortmann. 2009. The morphosyntax of varieties of english worldwide: A quantitative perspective. *Lingua*, 119(11):1643–1663.

Maite Taboada. 2016. Sentiment analysis: An overview from linguistics. *Annual Review of Linguistics*, 2.1.

Zhen Tao, Dinghao Xi, Zhiyu Li, Liumin Tang, and Wei Xu. 2024. Cat-llm: Prompting large language models with text style definition for chinese article-style transfer. *arXiv preprint arXiv:2401.05707*.

R.W. Van Hout and A. R. Vermeer. 2007. *Comparing measures of lexical richness*. Cambridge University Press.

Haiyan Zhou, Yue Jiang, and Letao Wang. 2023. Are daojing and dejing stylistically independent of each other: A stylometric analysis with activity and descriptivity. *Digital Scholarship in the Humanities*, 38(1):434–450.

Sanja Štajner and et al. 2012. What can readability measures really tell us about text complexity. *Proceedings of workshop on natural language processing for improving textual accessibility*.