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

DELPH-IN 2025/ACL 2025

Olga Zamaraeva, Dan Flickinger, Francis Bond, Carlos Gómez-Rodríguez

Universidade da Coruña/CITIC, Palacký University at Olomouc

July 7, 2025

Introduction: Comparing LLM-generated and human-authored text

- Are there systematic differences in structure between LLM-generated and human-authored news?
  - Growing area of research
  - Most studies focus on building classifiers
  - Some provide comparisons wrt vocabulary[1]
  - Fewer compare with respect to grammatical features[2]
- This study:
  - Systematic comparison wrt grammatical features
  - based on independent linguistic theory

1 Sandler et al. 2024; Juzek and Ward 2025
2 Muñoz-Ortiz et al. 2024; Sardinha 2024

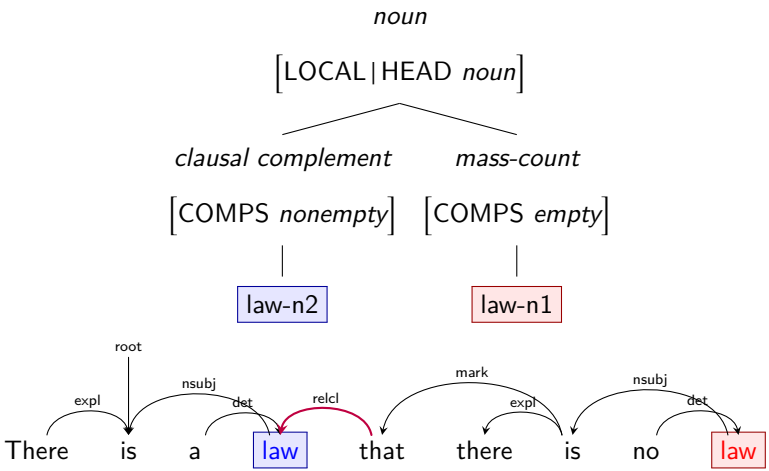
HPSG syntax theory

- Head-Driven Phrase Structure Grammar (HPSG)[3]
  - a theory of syntax based on constraint unification
  - developed by linguists independently of NLP tasks
  - well-formed sentences are feature structures where all constraints unify
- DELPH-IN HPSG:
  - Specific formalism and implementation
  - https://github.com/delph-in/docs/wiki/
- English Resource Grammar[4]
  - 94% accuracy on WSJ, Wikipedia, and more
- High consistency and precision

3 Pollard and Sag 1994

4 Flickinger 2000, 2011

HPSG hierarchy (simplified): Lexical types, abstraction over vocabulary



Experimental setup

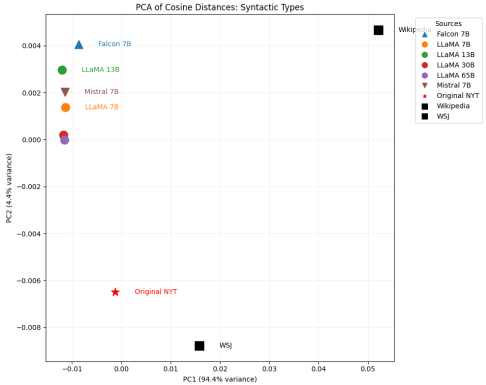
- Re-use the data from Muñoz-Ortìz et al. (2024)[5]
  - Original New York Times articles
  - Texts generated by LLMs with first 3 words of the original as prompts
- Parse all data with the ERG
- Study distributions of HPSG types (syntactic and lexical)

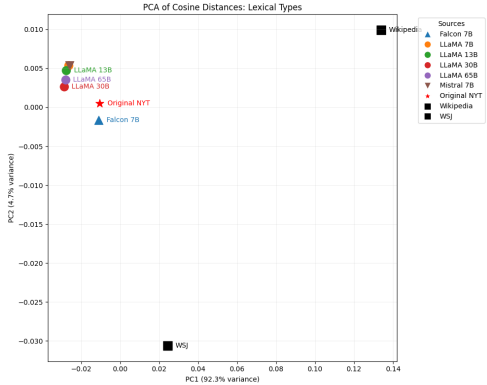
Dataset	# Sent. in dataset	Model size	Training tokens	Data sources
LLaMa	37,825	7B	1T	English CommonCrawl (67%), C4 (15%), GitHub (4.5%), Wikipedia (4.5%), Gutenberg and Books3 (4.5%), ArXiv (2.5%), Stack Exchange (2%)
	37,800	13B	1T	
	37,568	30B	1.5T	
	38,107	65B	1.5T	
Falcon	27,769	7B	1.5T	RefinedWeb-English (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

5 Muñoz-Ortiz et al. 2024

Results: More robust differences in syntax than in the vocabulary

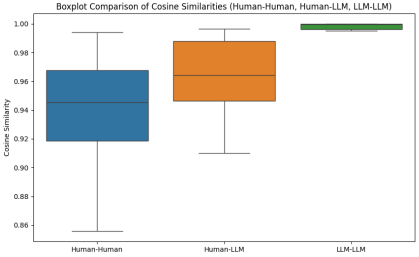
- Cosine similarity of datasets represented by type distribution:
  - Syntactic differences persist across genre
  - Genre appears more important than author's nature in terms of lexical types (vocabulary)

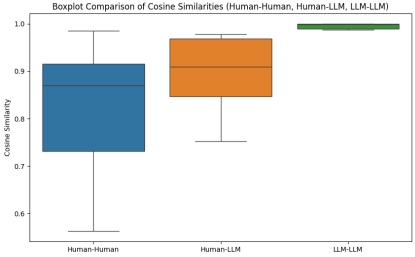




Results: Humans differ more from each other than from LLMs

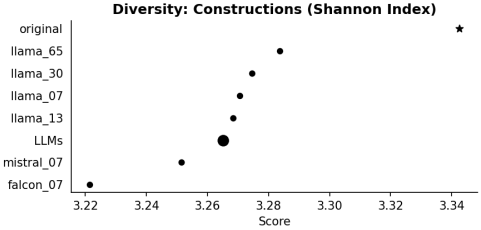
- LLMs are an 'average human author'
  - Makes sense, since LLM was trained on 'all' authors' texts
  - Humans vary even more with respect to lexical types they use (comared to syntactic constructions)

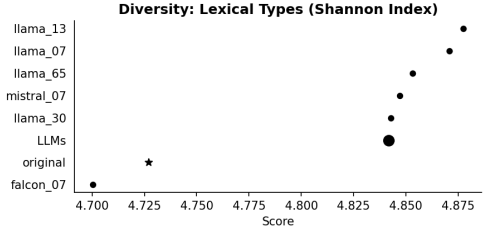




Results: Diversity index pattern shows opposite trends wrt syntax vs lexical

- Shannon diversity index:
  - Intriguing pattern reversal, especially given cosine similarity results
  - all LLMs together (large dot) are less diverse than each of them separately (except Falcon), but more than human authors





Examples of differences

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

Takeaways

- Linguistic theory developed independently from NLP is a robust evaluation/interpretation resource
- We can use them to analyze LLM outputs from a syntactic perspective and compare them to human-authored texts
- Differences in vocabulary distributions do not seem to persist across genres
- Differences in syntax seem to persist across genres
  - LLMs use more basic constructions and more relative clauses
  - Humans use more stylistically marked constructions
- Intriguing reverse patterns in diversity between syntactic and lexical types

Acknowledgments

We acknowledge the European Union's Horizon Europe Framework Programme which funded this research under the Marie Skłodowska-Curie postdoctoral fellowship grant HORIZON-MSCA-2021-PF-01 (GAUSS, grant agreement No 101063104); and the European Research Council (ERC), which has funded this research under the Horizon Europe research and innovation programme (SALSA, grant agreement No 101100615). We also 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), and Centro de Investigación de Galicia "CITIC", funded by the Xunta de Galicia through the collaboration agreement between the Consellería de Cultura, Educación, Formación Profesional e Universidades and the Galician universities for the reinforcement of the research centres of the Galician University System (CIGUS).

References I

Sandler, Morgan et al. (2024). "A Linguistic Comparison between Human and ChatGPT-Generated Conversations". In: *arXiv preprint arXiv:2401.16587*.

Juzek, Tom S 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*, pp. 6397–6411.

Muñoz-Ortiz, Alberto et al. (2024). "Contrasting linguistic patterns in human and LLM-generated news text". In: *Artificial Intelligence Review* 57.10, p. 265.

Sardinha, Tony Berber (2024). "AI-generated vs human-authored texts: A multidimensional comparison". In: *Applied Corpus Linguistics* 4.1, p. 100083.

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

Flickinger, Dan (2000). "On building a more efficient grammar by exploiting types". In: *Natural Language Engineering* 6.01, pp. 15–28.

— (2011). "Accuracy v. Robustness in Grammar Engineering". In: *Language from a Cognitive Perspective: Grammar, Usage and Processing*. Ed. by Emily M. Bender and Jennifer E. Arnold. Stanford, CA: CSLI Publications, pp. 31–50.