

# Archaeology of Intelligent Machines: Comparing Romanian Language Usage in Romania with Romanian Usage in the Diaspora

1st Semester of 2024-2025

## Authors

sara-ioana.ionescu@s.unibuc.com  
alexandru.nohai@s.unibuc.com  
medeea-maria.pericica@s.unibuc.com

## Abstract

This study focuses on observing the linguistic contact between Romanian and the majority language in various diaspora regions. By employing advanced natural language processing techniques such as tokenization, vectorization, and statistical analysis, the project identified nuanced patterns reflecting how Romanian interacts and adapts to the linguistic environments of host countries. The findings highlight socio-cultural dynamics and offer insights into the evolving linguistic identity of Romanian speakers abroad.

## 1 Introduction

The relationship between Romanian as written in Romania and in the diaspora represents a fascinating linguistic and cultural study. Sharing a common language base, these variations evolve under distinct cultural, geographical, and social influences. This project aims to capture and analyze these differences systematically.

This study aims to:

- Observe the linguistic contact between Romanian and the majority language in each region.
- Identify unique linguistic features and adaptations across contexts.
- Quantify stylistic and semantic differences using advanced NLP techniques.
- Analyze socio-cultural influences that shape language use in the diaspora.

We developed a comprehensive NLP pipeline for this analysis, incorporating text normalization, tokenization, diacritic restoration, and advanced statistical evaluations. The analysis is enriched by comparisons with previous studies and detailed corpus evaluations.

We chose this subject because we found the context of Romanian elections intriguing and wanted

to explore whether significant differences exist in how these events are covered in diaspora articles compared to those in Romania.

## Contributions

- **Alex:** Developed the NLP pipeline, including data preprocessing, tokenization, and statistical analysis.
- **Medeea:** Focused on data collection and visualization. Curated the corpus from diverse regional sources and generated graphs to highlight linguistic patterns.
- **Sara:** Conducted the interpretive analysis and comparative study. Analyzed the socio-cultural implications of linguistic patterns and drafted the findings and conclusions sections.

## 2 Approach

### Approach

To complete this project, we employed a series of statistical and natural language processing (NLP) techniques to analyze regional variations in written Romanian. Below, I detail the approach and methods used.

### Data Collection and Corpus Details

The code and dataset are hosted in a Git repository. <https://github.com/mariamedeea/Romanian-in-Different-Regions>

### Software Tools Used

- **Programming Language:** Python
- **Libraries:** pandas, NumPy, NLTK, scikit-learn, spaCy, and matplotlib
- **Environment:** Google Colab for computational processing and visualization

## Training and Processing Time

The processing time varied by task:

- Text preprocessing : 1.5 hour per region
- Feature extraction (bigram, trigram computation, TF-IDF): 30 minutes
- Generating visualizations: 15 minutes
- total NLP Pipeline execution: 15 minutes

## Machine Learning and Optimization Techniques

While this project was predominantly statistical, we employed TF-IDF (term frequency-inverse document frequency) for lexical analysis and cosine similarity for identifying characteristic words for regions (Figure 2).Preprocessing included stop-word removal, stemming, and lemmatization to improve text standardization.

## Evaluation Report

The evaluation focused on comparing linguistic patterns across regions. Key findings included:

- **Distribution of POS Tags:** Regions demonstrated distinct usage patterns, e.g., one region favored verbs while another favored nouns.
- **Anagram Analysis:** Highlighted unique morphological traits by region.
- **Bigram/Trigram Analysis:** Identified commonly co-occurring phrases and syntactic structures.
- **Characteristic Words for Regions:** Derived using TF-IDF, showing lexical uniqueness.
- **Loanwords:** Assessed for frequency and type, indicating cultural influences.
- **Average Text Length:** Regions with richer descriptions or narratives had longer texts.

## Visualisation

Below is a visualisation illustrating the findings:

## 3 Findings and Insights

### Regional Linguistic Patterns

- **Germany:**
  - Strong influence of geography, culture, and politics (e.g., *Renania, Bundestag*).

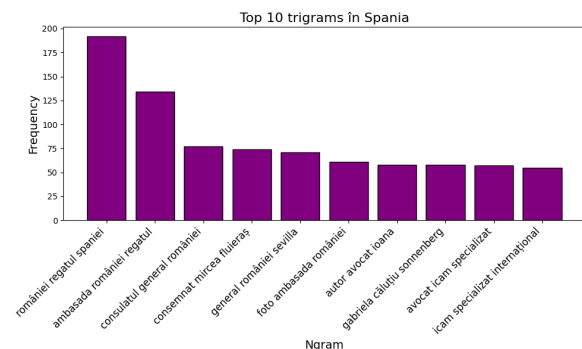
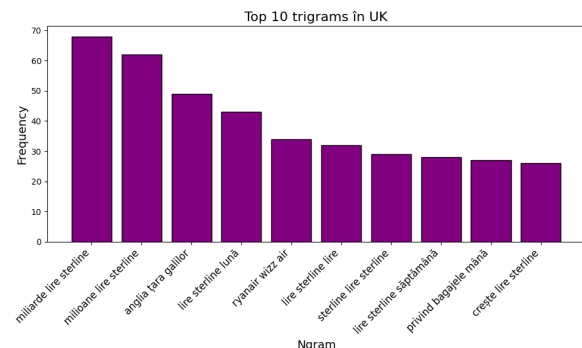
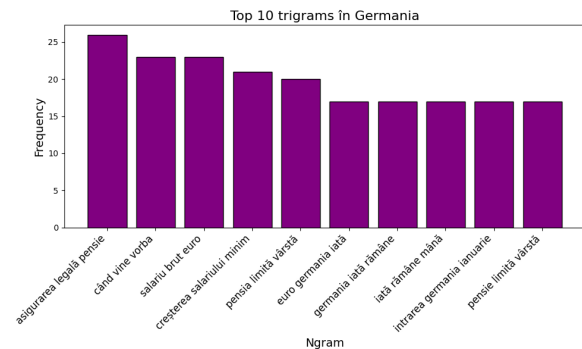
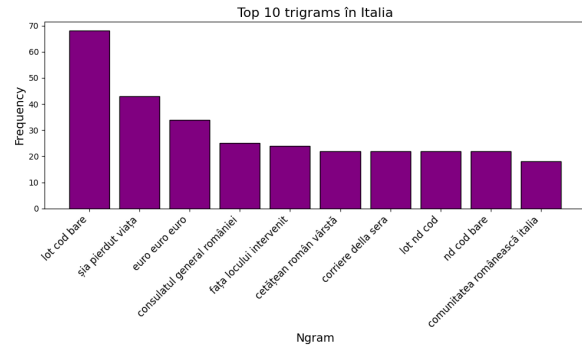


Figure 1: Trigrams

110	– Administrative topics are prominent,	<b>4 Limitations</b>	150
111	such as <i>Kindergeld</i> (child benefits).	While the findings are robust, several limitations	151
112	• <b>Romania:</b>	were noted:	152
113	– Dominance of media and news-related	• Dependence on word-level features, limiting	153
114	terms (e.g., <i>DCNews</i> , <i>Blinken</i> ).	contextual depth.	154
115	– Focus on global and local events.	• Corpus diversity was constrained by the avail-	155
116	• <b>Italy:</b>	ability of textual data from specific regions.	156
117	– Influence of local culture and administra-	• Advanced models such as transformers were	157
118	tion (e.g., <i>Modena</i> , <i>Carabinieri</i> ).	not utilized, which could provide richer in-	158
119	– Social and work-related topics are evi-	sights.	159
120	dent.	Future work will address these limitations by	160
121	• <b>Spain:</b>	expanding corpus diversity and incorporating state-	161
122	– Geographic and cultural connection (e.g.,	of-the-art NLP techniques.	162
123	<i>Canare</i> , <i>Picasso</i> ).	<b>5 Conclusions and Future Work</b>	163
124	– Translation and integration topics high-	This project successfully examined the linguistic	164
125	lighted.	adaptations of Romanian in diaspora contexts, high-	165
126	• <b>UK:</b>	lighting the socio-cultural dynamics influencing	166
127	– Economic and administrative focus (e.g.,	language use. Key takeaways include:	167
128	<i>GBP</i> , <i>HMRC</i> ).	• Romanian exhibits significant adaptability, in-	168
129	– References to health services ( <i>NHS</i> ) and	fluenced by the majority language in host	169
130	geography.	countries.	170
131	<b>General Observations</b>	• Regional variations offer a window into cul-	171
132	• Distinct cultural and administrative terms re-	tural integration and identity.	172
133	fect local adaptation.	• NLP pipelines are effective in quantifying and	173
134	• Media and online influence are significant	visualizing linguistic patterns.	174
135	across all regions.	Future directions include:	175
136	• Romanian diaspora integrates into local issues	• Incorporating social media data for more dy-	176
137	while maintaining ties to Romania.	namic analyses.	177
138	<b>Regional Linguistic Insights</b>	• Using transformers and contextual embed-	178
139	• <b>Romania:</b> Diverse vocabulary; focus on local	dings for deeper insights.	179
140	and global news.	• Expanding the study to spoken language anal-	180
141	• <b>Germany &amp; UK:</b> Practical terms related to	ysis.	181
142	economy and administration dominate.	<b>References</b>	182
143	• <b>Italy &amp; Spain:</b> Balance of cultural and social	• Susan Sanders, Tom Dotz, Tom Hoobyar	183
144	integration with Romanian identity.	(2013). "NLP: The Essential Guide"	184
145	<b>Overall Patterns</b>		
146	• Linguistic adaptation is visible through local		
147	influences in diaspora regions.		
148	• Romanian identity remains strong across all		
149	analyzed texts.		

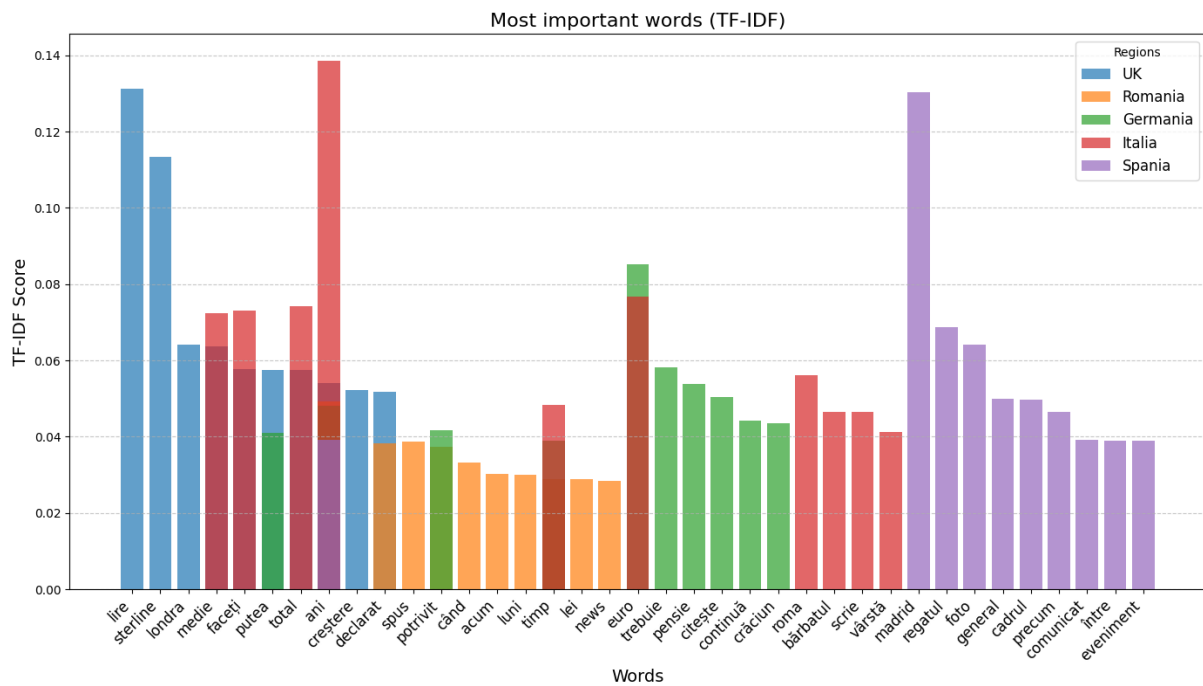


Figure 2: TF-IDF

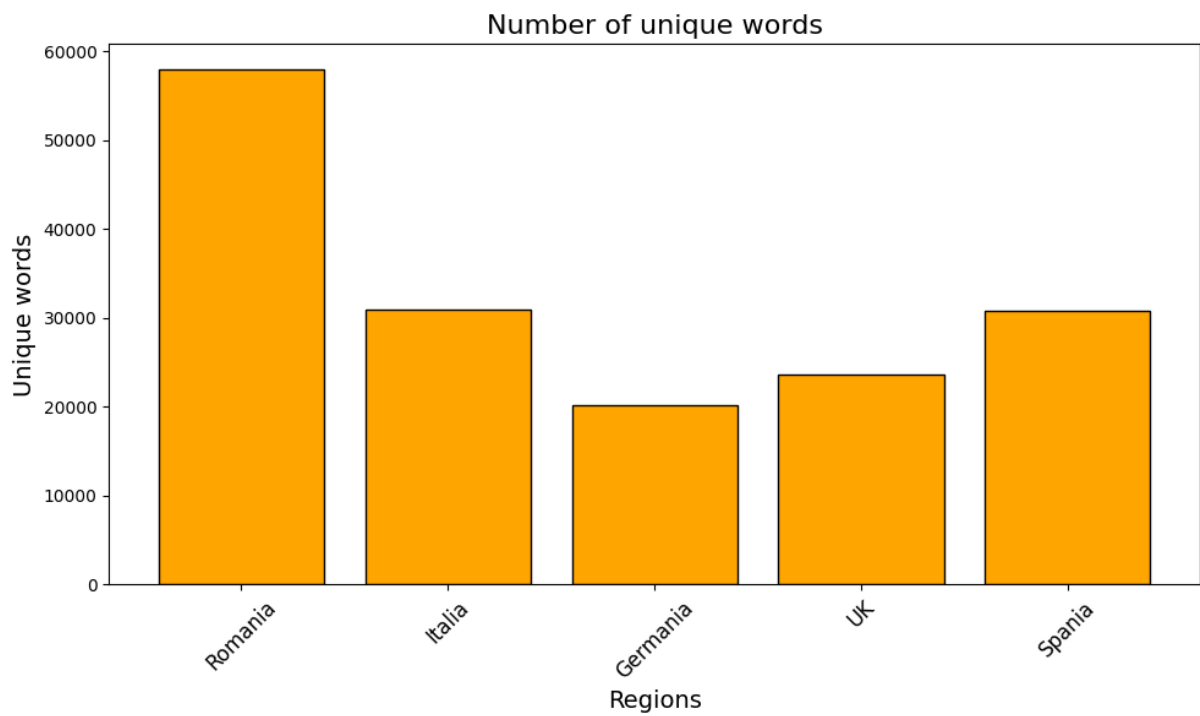


Figure 3: Unique Words