Comparing LLM-generated and human-authored news text using formal syntactic theory

Olga Zamaraeva[1] olga.zamaraeva
Dan Flickinger[2] danflick
Francis Bond[3] francis.bond
Carlos Gómez-Rodríguez[1] carlos.gomez

[1]Universidade da Coruña, CITIC (@udc.es)
[2]Independent Researcher (@alumni.stanford.edu)
[3]Palacký University at Olomouc, Department of Asian Studies (@upol.cz)

Abstract

This study provides the first comprehensive comparison of New York Times-style text generated by six large language models against real, human-authored NYT writing. The comparison is based on a formal syntactic theory. We use Head-driven Phrase Structure Grammar (HPSG) to analyze the grammatical structure of the texts. We then investigate and illustrate the differences in the distributions of HPSG grammar types, revealing systematic distinctions between human and LLM-generated writing. These findings contribute to a deeper understanding of the syntactic behavior of LLMs as well as humans, within the NYT genre.

1 Introduction

Studying linguistic properties of LLM-generated text and comparing it to human-authored text is a topic of growing interest in the field of natural language processing (NLP). Previous research has predominantly focused on training classifiers (is the text LLM-generated or no); a few studies include an analysis of differences in vocabulary distribution, use of dependency structures, or sentiment properties of the text (see § 2). In this study, we systematically analyze *grammatical* differences of LLM-generated vs. human-authored text through the lens of a formal syntactic theory developed for linguistic research independently of NLP.[1] Using a formal theory for analysis and evaluation is a way to overcome some of the biases that arise from using tools developed directly in the context of designing NLP tasks. We hope this will lead to further systematic discoveries about grammatical properties of LLM-generated text and how they differ from human-authored text. In this paper, we use the broad-coverage English Resource Grammar (Flickinger, 2000, 2011) to analyze texts in the New York Times genre.

2 Related work

Our study is concerned with the analysis of the grammatical properties of LLM-generated texts as compared to human-authored texts. Here, we review the literature with a similar focus. This leaves out of scope papers concerned with building classifiers or with sentiment and semantic analysis.

Muñoz-Ortiz et al. 2024 include a study of syntactic and vocabulary diversity in NYT-style news. They conclude that measurable differences can be detected, including at the level of grammar, and that human-authored texts exhibit more variety of vocabulary, shorter constituents, and more optimized dependency distances. Narayanan et al. (2024) use the Universal Sentence Encoder (USE: Cer et al., 2018) to compare human-authored and AI-generated code explanations and find statistical differences, though without linguistic analysis. Sandler et al. (2024) base the comparison on ChatGPT-human dialogues, using primarily lexical features, not syntactic, and find greater diversity in texts written by humans. Notably, they use dictionary-style features and not just raw vocabulary. So do Alvero et al. (2024), who compare college application essays (submitted in 2016-2017) with texts generated by GPT-3.5 and GPT-4. They find that human authors show more variety in e.g. verb usage. Juzek and Ward (2025) study the vocabulary of LLMs linking it to the increase of use in certain vocabulary items in scientific abstracts (e.g. the word 'delve'). Park et al. (2025) perform a statistical comparison by clustering linguistic features (this is necessary to obtain statistically significant results in the context of multiple comparisons be-

[1]Annotation schemes such as Universal Dependencies (UD: Nivre et al., 2016) or Penn Treebank (PTB: Marcus et al., 1993) are related to syntactic theory but they have been developed as guidelines for hand-annotating corpora specifically for NLP. As such, they are less detailed and consistent than a formal theory and less independent from NLP tasks themselves.

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 9041–9060
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

tween many features). They conclude that LLM-generated texts have a distinct statistical footprint from human-authored text. Shaib et al. (2024) compare strings of POS-tags, which they call "syntactic templates", finding that LLMs tend to repeat these templates more than humans do. Finally, several studies base the comparison on a set of linguistic features proposed for rhetoric styles by Biber (1991, 1995) and Biber and Conrad (2019). In particular, Reinhart et al. (2024) show that LLMs prefer certain grammatical constructions and thus struggle to match styles that do not employ them (according to Biber). The constructions include participial clauses, 'that'-subject clauses, nominalization, phrasal and clause coordination. Sardinha 2024 also uses the "Biber features". This study is perhaps the closest in spirit to ours, since it uses an independently developed linguistic framework and presents examples of the differences found.

3 Methodology

The central idea of our methodology is to apply formal syntactic theory to analyzing structural properties of texts generated by LLMs as compared to human-authored texts. We use the HPSG theory of syntax (§3.1), specifically its implementation as the English Resource Grammar (§3.2), the largest available implementation of a formal grammar in terms of its coverage over naturally occurring text (in any language and in any theory). While applying such methodology implies the investment in building the grammar, the HPSG theory and the formalism were developed precisely to be used for a wide variety of languages. The cross-linguistic applicability of the theory has been continuously tested in the context of the Grammar Matrix (Bender et al., 2002, 2010; Zamaraeva et al., 2022) and the AGGREGATION (Bender et al., 2020; Howell and Bender, 2022) projects.

3.1 HPSG

Head-driven Phrase Structure Grammar (HPSG: Pollard and Sag, 1994) is a formal theory of syntax that uses a fully explicit formalism, so it can be implemented on the computer in its entirety as a grammar which then maps sentences to complete structures automatically, while remaining fully consistent and interpretable. The theory represents syntactic structure and elements of the syntax-semantic interface (dependencies, quantifier scope, information structure) as a complex graph, which can also be visualized as an attribute-value matrix of features and their values (such as the feature HEAD having a value *noun*). HPSG assumes lexical types which can house multiple lexical entries, and, unlike raw vocabulary forms, lexical types contain information about syntactic properties of words.

The grammar as a whole (the lexicon included) is a hierarchy of types. Figure 1 shows a very small and simplified portion of the HPSG type hierarchy, with only two features (HEAD and COMPS, complement list). This part pertains to the lexicon and lexical types. The noun 'law' can behave in different ways syntactically, which motivates two lexical entries belonging to two different types (which may house other nouns as well). In §5 we report on how this word is one of the examples of differences in human-authored and LLM-generated texts that we examined. The real type hierarchy, such as the one in the ERG (§3.2), consists of hundreds of types with dozens of features, allowing us to examine grammatical properties of sentences in detail.

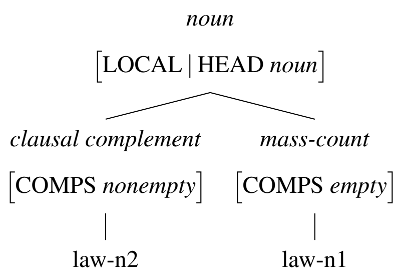

Figure 1: Part of the HPSG type hierarchy (simplified; adapted from ERG).

3.2 English Resource Grammar

The English Resource grammar is a grammar of English implemented in HPSG (Flickinger, 2000, 2011).[2] The ERG is continuously developed as part of the DELPH-IN open-source grammar engineering initiative.[3] It is a broad coverage precision grammar, meaning that it will parse 94%[4] of reasonably well-edited English text but is not expected to yield any structure for a sentence impossible in English. Since the grammar is precise and consistent, it can be used to automatically create precise and consistent treebanks. It has been shown that including such treebanks in the training data improves performance of various NLP systems (Lin et al., 2022; Hajdik et al., 2019; Chen et al., 2018;

[2] Regular releases: `https://github.com/delph-in/erg`
[3] `https://github.com/delph-in/docs/wiki`
[4] Per the 2025 release documentation

Table 1: Datasets: reproduced in full from Table 1 in Muñoz-Ortiz et al. 2024, plus the information on Redwoods.

Dataset	# Sent. in dataset	Model size	Training tokens	Data sources
LLaMa	37,825	7B	1T	English CommonCrawl (67%), C4 (15%),
	37,800	13B	1T	GitHub (4.5%), Wikipedia (4.5%),
	37,568	30B	1.5T	Gutenberg and Books3 (4.5%), ArXiv (2.5%),
	38,107	65B	1.5T	Stack Exchange (2%)
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

Buys and Blunsom, 2017). Some of the properties of the ERG are summarized in §4, Table 2. The grammar is implemented in the DELPH-IN Joint Reference Formalism (Copestake, 2002) and can be used with any DELPH-IN tools. We parsed the data with the latest version of the ERG[5] and ACE (Crysmann and Packard, 2012),[6] and then used the Pydelphin tools[7] along with packages such as Numpy (Harris et al., 2020), Pandas (McKinney, 2010), and scikit-learn (Pedregosa et al., 2011) to analyze the derivations by counting the occurrences of phrasal constructions, lexical (inflectional and derivational) rules, and lexical types, and studying the relative frequency distributions through cosine similarity and diversity metrics (see §5).

4 Data and generative models

To study the differences between LLM-generated and human-authored news texts, we use the dataset created by Muñoz-Ortiz et al. (2024). We choose this dataset for two main reasons: 1) by using news articles, we can make sure the LLMs did not have access to the corresponding human-authored articles at the time of training; 2) by reusing the dataset from a previous study, we enable comparisons of analyzing the data with UD and with the fully-fledged grammatical theory provided by HPSG. In addition, we used the Wall Street Journal (WSJ) and Wikipedia portions of the Redwoods Treebank (Oepen et al., 2004), an ERG-parsed corpus accompanying each release of the ERG. We use WSJ and Wikipedia to see which differences between human and LLM writing persist beyond the NYT style. We release the ERG-parsed LLM-generated data through GitHub.[8]

The 'NYT' datasets from Muñoz-Ortiz et al. 2024 include the original New York Times (NYT) article lead paragraphs and LLM-generated texts obtained from 6 different LLMs by prompting them with the headlines together with the first 3 words of the lead paragraph.[9] The original NYT human-authored data consists of the lead paragraphs for articles between October 1, 2023, and January 24, 2024 obtained with the NYT Archive API.[10] The LLMs they used were all released prior to October 1, 2023, and included various versions of LLaMA (Touvron et al., 2023), the 7B version of Falcon (Almazrouei et al., 2023), and the 7B version of Mistral (Jiang et al., 2023). Following Muñoz-Ortiz et al. (2024), we want to consider the influence of scaling (different LLaMas with the same architecture, training dataset and training setup, but different model size) separately from the other aspects that differentiate the LLMs (LLaMa vs Mistral vs Falcon). The properties of the datasets and the models used to generate them (where appropriate) are in Table 1. LLM-generated datasets have more sentences, but the sentences written by humans are longer (see Figure 3 in Muñoz-Ortiz et al. 2024).

The NYT data accounts for almost all syntactic and morphological rules registered in the grammar; for about 79% of the lexical types, and for about 61% of the lexical entries (Table 2).

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

[6] https://sweaglesw.org/linguistics/ace/download/ace-0.9.34-x86-64.tar.gz

[7] https://pydelphin.readthedocs.io/

[8] https://github.com/olzama/llm-syntax/releases/tag/1.0.0

[9] The LLM-generated data associated with Muñoz-Ortiz et al. 2024 can be found here: https://zenodo.org/records/11186264

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

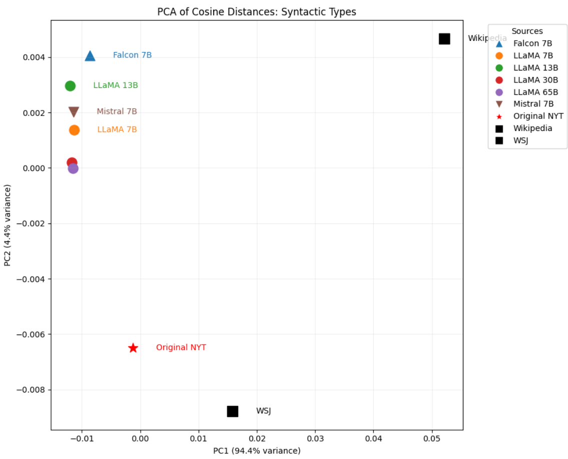

Figure 2: Cosine similarity: syntactic types

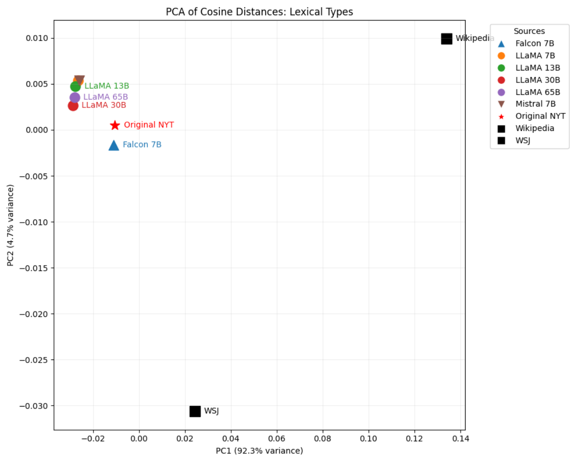

Figure 3: Cosine similarity: lexical types

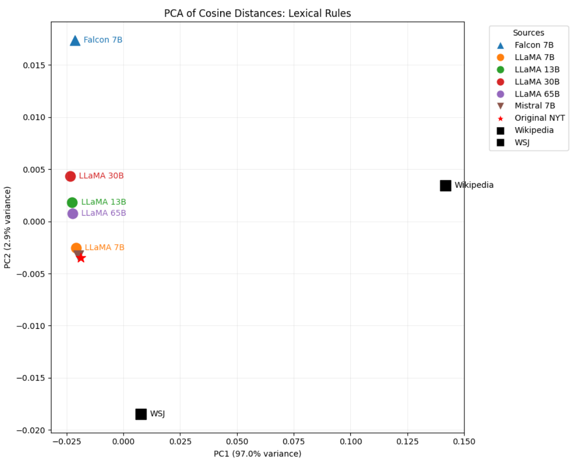

Figure 4: Cosine similarity: lexical rules

Construction type	ERG	Data
syntactic	298	289
lexical type	1,398	1,105
lexical entry	44,366	27,311
morphological rule	100	99

Table 2: Properties of the English Resource Grammar and the coverage of types by the NYT data

5 Results

We present the comparison of type distributions between the human-authored and LLM-generated data, including WSJ and Wikipedia data to see whether the differences persist across styles or genres. We look at cosine similarity of the construction distributions (§5.1) and at two diversity indices (§5.2). We look at syntactic and lexical types as well as lexical (morphological) rules separately.

5.1 Cosine similarity

We find that human authors and LLMs clearly differ in terms of their corresponding syntactic and lexical type distributions, and that this may persist across style and genre.[11] If we consider only syntactic and lexical types (Figures 2-3),[12] we see clearly that human-authored texts are distinct in their HPSG type distributions from the closely-clustered LLMs and furthermore, that human-authored NYT texts are more similar to WSJ (different style, same genre) than to Wikipedia (different genre). This is true for syntactic and lexical types, although with lexical types, Falcon is an outlier, and the effect of style and genre seems bigger. However, in terms of lexical (inflectional and derivational) rules, we observe that the distribution of human NYT authors is very similar to LLMs except Falcon. These findings align with what we see when we apply diversity metrics (§5.2). In this paper, we focus on the most salient differences between LLMs and human NYT authors, and investigating the intriguing role of lexical rules remains future work. One hypothesis is that the distribution of lexical rules is very closely tied to genre and style (and that the Falcon model is somehow special in this respect).

5.1.1 Frequent syntactic constructions

Among the frequent syntactic constructions (Figure 5; Appendix A), we see differences insensitive to genre[13] in the head-complement construction

[11]We use PCA projection to help visualize the differences in the 98-100% similarity range. The underlying data is provided in Appendix B.

[12]The data is not directly comparable, hence the scale differences.

[13]We have run the Mann-Whitney U-test for statistical significance for these comparisons. The p-values < 0.05 are listed

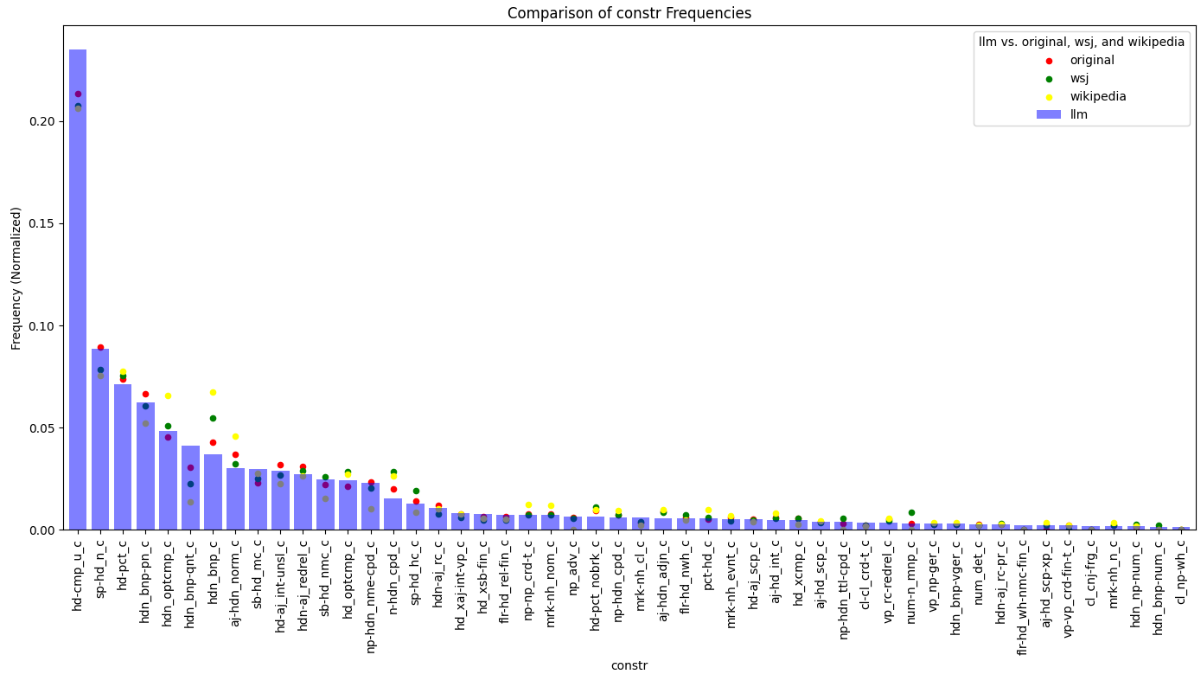

Figure 5: LLM use of syntactic constructions compared to human writers. Cases of particular interest are where the dots cluster closely together and are noticeably higher or lower than the blue bar representing LLM. Also of certain interest are cases where all the dots are higher or lower than the bar but not very close to each other.

(human authors use less of it in all human-authored datasets we examined), a couple of punctuation-related constructions (note that from the point of view of the ERG, punctuation is not only a token; it also matters how exactly it gets placed in the sentence, so, this is a syntactic matter), and the adjunct-head construction licensing double modification (e.g. *big old cat*). There might be something of note going on with bare noun phrases and noun compounds as well; the human authors appear to use them more; however the differences between styles (WSJ) and genre (Wikipedia) seem to be greater than the differences between human NYT writers and LLM-generated NYT-style news. Differences in punctuation have been observed (Muñoz-Ortiz et al., 2024); however the head-complement construction is a general grammatical feature which does not have a direct equivalent in the UD framework. In UD, there is the OBJ dependency, which refers to a dependency between a direct object and a verb, and is a concept from the syntax-semantic interface. A head-complement construction is a general syntactic construction that licenses constituents which combine a head element with its complement. The head does not need to be a verb (nouns and adjectives can have complements too, for example). In this study, we do not include further analysis of the differences in the use of head-complement constructions by LLMs and by human authors, but in future work, it would be interesting to see, for example, whether there is a difference in subconstituents or in the lexical types or entries forming the head-complement constituent itself.

in Appendix D. However, we perform a large number of comparisons, and when we apply FDR correction to the p-values, none of them come out as significant, which is not surprising given that we only have 9 datasets to compare.

5.1.2 Syntactic long tail

It is possible that some salient differences lie in the "long tail" of the distributions (not shown in Figure 5). The ERG is a unique resource to study this long tail, being a comprehensive representation of the English language which, while validated empirically, was developed with close attention to a wide range of phenomena, not only the most frequently occurring ones.

The following constructions occur only 0 or 1 times in a sample from human-authored NYT text, while similar size samples from the LLM-generated texts contain more than 10 instances: sequence of numbers; fragment lexical conjunction ("But!"); parenthetical modifier ("Some person (tall) was running away"); mass noun coordination ('sand and gravel');[14] modifier phrase formed from 'mea-

[14]Note the special syntactic properties of this construction,

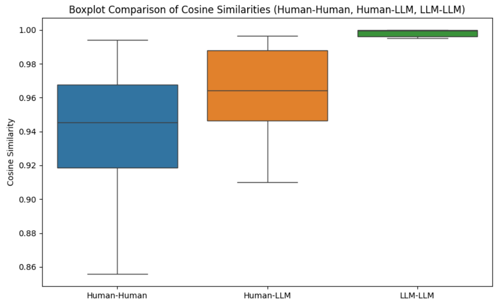

Figure 6: Humans vary more from one another than they do from an LLM, and LLMs vary little from each other.

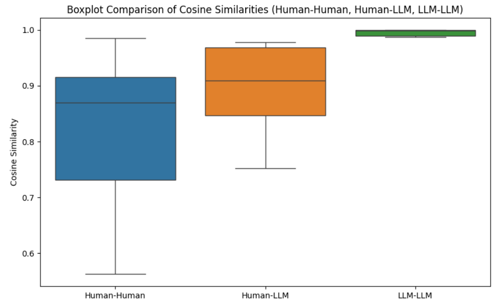

Figure 7: Human authors have particularly large variance when it comes to the lexical types they use

sure' nouns (*We slept the last mile*).

Humans use all of these long-tail constructions occasionally (which is how they came to be represented in the ERG in the first place); their not occurring in the NYT dataset could just be by chance. Future experiments with more data are needed. In the meantime, we show that HPSG analysis aligns with previous findings with UD (e.g. that current LLMs are known to favor numbers and measure-related vocabulary (Muñoz-Ortiz et al., 2024)), and identify constructions possibly typical for LLMs which have not previously been noted (§6.1).

5.1.3 Lexical (morphological) rules

We do not observe any differences of note in the LLM and human use of frequent lexical rules (inflectional and derivational morphology),[15] except in all human-authored datasets, plural nouns have been used with greater relative frequency than in the LLM-generated texts (but there is more variation between the genre/style). This shows once again the importance of separating morphological information from syntactic and lexical when analyzing language (cf. Bender and Good 2005).

5.1.4 Lexical entries and types

Human writers use roughly twice as many different lexical *entries* as each LLM taken separately (Table 3). This confirms previous findings that humans show more variation in vocabulary use (see §2). But if we combine all of the LLM-generated data and sample from it, this collective LLM author has a greater lexical diversity than the human authors. This calls for further investigation of what makes the collective LLM vocabulary more varied. As for lexical *types*, LLMs seem to have greater diversity in terms of just the number of unique lexical types they use in the sample (with the exception of falcon-7B). When we look at the specific lexical types accounting for these distinct footprints, we see that of the 66 types which do not occur in any of the LLMs, 43 belong to the bottom 10% in terms of frequency, 21 to the bottom 25%, and only 2 to the bottom 50%. The two frequent ones include a special kind of mass noun such as 'next' in 'The next is Kim', and the special kind of 'if' such as in 'The happy if confused customer left' (the customer was confused, but was happy nevertheless).

Model	Lexical Types		Lexical Entries	
	Not in	Only in	Not in	Only in
llama 7B	62	70	5,704	2,519
llama 13B	71	80	5,557	2,617
llama 30B	65	62	5,531	2,608
llama 65B	66	74	5,302	2,745
mistral 7B	73	76	5,809	2,353
falcon 7B	91	55	6,212	2,015
all llms	66	70	1,721	2,398

Table 3: Lexical types and entries found only in human-authored or only in synthetic data, sample 25K.

5.1.5 Individual author variance

In addition to looking at NYT human authors collectively, we are interested in how much they differ from each other and whether these individual differences are greater or not than the differences between humans and LLMs (Figures 6-7). We perform the comparison with 12 authors that have more than 100 sentences attributed to them in the NYT data. The comparison is again based on cosine similarity, where the vectors are construction/type frequencies normalized by total number

such as underspecified number agreement: *Sand and gravel has/have arrived.*

[15] The only infrequent rule of note is the one related to currencies ("A one-dollar book", where the rule is responsible for the special currency-related properties of the phrase "one dollar", as compared to any generic noun phrase).

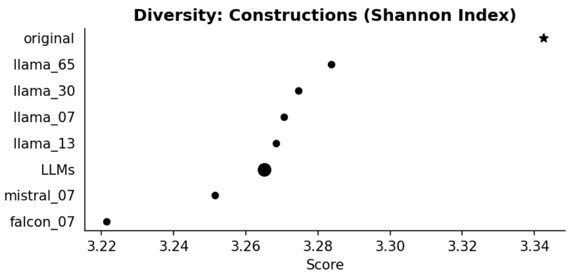

Figure 8: Construction Diversity (Shannon Index)

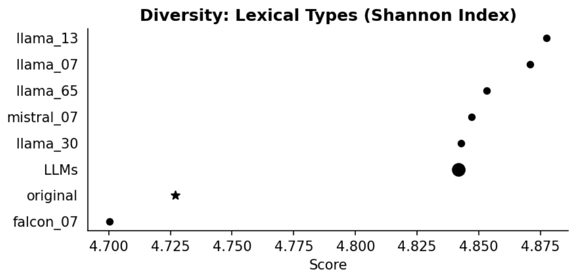

Figure 9: Lexical Type Diversity (Shannon Index)

of occurrences in the data. Here we include a comparison based on all the HPSG types together.

We find that human writers differ from each other more than a human author differs from an LLM, and LLMs differ very little from each other (Figure 6). If we look at lexical types, we see that humans vary particularly strongly in their use of lexical types, while LLMs have the same kind of small variance in this respect as they do in other types of constructions (Figure 7).[16]

As far as we know, our study is the first pairwise comparison of human authors and LLMs along detailed grammatical dimensions, and we show for the first time that a human-authored text is more similar to an LLM-generated text than to another human-authored text (by a different author). This makes sense if we see an LLM-generated text as "averaged" with respect to grammatical features that humans use in their language. This can also be seen in their increased use of the most general structures such as the head-complement phrase (Figure 5). Our results also confirm the previous observations that LLMs are very similar to each other in terms of the types of constructions that they use (see §2).

5.2 Diversity

To quantify diversity in the texts we applied two biodiversity measures that have become standard in stylometry and authorship attribution (McKinney, 2010; Stamatatos, 2009): **Shannon entropy** H and the **Gini–Simpson index** $1-\lambda$. The former captures the balance (evenness) of the distribution, while the latter is interpretable as the probability that two randomly drawn tokens belong to different types. Because both indices give the same rank orderings in our data (see Appendix C), we only discuss Shannon entropy here.

Constructions Figure 8 plots H for syntactic constructions. Human-produced texts ("original") are clearly the most diverse ($H = 3.342$), and all language-model outputs fall below that benchmark ($H = 3.221$–3.284).[17] The *largest* LLaMa model (65 B) is the closest to humans ($H = 3.284$), whereas the Falcon model is the least diverse ($H = 3.221$). Interestingly, when we pool every LLM output into a single corpus, its diversity *drops* slightly to $H = 3.265$. Aggregation adds a handful of rare constructions that were unique to individual models, but it also amplifies the high-frequency, general constructions that all models share, skewing the distribution and lowering overall entropy.

Lexical types The pattern reverses when we consider lexical types (Figure 9). Here, LLM outputs are more diverse than human-authored texts: the least diverse system (Falcon) scores $H = 4.700$, followed by the original human data at $H = 4.727$. All other LLMs surpass humans, with LLaMA-13B at the top ($H = 4.877$). These differences are statistically significant ($p < 0.01$). Investigating this pattern reversal is future work.

6 Examples of salient differences

6.1 Syntactic constructions

We have examined some of the constructions which are used noticeably more by human authors than by the collective LLM, or vice versa.[18] The constructions where the difference in relative frequency is most clear notably include the head-complement construction and the subject-head construction — the two most basic constructions forming any typical clause. Here we do not attempt to analyze the numerous examples of this kind of construction use

[16] Since we have more data for each LLM than for each human, we confirmed that we see similar distributions in a balanced dataset, if we sample randomly from the LLM data.

[17] A permutation test with 10,000 resamples confirms a reliable gap ($p < 0.01$).

[18] We have selected such constructions based on the statistical significance of the comparison between relative frequencies.

Construction	Ex	Humans	LLMs (avg)
Absolute VP	'As told, ...'	10	3.8
Double NP apposition	'an eye for detail, decades of a culture in transition'	11	5.2
Double appos. modifier	'accurate, but inadequate, descriptor'	12	5.6
Adjective-participle modifier	'right-handed', 'red-colored'	125	64.6
Bare NP coordination	'..., author and commentator, ...'	311	117
Paired marker	'Both this article and other discussions', 'not only...'	326	185
Adjective coordination	'emotional and spiritual'	390	625
Modifier clause appos.	'his critics, mostly unnamed'	826	434
Participial clause	'...having tried that,...'	1,736	1,116
Inverted adjunct	'Below are some of the facts...'	5	14.8
Clause-clause coordination	'which ones are and which ones aren't'	45	105
Filler-head non-question wh	'How best to proceed: [...]'	149	306
Questions	'How do you stay safe?'	268	428
Clause conjunction fragment	'But the observation suits him.'	939	2,076
Marker clause	'..., and that's a good thing'	2,891	5,660
Relative clauses	'...a vote that many in Europe have seen as a bellwether or support...'	4,929	6,721
Clause with extracted subject	'Chris Snow, [...], became an advocate for the victims of the disease.'	5,072	7,327
Subject-head	'The house passed the measure earlier this week.'	17,850	27,753
Quantity NP	'many in Europe'	23,611	40,881
Head-complement	'It's not acceptable for democracy'	164,806	224,529

Table 4: Examples of selected syntactic constructions which seem to have noticeably different frequency in human-authored and in LLM-generated data (25K sentence sample)

(leaving it to future work) but nonetheless include an example from the corpus for each (Table 4).

Table 4 aligns with some of the previous findings (Muñoz-Ortiz et al. 2024 and Sardinha 2024, among others), namely that LLMs tend to use more quantity-related words and phrases; that LLM-generated texts have more structures which can be classified as a generic 'verb phrase' (VP) or 'sentence' (S), which in our analysis would correlate with the higher frequencies of head-complement and head-subject constructions; that LLMs tend to use more clause coordination; and that human authors tend to produce more prepositional phrases in the NYT-style writing. However, we do not confirm the finding of Sardinha (2024) that LLMs use more participial modifiers; in our data, humans use it more. In addition, we can hypothesize several other systematic differences using the ERG elaborate syntactic type hierarchy. According to our analysis, the LLMs collectively tend to use more relative clauses and questions, more clause chains, more clauses with extraposed subjects, and more extraposed adjuncts. In contrast, human authors use more stylistic devices such as participial modifiers, full clause modifiers, double adjective apposition, coordinated prepositional phrases, coordinated adjective modifiers, double noun phrase apposition, and the so-called absolute verb phrase. In summary, human authors use more of the lower frequency, stylistically special constructions.

6.2 Lexical types and lexical entries

There are many differences between the lexical footprints of LLM-generated and human-authored text in terms of low-frequency items. If the word is both low frequency and belongs to a lexical type which does not have many members, it is hard to say whether its use is just an accident or could be informative. Therefore, we focus on items which are high frequency but occur only in human-authored or only in LLM-generated data (Tables 5-6).[19]

We take advantage of the ERG lexical type hierarchy and look at how the lexical entries which seem to distinguish LLM-generated text from human-authored text can be grouped together in grammatical terms. One example of the lexical entries found only in human-authored data is 'law_n2' (with a clausal complement). This lexical entry is present in the ERG lexicon along with the mass-count noun 'law_n1' and belongs to a different lexical type. The word 'law' certainly occurs in LLM-generated data as well, but only as the mass-count noun. We find that only in the human-authored data is this word used as something that can take a clausal complement, e.g. 'There is a law that...'. This is the kind of distinction that we are looking for in our study; if we did not have the ERG lexicon

[19]We must note that such differences can always be attributed to sampling. Obviously, a human writer can use any of the items from Table 6, and it is trivial to have an LLM produce any of the things from Table 5.

Lex. entry	occurr.	example
OOV verb	178	'twerk', 'steamroll'
risk_n3	144	'at your own risk'
haven't	88	'If you haven't already...'
night_def	82	'spend the night'
a_per_p	81	'a night', 'a barrel'
see_imp	69	'See the results...'
including_pp	65	'...including on April 17'
yet_conj	64	'...yet there it is'
dozen_a1	62	'a couple dozen pages'
winter_n1	61	'With winter approaching,..'
down_vmod	59	'walk/skip/sprint down'
almost_deg2	56	'almost always'
present_v1	51	'A puzzle presented to students'
over_pp	50	'The wait is over.'
black_n2	50	'growing up Black'

Table 5: Frequent (top 15) lexical entry usages unique for the human-authored dataset

Lex. entry	occurr.	example
ellipsis	202	'She was 86...'
and_or_conj	156	'SF/SPCA'
like_comp	125	'It looks like the case...'
num_ne	119	'28th of July, 1966'
square_brack	117	'...using [the law]'
time_ne	100	'January 31st, 2019 5:34 pm'
please_root	100	'Please write to corrections'
be_nv_is_cx_3	96	'That's why we did it.'
then_adv	82	'by/since then'
fact_n2	81	'The fact that...'
wasn't	81	'It wasn't that loud'
clear_a2	70	'It is not clear how.'
OOV noun	76	'Anwar al-Awlaki'
won't	70	'He won't care...'
realize_v2	69	'I realized that...'

Table 6: Frequent lexical entry usages unique for the LLaMa 65B-generated dataset

at our disposal, we could overlook the distinction.[20]

One of the main things that we see in Table 5 is that the real (human) authors of the NYT use more informal language even though they are following a style guide. A LLM certainly could also use expressions like 'a couple dozen' and 'haven't', and in fact it does use 'won't' and 'that's' (Table 6), but overall each LLM seems to be more consistently adhering to the style of the prompt. Another trait of the human-authored data is more direct/strong language, such as imperatives and expressions such as 'at your own risk'.[21] In contrast, the top frequent items unique for LLM-generated data contain entries belonging to numeric and punctuation types, in other words things related to formal presentation of the text. We note also the words 'fact' and 'clear' as generic but persuasive, and as such perhaps typical for LLM language (Table 6).

7 Conclusion

We present the first systematic comparison of LLM- and human-authored text through the lens of a formal grammatical theory (HPSG). We leverage the English Resource Grammar's explicit modeling of the principles of English syntax and lexicon, where detailed lexical types reflect the nuances of syntactic behavior of words.

Comparing to the previous study by Muñoz-Ortiz et al. (2024), which used the same dataset but employed the UD syntactic framework, our analysis through the lens of formal syntactic theory confirms the validity of its conclusions even at a finer-grained level. It also offers greater detail on specific constructions that distinguish LLM-generated text from human-authored news. Our study also reaches novel conclusions on the same dataset by comparing individual human authors between each other as well as to LLMs.

We find that overall, LLMs tend to be more similar to each other along these grammatical dimensions than to humans. We show the importance of separating syntactic analysis from morphological, and that in the use of morphological rules, LLMs and humans are strikingly similar within the NYT genre. We find that human authors show greater variation between each other than a human-LLM pair; an LLM appears as an "average" human author. Further investigation of this syntactic and lexical flattening should be the subject of future papers, now that we have laid the groundwork of methodology, presented our analytical tools, and identified specific HPSG types to look into.

Diversity indices show human-authored news as clearly distinct from all the LLMs (more diverse); however this is not so if we only look at lexical types. This opens up specific areas for future work.

We present some examples of constructions that occur more in human-authored than in LLM-generated news texts, and vice versa, confirming some but not other previous findings (such as the use of participial modifiers as more characteristic of LLM-generated text, which we do not confirm). Further experiments with various sampling techniques can provide further insight; in any case, using a resource such as the ERG is a way to ensure consistency and depth with respect to data analysis.

[20] Of course such distinctions should correlate with the differences in syntactic construction use.

[21] We have no ready explanation for why the word 'winter' (without the article) would only occur in human-authored data, or why the verb 'realize', in its most common usage, would happen to not occur there.

Limitations

There are many methodological limitations related to work with LLM-generated text. An LLM will generate a different text every time, and a lot depends on the prompt, and our resources in terms of generation are limited. Otherwise the main limitation here is that we only look at one genre (NYT-style news). We do include other types of data and our analysis of the overall distribution reflects this; however in our discussion of specific examples we still focus on the NYT-style data. Another limitation is that we only have large HPSG grammars for a handful of languages, and indeed only the ERG is big enough to cover 94% of news text, limiting the utility of our approach in comparisons of text in other languages. This is why our study is only about English.

Acknowledgments

We thank Ann Copestake and Emily M. Bender, and more generally the DELPH-IN community for the useful discussion related to the paper.

We acknowledge grants SCANNER-UDC (PID2020-113230RB-C21) funded by MICIU/AEI/10.13039/501100011033; 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 for minor copy-editing (e.g. thesaurus suggestions) and for visualization ideas. We have used GitHub copilot for code autocompletion.

References

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, et al. 2023. The Falcon series of open language models. *arXiv preprint arXiv:2311.16867*.

AJ Alvero, Jinsook Lee, Alejandra Regla-Vargas, René F Kizilcec, Thorsten Joachims, and Anthony Lising Antonio. 2024. Large language models, social demography, and hegemony: comparing authorship in human and synthetic text. *Journal of Big Data*, 11(1):138.

Emily Bender and Jeff Good. 2005. Implementation for discovery: A bipartite lexicon to support morphological and syntactic analysis. In *Proceedings from the annual meeting of the Chicago Linguistic Society*, volume 41, pages 1–16. Chicago Linguistic Society.

Emily M. Bender, Joshua Crowgey, Michael Wayne Goodman, Kristen Howell, Haley Lepp, Fei Xia, and Olga Zamaraeva. 2020. Aggregation: Building computational resources automatically from igt. Invited poster at Reflections on the Impact of DEL-funded Research Over Fifteen Years, LSA 2020, New Orleans, LA.

Emily M. Bender, Scott Drellishak, Antske Fokkens, Laurie Poulson, and Safiyyah Saleem. 2010. Grammar customization. *Research on Language and Computation*, 8:1–50.

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.

Douglas Biber. 1995. *Dimensions of register variation: A cross-linguistic comparison*. Cambridge University Press.

Douglas Biber and Susan Conrad. 2019. *Register, genre, and style*. Cambridge University Press.

Jan Buys and Phil Blunsom. 2017. Robust incremental neural semantic graph parsing. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1215–1226.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, et al. 2018. Universal sentence encoder for English. In *Proceedings of the 2018 conference on empirical methods in natural language processing: system demonstrations*, pages 169–174.

Yufei Chen, Weiwei Sun, and Xiaojun Wan. 2018. Accurate SHRG-based semantic parsing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 408–418, Melbourne, Australia. Association for Computational Linguistics.

Ann Copestake. 2002. Definitions of typed feature structures. In Stephan Oepen, Dan Flickinger, Jun-ichi Tsujii, and Hans Uszkoreit, editors, *Collaborative Language Engineering*, pages 227–230. CSLI Publications, Stanford, CA.

Berthold Crysmann and Woodley Packard. 2012. Towards efficient HPSG generation for German, a non-configurational language. In *COLING*, pages 695–710.

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.

Valerie Hajdik, Jan Buys, Michael W Goodman, and Emily M Bender. 2019. Neural text generation from rich semantic representations. In *Proceedings of NAACL-HLT*, pages 2259–2266.

Charles R Harris, K Jarrod Millman, Stéfan J Van Der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J Smith, et al. 2020. Array programming with numpy. *Nature*, 585(7825):357–362.

Kristen Howell and Emily M Bender. 2022. Building analyses from syntactic inference in local languages: An HPSG grammar inference system. *Northern European Journal of Language Technology*, 8(1).

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7B. *arXiv preprint arXiv:2310.06825*.

Tom S Juzek and Zina B Ward. 2025. Why does ChatGPT "delve" so much? Exploring the sources of lexical overrepresentation in large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6397–6411.

Zi Lin, Jeremiah Zhe Liu, and Jingbo Shang. 2022. Towards collaborative neural-symbolic graph semantic parsing via uncertainty. *Findings of the Association for Computational Linguistics: ACL 2022*.

Mitchell Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. University of Pennsylvania Department of Computer and Information Science Technical Report No. MS-CIS-93-87.

Wes McKinney. 2010. Data Structures for Statistical Computing in Python. In *Proceedings of the 9th Python in Science Conference*, pages 56 – 61.

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.

Arun Balajiee Lekshmi Narayanan, Priti Oli, Jeevan Chapagain, Mohammad Hassany, Rabin Banjade, Peter Brusilovsky, and Vasile Rus. 2024. Explaining code examples in introductory programming courses: LLM vs humans. In *AI for Education: Bridging Innovation and Responsibility at the 38th AAAI Annual Conference on AI*.

Joakim Nivre, Marie-Catherine De Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, et al. 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.

Stephan Oepen, Dan Flickinger, Kristina Toutanova, and Christopher D Manning. 2004. LinGO Redwoods. *Research on Language and Computation*, 2(4):575–596.

Mose Park, Yunjin Choi, and Jong-June Jeon. 2025. Does a large language model really speak in human-like language? *arXiv preprint arXiv:2501.01273*.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

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.

Alex Reinhart, David West Brown, Ben Markey, Michael Laudenbach, Kachatad Pantusen, Ronald Yurko, and Gordon Weinberg. 2024. Do LLMs write like humans? Variation in grammatical and rhetorical styles. *arXiv preprint arXiv:2410.16107*.

Morgan Sandler, Hyesun Choung, Arun Ross, and Prabu David. 2024. A linguistic comparison between human and ChatGPT-generated conversations. *arXiv preprint arXiv:2401.16587*.

Tony Berber Sardinha. 2024. AI-generated vs human-authored texts: A multidimensional comparison. *Applied Corpus Linguistics*, 4(1):100083.

Chantal Shaib, Yanai Elazar, Junyi Jessy Li, and Byron C Wallace. 2024. Detection and measurement of syntactic templates in generated text. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6416–6431.

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

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Olga Zamaraeva, Chris Curtis, Guy Emerson, Antske Fokkens, Michael Wayne Goodman, Kristen Howell, TJ Trimble, and Emily M Bender. 2022. 20 years of the Grammar Matrix: Cross-linguistic hypothesis testing of increasingly complex interactions. *Journal of Language Modelling*, 10(1):49–137.

Appendices

A English Resource Grammar types

Table 7 shows the construction types appearing in Figure 5 with an expanded name and an example. This is a slightly modified version of the English Resource Grammar documentation.

B Cosine similarities

Tables 8-10 present the data underlying Figures 2-4 in §5.

C Diversity Measures

Figure 10 shows the diversities of constructions, lexical types and lexical rules measured with both the Shannon Index (on the left) and Simpson Index (on the right), as discussed in Section 5.2. Scores for the original human-generated sentences are shown with a star ($\star$), LLMs with a dot ($\bullet$) and the combined LLMs with a larger dot (⬤)

We measured the significance of the difference between the original human-generated sentences and combined LLM sentences using a permutation test, sampled 10,000 times. All combinations had an observed p-value of less than 0.01, except for the Lexical Rules measured with the Simpson Index (which is less sensitive to outliers), with $p = 0.13$.

D Mann-Whitney U-test

In this Appendix, we report the HPSG types for which the difference in relative frequency comes out as statistically significant ($p \leq 0.05$; Tables 11-13). However, when we apply the FDR correction, none of these p-values remain below the 0.05 threshold. The definitions and examples for all of these HPSG types can be found in the English Resource Grammar files.[22] The examples of where these types come up in the NYT corpus can be found in the data associated with this paper.[23]

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

[23] https://github.com/olzama/llm-syntax/releases/tag/1.0.0

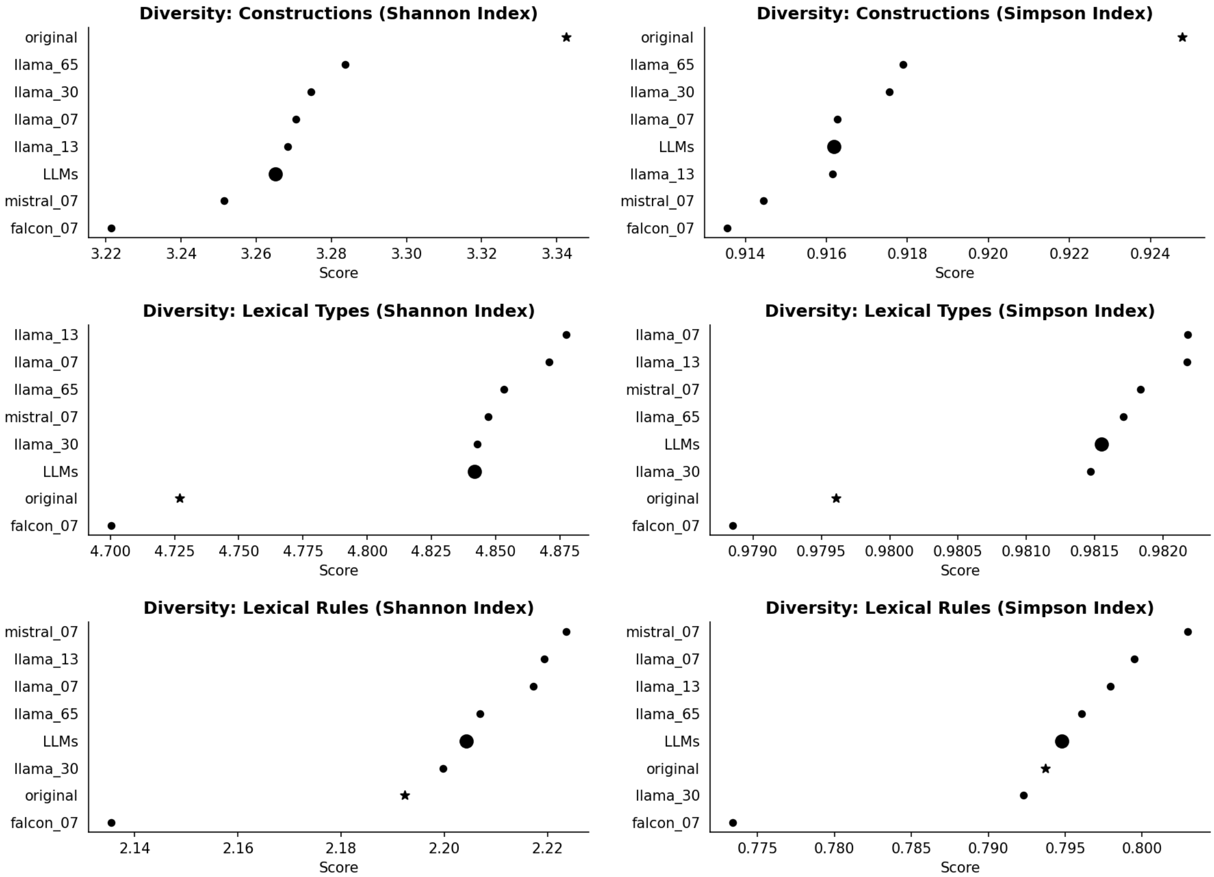

Figure 10: Diversity (Shannon and Simpson Indices)

Type Name	Definition	Example
sb-hd_mc_c	Head+subject, main clause	C arrived.
sb-hd_nmc_c	Hd+subject, embedded clause, subj has no gap	B thought [C arrived].
hd-cmp_u_c	Hd+complement	B [hired C].
hd_optcmp_c	Head discharges optional complement	B [ate] already.
hdn_optcmp_c	NomHd discharges opt complement	The [picture] appeared.
mrk-nh_evnt_c	Marker + event-based complement	B sang [and danced.]
mrk-nh_cl_c	Marker + clause	B sang [and C danced.]
mrk-nh_nom_c	Marker + NP	Cats [and some dogs] ran.
mrk-nh_n_c	Marker + N-bar	Every cat [and dog] ran.
hd_xcmp_c	Head extracts compl (to SLASH)	Who does B [admire] now?
hd_xsb-fin_c	Extract subject from finite hd	Who do you think [went?]
sp-hd_n_c	Hd+specifier, nonhd = sem hd	[Every cat] slept.
sp-hd_hc_c	Hd+specifier, hd = sem hd	The [very old] cat slept.
aj-hd_scp_c	Hd+preceding scopal adjunct	Probably B won.
aj-hd_scp-xp_c	Hd+prec.scop.adj, VP head	B [probably won].
hd-aj_scp_c	Hd+following scopal adjunct	B wins if C loses.
aj-hdn_norm_c	Nominal head + preceding adjnct	The [big cat] slept.
aj-hdn_adjn_c	NomHd+prec.adj, hd pre-modified	The [big old cat] slept.
aj-hd_int_c	Hd+prec.intersective adjunct	B [quickly left].
hdn-aj_rc_c	NomHd+following relative clause	The [cat we chased] ran.
hdn-aj_rc-pr_c	NomHd+foll.rel.cl, paired pnct	A [cat, which ran,] fell.
hdn-aj_redrel_c	NomHd+foll.predicative phrase	A [cat in a tree] fell.
hd-aj_int-unsl_c	Hd+foll.int.adjct, no gap	B [left quietly].
hd_xaj-int-vp_c	Extract int.adjunct from VP	Here we [stand.]
vp_rc-redrel_c	Rel.cl. from predicative VP	Dogs [chasing cats] bark.
hdn_bnp_c	Bare noun phrase (no determiner)	[Cats] sleep.
hdn_bnp-pn_c	Bare NP from proper name	[Browne] arrived.
hdn_bnp-num_c	Bare NP from number	[42] is even.
hdn_bnp-qnt_c	NP from already-quantified dtr	[Some in Paris] slept.
hdn_bnp-vger_c	NP from verbal gerund	Hiring them was easy.
np-hdn_cpd_c	Compound from proper-name+noun	The [IBM report] arrived.
np-hdn_ttl-cpd_c	Compound from title+proper-name	[Professor Browne] left.
np-hdn_nme-cpd_c	Compound from two proper names	[Pat Browne] left.
n-hdn_cpd_c	Compound from two normal nouns	The [guard dog] barked.
np_adv_c	Modifier phrase from NP	B arrived [this week.]
hdn_np-num_c	NP from number	[700 billion] is too much.
flr-hd_nwh_c	Filler-head, non-wh filler	Kim, we should hire.
flr-hd_wh-nmc-fin_c	Fill-head, wh, fin hd, embed cl	B wondered [who won.]
flr-hd_rel-fin_c	Fill-head, finite, relative cls, NP gap	people [who we admired]
vp-vp_crd-fin-t_c	Conjnd VP, fin, top	B [sees C and chases D.]
cl-cl_crd-t_c	Conjoined clauses, non-int, top	B sang and C danced.
np-np_crd-t_c	Conjoined noun phrases, top	[The cat and the dog] ran.
num-n_mnp_c	Measure NP from number+noun	A [two liter] jar broke.
cl_np-wh_c	NP from WH clause	[What he saw] scared him.
vp_np-ger_c	NP from verbal gerund	Winning money [pleased C.]
num_det_c	Determiner from number	[Ten cats] slept.
cl_cnj-frg_c	Fragment clause with conjunctn	And Kim stayed.
hd-pct_c	Head + punctuation token	B [arrived -] C left.
hd-pct_nobrk_c	Punctuation unrelated to bracketing	
pct-hd_c	Punctuation token + head	B arrived (today)

Table 7: Construction types and examples.

Model 1	Model 2	Cos
llama_30	llama_65	0.9999
llama_07	llama_13	0.9999
llama_07	mistral_07	0.9999
llama_13	llama_65	0.9998
llama_07	llama_65	0.9998
llama_13	mistral_07	0.9998
llama_13	llama_30	0.9998
llama_07	llama_30	0.9997
llama_65	mistral_07	0.9996
llama_30	mistral_07	0.9996
falcon_07	llama_30	0.9976
falcon_07	mistral_07	0.9972
falcon_07	llama_65	0.9972
falcon_07	llama_07	0.9966
falcon_07	llama_13	0.9966
llama_30	**original NYT**	0.9965
llama_65	**original NYT**	0.9964
falcon_07	**original NYT**	0.9958
llama_07	**original NYT**	0.9955
mistral_07	**original NYT**	0.9950
llama_13	**original NYT**	0.9950
wsj	**original NYT**	0.9949
llama_65	wsj	0.9908
llama_30	wsj	0.9907
wikipedia	wsj	0.9900
llama_07	wsj	0.9899
mistral_07	wsj	0.9894
llama_13	wsj	0.9891
falcon_07	wsj	0.9881
wikipedia	**original NYT**	0.9833
llama_65	wikipedia	0.9768
llama_07	wikipedia	0.9765
llama_30	wikipedia	0.9764
mistral_07	wikipedia	0.9763
llama_13	wikipedia	0.9745
falcon_07	wikipedia	0.9738

Table 8: Cosine similarity between LLM-generated and human-authored (*original NYT*) datasets; only syntactic constructions included.

Model 1	Model 2	Cos
llama_30	llama_65	0.9999
llama_13	llama_65	0.9999
llama_07	llama_13	0.9999
llama_13	llama_30	0.9999
llama_07	llama_65	0.9998
llama_07	llama_30	0.9998
llama_07	mistral_07	0.9997
llama_13	mistral_07	0.9996
llama_30	mistral_07	0.9995
llama_65	mistral_07	0.9995
falcon_07	llama_30	0.9984
falcon_07	llama_13	0.9982
falcon_07	llama_65	0.9980
falcon_07	llama_07	0.9978
falcon_07	mistral_07	0.9977
llama_30	**original NYT**	0.9976
llama_65	**original NYT**	0.9975
llama_07	**original NYT**	0.9969
llama_13	**original NYT**	0.9968
mistral_07	**original NYT**	0.9965
falcon_07	**original NYT**	0.9956
wsj	**original NYT**	0.9922
llama_07	wsj	0.9909
llama_13	wsj	0.9908
llama_65	wsj	0.9906
mistral_07	wsj	0.9906
llama_30	wsj	0.9897
falcon_07	wsj	0.9837
wikipedia	wsj	0.9724
wikipedia	**original NYT**	0.9600
llama_07	wikipedia	0.9579
mistral_07	wikipedia	0.9579
llama_65	wikipedia	0.9570
llama_30	wikipedia	0.9565
llama_13	wikipedia	0.9559
falcon_07	wikipedia	0.9506

Table 9: Cosine similarity between LLM-generated and human-authored (*original NYT*) datasets; only lexical type constructions included.

Model 1	**Model 2**	**Cos**
llama_13	llama_65	0.9999
llama_07	llama_65	0.9999
llama_30	llama_65	0.9999
llama_07	llama_13	0.9999
llama_13	llama_30	0.9998
llama_07	mistral_07	0.9998
llama_13	mistral_07	0.9997
llama_07	llama_30	0.9996
llama_65	mistral_07	0.9996
llama_30	mistral_07	0.9993
llama_65	**original NYT**	0.9990
llama_07	**original NYT**	0.9989
llama_30	**original NYT**	0.9989
llama_13	**original NYT**	0.9987
mistral_07	**original NYT**	0.9985
falcon_07	llama_30	0.9983
falcon_07	llama_13	0.9976
falcon_07	llama_65	0.9975
falcon_07	llama_07	0.9970
falcon_07	mistral_07	0.9966
falcon_07	**original NYT**	0.9962
wsj	**original NYT**	0.9932
llama_07	wsj	0.9923
mistral_07	wsj	0.9923
llama_65	wsj	0.9913
llama_13	wsj	0.9908
llama_30	wsj	0.9899
falcon_07	wsj	0.9822
wikipedia	wsj	0.9666
mistral_07	wikipedia	0.9476
wikipedia	**original NYT**	0.9474
llama_07	wikipedia	0.9464
llama_65	wikipedia	0.9427
llama_13	wikipedia	0.9418
llama_30	wikipedia	0.9393
falcon_07	wikipedia	0.9305

Table 10: Cosine similarity between LLM-generated and human-authored (*original NYT*) datasets; only lexical rule constructions included.

Table 11: Mann-Whitney U-test ($p \leq 0.05$) — Syntactic constructions

Frequent	
aj-hd_int_c	0.0238
aj-hdn_adjn_c	0.0238
aj-hdn_norm_c	0.0238
cl-cl_crd-t_c	0.0238
cl_cnj-frg_c	0.0476
cl_np-wh_c	0.0476
flr-hd_rel-fin_c	0.0238
flr-hd_wh-nmc-fin_c	0.0238
hd-aj_scp-pr_c	0.0238
hd-aj_vmod_c	0.0238
hd-cmp_u_c	0.0238
hd-pct_nobrk_c	0.0238
hd_xsb-fin_c	0.0238
hdn_bnp-qnt_c	0.0238
hdn_bnp_c	0.0238
mrk-nh_cl_c	0.0238
mrk-nh_n_c	0.0238
mrk-nh_nom_c	0.0476
n-hdn_cpd_c	0.0238
np-np_crd-t_c	0.0238
num_det_c	0.0476
sb-hd_mc_c	0.0238
vp_rc-redrel_c	0.0238
vp_sbrd-prd-prp_c	0.0238

Infrequent	
aj-hd_int-inv_c	0.0238
aj-hdn_crd-cma_c	0.0238
cl-cl_crd-int-t_c	0.0238
cl-np_runon_c	0.0238
cl_rc-inf-modgap_c	0.0476
cl_rc-inf-nwh_c	0.0476
flr-hd_nwh-nmc_c	0.0238
flr-hd_wh-mc_c	0.0476
flr-hd_wh-nmc-inf_c	0.0238
hd-aj_cmod-s_c	0.0476
hd-aj_vmod-s_c	0.0238
hd-hd_rnr-nb_c	0.0476
hd-hd_rnr-nv_c	0.0476
hd-hd_rnr_c	0.0238
hdn-aj_rc-asym_c	0.0238
hdn-aj_rc-propr_c	0.0238
hdn-aj_redrel-asym_c	0.0238
hdn-aj_redrel-pr_c	0.0238
hdn-np_app-dx_c	0.0238
hdn-np_app-mnp_c	0.0238
j-j_crd-att-t_c	0.0476
j-n_crd-m_c	0.0476
j-n_crd-t_c	0.0238

j_n-ed_c	0.0476
mrk-nh_atom_c	0.0238
n-hdn_cpd-pl-mnp_c	0.0431
n-hdn_cpd-pl_c	0.0238
n-j_j-cpd_c	0.0238
n-j_j-t-cpd_c	0.0238
n-n_crd-asym-t_c	0.0238
n-n_crd-div-t_c	0.0238
n-n_crd-im_c	0.0238
n-n_num-seq_c	0.0275
n-v_j-cpd_c	0.0238
np-np_crd-im_c	0.0238
np-np_crd-nc-m_c	0.0238
np_indef-adv_c	0.0476
np_nb-pr-frg_c	0.0238
num_prt-det-nc_c	0.0238
num_prt-of_c	0.0476
pp-pp_crd-im_c	0.0476
pp-pp_crd-t_c	0.0238
r_cl-frg_c	0.0476
sb-hd_q_c	0.0238
vp_sbrd-prd-pas_c	0.0238
vp_sbrd-pre-lx_c	0.0238
vp_sbrd-pre_c	0.0238

Table 12: Mann-Whitney U-test ($p \leq 0.05$) — Lexical types

Frequent	
aj_-_i-att_le	0.0238
aj_-_i-ord-one_le	0.0238
aj_pp_i-er_le	0.0238
aj_vp_i-seq_le	0.0238
av_-_dg-cmp-so_le	0.0238
av_-_dg-jo_le	0.0238
av_-_dg-sup_le	0.0238
av_-_i-vp-pr_le	0.0476
av_-_i-vp_le	0.0238
c_xp_but_le	0.0238
cm_np-vp_that_le	0.0238
cm_vp_to_le	0.0238
d_-_poss-my_le	0.0238
d_-_poss-our_le	0.0476
d_-_poss-their_le	0.0476
d_-_poss-your_le	0.0476
n_-_ad-pl_le	0.0476
n_-_c-ed-ns_le	0.0238
n_-_c-nocnh-cap_le	0.0238
n_-_c-ns_le	0.0238
n_-_c-time_le	0.0238
n_-_m-time_le	0.0476
n_-_m_le	0.0238

n_-_mc_le	0.0238
n_-_pn-sg_le	0.0238
n_-_pn-yoc-gen_le	0.0238
n_-_pr-dei-sg_le	0.0476
n_-_pr-he_le	0.0238
n_-_pr-i_le	0.0275
n_-_pr-it-x_le	0.0238
n_-_pr-it_le	0.0238
n_-_pr-me_le	0.0238
n_-_pr-rel-who_le	0.0238
n_-_pr-she_le	0.0238
n_-_pr-them_le	0.0476
n_-_pr-they_le	0.0238
n_-_pr-we_le	0.0238
n_-_pr-wh_le	0.0476
n_-_pr-you_le	0.0238
n_-_pr_le	0.0476
n_pp_c-ns_le	0.0238
n_pp_c-nsnc-of_le	0.0238
n_pp_c-pl_le	0.0238
n_pp_m_le	0.0476
n_vp_c_le	0.0238
p_cp_s_le	0.0476
p_np_i-ngap_le	0.0238
p_np_i-nm-poss_le	0.0476
p_np_ptcl_le	0.0476
pp_-_i-wh_le	0.0238
pt_-_bang_le	0.0238
pt_-_comma-informal_le	0.0238
pt_-_hyphn-rgt_le	0.0238
pt_-_period_le	0.0238
v_cp_fin-inf-q_le	0.0238
v_cp_prop_le	0.0238
v_np-cp_fin-inf_le	0.0238
v_np-pp_prop_le	0.0238
v_np-vp_bse_le	0.0238
v_np-vp_oeq_le	0.0238
v_np_be_le	0.0238
v_np_is-cx_le	0.0238
v_np_le	0.0238
v_np_poss_le	0.0238
v_np_was_le	0.0238
v_pp*-pp*_le	0.0238
v_prd_are-cx_le	0.0238
v_prd_been_le	0.0238
v_prd_being_le	0.0238
v_prd_is-cx_le	0.0238
v_prd_was_le	0.0238
v_prd_wre_le	0.0238
v_vp_has_le	0.0238
v_vp_have-f_le	0.0238
v_vp_seq_le	0.0238

v_vp_ssr_le	0.0238
Infrequent	
aj_-_i-att-er_le	0.0091
aj_-_i-one-nmd_le	0.0339
av_-_i-unk_le	0.0091
n_-_c-meas_le	0.0091
n_-_c-min_le	0.0091
n_-_m-hldy_le	0.0091
n_-_pn-abb_le	0.0339
n_-_pn-unk_le	0.0091
n_-_pr-her_le	0.0091
n_pp_c-dir_le	0.0091
pp_-_i-po-tm_le	0.0091

Table 13: Mann-Whitney U-test (p $\leq$ 0.05) — Lexical rules

Frequent	
n_det-mnth_dlr	0.0238
n_pl-irreg_olr	0.0238
n_pl_olr	0.0238
v_aux-cx-noinv_dlr	0.0238
v_j-nb-pas-tr_dlr	0.0238
v_n3s-bse_ilr	0.0238
v_nger-tr_dlr	0.0238
v_psp_olr	0.0238
Infrequent	
det_prt-of-agr_dlr	0.0476
j_enough-wc-nogap_dlr	0.0476
j_j-non_dlr	0.0238
j_j-un_dlr	0.0476
j_tough-compar_dlr	0.0238
n_n-hour_dlr	0.0238
v_aux-ell-ref_dlr	0.0476
v_aux-ell-xpl_dlr	0.0476
v_aux-sb-inv_dlr	0.0476
v_aux-tag_dlr	0.0238
v_j-nb-intr_dlr	0.0238
v_j-nb-pas-ptcl_dlr	0.0238
v_j-nme-tr_dlr	0.0238
v_v-pre_dlr	0.0476
v_v-re_dlr	0.0238
v_v-un_dlr	0.0238
w_mwe-3-wb_dlr	0.0219
w_mwe-wb_dlr	0.0476