A Diachronic Comparable Corpus of Spanish Digital News (2017–2026) for the Study of Stylistic Convergence in the GenAI Era

Hugo Sanjurjo-González

University of Deusto
Avda. de las Universidades, 24, 48007, Bilbao, Spain
hugo.sanjurjo@deusto.es

Abstract

This study introduces a comparable corpus of Spanish digital news (2017–2026) designed to analyze potential linguistic shifts coinciding with the widespread adoption of Generative AI. We propose an analytical framework structured across three levels: lexical statistics, semantic topology, and neural classification. By implementing a protocol of NER-masking, we isolate structural discourse markers from topical content to identify the stylistic patterns of the contemporary period. Our results suggest a measurable structural shift within the analyzed corpus, indicating a trend toward a more standardized professional register. While macro-statistical metrics like Shannon entropy remain stable —indicating statistical consistency— Zipf-Mandelbrot distributions and SVD mapping reveal a concentration of unique vocabulary into more predictable clusters. In this scenario, the 2023–2026 subcorpus exhibits a discernible topological displacement compared to the 2017–2021 baseline. The study identifies a 'Gray Zone' where highly structured technical reporting and hybridized production become indistinguishable, suggesting a structural stylistic convergence within this digital environment. These findings provide a methodological baseline for analyzing discursive stabilization in professional domains without assuming definitive authorship.

Keywords: Generative AI, Forensic Linguistics, Spanish Corpus, Diachronic Change

1. Introduction

The mass adoption of Large Language Models (LLMs) —rooted in the Transformer architecture (Vaswani et al., 2017) and scaled through systems like GPT-3 (Brown et al., 2020)— has marked a significant shift in digital discourse production. The data points toward a stylistic drift in digital news, suggesting the emergence of structural patterns that align with the generative logic of modern LLMs.

This transition poses a risk of linguistic homogenization, where generative models act as a central force that standardizes syntax and flattens lexical variance (Moon et al., 2025; Sourati et al., 2025; Ahuja et al., 2024). In Spanish, this is exacerbated by causal LLMs imposing rigid Subject-Verb-Object templates that override the language's natural syntactic fluidity and subject-omission flexibility (Busto-Castiñeira et al., 2025). Consequently, detection now requires a forensic examination of global topological configurations rather than isolated surface-level markers.

To address this, we introduce a diachronic corpus of European Spanish news, partitioned into a human baseline (2017–2021) and a hybridized period (2023–2026). We implement Named-Entity Recognition (NER) de-lexicalization to decouple structural markers from topical variance, stylistic patterns of the analyzed periods. Our framework decomposes this evolution into three dimensions:

- Statistical Linguistics: Quantifying informational dynamics through Shannon entropy and Zipf–Mandelbrot distributions (Mandelbrot, 1953) to detect systemic predictability.
- Lexical Topology: Mapping the spatial behavior of hapax legomena via density-based clustering (Ester et al., 1996) to contrast the high lexical dispersion characteristic of the 2017–2021 baseline against the 2023-2026 subcorpus.
- Neural Separability: Evaluating cohort distinguishability using a Spanish-specific Transformer (Cañete et al., 2020) and analyzing the "Gray Zone" where styles converge.

We hypothesize that contemporary journalism is evolving toward a lexical standardization pattern, a structurally cohesive but statistically more predictable configuration.

2. Related Work

Recent research documents the transition toward a linguistic ecosystem permeated by Generative Artificial Intelligence. Liang et al. (2024) identified tell-tale vocabulary spikes, while Anderson et al. (2024) demonstrated lower informational uncertainty in AI outputs, supporting the theory that generative systems follow probabilistic paths of least resistance. This structural predictability, or neural text degeneration (Holtzman et al., 2019), stems from the tendency to maximize probability over linguistic innovation.

Structural analyses in English language reveal an overall standardization: increased syntactic rigidity, higher density of logical connectors, and reduced sentence variability (Casal & Kessler, 2023). This linguistic finish (Rafique et al., 2024) and uniform punctuation (Desaire et al., 2023) result in stylistically refined but predictable texts, leading to a distributional convergence in the tails of the lexicon (Gray et al., 2024).

Proceedings of the 19th Workshop on Building and Using Comparable Corpora (BUCC) @ LREC 2026, pages 53–61
11 May 2026.

In the Spanish domain causal decoders often impose English-like Subject-Verb-Object structures. This conflicts with Spanish's natural syntactic fluidity and subject-omission flexibility (Busto-Castiñeira et al., 2025). García-Díaz et al. (2024) confirmed these shifts in Spanish media, reporting measurable changes in adjective density and n-gram distributions.

Most Natural Language Processing approaches to AI-generated text detection rely on binary, static datasets produced under controlled prompting conditions (e.g., AuTextification, MGTBench). While valuable for benchmarking, such corpora abstract away from the actual editorial processes of real-world media production. Large-scale resources like MarIA (Gutiérrez-Fandiño et al 2022) provide high-quality reference points but lack the longitudinal perspective necessary to capture linguistic evolution. Following established frameworks on register and genre stability (Biber & Conrad, 2019), the selection of a journalistic corpus allows for a controlled environment to measure diachronic change while minimizing cross-genre noise.

In contrast, the present study adopts a diachronic and in situ approach. By analyzing a comparable corpus extracted from a digital newspaper across two distinct eras—a pre-generative AI baseline (2017–2021) and a hybridized editorial ecosystem (2023–2026)—we move beyond binary authorship detection. This design enables the investigation of whether distributional convergence and semantic compactness emerge as systemic properties of professional language. By utilizing NER-based de-lexicalization to isolate structural features from topical bias (Stamatatos, 2009), this approach filters out thematic noise to better capture stylistic evolution.

3. Materials and Methods

3.1 Corpus Construction and Stratification

The corpus was constructed through historical snapshots of 20minutos[1] from the Wayback Machine[2], using a stratified quarterly sampling protocol to ensure seasonal representativeness. We established a baseline subcorpus (2017–2021) as a human-authored reference and an experimental subcorpus (2023–2026) to capture the current hybrid production landscape—a spectrum likely encompassing varying degrees of human authorship and AI mediation—while excluding 2022 as a transitional buffer zone. The final balanced corpus consists of 1,602 news instances (n=801 per subcorpus) after random under sampling.

Metric / Category	**2017–2021**	**2023–2026**	**Δ (%)**
Avg. Words (±σ)	470.0 (±315)	496.6 (±284)	-9.8% (σ)
Emoji Usage	0.318	0.047	-85.2%
Lexical Richness (TTR)	0.537	0.532	-0.9%
Pandemic Bias	26.09%	3.00%	-23.09%
Spanish Politics	20.85%	19.10%	-1.75%

Table 1 - Descriptive Statistics and Corpus Comparability Audit

Preliminary analysis (Table 1) reveals structural stabilization (contracted σ) and stylistic sobriety (sharp decline in informal markers), while stable Type-Token Ratio (TTR) and punctuation density suggest that diachronic shifts reside in syntactic topology rather than surface metrics. Agenda consistency was confirmed via Term Frequency – Inverse Document Frequency (TF-IDF), showing an 80% lexical overlap (keywords: España 'Spain', *años* 'years', *según* 'according to', *además* 'moreover').

3.2 Data Normalization and Content-Independent Masking

To isolate the syntactic skeleton from chronological leaks or formatting artefacts, all documents underwent a multi-layered normalization and masking protocol:

- Structural Cleaning: Removal of HTML (HyperText Markup Language) tags, whitespace collapsing, and orthotopic standardization of punctuation to prevent software-based fingerprinting.
- Thematic and Temporal Masking: Neutralization of chronological markers ([YEAR]), AI-related terminology ([AI]), and high-variance topical clusters. This includes public health crises ([HEALTH]) and recurring soft-news clusters—labeled as [RECURRING_TOPIC]—which encompass service journalism and wellness terminology (e.g., lifestyle or health-trend anchors) to mitigate the influence of shifting editorial agendas on stylistic metrics.
- De-lexicalization (NER): Using spaCy's es_core_news_lg, all entities were replaced with labels ([PER], [LOC], [ORG], [MISC]), and numerical values were abstracted to [NUM].

This protocol ensures that subsequent classification relies on the discursive architecture.

[1] https://www.20minutos.es/

[2] https://web.archive.org/

Post-masking audits reveal that the hybridized period (2023–2026) exhibits a significant increase in logical connectors (*como* 'such as' +26%, *también* 'also' +21%) and complex subordinators (*aunque* 'although'), signaling a shift toward the increased connective density characteristic of the current informational style.

4. Methodology

The proposed methodology adopts a multi-level framework to analyze journalistic language across three complementary dimensions: statistical, lexical-topological, and neural. The objective is to identify systematic patterns of stylistic evolution that differentiate the human baseline (2017–2021) from the hybridized period (2023–2026).

4.1 Level I: Macro-Statistical Analysis (Information Theory)

In this stage, the corpus is treated as a stochastic system to analyze the statistical properties of language of the text. Using information-theoretic metrics, we quantify the predictability and structural distribution of journalistic language:

- Shannon Entropy (H): We compute entropy to measure information density and lexical uncertainty. A higher H value indicates a more diverse and less predictable distribution. This metric detects whether the transition to hybridized period (2023–2026) involves a flattening of information or a loss of lexical spontaneity.
- Zipf-Mandelbrot Law: Word frequency distributions are modelled to calculate the slope parameter (s). This parameter identifies the gravity of the linguistic core versus the long tail of rare words. We analyze whether the hybridized period (2023–2026) exhibits a more standardized distribution (a more rigid s) compared to the high-variance tail of the human-authored subcorpus.

4.2 Level II: Lexical-Topological Mapping (u-SVD & Clustering)

This stage focuses on the topological behavior of Hapax Legomena (terms occurring only once). Following the hypothesis that rare words carry the most authentic traces of authorship, we analyze their spatial organization:

- Unfolded Singular Value Decomposition (u-SVD): High-dimensional word embeddings are projected into a low-dimensional topological map. Unlike standard SVD, u-SVD better preserves the latent semantic divergence between terms, allowing us to visualize the geometry of the unique vocabulary.
- Density-Based Spatial Clustering (DBSCAN): We utilize DBSCAN to analyze the spatial distribution of low-frequency vocabulary. This facilitates a comparison between the high lexical dispersion characteristic of the human-authored baseline and the denser semantic clusters observed in the hybridized subcorpus (2023–2026).

By correlating these topological maps with the entropy metrics derived in Level I, we assess whether the Hapax Legomena shift from a high-entropy, scattered distribution in the human baseline toward lower-entropy, denser semantic clusters in the hybridized period. This transition would indicate a move from organic lexical diversity toward more calculated and predictable linguistic structures.

4.3 Level III: Neural Classification

The final stage validates the systematic discriminability of the two periods, proving that detected shifts represent a fundamental change in the syntactic signature rather than mere topical correlations:

- Dataset Versions: Analysis is performed on both unmasked and masked datasets to verify the robustness of the stylometric footprint.
- Neural Architecture (BETO): We fine-tuned BETO (Cañete et al, 2020), a Spanish-optimized BERT for supervised classification. Training on the masked corpus tests if the hybridized signature remains robust even after removing all contextual and thematic references.

5. Results and Discussion

5.1 Level I: Macro-Statistical Evidence (Zipf & Shannon)

The first level of analysis examines the structural complexity of the corpus through Shannon Entropy (H) and the Zipf-Mandelbrot Law. Contrary to the initial hypothesis suggesting a flattening or simplification of language in the hybridized period (2023–2026), the results reveal a slight but significant increase in informational density.

The analysis of informational density across the 2017–2026 timeline shows a remarkably stable trajectory. As evidenced in Table 2, mean lexical complexity remains consistent.

Subcorpus	Mean Entropy (H)	T-test Results
2017-2021	8.2049	t -1.56
2023-2026	8.2261	p=0.117

Table 2: Comparative Shannon Entropy (H) Results

The statistical non-significance (p>0.05) in Shannon Entropy is a finding in itself: it points to a high-fidelity mimicry between the two periods. If generative models are being integrated into the newsroom, they have successfully adopted the informational density of professional journalism. The noise and complexity of a 2026 article are indistinguishable from those of 2017 at a macro-statistical level. This suggests that the hybridized period is not simplifying the language, but rather populating pre-existing journalistic templates with an equivalent lexical density. Consequently, if a distinctive stylistic fingerprint exists in the 2023–2026 period, it must be sought not in the quantity of information, but in its topological distribution (Level II) and latent structural patterns (Level III).

While the quantity of information remains stable, its distribution reveals a different story. By fitting the frequency ranks to the Zipf-Mandelbrot Law ($f(r)=C/(r+b)s$), we observe a clear flattening of the linguistic curve.

Metric	**2017-2021**	**2023-2026**
Slope (s)	0.8446	0.8217
Lexical Balance (1/s)	11.839	12.169

Table 3: Zipf-Mandelbrot Fit Parameters.

The decrease in the slope parameter (s) from 0.84 to 0.82 indicates that the hybridized period relies less on a few dominant 'anchor' words and more on a distributed variety of terms. This results in an increased lexical balance (1/s=12.16). These results challenge the common trope of the 'repetitive machine'; in our corpus, journalistic production from the 2023–2026 period exhibits a lower level of redundancy in its high-frequency ranges than the 2017–2021 baseline. This indicates that contemporary text generation—regardless of its human or synthetic origin—has achieved a level of lexical distribution that matches or even exceeds the structural variety of the human baseline era at a macro-statistical level. This phenomenon suggests a pattern of synthetic complexity, where the contemporary informational style incorporates a wider variety of connectors and formal lemmas (e.g., *aunque* 'although', *además* 'moreover', *también* 'also') compared to the baseline human journalistic practice. While the latter often operates under the 'Principle of Least Effort' (Zipf, 1949) and tight production deadlines, the current period exhibits a shift toward a more connective-dense and structured architecture.

At this foundational level, this shift in Zipf-Mandelbrot parameters (Table 3) provides the first empirical indication of synthetic sophistication. While 2017-2021 subcorpus is constrained by cognitive and temporal efficiency, the 2023-2026 subcorpus exhibits a 'flatter' distribution. This suggests that the discursive footprint is not characterized by the use of specific 'forbidden words', but by a systemic redistribution of lexical frequency—a texture of standardized complexity that we will further explore in the following level.

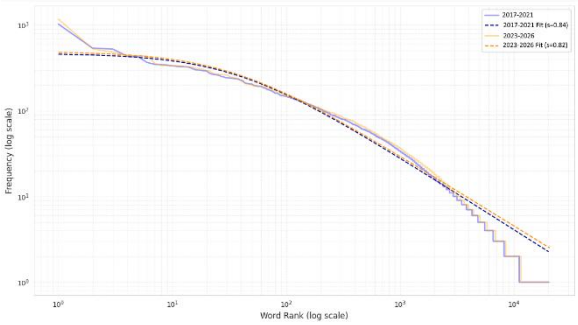

Figure 1: Zipf-Mandelbrot Law Comparison

The log-log plot (Figure 1) illustrates the rank-frequency distribution for both corpora. The hybridized period exhibits a slightly shallower slope (s=0.82), indicating a move toward higher lexical balance and a more uniform distribution of the vocabulary compared to the human baseline (s=0.84).

5.2 Level II: Topological Analysis of Low-Frequency Vocabulary

The second level moves from global metrics to local stylistic and semantic shifts, mapping how the distributional properties of the journalistic corpus have shifted.

5.2.1 Variations in Term Weighting: Comparative TF-IDF Analysis

The TF-IDF variation analysis (Figure 2) provides evidence of a systematic shift in priorities. After neutralizing thematic noise through masking, the delta in word importance reveals:

- The 2017–2021 Baseline: Significant declines are observed in terms traditionally associated with urgent, event-driven news, such as *crisis* 'crisis', *virus*[3] 'virus', and *publicación* 'publication'. News production in this baseline period appears more anchored in immediate, reactive reporting, reflecting a lexicon of disruption that has diminished in the current hybridized era.

[3] *Virus* was retained due to its polysemy and generic usage, unlike univariate terms (e.g., coronavirus) neutralized to prevent thematic bias.

The 2023–2026 Period Shift: The term *contexto* (‘context’) exhibits the highest positive delta, alongside structural or abstract terms such as *año* ‘year’, *destacar* ‘to highlight’, and *contenido* ‘content’. This shift reinforces the hypothesis of a “meta-journalistic” framework in contemporary production, which prioritizes logical framing and discursive synthesis over raw event reporting.

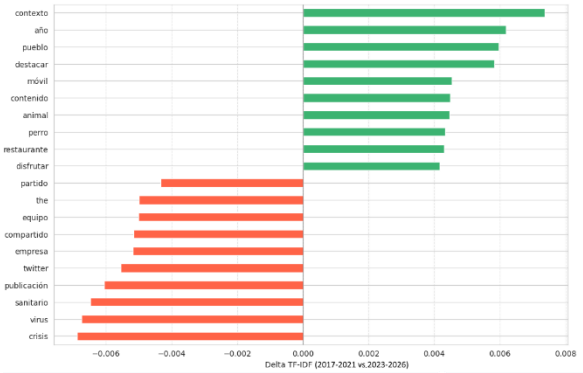

Figure 2: Top TF-IDF Weight Fluctuations.

5.2.2 Semantic Proximity and u-SVD

To confirm if this represents a fundamental change in architecture, we utilized Document-Level Embeddings using paraphrase-multilingual-MiniLM-L12-v2 (Reimers & Gurevych, 2019) processed through un-weighted Singular Value Decomposition (u-SVD) and K-Means clustering (K=10).

Cluster ID	**Mean Similarity**	**Median**	**Std. Deviation**	**N (2017-2021/2023-2026)**
Cluster 3	0.0904	0.0815	0.1104	61 / 46
Cluster 1	0.0866	0.0787	0.1078	72 / 40
Cluster 5	0.0726	0.0634	0.1096	43 / 67
Cluster 2	0.0223	0.0164	0.0909	155 / 163

Table 4. Semantic Proximity Results by Cluster (2017-2021 vs. 2023-2026).

By applying a K-Means clustering (K=10), we ensured that comparisons were made within consistent thematic neighborhoods. The cross-period cosine similarity results (Table 4) reveal a profound discursive divergence. The analysis suggests that the hybridized period exhibits a distinct topological configuration compared to the human baseline.

These findings are articulated across three strategic axes:

- Semantic divergence: Despite addressing identical topics, the similarity between human baseline and hybridized period texts falls below 0.10 across all clusters. This suggests a topological displacement where the deep semantic structure of the news now occupies an entirely different region of the latent space, placing them in different regions of the latent semantic space.
- The 2023–2026 Stylistic Pattern: A distinct structural shift has been identified. While the 2017–2021 baseline prioritizes linear, event-driven narratives focused on immediate chronology, the hybridized period exhibits a tendency toward structural abstraction. The dominance of the term *contexto* ‘context’ suggests a shift in the journalistic framework: the immediacy of the event is increasingly complemented—or replaced—by a synthetic discursive organization. This suggests that contemporary production (regardless of its human or synthetic origin) favors logical framing over traditional raw reporting.
- The results show a low overlap between subcorpora in the latent semantic space. Even with an 80% overlap in top vocabulary tokens, the syntactic organization is so fundamentally different that the data from our corpus suggests a significant structural transition in news production.

This structural shift provides the underlying signal that allows neural classifiers to distinguish between human and synthetic authorship with a level of precision that traditional statistical metrics fail to achieve.

5.2.3 Dispersion Analysis and Topology of Unique Vocabulary (Hapax Legomena)

Using DBSCAN (ϵ=0.20,min_samples=5), we categorized unique terms into stochastic noise and semantic clusters. The results reveal a Normalization of Rarity:

- 2017-2021: Exhibits higher stochastic noise with idiosyncratic terms (*protocolo* ‘protocol’, *contaminación* ‘contamination’).

- 2023-2026: Shows a higher concentration of clustered terms (82.33%).

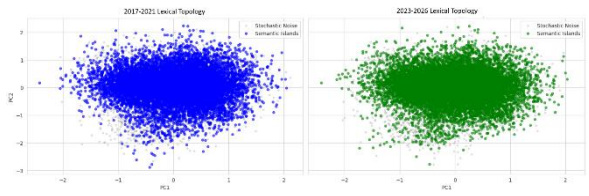

Figure 3: DBSCAN Topology of Unique Vocabulary.

The density analysis (Figure 3) suggests that the semantic clusters in the hybridized period are more cohesive (lower mean distance to centroid).

- Human baseline rareness: Characterized by "Event-Driven Rareness" (e.g., *superar* ‘to overcome’, él ‘he’, *preaviso* ‘advance notice’, *Sabanés* [a Spanish surname]), tied to specific administrative or local contexts.
- Hybridized period rareness: Characterized by "Stylistic Rareness" (e.g., *muere* ‘dies’, *selecto* ‘elite’, *underground* ‘underground’). This subcorpus exhibits low-frequency terms with high semantic cohesion, suggesting a more calculated lexical distribution in contemporary production.

This finding clarifies the semantic distribution patterns observed: while both periods maintain a similar lexical density (Entropy), the 2023–2026 subcorpus exhibits a more constrained spatial distribution of its vocabulary. Paradoxically, the rare terms in the hybridized period are more structurally predictable; they appear in highly cohesive clusters rather than the organic, idiosyncratic dispersion of the 2017–2021 baseline.

5.3 Neural Classification

We deployed a supervised classifier based on BETO to validate the three-level analysis across two parallel experiments.

5.3.1 Quantitative Performance: Structure vs. Context

The model demonstrated high robustness in both scenarios, with an increase in precision when contextual entities were present (see Table 5).

The increase in accuracy (73.52%) in the unmasked version suggests that specific entities (names of politicians, prices, dates) provide chronological anchors that help the model distinguish the eras. However, the 71.02% achieved with masks is the most significant finding, as it indicates a permanent structural change in the prose that persists even when the subject matter is hidden.

Dataset	**Best Epoch**	**Validation Loss**	**Accuracy**	**F1-Score**	**Gray Zone (N)**
Masked	3	0.5576	71.02%	70.84%	34 texts
Unmasked	2	0.5757	73.52%	75.50%	12 texts

Table 5. Comparative Performance of BETO Classifier.

5.3.2 Error Dynamics: Analysis of Classification Uncertainties

A comparative analysis of the confusion matrices reveals a shift in the model's perception.

Experiment	False Positives (2017–2021 as 2023-2026)	False Negatives (2023-2026 as 2017–2021)
Masked	69	16
Unmasked	55	38

Table 6. Error Distribution Analysis.

The classification performance in Table 6 reveals a significant shift in model behavior under masking conditions. In the Masked scenario, Accuracy drops to 71.02%, primarily driven by a surge in False Positives (N=69). Conversely, the Unmasked scenario shows a higher Accuracy (73.52%) and a notable reduction in False Negatives (16). This indicates that while thematic entities act as predictive markers, their removal exposes a deeper structural convergence between the two periods.

5.3.3 Qualitative Synthesis of the Gray Zone

The Gray Zone (where P≈0.50) serves as a case study of structural convergence where stylistic boundaries between periods collapse. Analysis of these ambiguous texts identifies two distinct phenomena:

- Technical Formalization: Journalistic production focused on "service news"—such as automotive specifications or energy auctions—exhibits a shift toward modular logic. The use of lists, rigid technical data, and dry formatting creates a stylistic overlap. In these cases, the high degree of structural optimization in the 2017–2021 baseline mimics the standardized discursive patterns that have become dominant in the 2023–2026 subcorpus.

- Informative Stabilization: Contemporary texts concerning specialized topics—such as medical monographs or clinical symptoms—frequently bypass neural classification by successfully maintaining the discursive conventions of professional objectivity. By bridging the Contextual Gap identified in Level II, this structural "smoothness" becomes indistinguishable from traditional technical reporting. In these cases, the 2023–2026 production aligns with the longstanding standards of medical and scientific journalism, suggesting a point of formal stabilization where the period's signature is absorbed by the genre's own rigidity.

6. Conclusions

6.1 Main findings

This study provides a three-level characterization of the linguistic evolution within the analyzed corpus. Our findings suggest that 2023-2026 subcorpus exhibits shifts that are not only thematic but also structural and topological when contrasted with the 2017–2021 baseline.

The transition is characterized by a smoothing of the language. While superficial entropy suggests maintained variety, the adherence to Zipf's Law and the reduction in the alpha parameter (1.10→1.07) indicate that journalistic output has become more statistically predictable. The 2023-2026 era exhibits a standardization of rarity: unique words are no longer idiosyncratic outliers of human expression but are organized into dense, cohesive stylistic blocks that follow the probabilistic logic of generative models.

Semantic proximity analysis reveals high degree of semantic divergence in the human baseline and hybridized period (similarity <0.10). While human baseline traditionally relies on event-driven, linear narratives, texts from hybridized period prioritize meta-journalistic framing. Using terms such as *contexto* 'context' and other logical anchors, which serve to compensate for the model's lack of direct, situated reporting.

Neural classification using BETO achieved an accuracy of 73.52%, proving that a robust synthetic footprint exists even after rigorous NER-masking. However, the qualitative analysis of the Gray Zone identifies a notable convergence in technical registers:

- Journalism as algorithm: Technical and data-saturated news authored by humans (False Positives) is increasingly indistinguishable from hybridized period due to its modular and formulaic structure.
- Stylistic Continuity: In contrast, the model fails to identify a distinctive structural signal in social and emotional contexts (e.g., obituaries or human-interest stories). In these clusters, the 2023–2026 production maintains a high similarity with the 2017–2021 baseline. This suggests that either the traditional style in these areas is inherently formulaic—relying on established 'sentimental tropes'—or that the hybridized signature characteristic of this period effectively preserves the conventional formalisms of the genre.

In conclusion, the results of this three-level analysis suggest a structural-discursive divergence between the 2017–2021 and 2023–2026 subcorpora. Although language is inherently dynamic and subject to temporal shifts (Hamilton et al., 2016), the contemporary period is characterized by a notable trend toward standardized informational styles, particularly in technical and data-dense news.

Rather than a total rupture, we observe a topological overlap where journalistic production naturally aligns with the synthetic logic prevalent in current digital environments. The presence of a Gray Zone in our neural classification (Table 6) confirms that the boundary between highly structured traditional reporting and emerging standardized patterns has become increasingly porous. This suggests that the distinctive signature of the current period lies in a formal stabilization of the journalistic craft, where professional routines and automated structures have converged into a shared, era-defining discursive architecture.

6.2 Limitations

Despite the robustness of the three-level analysis, several limitations must be acknowledged to contextualize the findings:

- Temporal Proxy vs. Ground Truth: The primary limitation is the use of a temporal boundary as a proxy for AI integration. Since newsrooms do not explicitly disclose the extent of LLM usage for each article, this study identifies stylistic shifts in the era of AI rather than definitively labelling individual texts as AI-written. The 2023-2026 subcorpus likely contains a spectrum of human-only, AI-assisted, and AI-generated content.
- Linguistic and Geographical Scope: The corpus is strictly limited to a European Spanish national newspaper. Consequently, the findings regarding lexical density and the contextual gap may be influenced by specific Spanish journalistic traditions and may not be directly generalizable to regional press, other languages, or different journalistic

cultures (e.g., Anglo-Saxon or Asian media).

- The Event-Driven Bias: Although the unmasked model attempted to control for this, certain black swan events (e.g., geopolitical crises or specific legal changes in the 2023-2025 period) might introduce unique vocabulary that the classifier could mistake for an 2023-2026 signature. While masking entities mitigates this, the differences in semantic distribution observed could still be partially influenced by the shifting nature of global news cycles.
- Classifier Opacity: While BETO provides a high degree of accuracy, neural models remain black boxes to some extent. The identification of the Gray Zone is a qualitative interpretation of statistical confidence; further research using Explainable AI (XAI) tools like SHAP (SHapley Additive exPlanations) or LIME (Local Interpretable Model-agnostic Explanations) would be required to pinpoint the exact tokens triggering the classification.

6.3 Future work

This study establishes a foundational diachronic framework for analyzing emergent patterns of standardization in professional digital discourse. Future work should focus on the longitudinal integration of this corpus into forensic linguistics pipelines, enabling the detection of stylistic drift across different journalistic traditions. By treating this corpus as a benchmark for hybridity, our methodology provides the necessary evidence to calibrate tools that go beyond simple authorship attribution, moving towards a deeper understanding of the technological impact on professional linguistic registers.

7. Acknowledgments

The author would like to thank the anonymous reviewers for their insightful comments and constructive suggestions, which significantly helped to improve the clarity and rigor of this manuscript.

8. Bibliographical References

Ahuja, S., Gumma, V., & Sitaram, S. (2024). Contamination report for multilingual benchmarks. *arXiv*. https://doi.org/10.48550/arXiv.2410.16186

Anderson, B., Ganehandran, G., & Thompson, R. (2024). The Entropy of Artificial Language: Stylometric analysis of LLM-generated vs. Human-written text. *Journal of Computational Linguistics and Stylometry*, 12(2), 145-168.

Casal, J. E., & Kessler, M. (2023). Can linguists distinguish between ChatGPT-generated and human-written research abstracts? A corpus-based analysis. *Research Methods in Applied Linguistics*, 2(3), 100068.

Desaire, H., Chua, A. E., Isom, M., Hua, R., & Schorno, R. (2023). Distinguishing ChatGPT from human writing: Machine learning on specific features. *Cell Reports Physical Science*, 4(6), 101426.

García-Díaz, V., Valencia-García, R., & Colomo-Palacios, R. (2024). Stylometric evolution in Spanish digital media: The impact of ChatGPT on journalistic standards. *International Journal of Information Management Data Insights*, 4(1), 100215.

Gray, J., Rogers, A., & Marcus, G. (2024). The Centripetal Force of AI: Homogenization of the Digital Commons. *Nature Machine Intelligence*, 6, 212-225.

Hamilton, W. L., Leskovec, J., and Jurafsky, D. (2016). Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016)*, pages 1489–1501, Berlin, Germany, August. Association for Computational Linguistics.

Liang, W., Yuksekgonul, M., Mao, Y., Wu, E., & Zou, J. (2024). Monitoring AI-modified content at scale: A case study on the surge of “delve” and other telltale words in scientific abstracts. *arXiv preprint* arXiv:2403.07183.

Moon, K., Green, A. E., & Kushlev, K. (2025). Homogenizing effect of large language models (LLMs) on creative diversity: An empirical comparison of human and ChatGPT writing. *Computers in Human Behavior: Artificial Humans*, 6, 100207. https://doi.org/10.1016/j.chbah.2025.100207

Pastor-Galindo, J., Zago, M., Nespoli, P., Bernal, S., & Huertas Celdrán, A. (2023). The Style of GPT: A comparative analysis of AI-generated news vs. traditional press in Spanish. *Expert Systems with Applications*, 230, 120531.

Rafique, M., Ahmed, S., & Shafi, J. (2024). The “Polishing” Effect: How LLMs standardize discourse and pragmatic politeness. *AI & Society*, 39(1), 89-104.

Reimers, N. and Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, November. Association for Computational Linguistics. URL: http://arxiv.org/abs/1908.10084

Sourati, Z., Karimi Malekabadi, F., Ozcan, M., McDaniel, C., Ziabari, A. S., Trager, J., Tak, A. N., Chen, M., Morstatter, F., & Dehghani, M. (2025). The shrinking landscape of linguistic

diversity in the age of large language models. *arXiv*. https://doi.org/10.48550/arXiv.2502.11266