More Aligned, Less Diverse?
Analyzing the Grammar and Lexicon of Two Generations of LLMs

Adrián Gude[1] **Roi Santos-Ríos**[1*] **Francis Bond**[3] **Dan Flickinger**[2]
Carlos Gómez-Rodríguez[1] **Olga Zamaraeva**[1]

{adrian.lopez.gude, roi.santos.rios, carlos.gomez, olga.zamaraeva}@udc.es
danflick@alumni.stanford.edu francis.bond@upol.cz

[1]Universidade da Coruña, CITIC [2]Independent Researcher [3]Palacký University, Olomouc

Abstract

This study contributes to a growing line of research in comparing LLM-generated texts with human-authored text, in this case, English news text. We focus in particular on the evaluation of syntactic properties through formal grammar frameworks. Our analysis compares two generations of LLMs in the context of two human-authored English news datasets from two different years. Employing the Head-Driven Phrase Structure Grammar (HPSG) formalism, we investigate the distributions of syntactic structures and lexical types of AI-generated texts and contrast them with the corresponding distributions in the human-authored New York Times (NYT) articles. We use diversity metrics from ecology and information theory to quantify variation in grammatical constructions and lexical types. We show that English news text has changed little in the given time frame, while newer LLMs display reduced syntactic and, especially, lexical diversity compared to older, non-instruction-tuned models. These findings point to future work in studying effects of instruction tuning, which, while enhancing coherence and adherence to prompts, may narrow the expressive range of model output.

1 Introduction

Large language models (LLMs) are increasingly compared to human writers across a growing range of linguistic and stylistic dimensions (e.g., Reinhart et al. 2025; Moon et al. 2025; Rashid et al. 2024). However, it remains unclear how such comparisons should be made and which dimensions best capture the differences. This limits our ability to draw robust conclusions about what makes human and machine writing distinct.

In this paper,[1] we take a step toward addressing this gap by examining diversity, both lexical and syntactic, as a consistently informative dimension for comparing human and LLM-generated text. Building on the formal grammatical framework of Head-Driven Phrase Structure Grammar (HPSG) and the English Resource Grammar (ERG) (Bender et al., 2002; Flickinger, 2011), we analyze variation in syntactic constructions and lexical types using diversity metrics drawn from ecology and information theory (Shannon and Simpson indices; Magurran 2004; Stamatatos 2009).

In contrast to prior related work (Zamaraeva et al., 2025; Muñoz-Ortiz et al., 2024), we compare two generations of LLMs and human-authored news writing. For the human side, we analyze New York Times (NYT) lead paragraphs from two distinct periods (2023 and 2025). For the LLM side, we compare a suite of base models trained prior to 2023 (LLaMA, Mistral, Falcon) with newer instruction-tuned models trained up to 2024 (Qwen 2.5, Mistral 7B v0.3, GPT-4o, LLaMa 3.3).

Our results show that English news text remains stable across time in all diversity metrics, suggesting a consistent balance of grammatical and lexical variety in professional news prose. In contrast, in LLM-generated text, we observe that both syntactic and lexical diversity decline substantially in newer, instruction-tuned models, with the effect being especially pronounced for lexical diversity. In addition, the newer LLM texts are easier to parse and take less computer memory to do so. Our findings suggest that, while instruction tuning is designed to improve the helpfulness and coherence of responses to natural-language prompts (Ouyang et al., 2022), it has the side effect of reducing the syntactic and lexical breadth of the outputs. Instruction tuned models generate outputs that are stylistically narrower and less varied than both humans and, interestingly, than earlier base models.

Overall, our study highlights diversity metrics as a robust, linguistically grounded way to track stylistic and grammatical shifts across model gen-

[1]Adrián Gude and Roi Santos-Ríos contributed equally to this work. Both authors should be regarded as joint first authors.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 38900–38911
July 2-7, 2026

erations, shedding light on how current tuning paradigms may trade off lexical variety for stylistic control.[2]

2 Related work

Muñoz-Ortiz et al. (2024) conducted a large-scale quantitative analysis contrasting texts generated by base (non-instruction tuned) LLMs with human-written news texts. Their results revealed differences across multiple linguistic dimensions, including morphological, syntactic, psychometric, and sociolinguistic aspects. These findings established a detailed baseline highlighting how human linguistic patterns remain more diverse and less homogenized compared to model-generated text.

Zamaraeva et al. (2025) further extended this line of research by conducting a comparison on the same data using Head-Driven Phrase Structure Grammar (HPSG) and the English Resource Grammar (ERG), providing a fine-grained perspective within an independent linguistic-theoretic framework. Their study showed that human-authored texts had greater grammatical but *less* lexical diversity than the LLM texts, and that human writers differed from each other more than each differed with respect to any LLM (in other words, that LLMs act as an "average" writer).

So far, little research has been done on the effects of instruction tuning and reinforcement learning from human feedback (RLHF) on LLMs' grammatical diversity. Padmakumar and He (2024) found that RLHF affected vocabulary type/token ratios of LLMs across LLMs, leading to more homogeneous texts. Another evaluation of the different stages of RLHF training showed that although RLHF improves out-of-distribution generalization compared to supervised fine-tuning, it significantly reduces output diversity measured through a combination of N-gram counting, semantic cosine similarity, and natural language inference metrics, revealing a tradeoff between adaptability and linguistic variety (Kirk et al., 2024; Shypula et al., 2025). To summarize, prior research suggests that, while modern LLMs have improved in fluency and instruction-following capabilities, these advances may come at the cost of lexical and stylistic diversity even in the new families of LLMs. Human texts, by contrast, still exhibit greater variety and more complex linguistic structures.

Our study contributes to this idea in several key ways. First, while previous work has compared a single generation of LLMs to human text or analyzed the effects of RLHF in isolation, we conduct a direct, diachronic comparison between two generations of LLMs: older base models and newer, instruction-tuned ones. Second, we mirror this generational approach on the human side by contrasting these models against human-authored news texts from two corresponding time periods. Furthermore, we use a linguistic-theoretic grammar framework, namely HPSG, to provide a way of analyzing syntactic and lexical properties of language that is independent from natural language processing tasks and thus should be more robust/generalizable. This framework allows us to look at lexical distributions in a systematic manner, beyond the surface information that vocabulary counts provide. HPSG lexical types are complex representations of word types that specify aspects of their syntactic behavior. We are not aware of a similar resource within, e.g., the UD framework; POS distinctions are too coarse. Last but not least, we employ independent diversity metrics from ecology.

3 Methodology

We use the HPSG-based framework utilized e.g. by Zamaraeva et al. (2025) and focus on diversity metrics from ecology and information theory to quantify changes in grammatical and lexical variability over time.

3.1 English Resource Grammar

The goal of this study is to provide an insight into how LLMs change over time with respect to the grammatical properties of their writing. For this purpose, it is not enough to look at just the vocabulary, and, while looking at dependency structures like Universal Dependencies (Nivre et al., 2016), as Muñoz-Ortiz et al. (2024) did is useful, we are interested in a comparison within a framework that is rooted in a formal linguistic theory and not inherently biased towards performance on NLP tasks. For this reason, we choose the DELPH-IN HPSG framework, following Zamaraeva et al. (2025). Head-driven Phrase Structure Grammar (HPSG; Pollard and Sag, 1994) is a theory of syntax that was developed by linguists theoretically and with empirical validation in mind, which is why the theory is associated with fully explicit formalisms that can be fully implemented on the

[2]Code available at: https://github.com/olzama/llm-syntax/

computer, allowing for rigorous validation of the theoretical claims made about syntax. DELPH-IN[3] is one such formalism that matured into a long-term international grammar engineering effort. In particular, the English Resource Grammar (ERG; Flickinger, 2000, 2011) has been in continuous development with regular releases,[4] reaching 94% average accuracy over a variety of English corpora. Importantly and in contrast to statistical parsers, the ERG is designed to provide structure for *all possible well-formed* English sentences and to *reject* strings that do not correspond to well-formed English utterances. This is crucial for our methodology, because we want to be able to compare LLMs between each other and to human writers with respect to *rare* linguistic phenomena, not only the frequent ones, and also with respect to sentences which were *not* parsed by the grammar for some reason that can be informative. The second property of the HPSG/ERG that is important to us is its structure: the grammar is represented as a clear hierarchy of *syntactic and lexical types*. The type definitions are detailed HPSG structures which specify sets of constraints which make a construction possible (such as: "this is a verb and it requires two complements of which one is a noun", to provide a simplistic example). Syntactic type definitions are used at parse time to "license" phrases bottom-up, until a complete sentence spanning the whole input string is built (such as noun phrases, verb phrases...), whereas lexical types are used by the parser only at the beginning of the parsing but provide very rich information about the specific constraints that need to be met. In this sense, lexical types are drastically different from vocabulary items which are just surface strings representing words. Finally, grammar-based exhaustive chart parsing provides us with a window into the sentences' *parsability*: how easy or difficult is it to parse a particular sentence, as a proxy measure of how complex it is.

3.2 Portability

While our experiments focus on English, the grammar-based approach itself can be used with any language (notably, also with low-resource languages). The HPSG framework is based on general linguistic theory and is not specific to English, and the tools used to develop these grammars are likewise language-independent. At the same time, each implemented grammar necessarily includes language-specific layers, which requires expert effort to develop. As a result, existing resources differ in their size and coverage, with English currently having the most mature and broad-coverage grammars. However, the required investment is comparable to that of training machine learning systems, which depend on significant data and computational resources that may not be equally available across languages or domains. We therefore see the methodology as broadly applicable, with its extension to other languages primarily dependent on the continued development of high-quality grammatical resources. The same applies across different genres, as once an adequate grammar is available for a language, the approach can be readily transferred to non-news domains with not much additional effort.

3.3 Diversity metrics

We use two diversity measures, the Shannon-Wiener diversity index (H, or Shannon Index), and the Simpson Diversity Index (Magurran, 2004).

Shannon: $H' = -\sum_{i=1}^{S} p_i \ln(p_i)$

Simpson: $D = 1 - \sum_{i=1}^{S} p_i^2$

The Shannon Index is widely used as an ecological measure of species diversity (Spellerberg and Fedor, 2003). It considers both the number of species (richness) and the evenness of their distribution, meaning a higher H value indicates greater diversity. The index is the same as the Shannon Entropy, and quantifies the uncertainty or "surprise" of predicting the next species in a community.

The Simpson Index (Simpson, 1949) measures the probability that two individuals (or tokens) randomly selected from a sample will belong to different categories (e.g., species, construction, lexical type, ...). It is less sensitive to low frequency phenomena than the Shannon Index.

4 Data and generative models

To study the differences between LLM-generated and human-authored news texts, we construct a new dataset similar in structure to the one used in Muñoz-Ortiz et al. (2024) and Zamaraeva et al. (2025). We use New York Times (NYT) lead paragraphs from February to May 2025. We assume that the LLMs under investigation, all trained on data up to 2024, could not have encountered these human-authored articles in training.

[3] https://delph-in.github.io/docs/home/Home/

[4] https://github.com/delph-in/erg/releases/tag/2025

Table 1: Datasets: reproduced in full from Table 1 in Muñoz-Ortiz et al. 2024, alongside the experiments done in Zamaraeva et al. 2025.

Dataset	# Sent. in dataset	Model size	Training tokens	Data sources
LLaMa	37,825	7B	1T	Not disclosed
	37,800	13B	1T	
	37,568	30B	1.5T	
	38,107	65B	1.5T	
Falcon	27,769	7B	1.5T	RefinedWeb-English (76%), RefinedWeb-Euro (8%), Gutenberg (6%), Conversations (5%) GitHub (3%), Technical (2%)
Mistral	35,086	7B	Not disclosed	Not disclosed
Original NYT	26,102	N/A	N/A	New York Times Archive, Oct. 1, 2023 - Jan. 24, 2024
Redwoods (WSJ)	43,043	N/A	N/A	Wall Street Journal sections 1-21
Redwoods (Wikipedia)	10,726	N/A	N/A	Wikipedia

Table 2: Datasets contributed with this paper.

Dataset	# Sent. in dataset	Model size	Training tokens	Data sources
Qwen 2.5	37,825	14B	18T	Not disclosed
	26,498	32B	18T	
	34,892	72B	18T	
LlaMa 3.3	39,306	70B	15T+	Not disclosed
Mistral v.0.3	33,840	7B	1T	Not disclosed
GPT-4o	50,544	Not disclosed	13T	Not disclosed
Original NYT	26,102	N/A	N/A	New York Times Archive, Feb. 1, 2025 - May. 31, 2025

```
system_prompt: "You are a professional
    journalist specializing in writing news.
    Follow the given structure."

user_prompt: "You will write a news lead
    paragraph using the inputs below.
  Inputs
  Headline: {headline}
  LeadThreeWords: {lead_three_words}
  Requirements - Mandatory
  Write one paragraph of several sentences (
      more than one, e.g. two-three (2-3)); no
       title, no bullets.
  Output format: the paragraph only, no
      preamble or labels."
```

Figure 1: Prompts used to generate news lead paragraphs from LLMs (system and user prompts).

Our new NYT dataset mirrors the structure of Muñoz-Ortiz et al. 2024. Human-authored texts consist of lead paragraphs downloaded from the NYT Archive API.[5] For each headline, we prompted our set of LLMs with the headline and the first three words of the lead paragraph to generate synthetic leads. Whereas Muñoz-Ortiz et al. (2024) used base models that could be prompted directly for text completion, the instruction-tuned models need a more explicit prompt in the form of instructions telling them to complete the paragraph, which is shown in Figure 1. The analyses we present refer exclusively to the human-written and LLM-generated lead paragraphs

The LLMs in Muñoz-Ortiz et al. (2024) included earlier systems (LLaMA, Falcon 7B, and Mistral 7B, all released prior to October 2023), while in this study we extend the design to more recent models: Qwen 2.5 (14B, 32B, 72B), LLaMA 3.3 (70B), Mistral 7B v0.3, and GPT-4o. Following Muñoz-Ortiz et al. (2024), we continue to distinguish scaling effects within a single architecture (e.g. different Qwen sizes) from differences due to model families.[6]

Dataset and model properties are summarized in Tables 1-2. The hyperparameters used for

[5] https://developer.nytimes.com/docs/archive-product/1/overview

[6] As a total, we performed 3 initial text generations with Qwen 2.5 32B to calibrate the outputs of the LLMs, then one execution per model, and lastly, another execution per model with the latest prompt, disclosed in Figure 1. The total number of executions amounts to 17.

all models are the following: temperature: 0.7, top_p: 0.92, top_k: 50, repetition_penalty: 1.05, max_new_tokens: 1000, num_return_sequences: 1, num_beams: 1. The average sentence length is in the range of 18-20 tokens for 2023 LLMs, while the newer 2025 models generate around 22-29 tokens, as shown in Table 7.

4.1 Scope

We deliberately restrict our experimental scope to a controlled news genre when comparing linguistic properties of LLM-generated and human-authored texts. This design choice follows prior work (Zamaraeva et al., 2025; Muñoz-Ortiz et al., 2024) and reflects both methodological and conceptual considerations. First, genre may exert an influence on linguistic structure, outweighing individual author effects (Biber, 1991); limiting genre variation therefore reduces confounds and allows clearer attribution of observed differences to generation source rather than discourse conventions. Second, focusing on one genre enables more reliable measurement of fine-grained linguistic properties, which may otherwise be obscured by cross-genre heterogeneity. Finally, while large-scale data generation across multiple genres would be desirable, it is computationally and financially costly, and may incentivize breadth at the expense of analytical depth. Our scope prioritizes internal validity and interpretability, providing a principled basis for future work to test the generality of these findings across genres and domains.

5 Results

In the following subsections we compare the grammar distributions in the human-authored and the LLM-generated datasets. The distributions were obtained with the English Resource Grammar (§3.1). We provide a list of the most distinctive syntactic and lexical types, as well as examples of where they are found in the data. However, the main point is that the datasets produced by people and by older and newer LLMs can be distinguished at the level of grammatical types **distributions**, taken as a statistical snapshot. Examples are meant to be illustrative but not necessarily explanatory. Any dataset can contain any instance of any syntactic or lexical type.

5.1 Syntactic types: Humans and LLMs

Table 3 shows that, compared to LLM-generated texts, human-authored English news text makes frequent use of constructions that help bind text to concrete events, locations, and temporal frames. In particular, human-authored news texts show higher use of clause-embedding and attribution structures, a tendency that correlates with reportive verbs requiring subordinate clauses (such as 'said' in 'Critics said...', Table 4).

Figure 2 shows that humans are clearly more diverse in their use of syntactic constructions than all LLMs.[7] In the case of the human texts, NYT-authored texts retrieved from 2025 are slightly less varied than the 2023 texts, but very close (seen also in terms of the Simpson index; Figure 3).

5.2 Syntactic types: 2023 and 2025 LLMs

Table 5 gives some examples of the distributional differences between the LLMs from 2023 and the LLMs from 2025. Notably, the newer 2025 LLMs are **not** more diverse than 2023 models; in fact, they form a **lower**-diversity band, despite being larger and trained on more recent data. This appears to be the result of the newer LLMs avoiding specifics such as names, dates, measures, etc., which influences not only lexical but also syntactic distribution. In this sense, older models were more like human writers.

In 2023, the constructions that contribute the most to the differences in diversity in comparison with 2025 are adjective-headed phrases and bare noun phrases (Table 5). In contrast, the 2025 models favor constructions that involve modification and coordination. These include noun phrases with modifiers, participial subordinate phrases, and coordinated noun structures. All of these add volume to the output but not necessarily content. Although bare noun phrases (common in the use of proper names) are present, they are less characteristic of the overall distribution, ranking 4th instead of 2nd (Table 5). Overall, these results suggest that newer LLMs are more careful about outputting factually incorrect information, which, curiously, can be detected at the level of syntax.

5.3 Lexical types: Humans and LLMs

Figure 4 shows an interesting difference with respect to lexical types. While syntactically, human authors of English news were clearly the most diverse, when it comes to lexical types, LLM outputs from 2023 show the highest diversity, surpassing all other corpora. Human-authored English news

[7] Shannon indices were computed using maximum bootstrap = 10,000

Construction	Preferred by	Example Sentence	Constituent
Nominal head + preceding adjunct	LLM	Phyllis Dalton, the Oscar-winning costume designer known for her meticulous work on historical epics, passed away at the age of 99.	historical epics
Subordinate pred phrase from participial VP	LLM	The Learjet was carrying a pediatric patient when it crashed, killing all on board.	killing all on board
Head + following scope adjunct	LLM	Speaker Kevin McCarthy began the final day before the shutdown, facing dim prospects of passing a funding measure.	facing dim prospects
Bare NP	LLM	Tariff strategy could harm economic growth and job creation.	job creation
Proper-name bare NP	Humans	Midterm elections reshaped the balance of power in Washington.	Washington
Bare NP from quantified daughter	Humans	Flights raised concerns over those being returned to Honduras.	those being returned to Honduras
Fragment NP	Humans	Chinese couple assembled building blocks: graduate degrees and careers.	graduate degrees and careers
Measure NP	Humans	Near-miss incidents occurred over a few years.	few years

Table 3: Syntactic constructions contributing the most to statistical differences between 2025 LLMs and English news text.

Lex types	Preferred by	Example Sentence	Constituent
Adjective (intersective, non-comparative)	LLM	A look back at her career reveals a dedication to authenticity and detail that brought eras long past vividly to life on screen.	past
Transitive verb with NP complement	LLM	Critics argue this action undermines collective bargaining rights and could lead to increased tensions with federal employee unions.	undermines
Count noun (lexical)	LLM	A man was wounded and a nine-year-old boy was fatally shot during an incident in Newark, according to officials.	man
Count noun with non-specific reference	LLM	The unrest has displaced thousands and raised concerns over the control of vital resources such as cobalt and copper.	concerns
Proper noun (generic)	Human	Internal government reports have revealed that Reagan National Airport experienced multiple near-miss incidents over the past few years.	Airport
Adj (intersective, coordinated, generic)	Human	A farm in rural Oregon has made headlines with an extraordinary offer: 40,000 pounds of fresh salmon, completely free of charge.	40,000
Verb (PP + finite CP complement, implicative)	Human	Critics said the president's characterization was not only inaccurate but also deeply hurtful to individuals with disabilities.	said
Proper noun (referential)	Human	More than a dozen prosecutors at the Washington U.S. Attorney's Office have been dismissed, according to sources familiar with the matter.	Washington

Table 4: Lexical types contributing the most to statistical differences between 2025 LLMs and English news text

Construction	Year	Example Sentence	Constituent
Nominal head + preceding adjunct	2025	Phyllis Dalton, the Oscar-winning costume designer known for her meticulous work on historical epics, passed away at the age of 99.	historical epics
Subordinate pred phrase from participial VP	2025	The Learjet was carrying a pediatric patient when it crashed, killing all on board.	killing all on board
Head + following scope adjunct	2025	Speaker Kevin McCarthy began the final day before the shutdown, facing dim prospects of passing a funding measure.	facing dim prospects
Bare NP	2025	Tariff strategy could harm economic growth and job creation.	job creation
Adjective–Headed Normal Construction	2023	"I am concerned about how they are using [the law]," Cardin said.	they
Bare NP	2023	Senator Ben Cardin told Al-Monitor that he will oppose the release of the remaining $650 million in military aid to Egypt.	Senator Ben Cardin
Subordinated Predicative VP (Participial)	2023	It was 40 degrees outside, but people were lined up wearing shorts and flip-flops.	but people were lined up wearing shorts and flip-flops.
Head–Adjective Small Clause (Predicative)	2023	When the Supreme Court returns on Monday from its summer recess, it will be without Justice Antonin Scalia.	Justice Antonin Scalia.

Table 5: Syntactic constructions contributing the most to the statistical differences between the 2023 and 2025 LLMs.

Lex types	Year	Example Sentence	Constituent
Adjective (intersective, non-comparative)	2025	A look back at her career reveals a dedication to authenticity and detail that brought eras long past vividly to life on screen.	past
Transitive verb with NP complement	2025	Critics argue this action undermines collective bargaining rights and could lead to increased tensions with federal employee unions.	undermines
Count noun (lexical)	2025	A man was wounded and a nine-year-old boy was fatally shot during an incident in Newark, according to officials.	man
Count noun with non-specific reference	2025	The unrest has displaced thousands and raised concerns over the control of vital resources such as cobalt and copper.	concerns
Utterance particle	2023	I don't see any signs that inventories are excessive.	I
Pronoun (personal, first person singular, it)	2023	"It's not acceptable for a democracy."	It
Pronoun (personal, second person, you)	2023	"If you ran for president," he wondered, "would you be able to win in Iowa?"	you
Pronoun (personal, first person plural, we)	2023	There was a time when we could blame our problems on Bush, but no more.	we

Table 6: Lexical types contributing the most to the statistical differences between the 2023 and 2025 LLMs.

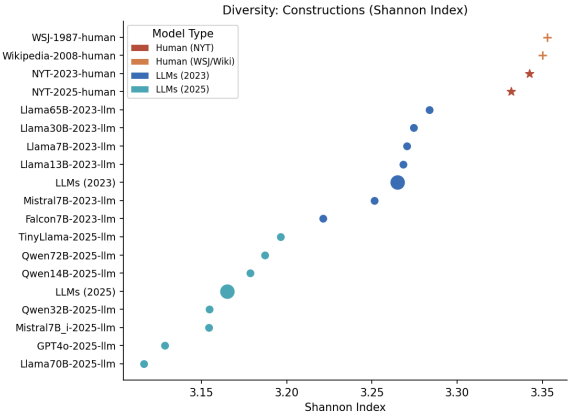

Figure 2: Syntactic construction diversity measured using the Shannon Index. Higher values indicate a more varied distribution of syntactic constructions. On the Y-axis, each point corresponds to a model name.

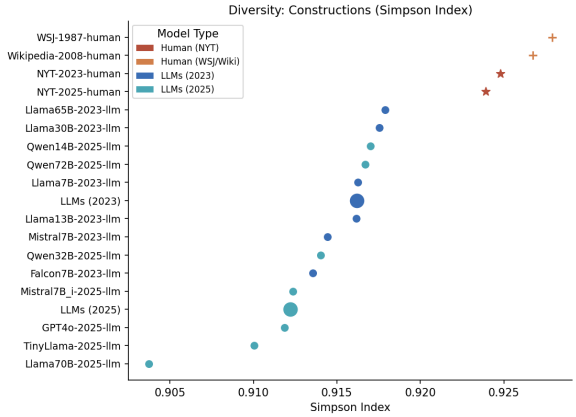

Figure 3: Syntactic construction diversity measured using the Simpson Index. Higher values indicate a more varied distribution of syntactic constructions. On the Y-axis, each point corresponds to a model name.

texts sit in the middle, and LLMs from 2025 rank the lowest. This indicates that, despite being larger and trained on more recent corpora, the newer systems employ a lesser variety of vocabulary groupings characterized by certain syntactic behavior (an example of such a group would be mass nouns, or clause-embedding verbs, etc). This surprising result calls for further investigation, possibly in the dimension of training paradigms, post-training and alignment. The same ranking is observed with Simpson indices (Figure 5). Humans, meanwhile, remain a stable reference, showing similar lexical type diversity as before, with a particularly distinctive use of proper names (Table 4).

5.4 Lexical types: 2023 and 2025 LLMs

The lexical-type ranking in Table 6 shows the change in the lexical types contributing the most to the diversity of the LLMs' distributions between 2023 and 2025. In 2023, the lexical types contributing the most to the diversity are utterance particles and personal pronouns. This suggests that the models' output is oriented toward conversational framing and speaker reference, with less emphasis on factual content. By contrast, 2025 models have a distinctive distribution of the very 'basic' lexical types: adjectives, common nouns, and transitive verbs. This may be a feature of a generic style that avoids specifics.

5.5 Punctuation

When we split the lexical types into punctuation and non-punctuation, in Figures 6 and 8 with Shannon index and Figures 7 and 9 with Simpson index, we see a very clear distinction: the 2025 models are much less diverse in their usage of punctuation.

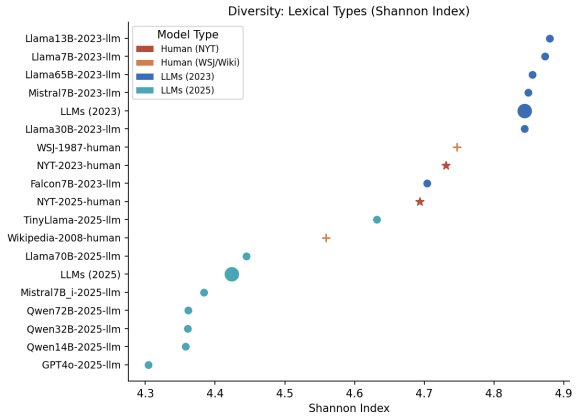

Figure 4: Lexical type diversity measured using the Shannon Index. Higher values indicate a more varied distribution of lexical constructions. On the Y-axis, each point corresponds to a model name.

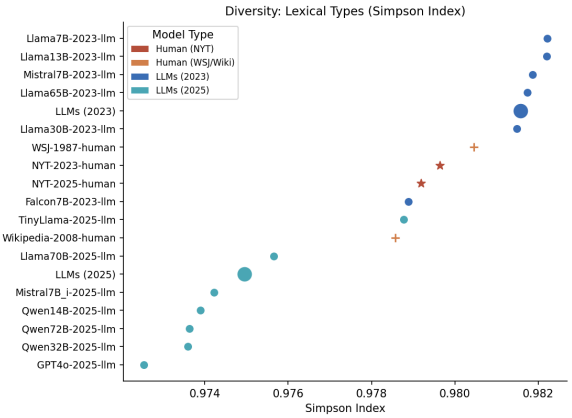

Figure 5: Lexical type diversity measured using the Simpson Index. Higher values indicate a more varied distribution of lexical constructions. On the Y-axis, each point corresponds to a model name.

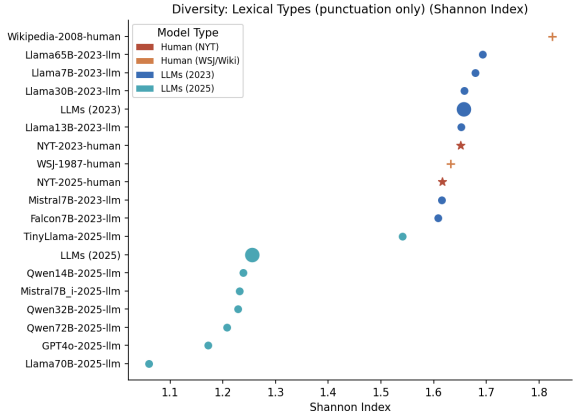

Figure 6: Lexical type diversity measured using the Shannon Index considering only punctuation. On Y-axis, each point corresponds to a model name.

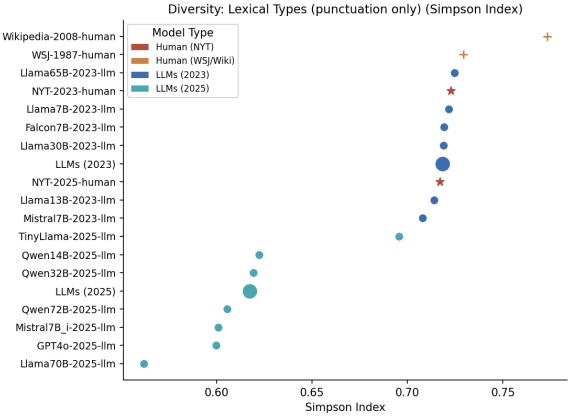

Figure 7: Lexical type diversity measured using the Simpson Index considering only punctuation. On Y-axis, each point corresponds to a model name.

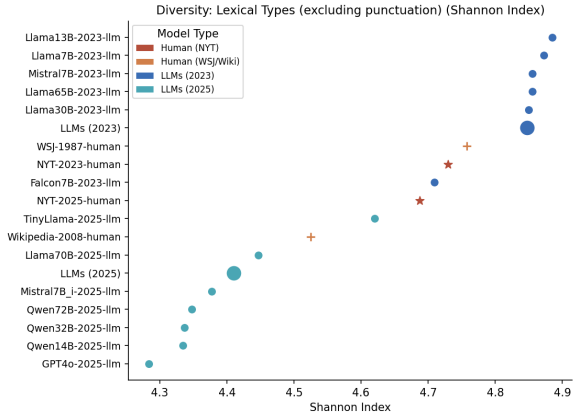

Figure 8: Lexical type diversity measured using the Shannon Index excluding punctuation. On Y-axis, each point corresponds to a model name.

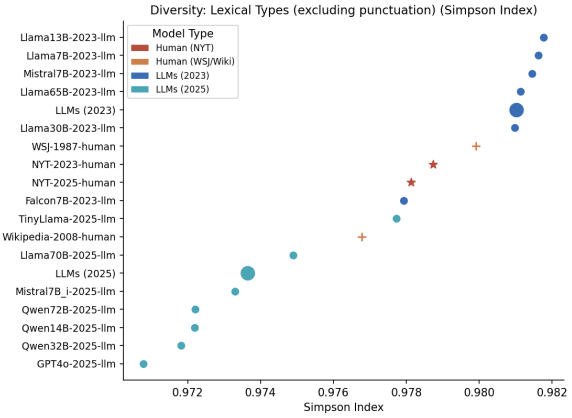

Figure 9: Lexical type diversity measured using the Simpson Index excluding punctuation. On Y-axis, each point corresponds to a model name.

Some anomalies are seen in GPT-4o: it generates very few em-dashes or semicolons, even though they are common in journalists' writing and they have been described as common in LLM-generated

text (Srivastava, 2025). It also generates very few quotations (correlating with its diminished use of the reportive verbs such as 'said'), whereas the rest of the models have shown no special behaviors like these. For example, the WSJ has many sentences like this: *"In Asia, as in Europe, a new order is taking shape," Mr. Baker said.*, but these are very rare in the output from GPT-4o. We hypothesize that this is a response to post-training aimed at reducing hallucinations and non-factual output. Still, this is a significant difference with what we expect from newspaper text.

5.6 Text length and parsability

The length of the generated texts is also different between the 2023 and 2025 models. More specifically, the 2025 models consistently generate longer sentences than both human-authored English news sentences and the sentences produced by 2023 models. In 2023, human sentences were about 10–20% longer than LLM sentences. In contrast, newer models produce sentences that are 15–30% longer than sentences written by human authors. The 2025 systems also reduce short sentences (1–15 tokens) by factors of 9 to 30 and cut non-sentence fragments by a factor of five. Despite this, the ERG requires **fewer** resources (time and space) to parse the sentences generated by 2025 models. Despite generating longer sentences, newer LLMs do not appear to create more complex structures. Instead, they are easier for the ERG to parse (recall the ERG is a deterministic, exhaustive search chart parser, which can run out of time or space to parse a sentence). Human-authored English news sentences from both 2023 and 2025 reach about 94–95% parse success, while LLM texts exceed 96% and approach 99% for the newest models. Even with substantial length increases, 2025 LLM sentences remain highly parsable, reflecting structural regularity that aligns with the best-understood and streamlined parts of the grammar, unlike human-authored English news text, which triggers phenomena that may be less understood or objectively challenging to parse. This means that the that the text generated by the newer instruction-tuned LLMs is easier for the ERG to parse than both earlier LLM outputs and English human-authored news text. This finding motivates future work exploring how specific grammatical structures correlate with parsability. More details about time and space required to parse each dataset are Appendix A.

6 Conclusion

In this work, we compared two generations of LLMs with two temporal samples of human-authored news writing, studying their syntactic structure within a formal linguistic framework. By applying diversity metrics from ecology and information theory to distributions of grammar constructions in LLM-generated and human-authored writing, we provided an interpretable view of how model-generated text changed over time relative to human language production. Our findings show that while English news text remains stable in both syntactic and lexical diversity, LLMs exhibit a shift: newer instruction-tuned systems produce text that is syntactically and lexically less diverse.

Despite producing longer sentences compared to human writers and to older LLMs, the 2025 instruction-tuned LLMs display reduced lexical variety and lower constructional diversity. The newer models also generate text that is easier to parse, suggesting increased predictability. Together, these trends indicate that the more recent instruction-tuned LLMs are constrained to a less expressive language space, yielding outputs that are more formulaic and less varied than English news text.

Our results confirm with two independent frameworks — ecology diversity metrics combined with a linguistic-theoretic account of grammar — that larger and newer models are not closer to human linguistic behavior, at least in the domain of professional news writing. Instead, even though their fluency seems to have improved, they show a growing divergence in lexical and grammatical diversity from humans. Future work could examine how the instruction-tuning of the models affects their ability to generate more diverse texts, and if with the right procedures, we could enhance these linguistic capabilities that they seem to have traded off.

Limitations

There are some limitations in terms of methodology based on the nature of the study. First, working with LLMs, which are non-deterministic models, introduces variability in the generated outputs, as results depend strongly on both the prompt design and the specific model used.

Second, this study focuses primarily on NYT-style news articles, which do not fully represent the broader spectrum of writing styles.

Third, the analysis of more recent LLMs is constrained by hardware limitations. Due to the large

size of some of the models considered, quantization is applied during the inference phase when generating synthetic data. Specifically, we apply 4-bit quantization to the largest models (LLaMA 65B and Qwen 2.5 72B). In addition, LLaMA 65B is also evaluated under 8-bit quantization, while the remaining models are used without quantization.

Finally, another limitation is the scarcity of large-scale HPSG grammars. Currently, only a few languages have wide-coverage implementations, and among them the English Resource Grammar is the only one extensive enough to parse roughly 94% of news text. Consequently, the present study is necessarily restricted to English.

Acknowledgments

We acknowledge grants GAP (PID2022-139308OA-I00) funded by MICIU/AEI/10.13039/501100011033/ and ERDF, EU; LATCHING (PID2023-147129OB-C21) funded by MICIU/AEI/10.13039/501100011033 and ERDF, EU; and TSI-100925-2023-1 funded by Ministry for Digital Transformation and Civil Service and "NextGenerationEU" PRTR; as well as funding by Xunta de Galicia (ED431C 2024/02).

CITIC, as a center accredited for excellence within the Galician University System and a member of the CIGUS Network, receives subsidies from the Department of Education, Science, Universities, and Vocational Training of the Xunta de Galicia. Additionally, it is co-financed by the EU through the FEDER Galicia 2021-27 operational program (Ref. ED431G 2023/01).

We have used ChatGPT and Gemini for minor copy-editing (e.g. thesaurus suggestions) and for visualization ideas. We have used GitHub copilot for code autocompletion and Claude Code for final refactoring.

References

Emily M. Bender, Dan Flickinger, and Stephan Oepen. 2002. The Grammar Matrix: An open-source starter-kit for the rapid development of cross-linguistically consistent broad-coverage precision grammars. In *Proceedings of the Workshop on Grammar Engineering and Evaluation at the 19th International Conference on Computational Linguistics*, pages 8–14, Taipei.

Douglas Biber. 1991. *Variation across speech and writing*. Cambridge University Press.

Dan Flickinger. 2000. On building a more efficient grammar by exploiting types. *Natural Language Engineering*, 6(01):15–28.

Dan Flickinger. 2011. Accuracy v. robustness in grammar engineering. In Emily M. Bender and Jennifer E. Arnold, editors, *Language from a Cognitive Perspective: Grammar, Usage and Processing*, pages 31–50. CSLI Publications, Stanford, CA.

Robert Kirk, Ishita Mediratta, Christoforos Nalmpantis, Jelena Luketina, Eric Hambro, Edward Grefenstette, and Roberta Raileanu. 2024. Understanding the effects of rlhf on llm generalisation and diversity. *Preprint*, arXiv:2310.06452.

Anne E. Magurran. 2004. *Measuring Biological Diversity*. Blackwell Publishing, Oxford.

Kibum Moon, Adam E Green, and Kostadin Kushlev. 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*, page 100207.

Alberto Muñoz-Ortiz, Carlos Gómez-Rodríguez, and David Vilares. 2024. Contrasting linguistic patterns in human and LLM-generated news text. *Artificial Intelligence Review*, 57(10):265.

Joakim Nivre, Marie-Catherine De Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, and Reut Tsarfaty. 2016. Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.

Vishakh Padmakumar and He He. 2024. Does writing with language models reduce content diversity? *Preprint*, arXiv:2309.05196.

Carl Pollard and Ivan A. Sag. 1994. *Head-Driven Phrase Structure Grammar*. Studies in Contemporary Linguistics. The University of Chicago Press and CSLI Publications, Chicago, IL and Stanford, CA.

Md Mamunur Rashid, Nilsu Atilgan, Jonathan Dobres, Stephanie Day, Veronika Penkova, Mert Küçük, Steven R Clapp, and Ben D Sawyer. 2024. Humanizing ai in education: A readability comparison of llm and human-created educational content. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 68, pages 596–603. SAGE Publications Sage CA: Los Angeles, CA.

Alex Reinhart, Ben Markey, Michael Laudenbach, Kachatad Pantusen, Ronald Yurko, Gordon Weinberg, and David West Brown. 2025. Do llms write like humans? variation in grammatical and rhetorical styles. *Proceedings of the National Academy of Sciences*, 122(8):e2422455122.

Alexander Shypula, Shuo Li, Botong Zhang, Vishakh Padmakumar, Kayo Yin, and Osbert Bastani. 2025. Evaluating the diversity and quality of llm generated content. *arXiv preprint arXiv:2504.12522*.

E. H. Simpson. 1949. Measurement of diversity. *Nature*, 163:688.

Ian F. Spellerberg and Peter J. Fedor. 2003. A tribute to claude shannon (1916–2001) and a plea for more rigorous use of species richness, species diversity and the 'shannon–wiener' index. *Global Ecology and Biogeography*, 12(3):177–179.

Rajesh Srivastava. 2025. How LLMs turned the em dash (—) into a villain - Technical nuances.

Efstathios Stamatatos. 2009. A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*, 60(3):538–556.

Olga Zamaraeva, Dan Flickinger, Francis Bond, and Carlos Gómez-Rodríguez. 2025. Comparing LLM-generated and human-authored news text using formal syntactic theory. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9041–9060, Vienna, Austria. Association for Computational Linguistics.

Appendices

A Parsability

Tables 7 and 8 show that newer instruction-tuned LLMs produce texts that are easier to parse than both human-authored English news text and earlier LLM outputs.

Profile 2023 (RAM limit 21G)	Items	Parsed %	Length toks/S	Short %	Frgmt %	Time sec/S	Space Gb/S	>Limit %
nyt-2023-human	*26092*	*93.4*	*22.33*	*33*	*13*	*11.5*	*2.3*	*4.3*
falcon07-2023-llm	27769	97.7	18.62	37	10	4.4	0.9	1.1
llama07-2023-llm	37825	96.4	19.42	35	12	6.0	1.2	1.8
llama13-2023-llm	37800	97.1	18.60	38	13	5.1	1.0	1.4
llama30-2023-llm	37568	96.9	18.17	39	12	4.9	1.0	1.3
llama65-2023-llm	38107	96.4	18.76	37	12	5.7	1.1	1.7
mistral7b-2023-llm	35086	97.3	18.36	38	13	4.7	0.9	1.1
2025 (RAM limit 31G)								
nyt-2025-human	*24053*	*94.8*	*22.19*	*32*	*12*	*12.2*	*2.4*	*2.6*
qwen14-2025-llm	26498	98.2	25.63	4	2	7.2	1.6	0.6
qwen32-2025-llm	34892	98.4	25.62	3	2	6.7	1.3	0.5
qwen72-2025-llm	34614	98.5	26.05	2	2	6.0	1.2	0.3
llama70-2025-llm	39306	97.9	29.87	1	2	9.5	2.0	0.9
gpt4o-2025-llm	50544	98.5	25.99	1	2	4.6	1.0	0.2
mistral7i-2025-llm	39708	98.3	25.87	4	2	7.4	1.4	0.6

Table 7: Parsing statistics for human-authored (in italics) and LLM-generated news datasets from 2023 and 2025. Each row reports the percentage of sentences successfully parsed by the ERG, average sentence length in tokens, proportion of short sentences (≤ 15 tokens), proportion of fragments, mean CPU time and memory per sentence, and proportion of sentences exceeding resource limits.

Profile	Time (CPU-seconds/sent)				Space (Gbytes/sent)			
(length in tokens)	31-35	36-40	41-45	46-50	31-35	36-40	41-45	46-50
nyt-2023-human	*13*	*28*	*48*	*67*	*2.6*	*5.2*	*9.1*	*13.1*
falcon07-2023-llm	11	24	44	68	2.2	4.5	8.3	13.6
llama07-2023-llm	14	26	50	68	2.7	5.0	9.5	13.4
llama13-2023-llm	13	28	50	65	2.6	5.3	9.5	13.3
llama30-2023-llm	14	28	51	69	2.6	5.3	9.4	13.3
llama65-2023-llm	14	30	50	67	2.7	5.5	9.4	13.1
mistral7b-2023-llm	13	27	48	73	2.5	5.1	8.8	14.1
nyt-2025-human	*13*	*30*	*52*	*84*	*2.7*	*5.9*	*10.1*	*16.2*
qwen14-2025-llm	10	25	49	77	2.3	5.1	9.8	15.3
qwen32-2025-llm	10	24	50	78	2.0	4.4	8.7	14.0
qwen72-2025-llm	9	21	42	67	1.8	4.0	7.6	12.1
llama70-2025-llm	8	17	38	61	1.7	3.5	7.3	11.7
gpt4o-2025-llm	8	17	36	54	1.6	3.3	6.5	9.8
mistral7i-2025-llm	10	25	50	81	2.0	4.5	8.7	14.0

Table 8: Average parsing cost by sentence-length bin for human (in italics) and LLM-generated texts. CPU time and memory consumption per sentence (in seconds and GB, respectively) for sentences binned by length (31–50 tokens).