Using Syntactic Theory to Study Properties of Human-Authored vs. LLM-generated News Text

Olga Zamaraeva[1], Ardián Gude[1], Roi Santos-Rios[1], Francis Bond[2], Dan Flickinger[3], and Carlos Gómez-Rodríguez[1]

[1] Universidade da Coruña, CITIC, Spain
olga.zamaraeva/carlos.gomez/adrian.lopez.gude/roi.santos.rios@udc.es
[2] Palacký University at Olomouc, Czechia
francis.bond@upol.cz
[3] Stanford University (retired), U.S.A.
danflick@alumni.stanford.edu

Abstract

This paper summarizes two recent studies that compare human-authored and large language model (LLM)-generated news texts using linguistic grammar theory.[1] By reusing a compiled symbolic grammar rather than training or invoking additional neural components, the pipeline provides interpretable measurements with modest and predictable compute. Within this energy-efficient setup, we summarize evidence that instruction-tuned LLMs occupy a narrower region of grammatical space than human journalists and earlier base models, while also being systematically easier to parse.

1 Introduction

Linguistic theory with fully explicit, stable formalisms offers compact, interpretable measurement with respect to linguistic properties of texts. In particular, Head-driven Phrase Structure Grammar (HPSG) [3], instantiated for English by the English Resource Grammar (ERG) [1], provides structured analyses of phrase and lexical types without the need to train neural parsers. The resource is compiled once and reused; updates require no computationally expensive procedures, yielding a pipeline whose footprint grows slowly and supports theory-grounded diagnostics sensitive to syntactic and lexical choices.

A central goal of this work is to situate LLM-produced language relative to human writing, specifically in the domain of New York Times-style news. We compare similarity and diversity of the grammatical and lexical repertoire across two generations of LLMs (2023 and 2025) and two corresponding collections of human-authored lead articles. Our experiments are relevant to the issues of stylistic homogenization, stability of journalistic conventions under model assistance, and possible trade-offs between expressiveness and regularity.

We summarize recent work comparing human news leads with LLM generations from the same headlines and initial cues: (i) a cross-sectional comparison of human leads and contemporaneous model outputs [4]; and (ii) a generational comparison of earlier base models and newer instruction-tuned models against time-matched human baselines [2]. Our findings show that human writers maintain similar characteristics with respect to the variety of grammar and lexical features which is distinct from LLM's inventories. Furthermore, human writers are more diverse with respect to both syntactic and lexical features than the newer generation of LLMs, while older LLMs are less diverse than humans syntactically but more diverse lexically. We hypothesize that this is the effect of instruction tuning.

[1]We thank the Spanish Ministry for Digital Transformation and Civil Service and 'NextGenerationEU'/PRTR under for Grant TSI-100925-2023-1.

2 Methods

We pair New York Times lead paragraphs with model-generated leads prompted from the same headlines and initial tokens, using two non-overlapping collection windows that bracket model generations (earlier base models and later instruction-tuned models) [4, 2]. All texts are parsed with the ERG, yielding interpretable structures grounded in HPSG [3, 1]. From successful parses we derive grammar-informed feature spaces: normalized counts of phrasal constructions and lexical types.

We compute standard diversity indices over construction and lexical-type distributions (entropy-style measures alongside a complementary index less sensitive to the long tail, such as Shannon and Simpson indices) [2]. For group comparison, normalized type-frequency vectors support simple similarity analyses that capture both separation (humans vs. models) and dispersion (inter-author vs. inter-model spread) [4]. Finally, the parser provides an operational profile—coverage, runtime, memory, and errors related to exceeding resource limits, giving an indirect probe of regularity and ambiguity in the texts.

3 Results and Conclusion

Across both data collection windows, LLM generations produced by instruction-tuned systems are systematically easier to parse than both earlier base models and human leads: runtime and memory per sentence drop, and hard cases become rare — and this is despite the fact that LLM-generated sentences become significantly longer on average compared to both human-written and earlier LLM-generated sentences. This pattern indicates more regular, grammar-friendly outputs and fewer edge phenomena in instruction-tuned generations [2].

Humans are consistently more diverse than LLMs in *syntactic* terms across both collection windows. The *lexical* ordering differs: earlier base models attain the highest lexical-type diversity, human sets lie below them, and newer instruction-tuned models are lowest and cluster most tightly. The lexical narrowing of newer models co-occurs with higher parsability and suggests an effect of instruction tuning, though testing this hypothesis remains future work [4, 2].

We conclude that a theory-driven pipeline can be energy-efficient without sacrificing analytical resolution. In this setting, we discover that instruction-tuned LLMs appear both more uniform and more grammar-regular than human leads and earlier base models. Human texts, by contrast, preserve a balance of common scaffolds and rarer constructions that maintain higher diversity and greater inter-author dispersion.

References

[1] Dan Flickinger. On building a more efficient grammar by exploiting types. In *INLG/EACL Workshop on Efficient Processing with HPSG*. 2000.

[2] Adrián Gude, Roi Santos-Rios, Francis Bond, Dan Flickinger, Carlos Gómez-Rodríguez, and Olga Zamaraeva. More aligned, less diverse? Analyzing the grammar and lexicon of two generations of LLMs. Under review.

[3] Carl Pollard and Ivan A. Sag. *Head-driven Phrase Structure Grammar*. University of Chicago Press and CSLI Publications, Chicago, IL and Stanford, CA, 1994.

[4] Olga Zamaraeva, Dan Flickinger, Francis Bond, and Carlos Gómez-Rodríguez. Comparing LLM-generated and human-authored news text using formal syntactic theory. In *Proceedings of ACL-2025*, 2025.