Evaluating Vocabulary Usage in LLMs

Matthew Durward and **Christopher Thomson**
University of Canterbury
matthew.durward@pg.canterbury.ac.nz
christopher.thomson@canterbury.ac.nz

Abstract

In the rapidly evolving educational technology landscape, the potential applications and limitations of AI-generated content need greater scrutiny. This study explores the authenticity of AI-generated texts by comparing vocabulary usage between human-authored texts and those generated by AI across different registers, specifically in news and creative writing. Employing Vocabulary-Management Profiles (VMPs) for structural analysis and word keyness analysis to evaluate vocabulary frequency and dispersion, we reveal distinct patterns of text production. Our results demonstrate variation in vocabulary usage between human and AI-generated texts across registers, and shows how VMPs capture these differences effectively. These findings highlight the challenges Large Language Models (LLMs) face in mimicking human text generation and open some new avenues for examining characteristics of vocabulary use relevant to applications in education.

1 Introduction

We are navigating a transformative era, marked significantly by the integration of AI technologies into various aspects of daily life. This is particularly evident in the realm of language learning, where Large Language Models (LLMs) have become instrumental. LLMs find application across diverse sectors including education, healthcare, and research, showcasing their versatility and impact (Hosseini et al., 2023). The role of LLMs in language acquisition and written composition deserves special attention; it is claimed they offer substantial benefits to learners through personalized learning experiences, interactive prompts for questions and examples, and feedback on writing (Dao, 2023). This highlights the potential of LLMs to enhance the efficacy of language learning strategies significantly.

While the potential is certainly undeniable, a factor that is worth addressing is whether texts produced by LLMs - particularly in the form of examples generated in a learning environment - accurately represent what a learner is likely to observe in a real-world scenario. In particular, we aim to gain a better understanding of whether LLMs generate text with respect to different registers in a fashion similar to humans.

Previous research (AlAfnan and MohdZuki, 2023; Gómez-Rodríguez and Williams, 2023) provides some insight into the perceived 'style' and characteristics of LLM production. We narrow our scope to focus on attributes related to discourse and vocabulary, two adjacent concepts that we expect to differ by register. In particular, we are interested in how vocabulary is deployed within the structure of texts. Anecdotally, LLM text is often described as 'generic' or 'bland' in tone. Thus, we were motivated to understand the extent to which such differences are linked to lexical diversity and the rate, or sequencing, with which new vocabulary is introduced. To achieve this, we investigate human authored and machine generated texts through Vocabulary-Management Profiles (VMP). In their simplest interpretation, VMPs provide a linear representation, that can be graphically illustrated, representing the rate of newly introduced vocabulary through the progression of a text.

2 Related Work

2.1 AI text for Language Acquisition and Development

Language learners and teachers are enthusiastic about LLMs, but research on their pedagogical uses is still in infancy. Researchers Kostka and Toncelli (2023) highlight the opportunities of these systems in an English Language Learning setting and advocate for cross-collaboration between educators, students, and developers. We are at a point where LLM systems are being adopted in an ever-growing manner, and efforts are needed to

Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications, pages 266–282

understand what differentiates AI-generated text from authentic human-authored text and what consequences may flow from these differences in an educational setting. For instance, Vaccino-Salvadore (2023) outline areas of concern and ethical considerations when bringing systems like ChatGPT into the classroom for language learning, especially the bias and diversity constraints inherent in these systems. We must remember that systems like ChatGPT and similar LLMs derive their training data largely from the internet, potentially reproducing or amplifying cultural and linguistic biases, replicating dominant themes and linguistic patterns, and reducing the diversity of language compared to what is actually in use around the world (Ray, 2023).

Bringing LLMs into a language learning setting involves many considerations. While research (Baskara and Mukarto, 2023) highlights the potential benefits, such as the personalization, or generation, of authentic learning materials, more work on how these systems differ from human-generated text would be beneficial. We also note that recent work on the degradation of LLMs trained on synthetic data such as (Shumailov et al., 2023) and (Guo et al., 2023) suggests that small reductions in quality or diversity of learning materials can, if propagated, be catastrophic for language models. We must understand how text generated by LLMs differs from human-authored text to evaluate these synthetic materials properly for human language instruction.

2.2 Vocabulary Management Profiles

Vocabulary-management profiles provide a method for measuring the rate at which new vocabulary is introduced throughout a text and a convenient means of representing this graphically (Youmans, 1991, 1994). Prior to developing VMPs, Youmans's (1990) worked on conceptually similar graphic representations through the broader application of type-token vocabulary curves (TTVC), their derivations, and their estimation of vocabulary size. Type-token modelling- examining the ratio of the number of unique word (types) and number of collective words (tokens), are readily examined in the field of linguistics (Mitchell, 2015) and the area of language development and acquisition (Jarvis, 2013). Token curves approximate lexical diversity (LD) progression over time (or length of a text), but compared to VMPs, they offer a more generalized indication of lexical usage within a text. On observing a TTVCs, we can relate curves with a more

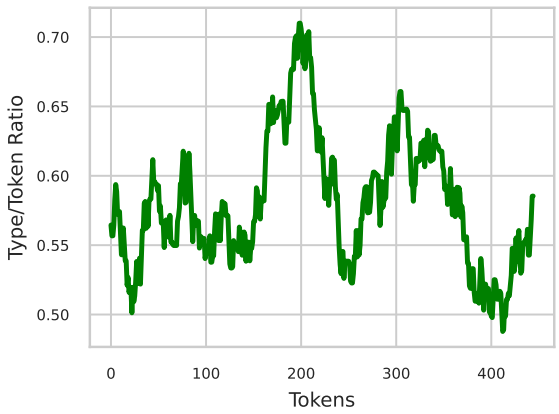

Figure 1: VMP Curve for a human-authored creative writing sample. The Type/Token Ratio (y-axis) averages individual word ratios of a moving window (set at 51 here) across a text, plotted against the token sequence (x-axis) of the text sample

pronounced increase in lexical diversity, whereas shallower curves denote a lower lexical diversity. VMPs improve on these earlier methods by observing the number of new (word-) types that occur in a moving window across the text. The difference is that VMPs aim to move beyond a static lexical assessment of the text as a whole and instead observe local patterns in the sequence of vocabulary use. Relating these structures to narrative or text structure, different slope trajectories can indicate factors signalling boundary points or "a new turn in the story" (Stubbs, 2006, p. 142). These turns can reveal the author's stylistic attributes and narrative structure. Indeed, they display a necessary storytelling component, balancing the inclusion of new words to help progress a story with the repetition of older words that help ensure text cohesion (Stubbs, 2006). Close analysis of these structural changes can expose how a writer navigates changes in topic and style or how diegesis relates to exegesis (Clement, 2013).

The advantage of investigating text in this manner is that it provides a structure resembling that found more traditionally in time-series data, which enables a flexible perspective of the scope of a text, peering into not only global trends but also narrowing the field into patterns that when they emerge can provide insight into representations of different groups of text. By observing an individual text under the lens of a VMP, relationships emerge as to the dynamics indicated through respective peaks and valleys "signaling the ebb and flow" of information in texts (Youmans, 1991, p. 4). Youmans sug-

gests that vocabulary used less frequently towards the end of moving intervals is often associated with introducing new topics. In contrast, vocabulary used more recently is likely to indicate the continuation of an existing topic. Evans (2020) refers to this as fractal patterning and attributes these formulations as evidence for the nested dynamics of self-similar attributes found from a global reference such as a novel with self-reciprocating emerging patterns of peaks and valleys, demarcated between narrowing orders of magnitude, as in sections, chapters, and paragraphs.

The degree of effectiveness of VMPs will vary with different applications. McKenny (2003) applied VMPs to ELL essays and observed their capacity to identify texts that follow stylistic choices of including new information as an inspirational factor for concluding remarks. However, because VMPs generalise patterns in the introduction or regularity of vocabulary use, they cannot substitute for context-specific qualitative analysis. For example, Meyer and Cooney (1994) found that VMPs benefit textual analysis by providing insight into the use of new and known information as measured by vocabulary but express limitations, particularly in the case of contextual usage, or *how* a word is used. This acknowledgement aligns with McKenny's (2003) positing the need for clear objectives when generalising about VMPs.

2.3 Word Keyness and Dispersion

In an educational context, the concept of 'Keyness' is closely aligned with creating word lists for language learning, emphasizing the strategic selection of high-frequency vocabulary. Nation (2006) highlights the value of these lists in planning vocabulary learning, a notion supported by further research (Nation, 2011). Vocabulary selection, tailored to learners' needs and specific domains like academia, plays a pivotal role. Such specialized lists, as Nation (2006) notes, are highly generalizable, proving effective across disciplines and extending to journalistic language. This effectiveness is further confirmed in English for Specific Purposes (ESP), showcasing the utility of targeted vocabulary strategies (Đurović, 2023).

To elucidate the variation in lexical usage across our corpora from a broad perspective, we implement two methods to measure a word's *keyness*. First, concerning word frequency, then second, we employ dispersion measures to discern between AI-generated and human-generated text. Dispersion assesses how evenly or unevenly a word is distributed within a corpus. Aiming to identify keywords that differentiate humans from AI-generated text, relying solely on frequency lists may fall short of offering a comprehensive understanding of vocabulary usage. Dispersion offers more profound insights into lexical patterns, as our analysis spans diverse sources (i.e., human and AI) and various genres (news versus creative writing). Given prior studies on source and genre variation (Biber, 1987; Kruger and Rooy, 2018), dispersion effectively provides a holistic view of vocabulary disparities. In contrast, the VMP analysis yields a more detailed, text-specific exploration of these differences.

Keyness broadly reflects a word's presence and significance in a corpus relative to its size, highlighting the word's distribution and importance (Jeaco, 2023). It is closely linked with dispersion, helping identify core vocabulary differences between corpora. Building on Egbert and Biber (2019)'s work on incorporating dispersion in keyword analysis, we apply Gries's (2021) method for a nuanced assessment of keyness. This method evaluates a word's frequency and dispersion to determine its unique role across corpora more accurately, avoiding biases introduced by frequency-based measures, such as the log-likelihood ratio. This approach enables a detailed comparison of vocabularies, offering insights into distinctive lexical patterns (Gries, 2021).

Our research aligns with the goals of authentic material matching used in a language learning context. Briefly, while there are competing notions, authenticity is described here as genuine language used in writing to communicate a meaningful message to a real audience, encompassing a wide variety of language (Gilmore, 2007; Morrow, 1977). There are numerous ways of measuring aspects of authenticity concerning discourse and lexical diversity, such as register variation multi-dimensional analysis (Biber, 2014) or linguistic feature extraction (Lee and Lee, 2023). By restricting our focus to vocabulary, we can disseminate variation in a manner easily processable by educators and learners. Often, overly complex systems with a multitude of features can add dimensions of entanglement, making it difficult for users to interpret results. VMPs are positioned to provide graphical representations that provide indications of the rate of introduced vocabulary where patterns are visually identifiable and computationally measurable. Through VMP and Keyness analysis, we can ex-

tract vocabulary information that can be conveyed intuitively, making identifying patterns readily understood by a spectrum of users.

3 Method

- RQ1: Are there distinguishable patterns of vocabulary usage across different text sources and registers?
- RQ2: In what ways does analyzing texts through VMPs uncover structural differences across sources and registers?
- RQ3: How do frequency and dispersion-based keyness analyses reveal vocabulary patterns across various text sources and registers?

We aim to investigate vocabulary usage through two distinct lenses. By implementing Vocabulary-Management Profiles (VMPs), we evaluate structural differences in writing patterns, shedding light on how texts from various sources and registers unfold. Then, through *word* keyness analysis, we qualitatively examine words associated with specific sources and registers to grasp salient differences through vocabulary usage better.

3.1 Data

This study investigates text under two dimensions of consideration: the source of the data (i.e. human or AI) and the register (i.e. news or creative). Data was retrieved from the DeepfakeTextDetect [1] dataset. Further details regarding the compilation of the initial dataset can be found in (Li et al., 2023). This combined dataset comprises eight different registers and text generated from 27 LLMs. We create a subset of extracted text from LLM sources, OpenAI gpt-turbo-3.5 and Meta LLaMA 65B, along with their human-generated counterparts. The length of a text plays a role in the observations of VMPs. Simply put, the longer a text is, the more observations can be extracted. Unfortunately, LLM prompts often generate texts well below what humans produce. We selected texts within a range of 400 to 500 tokens to strike a balance. Token counts are obtained after a preprocessing stage where words are converted to lowercase and punctuation is removed. Additionally, Youmans's (1991) found that further preprocessing modifications, such as removing affixes and conflating synonyms, have a minimal impact on the graphical representations of English discourses, so we refrained from any lemmatization or stemming procedures. This yielded 120 texts for each unique register/source combination, totalling 720 individual texts for analysis. Details of the corpus can be viewed in Table 1.

Source	Mean	Standard Deviation
creative_65B	448.70	29.62
creative_gpt	430.95	23.66
creative_human	450.47	28.75
news_65B	450.52	30.29
news_gpt	427.64	25.13
news_human	448.75	28.29

Table 1: Mean and Standard Deviation of Word Counts (in tokens) by Source.

3.2 Vocabulary Management Profiles

As discussed, VMPs can be thought of as a moving window through the progression of a text that measures the rate of newly introduced vocabulary. Youmans developed three methods for calculating VMPs (Youmans, "How to generate VMP 2.2s"); we use the VMP 2.2 method here. Our `vmp` function takes three parameters: `delta_x`, which is the size of the moving window; `cleaned_tokens`, a list of preprocessed tokens from the text; and for convenience, we specify `half_delta_x`, the middle value of the moving window, to be used for plotting. The function operates by sliding a window across the `cleaned_tokens`, giving each new vocabulary word a score of 1.0, and for each repeated word, determining a score using the following calculation: "(Number (index) of Current Word - Number (index) of Previous Occurrence - 1)/(Total tokens in the Text - 1)" (adapted from Youmans, "How to generate VMP 2.2s"). To ensure scores for the start of the text are consistent, the moving window centred on token 1 of the text 'wraps around' so that its first half covers the end of the text. This way, the VMP 2.2 measures vocabulary use at a 'second pass' through the text. (Youmans "How to generate VMP 2.2s").

Some considerations worth noting are the user-defined parameters. First, how a user wishes to treat common words and other preprocessing. We are interested in VMPs as a potential measure of stylistic and structural/topical changes, so we present results with common words retained (commonYes) and without (commonNo). Beyond this we have set aside investigation of the effects of different

[1] https://huggingface.co/datasets/yaful/DeepfakeTextDetect

kinds of preprocessing in the current study. An important parameter is the `delta_x` value. This value corresponds with the window size moving over each text. While Youmans (1991) suggests that longer `delta_x` values would be better suited for long-term patterns, it is also observed as having a smoothing effect on the trend of `delta_y` through a text. We examine a range of window sizes and suggest some additional smoothing techniques. A package to generate VMPs can be found at `github.com/matthewdurward/vmp`.

3.3 Keyness: Frequency and Dispersion

We investigate two independent properties related to vocabulary, *keyness* as it relates to frequency and again as it relates to dispersion. The combining of both is what Egbert and Biber (2019) describe as *key* keywords or words that demonstrate the collective power of both elements. To calculate keyness concerning frequency, we apply Gries's (2021) adaptation of Kullback-Leibler (KL) divergence to capture a word's association with a corpus. Equation (1) presents a generalized form of the Kullback-Leibler divergence, D_{KL} used to evaluate the extent of divergence between the conditional probabilities by observing a specific *word* in two corpora, compared to the overall probabilities within those corpora.

Equation (2) provides the calculation in application that measures how one probability distribution diverges from a second, expected probability distribution. In the context of text analysis, D_{KL} can be used to compare the distribution of word frequencies in one corpus or document (the "target") against another (the "reference"). A higher value of D_{KL} indicates a more significant divergence between the two distributions. If the divergence is zero, the two distributions are identical.

$$D_{KL}(p(\text{corpus} \mid \textit{word}) \,||\, p(\text{corpus})) \tag{1}$$

$$\left(a \times \log_2 \frac{a}{e}\right) + \left(b \times \log_2 \frac{b}{f}\right) \tag{2}$$

$$a = \frac{\text{Occ. of } \textit{word} \text{ in Target corpus}}{\text{Total Occ. of } \textit{word} \text{ in Target + Reference}} \tag{3}$$

$$b = \frac{\text{Occ. of } \textit{word} \text{ in Reference corpus}}{\text{Total Occ. of } \textit{word} \text{ in Target + Reference}} \tag{4}$$

In Equation (2), a signifies the relative frequency of a specific word in our target corpus (e.g., human news) compared to its presence in both the target and reference corpora (e.g., human news + GPT news) as illustrated in (3). Conversely, b indicates this word's relative frequency within the reference corpus (e.g., GPT news), also in relation to the combined target and reference, shown in (4). The variables e and f represent the proportion of all words in the target and reference corpora, respectively, to the total word count across both. The sign, or direction, of D_{KL} for frequency remains positive when the *word* in question prefers the Target corpus ($a > b$) and set to negative when the *word* prefers the Reference corpus ($b > a$). Thus, D_{KL} for frequency provides two aspects of consideration, the magnitude or strength of divergence and the direction of favorability for a corpus. Essentially, equation (2) quantifies how the distribution of a particular word differs between two textual datasets, helping to ascertain its distinctiveness or prevalence within one corpus as opposed to the other.

To compute dispersion, we adopt the methodology outlined by (Gries, 2021), utilizing the D_{KL} calculation previously employed to assess keywords for frequency. This method now serves as an analytical tool to gauge the distribution of a word's occurrence across different corpus segments, contrasting its distribution in one part of the corpus (target or reference) with the other parts. Applying this information-theoretic metric allows us to evaluate the frequency and spread of lexical items, providing nuanced insights into their usage patterns within and across corpora.

A normalization step is applied, $1 - e^{-D_{KL}}$, to transform the Kullback-Leibler divergence, D_{KL}, which can potentially range from 0 to ∞, into a value that falls within the closed interval [0, 1]. This transformation ensures that the dispersion measure is bounded and interpretable. Lower values of the normalized dispersion indicate less divergence from the expected distribution, whereas values closer to 1 suggest greater divergence.

3.4 Transformation and Dynamic-Time Warping

We revisit VMPs as a method for textual analysis, treating texts as time-series data to explore their dynamics using time-series analysis methodologies. To understand the stylistic and lexical variations across different sources and registers, we applied

Dynamic Time Warping (DTW). DTW functions as a measure of distance between two distinct text VMPs, with the progression of words representing time, and the type/token ratio of the VMPs serving as the unit of measurement. Recognizing the challenge posed by the variability and noise in raw VMPs, we preprocessed the data with wavelet denoising and Gaussian smoothing. This approach, employing the 'db1' wavelet for denoising and a sigma of 2 for smoothing, effectively minimized noise and highlighted long-term trends without sacrificing the VMPs' core characteristics.

These preprocessing steps clarified the VMPs for better interpretability and enhanced their analysis with DTW, allowing us to identify both subtle and pronounced differences in vocabulary usage. This nuanced examination, facilitated by signal processing techniques, affirms VMPs' utility for our context. Figure 2 illustrates an example of the comparative analysis of two separate pairs of individual text VMPs marked as similar (A) and dissimilar (B) using DTW. We calculated pairwise DTW distances of extracted VMPs both within the same register/source (such as creative/human) and between different sources but the same register (for example, news/65B compared to news/human). From these calculations, we derived distance matrices that were transformed into self-similarity measures. These measures are scaled between 0 and 1, where values closer to 1 indicate a higher similarity between a specific pair of Vocabulary Management Profiles (VMPs).

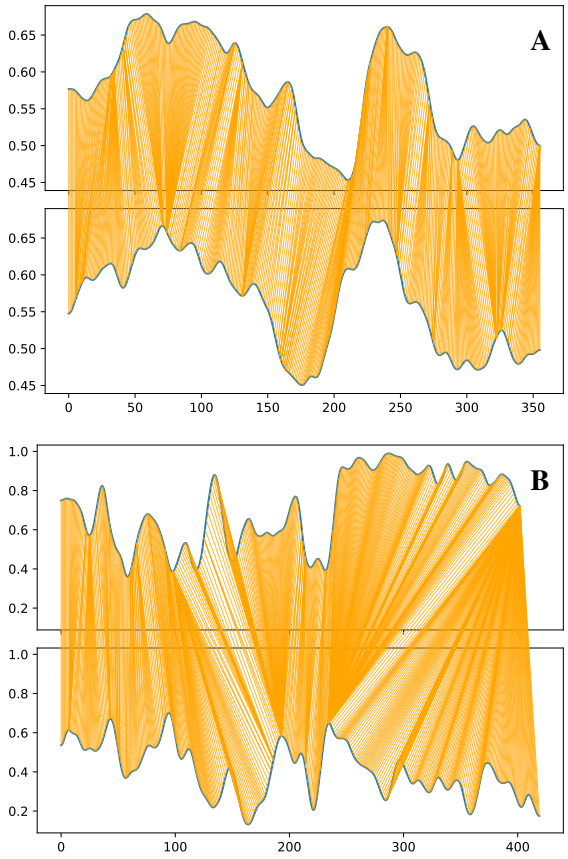

Figure 2: Dynamic Time Warping (DTW) visualizations illustrating the variability in VMP profiles for two pairs of example texts. Image (A) depicts a warping path with the minimal DTW distances, which suggests a closer similarity between 'creative-human' and 'creative-gpt' sample text sequences using a window of size 51. Image (B) presents a warping path with maximal DTW distance, where the orange lines exhibit more deviations, indicating substantial differences. This example uses a window size of 11. in the temporal patterns of 'news-human' and 'creative-65B' sample text sequences over a window of size 11. These paths reflect the level of adaptation required to align the sequences, with a more veritcal path implying less adjustment and a deviated path indicating more significant temporal distortion.

3.5 VMP Characteristic Features

Our analysis covers the interaction between different registers and sources, examining various conditions such as window sizes and the inclusion of common words. In this context, DTW serves as a method to quantify structural similarities. Beyond DTW, we further investigate time-series characteristics of the VMP themselves. To quantitatively assess the observed disparities, we employed three specific time series characteristic features: `DN_mean`, `DN_Spread_Std`, and `MD_hrv_classic_pnn40`. These features were derived using the *catch22* package, details of which can be found in the repository[2].

`DN_mean` computes the average Type/Token ratio across the series, serving as a measure of lexical diversity within the text. Higher values indicate a greater variety of words used. `DN_Spread_Std` measures the spread of the Type/Token ratios around the mean, quantifying the variability in lexical diversity across different text segments. Lastly, `MD_hrv_classic_pnn40` denotes the proportion of significant incremental changes within the series, effectively capturing the frequency and magnitude of fluctuations in lexical diversity. A higher value suggests more pronounced and rapid shifts in the Type/Token ratios, reflecting erratic changes in the VMP. Further details of features and extraction methods are described in (Lubba et al., 2019).

[2] https://github.com/DynamicsAndNeuralSystems/catch22

4 Results and Discussion

Initial observations comparing VMPs, as illustrated in Figure 3, reveal notable differences across all conditions of varying window sizes for both registers, represented in scenarios of excluding common words (Figure 6) and including common words (Figure 7).

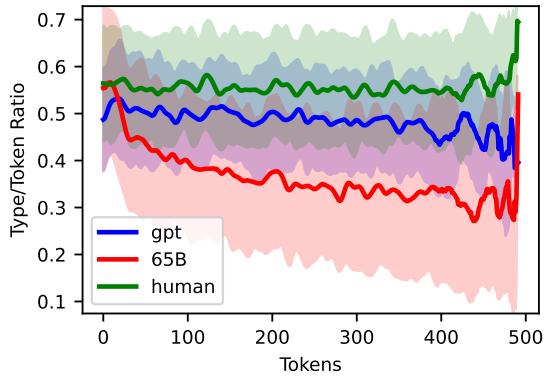

Figure 3: Sample VMP plot for the creative register texts by source including common function words with a window size of 11. It displays mean lines for the source groups, with variability indicated by shaded areas representing one standard deviation from the mean.

4.1 VMP Characteristics

To answer RQ1 and RQ2, statistical examination across various conditions, including common words and window size, has revealed significant distinctions among group sources for each of our tested VMP characteristic features. Our findings reported in this section exhibit highly significant p-values ($p < 0.00001$) for Kruskal-Wallis tests. Therefore, we emphasize the results with the most robust effect size, eta squared. From a broad perspective, human writing can be generalized as having higher lexical diversity represented through higher `DN_Mean` scores and higher consistent variability as demonstrated in `DN_Spread_Std`. Conversely, 65B demonstrates more sporadic episodes in texts with a generally higher `MD_hrv_classic_pnn40`. Notably, the most significant effect sizes were predominantly found in the news register, particularly for a window size of 25. For the feature `DN_Mean` in the commonNo category, a significant effect size of 0.4394 underlines a marked distinction primarily between 65B and human-generated texts, as well as between human and gpt variants. This difference points to the human-generated texts generally having higher Type/Token ratios than their counterparts.

Analyzing the `DN_Spread_Std` feature within the context of news content, particularly for the Delta 9, commonNo condition, provides insight into the variability of textual production across different sources. The effect size of 0.1955 indicates substantial variability differences among the groups, particularly between GPT and human VMPs. Posthoc comparisons further elucidate the nature of these differences: while both comparisons involving the 65B model (against GPT and human) showed significant results, indicating 65B's distinct variability profile, the direct comparison between news_gpt and news_human did not reach statistical significance (p=0.3882). The `MD_hrv_classic_pnn40` feature further highlighted significant disparities, most notably in the news content for Delta 35, commonNo, with an effect size of 0.0815, particularly evident in comparing news for 65B and GPT.

4.2 DTW Based Similarity for VMPs

To provide a broader perspective on the variations in distributions of VMPs, we transformed DTW distance scores between pairs of VMPs into self-similarity scores. This approach facilitates a comparative analysis of textual characteristics across different registers and sources, visualized in Figure 4 and further detailed by condition in Figure 8. Our analysis reveals that human-generated texts, particularly in the news register without common words and with a window size of 25, consistently demonstrate the highest values for our tested features, underscoring the distinctiveness of human linguistic patterns compared to those generated by AI models such as GPT and 65B.

To assess the statistical significance of observed differences between the creative and news registers within each source, we conducted Mann-Whitney U tests. Given the multiple comparisons made, we applied the Bonferroni correction. Our results showed highly significant differences between registers for all sources, with all adjusted $p < 0.00001$, demonstrating robust disparities. While it was expected that there would be differences between registers for our source, our attention relates to the effect size of our comparisons.

The effect size for these differences was quantified using the rank biserial correlation, which emphasizes the direction and magnitude of disparity between registers of the same source. This ap-

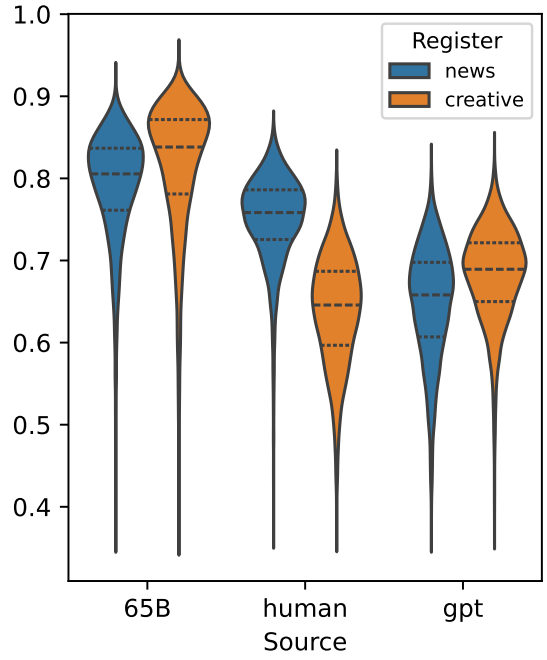

Figure 4: Distribution of DTW converted self similarity scores by source type for VMPs with a window size of 11, using all vocabulary. The violin plots illustrate score distributions across creative and news registers for 65B, human, and GPT sources, emphasizing the variations within each source and the distinctions between registers. Higher values indicate greater self similarity.

proach highlights each register and source's distinct linguistic features and VMP characteristics. As we can see in Figure 4, which shows results for the condition of a window size of 11 and no filtering of common words (CommonNo), we note that there are noticeable differences between news and creative VMPs for our sources.

A noteworthy observation pertains to the self-similarity within registers for each source group. Specifically, the human source group exhibits greater self-similarity and a more concentrated distribution for the news register, in contrast to a broader distribution and lower self-similarity for the creative register. Conversely, the 65B and GPT sources exemplify an opposite trend, with variations in self-similarity and distribution patterns. Table 2 indicates the most pronounced disparities are observable within the human source category, which consistently demonstrates the most substantial effect sizes, as denoted by r, thereby indicating a distinct separation between the news and creative registers. This distinction underscores the efficacy of text VMPs in differentiating between registers. Another focal point is the directional tendency of the correlation.

Source	r **(CommonYes)**	r **(CommonNo)**
	Interval: 9	
65B	-0.076	-0.317
Human	**0.945**	**0.654**
GPT	0.038	-0.410
	Interval: 11	
65B	-0.065	-0.291
Human	**0.941**	**0.827**
GPT	-0.010	-0.299
	Interval: 25	
65B	0.005	-0.077
Human	**0.657**	**0.586**
GPT	0.039	-0.186
	Interval: 35	
65B	0.017	-0.042
Human	**0.543**	**0.395**
GPT	-0.007	-0.128
	Interval: 51	
65B	0.058	0.019
Human	**0.464**	**0.334**
GPT	-0.053	-0.171

Table 2: Effect Sizes by Source and Condition. Note: r denotes the rank biserial correlation used as the effect size measure. Greater deviation from zero equates to larger disparity between registers for a particular source. Negative values indicate an opposite direction in polarity between registers for a source compared to human VMPs.

As Table 2 indicates, a discernible relationship exists with the window size employed for calculating text VMPs. It is important to note that larger window sizes correlate with identifying broader patterns within a text, whereas smaller windows are sensitive to finer-grained distinctions. A consistently higher effect size is attributed to the human source throughout the range of window sizes tested, indicating a more pronounced differentiation capability. Notably, after an initial increase from a window size of 9 to 11, the effect size for the human source gradually declines towards a window size of 51. In contrast, the 65B and GPT sources demonstrate comparatively weaker effect size strengths and fluctuate in directional tendency across varying window intervals. Comparing results with and without common words removed suggests that the more apparent register differentiation in human writing is consistent when considering both lexical and grammatical words.

4.3 Word Keyness

To answer RQ3, we extracted the top 100 *key* keywords by applying Equation (2) to distinguish between our corpora demonstrated in Table 3 for creative and Table 4 for news. Upon first inspection, there appears to be a notable propensity to use colourful language in the form of profanity, which is evident in human creative writing but absent in creative output from GPT. However, this becomes less apparent when comparing humans to 65B. Comparing the creative register, we notice a pronounced affinity towards darker thematic language expressed in human writing. Words such as: *bloody, die, torture, cry,* and *hate* are clear exemplars of this notion represented in human samples to GPT vocabulary usage. Conversely, GPT utilizes what can be described as more optimistic language, examples including: *succeeded, grateful, determined*. Some of these variations resonate between humans and 65B, but to a lesser extent. Words of interest would be aggressive or action words, such as: *threat, slammed, battle* indicating themes of conflict, whereas 65B demonstrates a polarity with humans through positive words of emotional tone, as in: *team, community, friendship*. Pulling back, we also see contractions, through the letter *d*, for human writing and when coupled with the presence of pragmatic markers *oh, uh, ah*, we can speculate on stylistic cues used by humans to signal variation in character speech, an aspect less prominent in our AI samples. Diverting our attention towards the news register, the LLMs tend to have more abstract and longer words, whereas humans tend to use more concrete and shorter words. A caveat to note here is that many of the human keywords relate to reports of events (sports results, financial results). We speculate that LLMs do not generate these (or not as much) because to do so they must start inventing specific facts. So, the keywords might reflect how LLMs are tuned to avoid levels of detail about the world that they cannot accurately emulate.

5 Conclusion

This study combined more seasoned and newer approaches for evaluating vocabulary usage between human and AI-generated texts. We noted structural differences in text sources, particularly in how VMPs respond to our research queries about discernible vocabulary patterns. Using distributional moment features like mean and standard deviation; we pinpointed statistical disparities between groups under various conditions, such as window size and vocabulary inclusion. By converting DTW distances into self-similarity measures, we observed marked differences in distributions by register for specific sources. These measurable variations underscore the distinct structural patterns of VMPs generated from different sources. Further investigation, particularly in response to RQ3, uncovered specific vocabulary that served as key indicators of thematic variations related to emotional tone. Understanding these variations can help educators and language learners select materials that best align with their learning objectives. We envision a scenario where aspects of LLM-produced text with lower mean VMPs could be combined with derived word keyness features to seek out text samples that incorporate desired vocabulary and appropriate repetition, an advantage for learning new vocabulary.

Limitations

This study takes a nuanced view of using Large Language Models (LLMs) in language learning settings. We do not oppose their use, as we recognize that there is support for such applications, and their use should align with educators' and learners' educational goals and objectives. However, we also note limitations in text selection. We acknowledge that register can be a fluid quality, and variations within a register may not be fully captured by the data used in our analysis. Moreover, although our dataset is balanced in terms of sample count, achieving a perfect balance in token length poses challenges. While truncating texts is a feasible approach, it's crucial to consider that details at the end of passages may reveal unique attributes of the sources.

AI-generated text was derived from default configurations. While adjusting parameters such as temperature or top-p could influence outcomes, we opted to examine versions which users will most likely encounter in educational settings. Our goal was to establish a baseline understanding of unaltered text production by LLMs, with plans to investigate the impact of varying parameters in future research. Gaining a deeper understanding of the production limitations of both sources can guide future research towards making LLMs more representative of human language. This insight can also effectively leverage LLMs' potential advantages.

References

Mohammad Awad AlAfnan and Siti Fatimah MohdZuki. 2023. Do Artificial Intelligence Chatbots Have a Writing Style? An Investigation into the Stylistic Features of ChatGPT-4. *Journal of Artificial Intelligence and Technology*.

Risang Baskara and Mukarto. 2023. Exploring the Implications of Chatgpt for Language Learning in Higher Education. *Indonesian Journal of English Language Teaching and Applied Linguistics*, 7(2):343–358. ERIC Number: EJ1391490.

Douglas Biber. 1987. A Textual Comparison of British and American Writing. *American Speech*, 62(2):99–119. Publisher: [Duke University Press, American Dialect Society].

Douglas Biber. 2014. Using multi-dimensional analysis to explore cross-linguistic universals of register variation. *Languages in Contrast*, 14(1):7–34.

Tanya Clement. 2013. Text Analysis, Data Mining, and Visualizations in Literary Scholarship.

Xuan-Quy Dao. 2023. Performance Comparison of Large Language Models on VNHSGE English Dataset: OpenAI ChatGPT, Microsoft Bing Chat, and Google Bard. ArXiv:2307.02288 [cs].

Jesse Egbert and Douglas Biber. 2019. Incorporating text dispersion into keyword analyses. *Corpora*, 14(1):77–104.

D. Reid Evans. 2020. On the fractal nature of complex syntax and the timescale problem. *Studies in Second Language Learning and Teaching*, 10(4):697–721.

Alex Gilmore. 2007. Authentic materials and authenticity in foreign language learning. *Language Teaching*, 40(2):97–118. Publisher: Cambridge University Press.

Stefan Th Gries. 2021. A new approach to (key) keywords analysis: Using frequency, and now also dispersion. *Research in Corpus Linguistics*, 9(2):1–33. Number: 2.

Yanzhu Guo, Guokan Shang, Michalis Vazirgiannis, and Chloé Clavel. 2023. The curious decline of linguistic diversity: Training language models on synthetic text.

Carlos Gómez-Rodríguez and Paul Williams. 2023. A Confederacy of Models: a Comprehensive Evaluation of LLMs on Creative Writing. ArXiv:2310.08433 [cs].

Mohammad Hosseini, Catherine A. Gao, David Liebovitz, Alexandre Carvalho, Faraz S. Ahmad, Yuan Luo, Ngan MacDonald, Kristi Holmes, and Abel Kho. 2023. An exploratory survey about using ChatGPT in education, healthcare, and research. Pages: 2023.03.31.23287979.

Scott Jarvis. 2013. Capturing the Diversity in Lexical Diversity. *Language Learning*, 63(s1):87–106.

Stephen Jeaco. 2023. How can we communicate (visually) what we (usually) mean by collocation and keyness?: A visual response to Gries (2022a). *Journal of Second Language Studies*, 6(1):29–60.

Ilka Kostka and Rachel Toncelli. 2023. Exploring Applications of ChatGPT to English Language Teaching: Opportunities, Challenges, and Recommendations. *Teaching English as a Second or Foreign Language–TESL-EJ*, 27(3).

Haidee Kruger and Bertus Van Rooy. 2018. Register variation in written contact varieties of English: A multidimensional analysis. *English World-Wide. A Journal of Varieties of English*, 39(2):214–242.

Bruce W. Lee and Jason Lee. 2023. LFTK: Handcrafted Features in Computational Linguistics. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 1–19, Toronto, Canada. Association for Computational Linguistics.

Yafu Li, Qintong Li, Leyang Cui, Wei Bi, Longyue Wang, Linyi Yang, Shuming Shi, and Yue Zhang. 2023. Deepfake text detection in the wild.

Carl H. Lubba, Sarab S. Sethi, Philip Knaute, Simon R. Schultz, Ben D. Fulcher, and Nick S. Jones. 2019. catch22: CAnonical Time-series CHaracteristics. *Data Mining and Knowledge Discovery*, 33(6):1821–1852.

John McKenny. 2003. Seeing the wood and the trees: Reconciling findings from discourse and lexical analysis. In *Paper at Corpus Linguistics 2003 Conference, University of Lancaster. University Centre for Computer Corpus Research on Language (UCREL). Technical Papers*, volume 16.

Jim Meyer and Brendan Cooney. 1994. The paragraph: Towards a richer understanding. *Work Papers of the Summer Institute of Linguistics, University of North Dakota Session*, 38(1).

David Mitchell. 2015. Type-token models: a comparative study. *Journal of Quantitative Linguistics*, 22(1):1–21. Publisher: Routledge _eprint: https://doi.org/10.1080/09296174.2014.974456.

Keith Morrow. 1977. Authentic texts and esp. *English for specific purposes*, 13:17.

I. Nation. 2006. How Large a Vocabulary is Needed For Reading and Listening? *The Canadian Modern Language Review*, 63(1):59–82. Publisher: University of Toronto Press.

I. S. P. Nation. 2011. Research into practice: Vocabulary. *Language Teaching*, 44(4):529–539.

Partha Pratim Ray. 2023. ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems*, 3:121–154.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Yarin Gal, Nicolas Papernot, and Ross Anderson. 2023. The curse of recursion: Training on generated data makes models forget.

Michael Stubbs. 2006. READING D: Exploring 'Eveline' with computational methods. In Sharon Goodman and Kieran O'Halloran, editors, *The art of English: literary creativity*, pages 138–144. Palgrave Macmillan ; In Association with the Open University, Basingstoke [England] ; New York : Milton Keynes, UK. OCLC: 63705884.

Silvia Vaccino-Salvadore. 2023. Exploring the Ethical Dimensions of Using ChatGPT in Language Learning and Beyond. *Languages*, 8(3):191. Publisher: MDPI AG.

Gilbert Youmans. 1990. Measuring Lexical Style and Competence: The Type-Token Vocabulary Curve. Accepted: 2009-02-05T21:48:19Z Publisher: Northern Illinois University.

Gilbert Youmans. 1991. A New Tool for Discourse Analysis: The Vocabulary-Management Profile. *Language*, 67(4):763–789. Publisher: Linguistic Society of America.

Gilbert Youmans. 1994. The Vocabulary-Management Profile: Two Stories by William Faulkner. *Empirical Studies of the Arts*, 12(2):113–130. Publisher: SAGE Publications Inc.

Zorica Đurović. 2023. Frequency or Keyness? *Lexikos*, 33.

A Appendix

Algorithm 1 Calculate Dispersion of a *word* in a Corpus

Require: A corpus divided into N parts, the *word* in question, $N \geq 1$

Ensure: Dispersion value of the *word* in range [0,1]

0: Let N be the number of parts, $N = 10$
0: Initialize array F to store the frequency f_i of the *word* in each part i
0: Initialize array S to store the size s_i of each part i
0: Initialize D_{KL} to 0
0: **for** $i = 1$ to N **do**
0:
0: $p_i \leftarrow \dfrac{f_i}{\sum_{j=1}^{N} f_j}$
0:
0: $q_i \leftarrow \dfrac{s_i}{\sum_{j=1}^{N} s_j}$
0:
0: **if** $p_i > 0$ **then**
0: $D_{KL} \leftarrow D_{KL} + p_i \times \log_2 \left(\frac{p_i}{q_i}\right)$
0: **end if**
0: **end for**
0: Dispersion $\leftarrow 1 - e^{-D_{KL}}$
0: **return** Dispersion =0

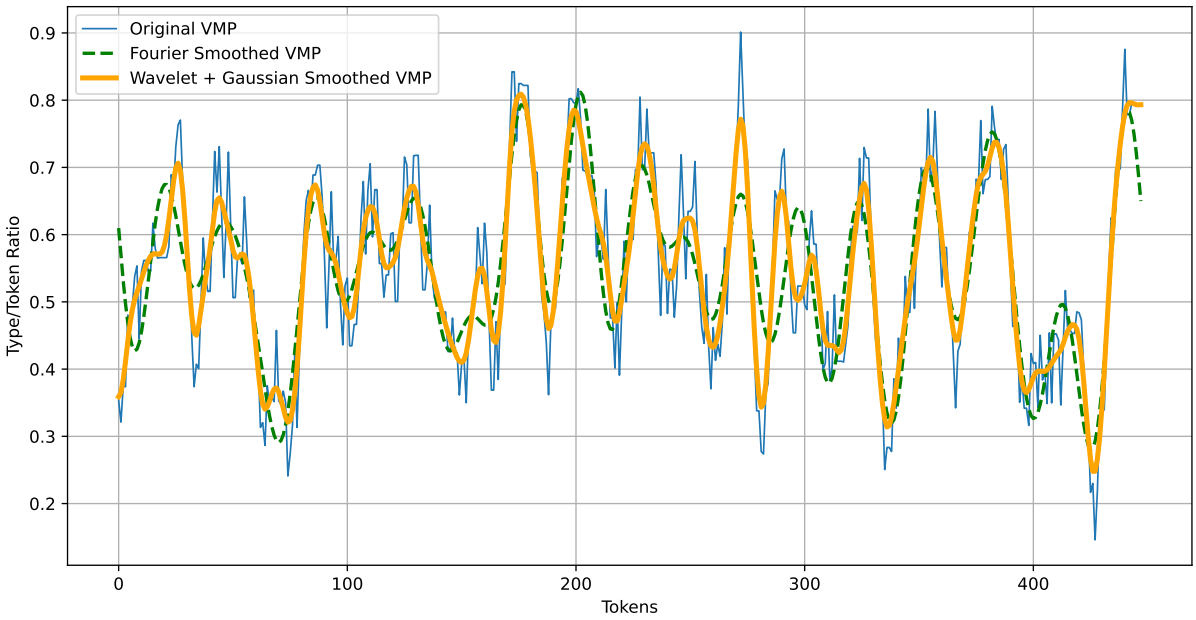

Figure 5: Smoothing transformation of news text sample from a human source for commonNo vocabulary condition with a window delta value 9. The blue solid line represents the original VMP data exhibiting natural variability and noise. The green dashed line shows the VMP data after Fourier transform-based smoothing, which reduces high-frequency fluctuations while preserving the main signal trend. The orange solid line, bolder for emphasis, displays the VMP data subjected to a two-stage smoothing process involving wavelet denoising followed by Gaussian smoothing, offering a balance between noise reduction and signal integrity preservation.

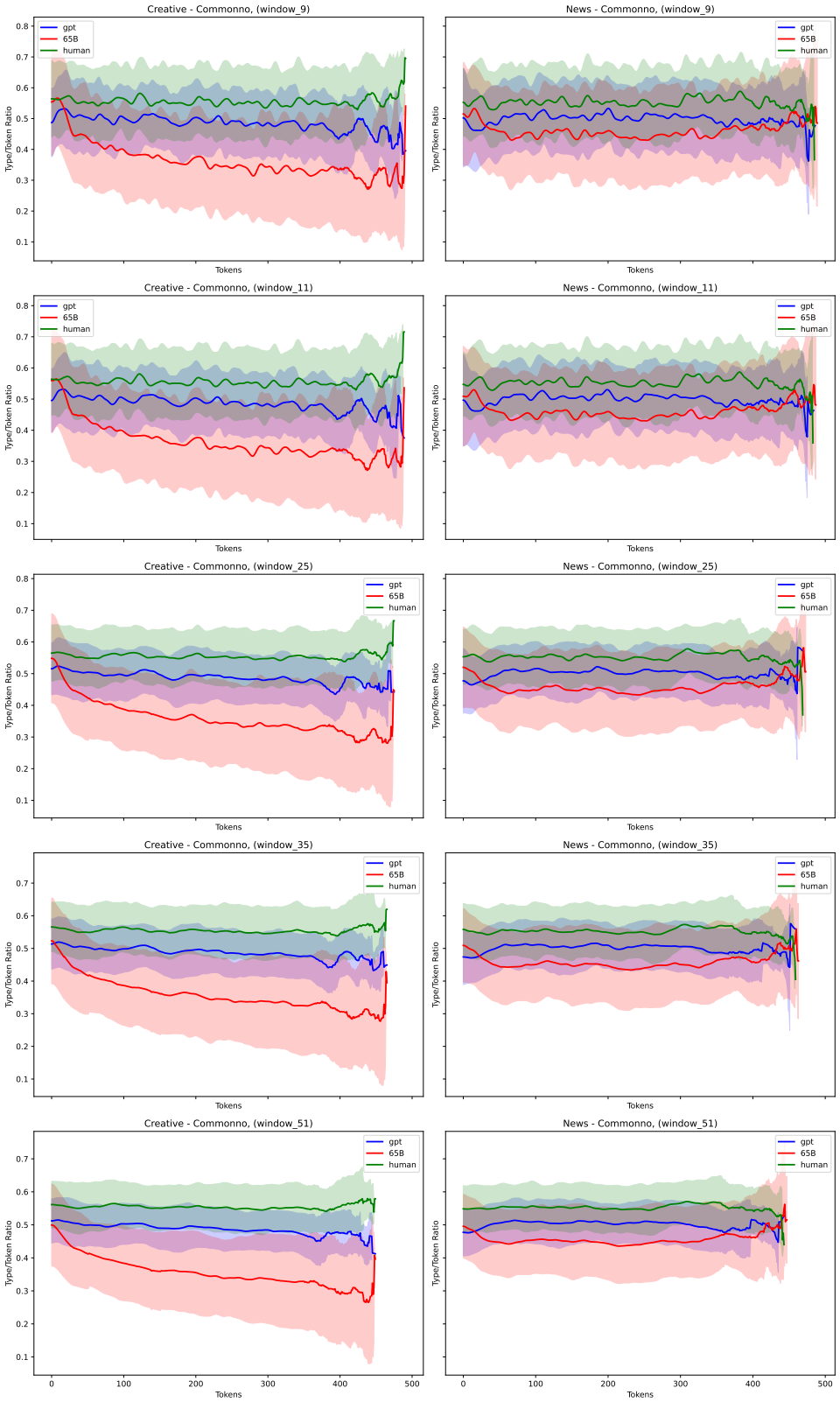

Figure 6: VMP commonNo

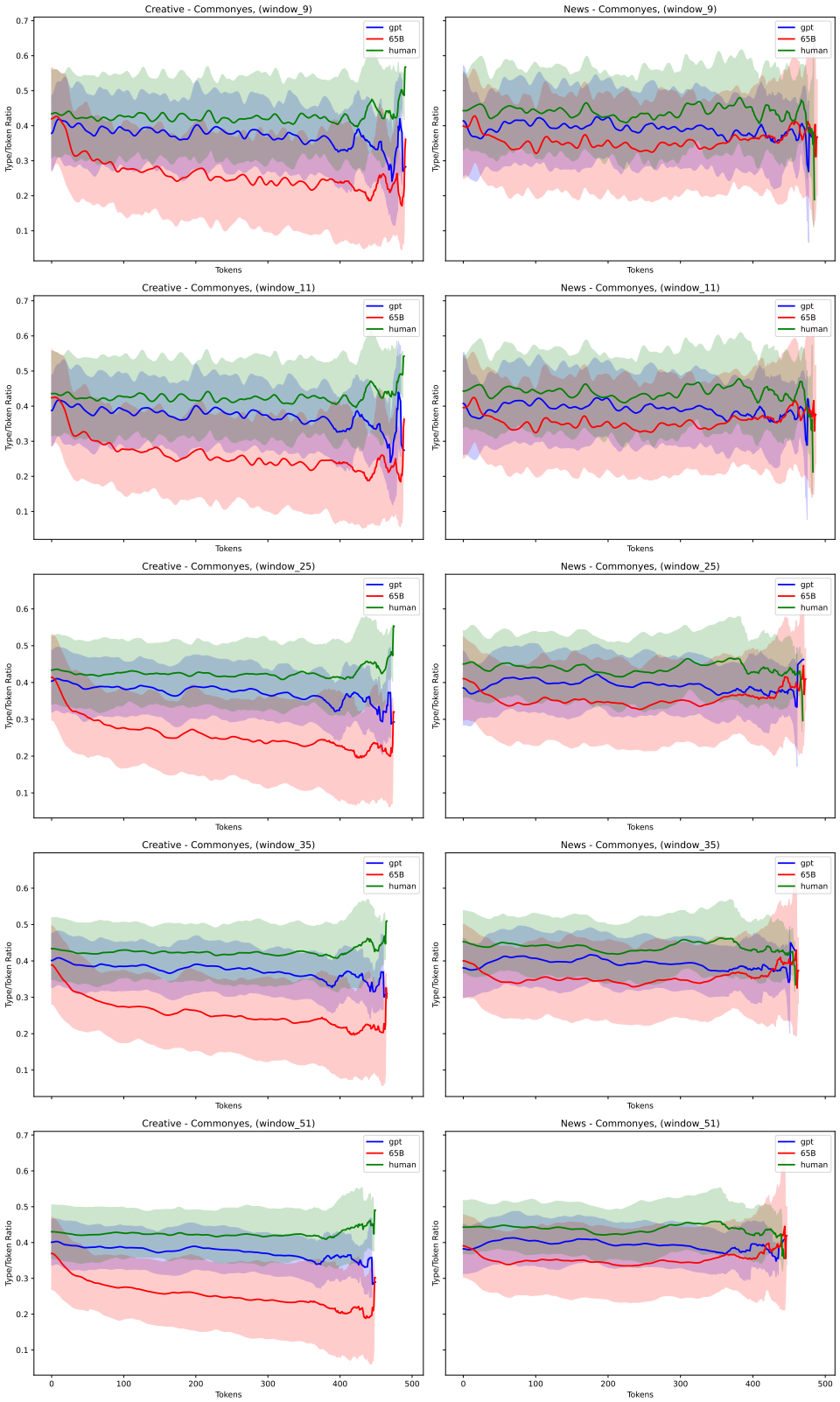

Figure 7: VMP commonYes

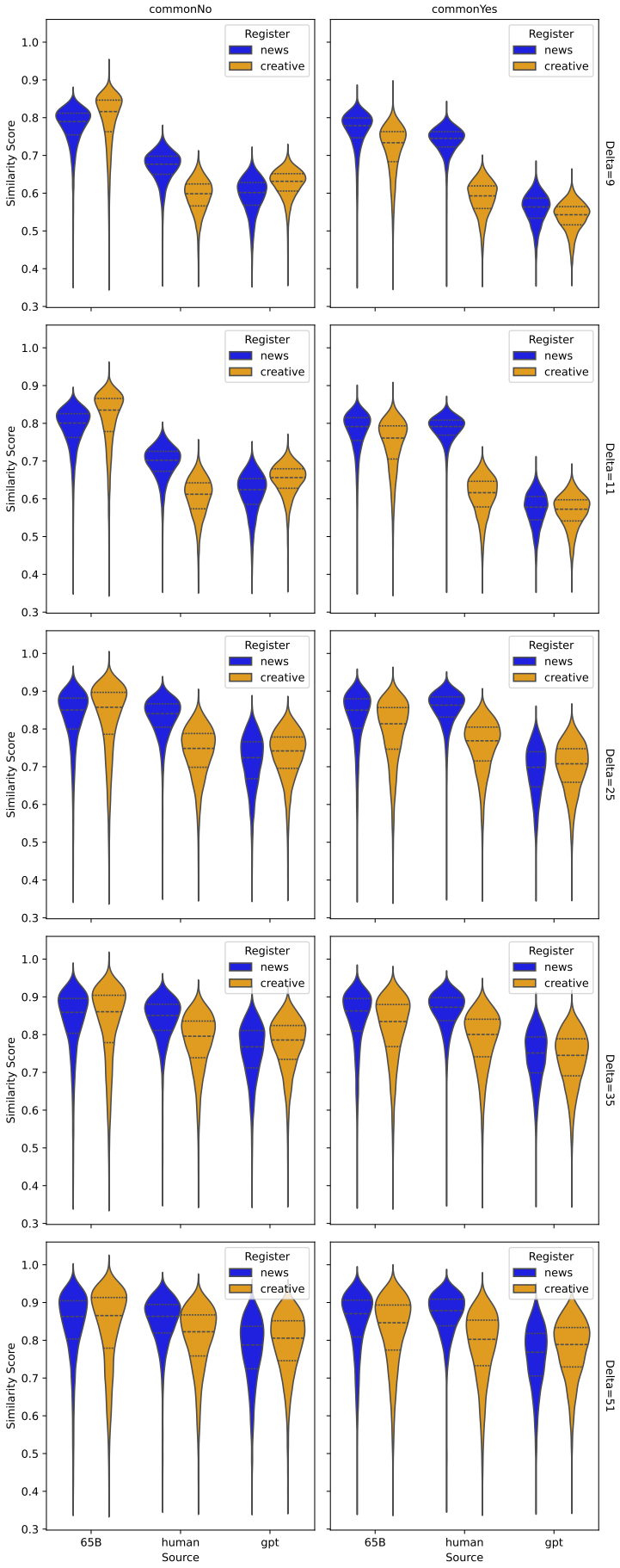

Figure 8: Self-similarity distribution plots of different sources for varying window sizes

Top 100 *Key* Keywords for Creative Register
Creative Target-human
ll, looks, mouth, oh, wo, says, pretty, couple, seems, anyway, fuck, hell, stuff, damn, office, ca, shit, paper, women, bastard, gets, picture, shouted, fingers, shoulder, definitely, ate, ah, guess, please, d, hate, knows, orange, sit, god, cry, torture, fault, die, na, edit, click, direction, deserve, relationship, breathing, normally, honest, flesh, son, jacket, eight, suppose, send, chair, cabinet, y, till, smoke, pile, armor, reaching, inevitable, starving, kicked, fed, hated, realise, shouting, chin, somewhere, hanging, kinda, scratch, gon, muster, nope, alcohol, bloody, blood, roll, slammed, dollars, yearold, decent, lights, accent, cheek, sits, bathroom, gotten, deserved, asleep, tear, writing, uh, literally, hall, obviously
Creative Reference-gpt
named, fascinated, grateful, tirelessly, determined, shared, excitement, significant, accepted, completed, relieved, consequences, lily, overjoyed, hesitant, sophie, practicing, including, respect, protect, differences, overwhelmed, intricate, eagerly, welcomed, skeptical, traveled, thrilled, unique, hugged, detail, opportunity, villagers, chatted, achieved, gathering, longed, mattered, approach, spreading, genuinely, expert, deserted, focused, catching, colony, choosing, importance, promising, mesmerized, frustrated, defend, insects, grew, noticed, rush, impressed, series, challenging, thrill, rebuild, value, succeeded, dense, behavior, lush, warriors, puzzle, intrigued, became, lilys, alex, determination, jacks, granted, technology, weapons, crops, team, gaining, decision, insight, peculiar, crucial, particularly, tool, dire, mortal, practiced, equally, routines, facility, frustration, mustered, grueling, industry, forests, judged, impending, sunny
Creative Target-human
deep, soul, seemed, slightly, perhaps, shit, powers, bastard, rise, ago, warm, address, count, swear, absolutely, further, thousand, though, impact, torture, odd, discovered, whenever, frozen, million, heading, normally, existence, sea, carry, appeared, necessary, battle, reality, flesh, definitely, century, similar, entered, jacket, eight, seven, data, cabinet, y, rushing, till, armor, dull, reaching, relief, inevitable, starving, clear, kicked, actual, brings, realise, space, souls, instant, blanket, kinda, smart, slow, muster, tightly, placed, causing, hands, somehow, threat, slammed, progress, landed, pressed, surely, stars, gold, silly, wet, bodies, gun, seeking, uh, advanced, literally, humanity, hundred, faster, advance, officers, pure, masters, leader, disgusting, intelligence, breath, particular, master
Creative Reference-65B
ruin, example, protect, couch, posted, pick, neighbors, labels, blow, upset, woods, scary, particularly, oven, writer, treat, ridiculous, suggested, jealousy, department, services, talks, relieved, pray, cleaned, react, financial, candy, persons, october, horny, depressed, glad, policy, music, thankful, hmm, levels, recover, ages, accomplish, cream, creepy, dads, feeding, filed, necklace, repairing, hugs, easter, nerve, ideas, liable, operating, nobodys, including, areas, sentences, hugged, blowing, kissing, cases, acting, concentrate, shadowy, rules, teaches, cooking, player, fund, jump, students, widened, filing, respect, christian, mix, investigate, explaining, curb, tubes, rural, recipe, airport, costs, fishing, backyard, lakes, tragic, statements, stabbing, expressing, crook, rode, sisters, borrowed, sobs, todays, amazon, dance

Table 3: Keyword Summary for Creative Register

Top 100 *Key* Keywords for News Register
News Target-human
psm, died, mr, main, ps, wednesday, parents, probably, d, told, added, near, spokesman, travelling, happened, eight, deputy, monday, thursday, mrs, talk, playback, radio, ms, morning, huge, apparently, march, records, single, either, chief, county, editor, weather, professor, captain, consumer, psbn, appeared, going, boss, refused, go, me, april, rangers, labours, accepted, twitter, crown, strongly, det, backing, possibly, internationally, brother, linked, partner, insurance, mps, achieved, communications, pictures, advised, loan, might, tv, recognised, flat, insisted, brilliant, absolutely, evening, nice, afford, strikes, afternoon, voted, sat, door, targets, staged, chris, obviously, innings, broke, estimates, bst, troops, injury, stephen, christmas, jail, four, pretty, pupils, stopped, scottish, ibrox
News Reference-gpt
conclusion, importance, culture, ultimately, practices, shape, behavior, risks, attention, argue, criticized, navigate, experiences, essential, efforts, deeply, traditional, stranger, dynamic, impossible, consumption, highlighting, thrilled, inspire, accountability, tech, ceo, arguing, unique, individuals, growing, ability, promising, towards, noted, alike, remains, organization, volatile, diagnosed, defense, resilient, proven, likes, uncertain, inspiration, unexpected, combat, effects, tasked, observers, examples, dedicated, opponents, ensuring, organized, guidance, topic, transition, responsibility, stable, handling, dedication, tirelessly, investigations, discrimination, muchneeded, implement, accessible, gender, controversy, significant, highprofile, emissions, takes, engaging, collaboration, transparent, remarks, uncertainty, recognized, laws, disputes, scandal, wellknown, ethical, achieving, cultural, create, spread, pandemic, equalizer, caution, ramp, cautious, component, effectively, scandals, strain, disrupt
News Target-human
speaking, revealed, troops, fans, warned, psm, regular, followed, pitch, deputy, september, ps, powers, radio, might, adding, losing, deals, prove, parent, eventually, independently, suggested, average, quarter, premiership, aged, rising, rugby, wanted, marks, african, bbc, labours, chose, praised, latter, backing, armed, internationally, monthly, eyes, sheffield, historic, loan, disruption, cold, unbeaten, recognised, flat, insisted, crowd, outcome, mistake, evening, strikes, proper, staged, obviously, operation, retain, complex, standing, celtic, lose, ownership, employers, games, favour, nottingham, sundays, euro, sparked, ali, stake, commit, mile, mutual, responding, dealt, length, appalling, militants, sit, defended, institution, indicated, contributed, automatically, quoted, rebuild, clearly, southeast, broken, subsequently, scores, ira, formal, cancelled, sofa
News Reference-65B
researchers, applications, center, delivers, ceo, base, method, mexico, photos, billion, developers, promising, organized, manifesto, awarded, apples, amazon, operated, organization, developer, deeply, development, effects, displayed, capabilities, export, author, episode, distributed, browser, stimulate, chapter, determine, forget, netherlands, flag, detective, object, restrictions, ties, caring, surveillance, spread, organizations, manipulate, dedicated, residential, factories, integrated, cruz, perfect, entry, tagged, takes, earth, gentleman, guy, cultural, approved, library, stable, apple, ability, trained, illinois, patterns, tags, updated, federation, consulting, vulnerability, sensitive, acquisition, useful, crew, airports, implemented, physics, tool, humans, interact, algorithms, valley, please, virtual, algorithm, materials, located, threats, historical, tools, fuel, australias, experiences, movies, manage, afraid, experiment, string, asks.

Table 4: Keyword Summary for News Register