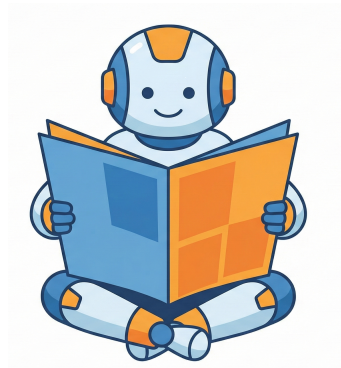
Lexical Diversity in News Texts

Pre-ChatGPT vs. Post-ChatGPT Classification

Topic 5 | NLP and IE 2025WS

16.01.2026

Jozef Janus | **Group 5**



Outline

1. Introduction
2. Milestone 1: Data & Preprocessing
3. Milestone 2: Baseline Methods
4. Results & Analysis
5. Discussion & Conclusion

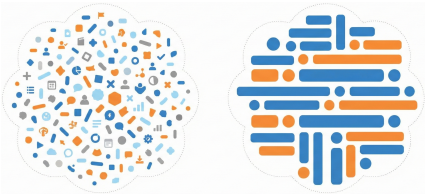
Introduction

What is Lexical Diversity?

Definition: A metric quantifying the variety of vocabulary used in a text.

Measure	Description
TTR	Type-Token Ratio (unique/total)
HD-D	Hypergeometric Distribution D
MTLD	Measure of Textual Lexical Diversity
VocD	Vocabulary Diversity

- **High Diversity:** Rich, varied vocabulary.
- **Low Diversity:** Repetitive, simple vocabulary.



Motivation: Why Pre- vs. Post-ChatGPT?

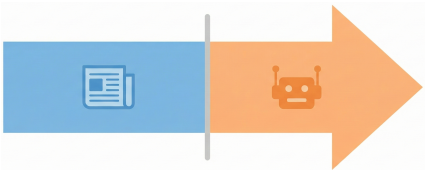
Context

ChatGPT Public Launch: **November 30, 2022**

Core Questions:

- Has AI-assisted writing permeated news media?
- Are there measurable shifts in vocabulary richness?
- Is writing style a temporal marker?

Implications: AI detection, journalism integrity, and longitudinal linguistic drift.



Research Question

Primary Investigation

*"Can lexical diversity and related features reliably predict whether a text was written **pre**- or **post**-ChatGPT?"*

The Task: Binary Classification

- **Class 0 (Pre):** Published Jan 2019 – Nov 2022
- **Class 1 (Post):** Published Dec 2022 – Dec 2024

Methodology Overview

1. **Data Collection:** Leipzig Corpora (English + Czech).
2. **Preprocessing:** Article reconstruction, cleaning, and labeling.
3. **Feature Engineering:** TF-IDF vectors + 7 Lexical Diversity metrics.
4. **Modeling:** 3 Rule-based baselines + 6 Machine Learning models.
5. **Evaluation:** F1-Score, Accuracy, and Error Analysis.

Milestone 1: Data & Preprocessing

Data Source: Leipzig Corpora Collection

Why this dataset?

- Consistent formatting across languages.
- News domain ensures professional writing standards.
- Explicit publication years available.

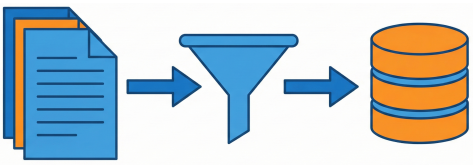
Scope (8 Archives):

Language	Years Included
English	2019, 2020, 2023, 2024
Czech	2019, 2020, 2023, 2024

From Sentences to Articles

The Reconstruction Pipeline:

1. **Parse:** Extract `sentences.txt`, `sources.txt`, and mapping files.
2. **Group:** Aggregate sentences by Source ID to rebuild full articles.
3. **Clean:** Normalize whitespace and artifacts.
4. **Filter:**
   - Min Length: 100 characters
   - Min Tokens: 20 tokens
5. **Label:** Assign binary class based on Nov 2022 cutoff.



Dataset Statistics

Global Stats:

Metric	Count
Total Articles	1,214,965
English	95.5%
Czech	4.5%
Pre-ChatGPT	42.6%
Post-ChatGPT	57.4%

Distribution by Year:

Year	Count	Class
2019	251k	Pre
2020	266k	Pre
2023	351k	Post
2024	345k	Post

Split: 80% Train / 20% Test (Stratified)

Milestone 2: Baseline Methods

Feature Engineering

1. TF-IDF (Term Frequency-Inverse Document Frequency) Vectors

- Max features: 5,000 | N-grams: (1, 2)
- Filtering: min_df=2, max_df=0.95

2. Lexical Diversity Features (7 scalars)

Feature	Note
`ttr`	Type-Token Ratio (Basic diversity)
`hdd`	Robust to length (42 draws)
`mtld`	Measure of Textual Lexical Diversity
`vocd`	Vocabulary Diversity
`stats`	Token count, Unique count, Avg length

Experimental Setup: Models

Rule-Based Baselines:

- **Date Threshold:** (Sanity check, 100%)
- **Text Length:** > 500 chars $\rightarrow$ Post
- **Diversity:** TTR $> 0.5 \rightarrow$ Post

Machine Learning Models:

- **Logistic Regression** (SGD[a])
- **Naive Bayes** (Incremental)
- **SVM**[b] (Hinge Loss)
- **PyTorch MLP**[c] (2 Hidden Layers)

[a] Stochastic Gradient Descent
[b] Support Vector Machine
[c] Multi-Layer Perceptron

Hardware & Environment

Compute Stack:

- **GPU:** 2× NVIDIA RTX 4090 (24GB)
- **CPU:** 32 Cores
- **RAM:** 128GB

Software Stack:

- Python 3.10
- PyTorch + CUDA
- scikit-learn + cuML
- Weights & Biases (Tracking)

Results & Analysis

Baseline Performance

Rule-Based Methods are insufficient.

Method	Accuracy	F1-Score
Date Threshold	100.00%	100.00%
Lexical Diversity (TTR)	57.40%	41.91%
Text Length	42.68%	27.40%

Note: TTR performs only slightly better than a random guess, while Text Length is anti-correlated.

Machine Learning Results

Key Finding

Lexical features improved accuracy by only **+0.03%** over TF-IDF alone.

Model	Features	Accuracy	F1
Logistic Reg.	**TF-IDF + Lexical**	**62.09%**	56.89%
Logistic Reg.	TF-IDF	62.06%	56.83%
Naive Bayes	TF-IDF + Lexical	61.23%	56.27%
PyTorch MLP	TF-IDF	60.17%	**59.73%**
SVM	TF-IDF	60.12%	48.96%

The Class Imbalance Problem

Best Model: Logistic Regression (TF-IDF + Lexical)

	Pred Pre	Pred Post
True Pre	24,611	78,922
True Post	13,189	126,271

- **76%** of Pre-ChatGPT texts are misclassified as Post-ChatGPT.
- Models default to the majority class ("Post").
- Indicates lack of distinctive signal in the text.

Qualitative Analysis

Texts are linguistically indistinguishable.

> *"He's also been inundated with repair requests from 'people who are digging out old bikes they haven't used in 10 years.'"*
>
> *(2020, Pre-ChatGPT → Predicted Post)*

> *"Dluh kanadského Cirque du Soleil se už vyšplhal na téměř 25 miliard korun..."*
>
> *(2020, Czech → Predicted Post)*

Conclusion: The difficulty is consistent across both English and Czech.

Discussion & Conclusion

Limitations & Future Work

Current Limitations

- **Data Gap:** Missing 2021-2022 (the transition period).
- **Imbalance:** 57% Post-ChatGPT bias.
- **Features:** TF-IDF misses deep semantic meaning.

Future Directions

- **Embeddings:** BERT/RoBERTa for semantic context.
- **Structure:** Analyze discourse and sentence complexity.
- **Sampling:** Use SMOTE to balance classes.

Final Conclusion

Summary: We processed **1.2M articles** across two languages and evaluated 9 models using lexical diversity metrics.

Research Answer

No. Lexical diversity features alone cannot reliably predict whether a news text is Pre- or Post-ChatGPT.

The marginal accuracy gain (+0.03%) suggests that news writing style has not shifted significantly at the lexical level.

Questions?

Resources & Reproducibility

Code: `https://github.com/Qq1b3/nlp-lexical-diversity`

Results: `https://tucloud.tuwien.ac.at/index.php/s/28D4SWoRYTqfj2e`

Tracking: `wandb.ai/jojino-tu-wien`