Comparing Romanian from Romania with Romanian from Moldova

Balaceanu Mihai, Bucur Stefan,
Constantin Tudor

Part 1

What articles exist on the topic and what has been done so far

Linguistic & Historical Research

- Most linguists view 'Moldovan' as a regional variety of Romanian.
- Distinction is mainly political (Soviet-era language policy).
- Differences: minor lexical & phonetic variation; Russian influence in Moldova.
- Key source: 'The Moldovan Dialect' (Philologia, 2017).

NLP Text Resources

- MOROCO: 33k+ labeled news samples (RO vs MD).
- Used for dialect classification & transfer learning.
- Links: aclanthology.org/P19-1068, github.com/butnaruandrei/MOROCO

Transformer-Based Studies

- VarDial 2020: monolingual Romanian models outperform multilingual ones.
- Shows detectability of subtle RO vs MD variation.
- Link: aclanthology.org/2020.vardial-1.18

Speech / Audio Datasets

- RoDia: Romanian regional dialect speech dataset.
- MoRoVoc: 93+ hours RO & MD spoken dataset.
- Useful for accent detection & ASR robustness.

What Research Has Achieved

- ML systems can distinguish RO-Ro and RO-Md text & speech.
- Established benchmark datasets for both modalities.
- Demonstrated measurable linguistic signals.

Part 2

What data we will use + EDA

Dataset

We will be using the following dataset: https://github.com/RoTak00/roro-analiza/releases/download/dataset-cleaned-1.0.0/data-cleaned.zip

Then, we will extend the crawling part, gather more texts.

EDA - Intermediate results

Top 15 Moldova-specific phrases (RO-MD):

	phrase	importance
188729	raionul gpe	0.885848
118355	la moment	0.489417
4833	a menționat	0.459379
214633	transmite ipn	0.456045
79598	din raion	0.452234
130206	menționăm că	0.441324
149292	or gpe	0.439221
77144	din arhiva	0.412716
18252	amintim că	0.380413
149443	orașul gpe	0.369144
136839	ne a	0.367415
231925	în gpe gpe	0.364790
79603	din raionul gpe	0.351955
33388	care este	0.346194
186140	publice locale	0.344644

Top 15 Romania-specific phrases (RO-RO):

	phrase	importance
159608	pe raza	-0.304400
7261	a transmis	-0.304652
117723	la gpe gpe	-0.307075
34722	care să	-0.313329
9574	acest lucru	-0.316369
230145	în ciuda	-0.317025
1960	a declarat	-0.339569
235263	în vedere	-0.345439
115354	județului gpe	-0.406027
104526	gpe gpe gpe	-0.417952
235552	în zona	-0.437551
118531	la nivelul	-0.472867
78707	din județ	-0.476181
190323	regiunea gpe	-0.656421
115214	județul gpe	-0.681373

Most common words (length > 3):

	count	percentage
cadrul	12743	0.222750
timp	11601	0.202788
mare	11339	0.198208
când	10974	0.191828
trebuie	10239	0.178980

Most common short words (length <= 3):

	count	percentage
și	229530	33.385793
ani	22912	3.332616
s-a	17458	2.539316
lei	12319	1.791834
loc	9874	1.436201

```
================================================
EXPLORATORY DATA ANALYSIS
================================================

1. Loading data from data-cleaned folder...
------------------------------------------------
✓ Total files loaded: 29620

Categories found:
  - judete: 16983
  - raioane: 6071
  - int_istoric: 4144
  - int: 2422

Top 15 regions found:
  - RepMoldova: 6071
  - Moldova: 5692
  - Oltenia: 4120
  - Ucraina: 3010
  - Muntenia: 2526
  - Ardeal: 1467
  - Serbia: 1134
  - Banat: 1124
  - Dobrogea: 965
  - Spania: 723
  - Canada_EN: 641
  - Germania: 500
  - UK: 499
  - Bucovina: 428
  - Crisana: 386
```

```
================================================================================
2. TEXT CONTENT ANALYSIS
================================================================================

Title Statistics:
  Average length: 81 characters
  Median length: 79 characters
  Min length: 0 characters
  Max length: 899 characters

Content Statistics:
  Average length: 2215 characters
  Median length: 1649 characters
  Min length: 20 characters
  Max length: 332006 characters

Word Count Statistics:
  Average title words: 13 words
  Average content words: 343 words

================================================================================
3. DATA COMPLETENESS
================================================================================
  Titles present: 29615/29620 (100.0%)
  Content present: 29620/29620 (100.0%)
  Metadata present: 29620/29620 (100.0%)

================================================================================
4. METADATA ANALYSIS
================================================================================
Original file extensions found:
  .html: 29620

================================================================================
5. LANGUAGE COVERAGE
================================================================================
  ✓ Romanian (Romania) - ro-RO: 23,549 articles
    - Categories: 'judete' (16,983), 'int' (2,422), 'int_istoric' (4,144)
  ✓ Romanian (Moldova) - ro-MD: 6,071 articles
    - Category: 'raioane' (from RepMoldova region)
```

```
============================
6. STATISTICS BY CATEGORY
============================

int:
  Articles: 7084
  Avg length: 2335 characters
  Avg words: 363 words

int_istoric:
  Articles: 4144
  Avg length: 2192 characters
  Avg words: 335 words

judete:
  Articles: 16983
  Avg length: 2300 characters
  Avg words: 355 words

raioane:
  Articles: 6071
  Avg length: 1845 characters
  Avg words: 284 words
```

```
============================================================
8. FINAL SUMMARY
============================================================

Dataset Overview:
  Total articles: 29,620
  Total unique regions: 18
  Total categories: 4

Content Characteristics:
  Average article: 2215 characters (~343 words)
  Longest article: 332,006 characters
  Articles present: 29620/29620 (100.0%)

Language Mix:
  ✓ Romanian (Romania) - ro-RO: 23,549 articles (79.5%)
  ✓ Romanian (Moldova) - ro-MD: 6,071 articles (20.5%)
```

Part 3

What models we will be using and the calculation requirements

Language Model Selection

1. Generative Models (Causal LMs)

- RoGPT2-base (~124M params) & RoGPT2-medium (~354M params)
  - *Source:* Hugging Face (Readerbench).
  - *Dataset:* Trained on OSCAR, Wiki-Ro, and Romanian news.
  - *Role:* Primary tools for calculating Perplexity (PPL).
  - *Hypothesis:* Lower PPL on RO-Ro (Standard) vs. Higher PPL on RO-Md (Moldovan) indicates dialectal "surprise."
- GPT-Neo Romanian (~780M params)
  - *Source:* Dumitrescu Stefan (Hugging Face).
  - *Role:* Acts as a high-capacity "Upper Bound" baseline to verify if perplexity patterns persist in larger models.

2. Masked Language Models (Encoders)

- BERT-base-romanian-cased-v1 (~124M params)
  - *Role:* Used for Pseudo-Perplexity (Salazar et al.) and extracting sentence embeddings.
  - *Application:* Downstream clustering and binary classification (RO-Ro vs. RO-Md).

Morpho-Syntactic Analysis Pipelines

Objective: Evaluate how standard tools (trained on Standard Romanian) degrade when processing Moldovan sub-dialects.

Selected Frameworks:

- UDPipe
  - Lightweight, trained on the Romanian Reference Treebank (RRT).
  - Provides: Tokenization, Lemmatization, PoS Tagging, Dependency Parsing.
- NLP-Cube
  - Neural framework (RNN-based) known for high accuracy in CoNLL shared tasks.
- Stanza (Stanford NLP) / RELATE
  - State-of-the-art performance; uses the TEPROLIN platform logic.
  - *Why this matters:* Stanza is currently considered the academic standard for accuracy in Universal Dependencies.

Computational Requirements

1. Generative Models (RoGPT2 & GPT-Neo)

- RoGPT2-base / medium (124M – 354M parameters)
  - Need: Moderate GPU (8GB VRAM).
  - Context: Efficient enough for standard research GPUs; requires ~1.5GB just to load, plus overhead for processing.
- GPT-Neo-Ro (780M parameters)
  - Need: High-Performance GPU (12GB – 16GB VRAM).
  - Context: This is the "heavy lifter." It requires significant memory to calculate perplexity on long texts without crashing.

2. Masked Models (BERT / DistilBERT)

- BERT-base-romanian (124M parameters)
  - Need: Low Resource (Standard GPU or CPU).
  - Context: Very lightweight. Can run easily on any standard laptop or Google Colab free tier.

3. Parsing Pipelines (UDPipe, Stanza, NLP-Cube)

- Need: High System RAM (16GB – 32GB) & Multi-core CPU.
- Context: These tools do not rely heavily on the GPU. Instead, they need strong CPU performance and plenty of RAM to load the linguistic dictionaries and process the text corpora.

Part 4

Evaluation methods

Evaluating Language Models

Primary Metric: Perplexity (PPL)

- Definition: A measurement of how well a probability model predicts a sample. Low PPL = High confidence; High PPL = Confusion/Surprise.
- Method: We will calculate the PPL of RoGPT2 and GPT-Neo on the test subsets of both the RO-RO (Romania) and RO-MD (Moldova) datasets.
- Hypothesis Validation:
  - We expect a Baseline PPL for RO-RO (since the models were trained on standard Romanian).
  - We expect a Higher PPL (Delta) for RO-MD, indicating dialectal divergence (lexical or syntactic differences).

Secondary Metric: Cross-Entropy Loss

- Used to track model convergence during any potential fine-tuning steps.

Evaluating Classification & Clustering (BERT)

Binary Classification Metrics (Ro vs. Md)

- F1-Score (Macro & Weighted): To balance precision and recall, ensuring the model isn't just predicting the majority class (RO).
- Confusion Matrix: To visualize exactly how many RO-MD texts are misclassified as RO-RO (False Negatives).

Embedding Analysis (Clustering)

- Method: Extract sentence embeddings using BERT-base-romanian.
- Visualization: Use t-SNE or PCA (Principal Component Analysis) to project the high-dimensional vectors into 2D space.
- Success Criteria: Visual separation (distinct clusters) between RO and MD data points would indicate distinct linguistic features.