

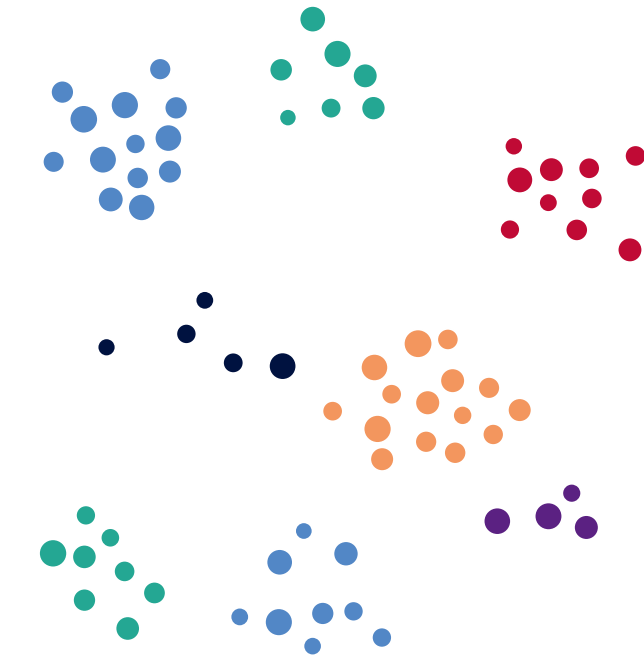
EMNLP 2026 · Oral Presentation · Budapest, Hungary

Measuring Data Diversity with Embeddings

A Taxonomy and Benchmark for Diversity Measures

Cantao Su · Anna Wegmann · Esther Ploeger · Dong Nguyen

NLP & Society Lab · NLP Research Group · AI & Data Science Division,
Utrecht University, The Netherlands



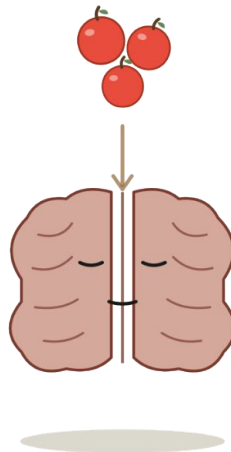
Universiteit Utrecht

Why Data Diversity Matters

Diversity
drives
capability

Narrow data

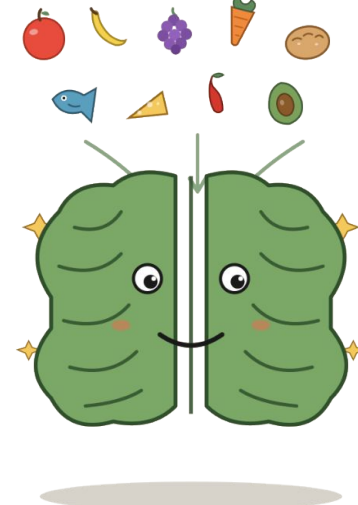
limited diet



narrow capability

Diverse data

diverse diet



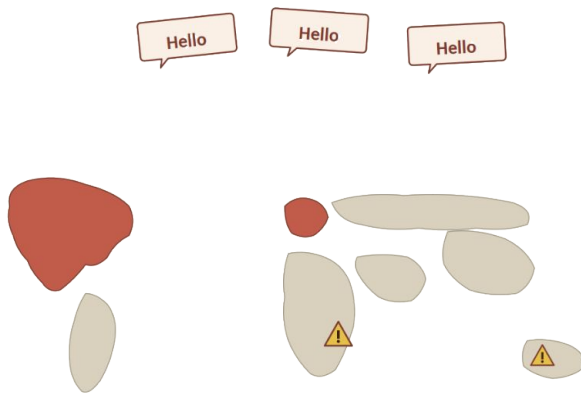
stronger capability

Why Data Diversity Matters

Diversity
improves
fairness

Western-centric

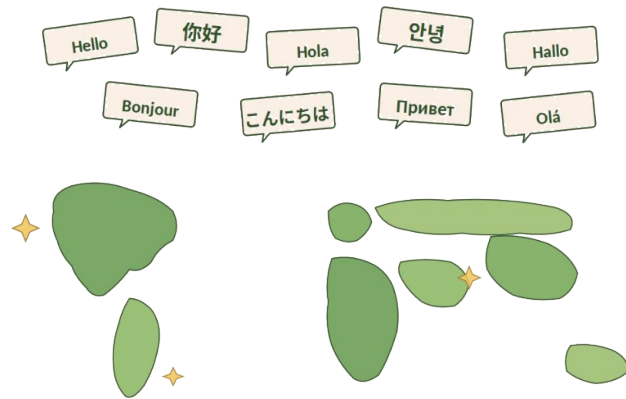
one voice · uneven coverage



biased & exclusionary outputs

Diverse & inclusive

many voices · global coverage



fair, inclusive, less toxic outputs

Why Data Diversity Matters

Diversity
meet
user
expectations

Repetitive outputs

falls short of expectation



"Suggest a way to greet someone."



"Hi there!"

"Hi there!"

× 3

"Hi there!"

same response, every time

Diverse outputs

meets the baseline expectation



"Suggest a way to greet someone."



"Hi there!"

"Hey, how's it going?"

"Good morning!"



varied responses across re-prompts

Measuring Diversity also Matters



Study its impact

on model performance,
fairness and robustness



Guide data collection

what to collect, filter or
select; which candidate
corpus to use

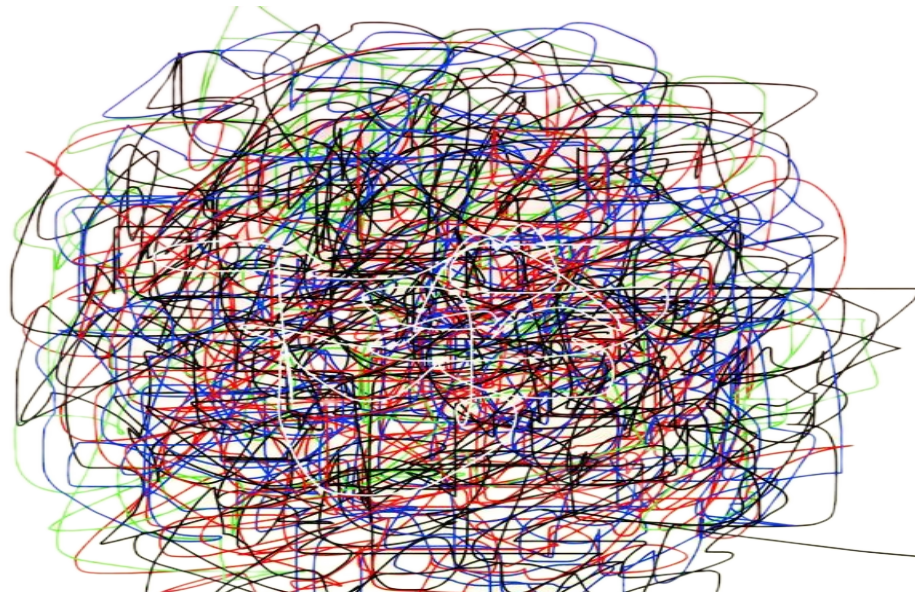


Evaluate outputs

are LLM generations or
synthetic data diverse
enough?

All of these rely on measures we can trust, yet the measures themselves are rarely evaluated

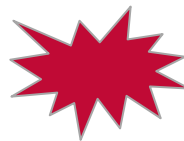
But the Landscape is Fragmented



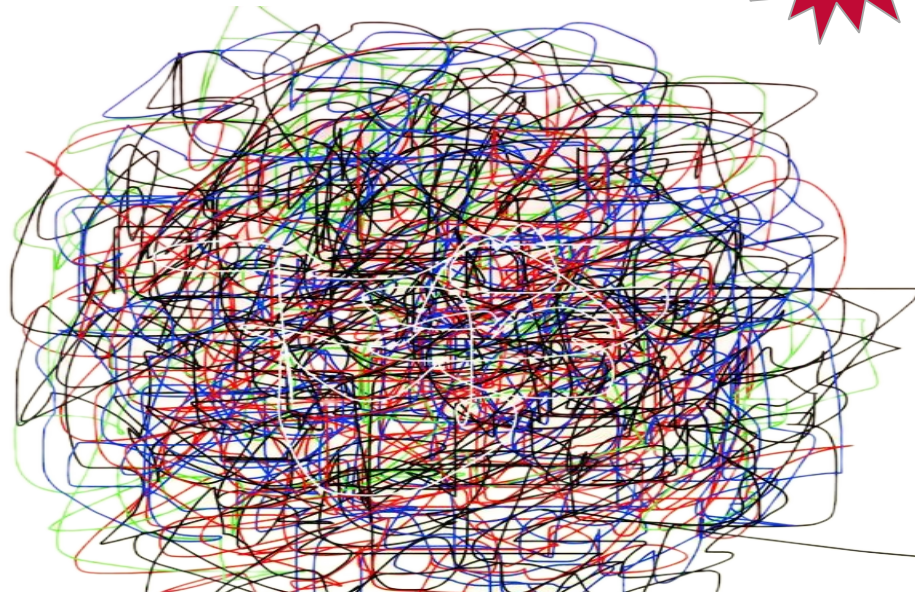
1. Diversity is overloaded
 - lexical
 - syntactic
 - semantic
 - stylistic



But the Landscape is Fragmented



Specify explicitly what diversity you mean!

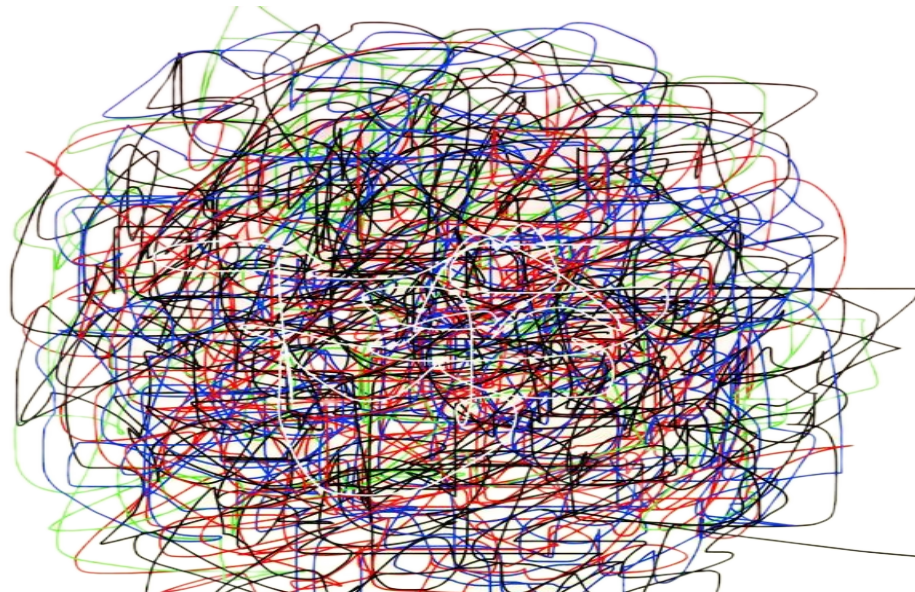


1. Diversity is overloaded

- lexical
- syntactic
- semantic
- stylistic



But the Landscape is Fragmented



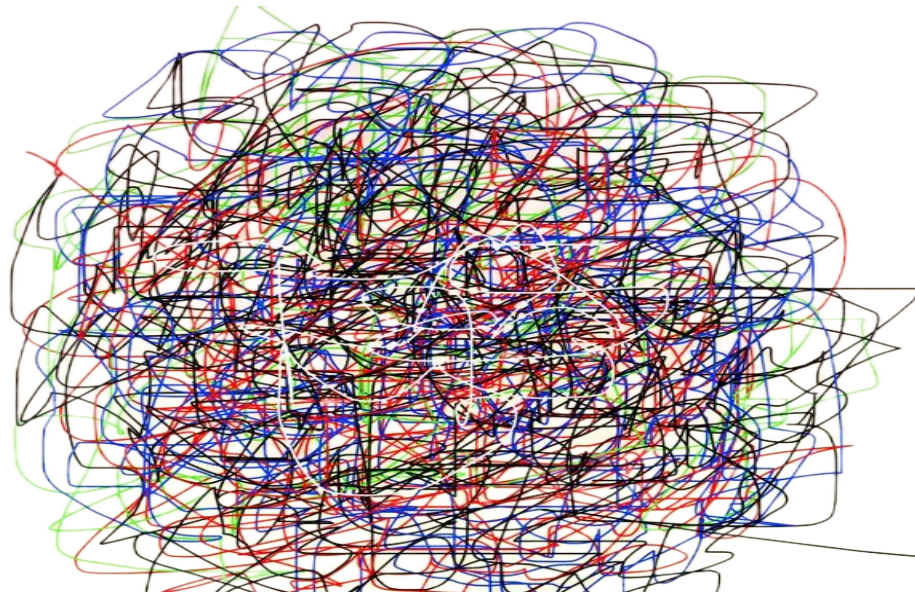
1. Diversity is overloaded

2. Inconsistent naming

- Same computation, different names
- Same name, different computations
- Similar ideas, different implementations



But the Landscape is Fragmented



1. Diversity is overloaded
2. Inconsistent naming
3. Quality is unsure
rarely evaluated:
which one should we use?



Toward Systematic Evaluation

A fragmented landscape

measures scattered across the literature



A unified taxonomy

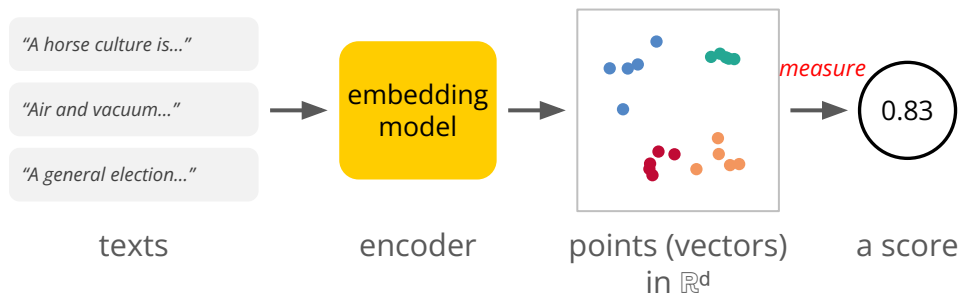
*four families
organized by underlying operation*

Distance	Geometry	Distribution	Graph
			

A taxonomy



Focus on embedding-based methods



The embedding model decides *what* diversity we measure (semantic, stylistic, ...)

Distance-based
Aggregates pairwise distances; no constructed structures.

- Mean Pairwise Dist.
- Sum Pairwise Dist.
- Energy
- Diameter
- Bottleneck
- Sum Diameter
- Sum Bottleneck
- Chamfer Dist.

Geometry-based
Quantifies spatial properties of embeddings.

- Convex Hull Vol. (2D)
- Span (Centroid)
- Span (Medoid)
- Cluster Inertia
- Geo. Mean of Std.
- Log-Determinant

Distribution-based
Quantifies distributional properties of embeddings.

- Vendi Score
- Rényi Kernel Ent.
- DCScore
- Bins Entropy (UMAP)
- MagArea

Graph-based
Summarizes a graph built over the embeddings.

- MST Dispersion
- Graph Entropy

1. Clean and organize $\Rightarrow$ Our taxonomy

1. Clean and organize $\Rightarrow$ Our taxonomy
2. Which measure should I use?

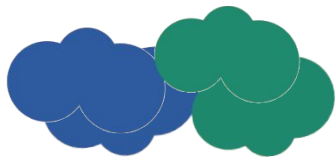
1. Clean and organize $\Rightarrow$ Our taxonomy
2. Which measure should I use?

Test $\Rightarrow$ our Benchmark *EmbDivBench*

Concept of Diversity

The classic view of diversity

fewer categories



low diversity

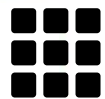


more categories



high diversity

Concept of Diversity is Evolving



Variety



low

high



Balance

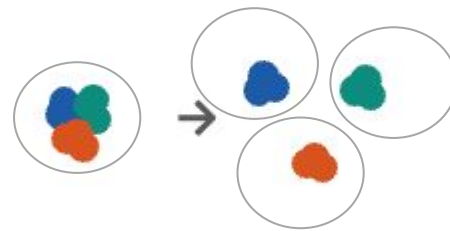


low

high



Disparity



low

high

Concept of Diversity is Evolving, but NLP falls behind

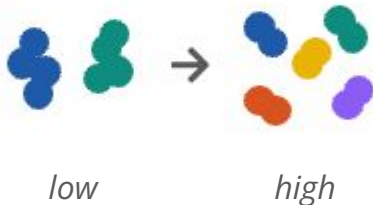
What most NLP work still focuses on



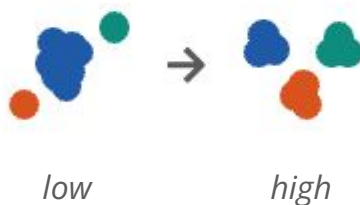
What Our Benchmark Brings

What most NLP work still focuses on

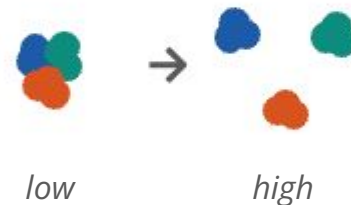
 Variety



 Balance

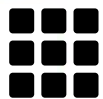


 Disparity



Vary One Component, Fix the Others

For each component: **five** datasets in which only that component increases



Variety



low

high



Balance

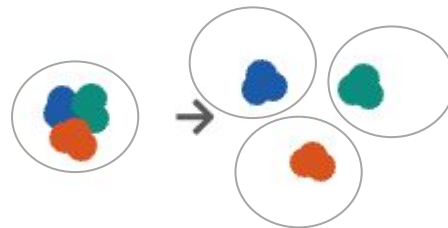


low

high



Disparity



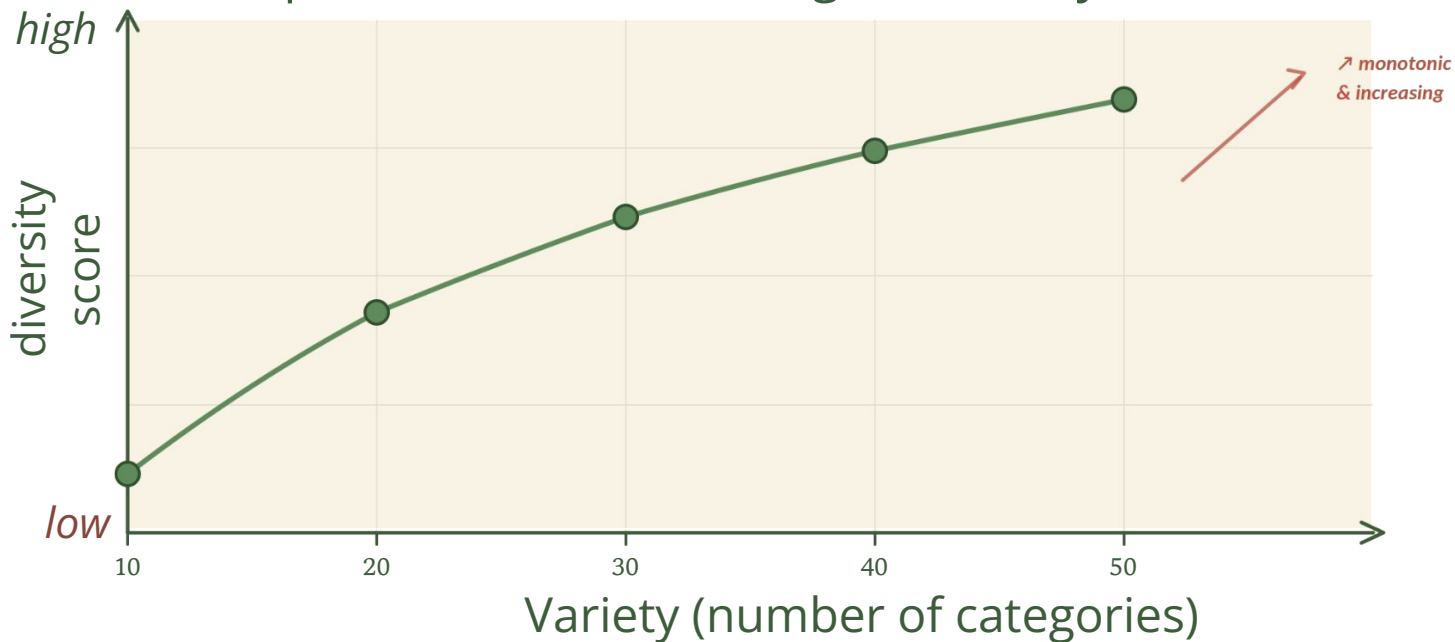
low

high

A good measure should respond like

A good measure ranks the five datasets in the right order (Spearman ρ)

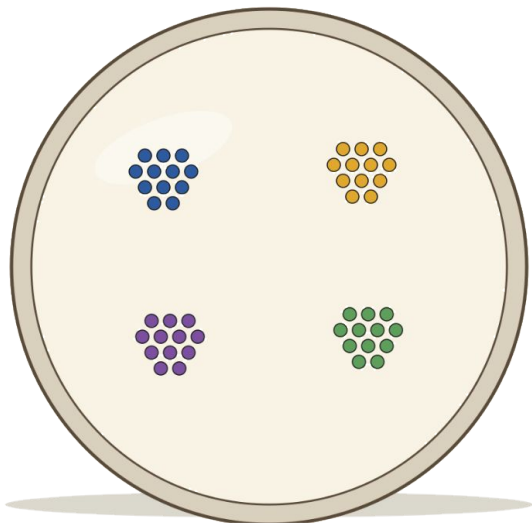
Expected behaviour along the variety axis



Two Different Tiers of Data

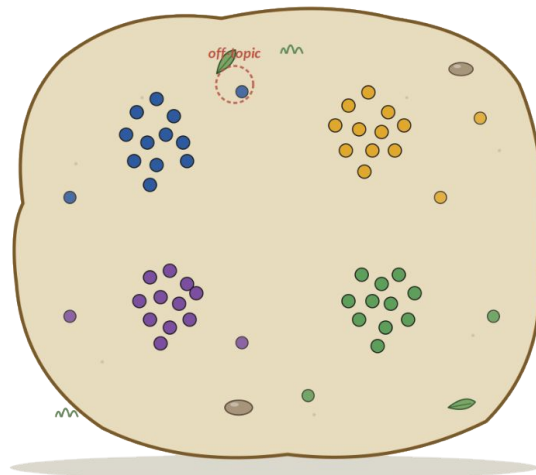
Synthetic Data

numeric vectors sampled from GMM
exact control



Wikipedia

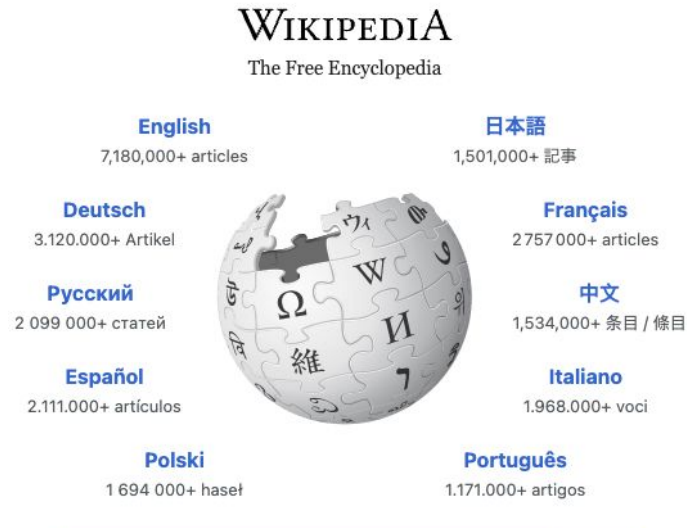
real text
approximate control



Wikipedia

Embedding-based Semantic Diversity

- with topic labels
- hierarchically structured



Wikipedia

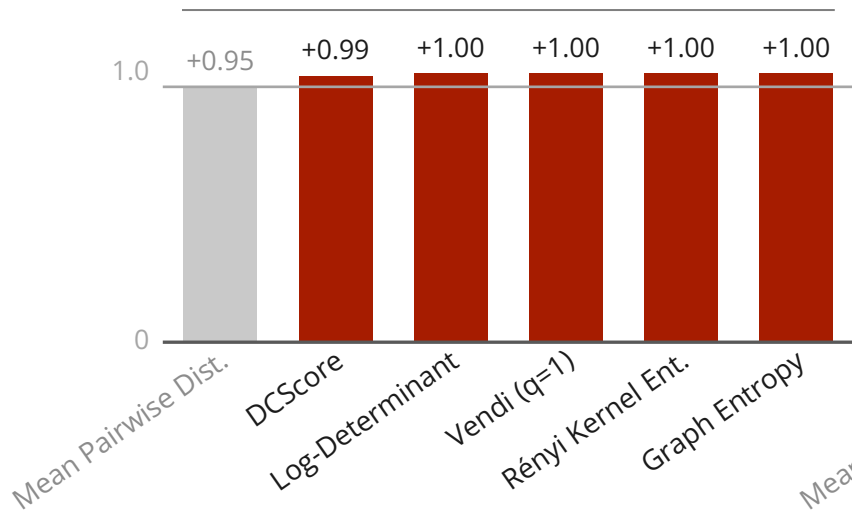
- with topic labels
 - hierarchically structured
- > proxy for semantics
 -> domains and topics
 (supercategories) (categories)

Domain	Topic	Example sentence
Government by country	National legislatures	A general election is an electoral process to choose most or all members of a governing body at the same time. After being elected the Prime Minister appoints the cabinet ministers and announces them to the Riksdag.
	Parliaments by country	Article 16.4.2° requires that the newly elected Dáil Éireann must convene no later than 30 days after the election.

Mean Pairwise Distance is not the best

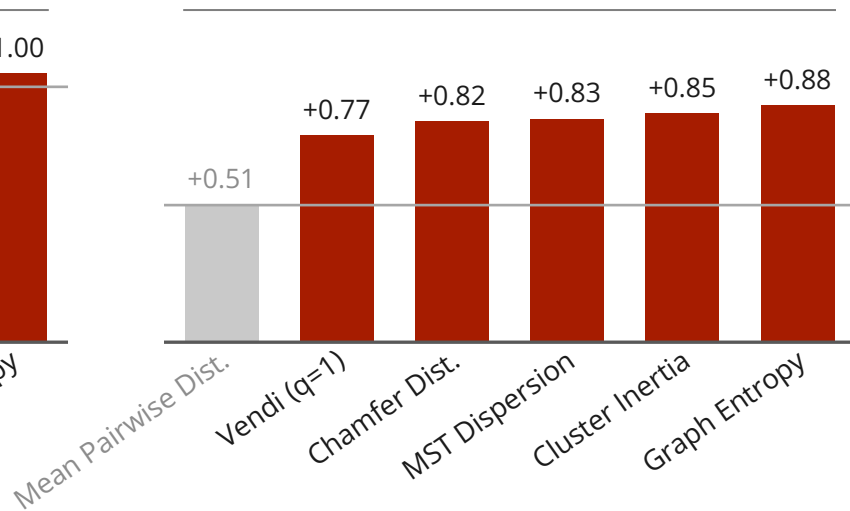
Mean Spearman ρ over variety, balance and disparity (5 seeds)

Synthetic data (GMM)



 Mean pairwise distance
 Most common in prior work

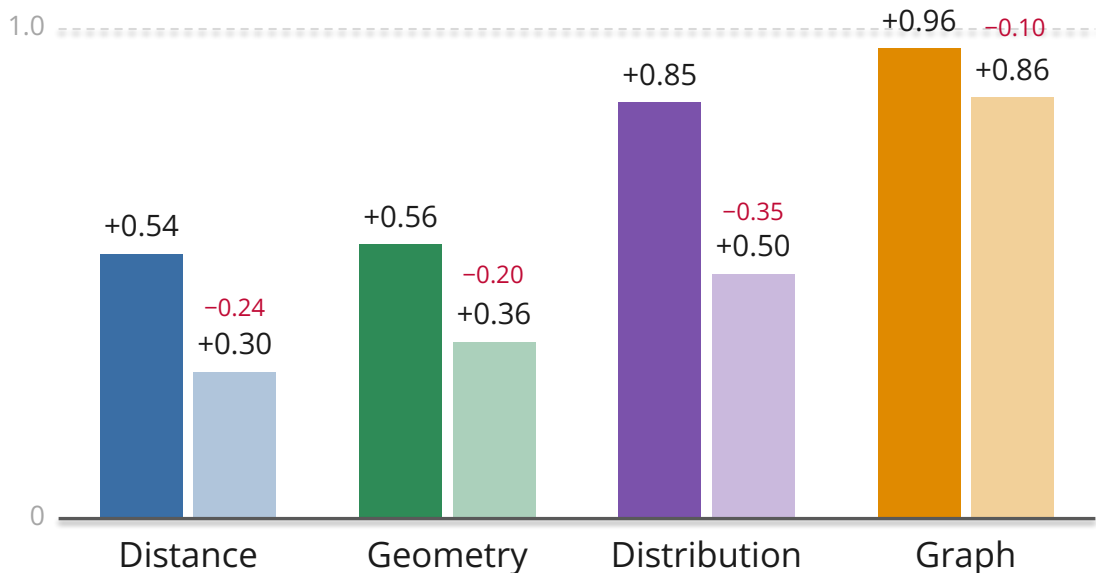
Wikipedia



 Several alternatives do better on both tiers

Real text is much more challenging

Family mean Spearman ρ , synthetic vs. Wikipedia (5 seeds)



■ Synthetic data (GMM)

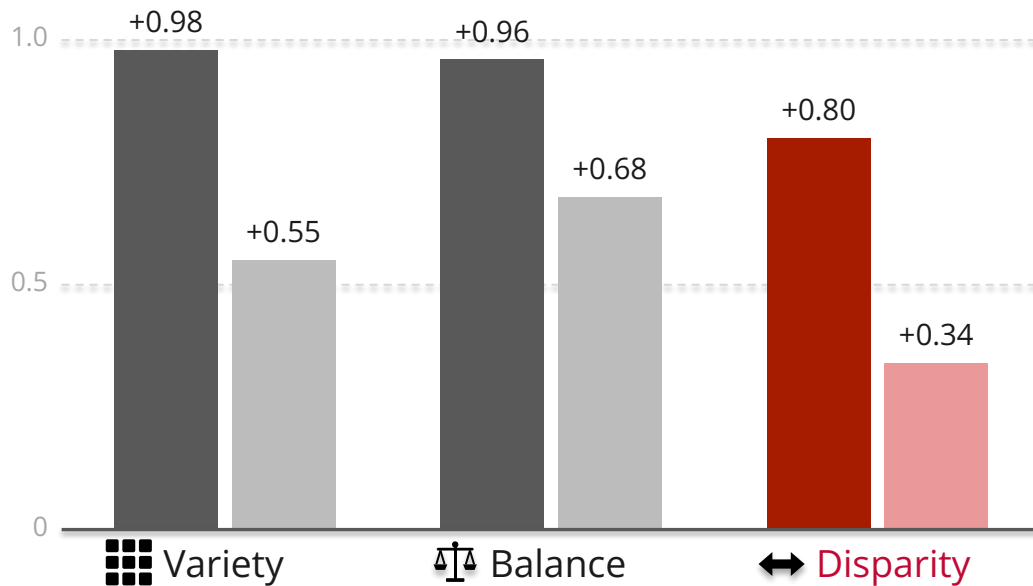
■ Wikipedia

Every family drops on real text

Graph-based drops least

Disparity is the hardest component

Median Spearman ρ across measures, per component (5 seeds)



■ Synthetic data (GMM)

■ Wikipedia

Lowest on both tiers

Lowest mean and median ρ of the three components

On real text, variety and balance swap places



ATTENTION NEEDED

Practical Takeaways

1. Common choices are not always the best.
2. Match the encoder to the dimension you want to measure (e.g., a style encoder misses semantic diversity)
3. Choose the measure to match the components of interest.

Shortlists

4. For component-specific:

- **Variety**: chamfer distance, MST dispersion.
- **Balance**: graph entropy.
- **Disparity**: MST dispersion.

5. No idea which component you care about?

- Graph entropy
- Vendi Score ($q=1$)

Thank You

Happy diversity measurement



Paper



Data & code



emb-diversity
package



NLP and Society Lab



Universiteit Utrecht

LinkedIn

