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:

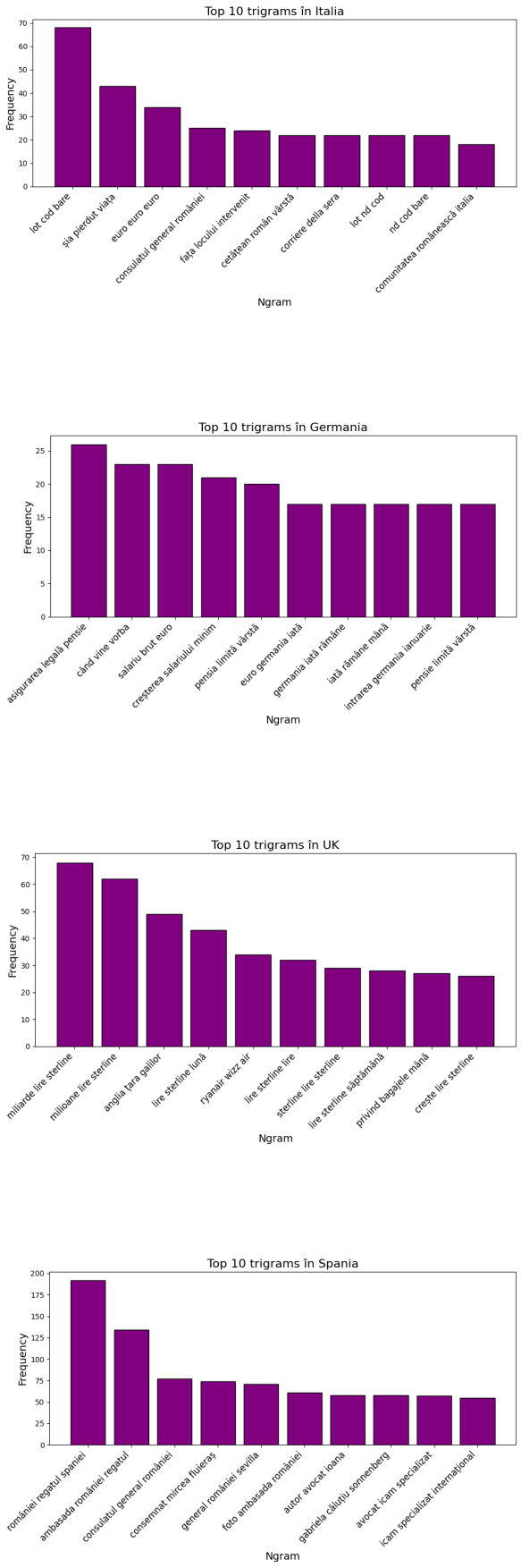

Figure 1: Trigrams

3 Findings and Insights

Regional Linguistic Patterns

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

  - Administrative topics are prominent, such as *Kindergeld* (child benefits).

- **Romania:**
  - Dominance of media and news-related terms (e.g., *DCNews*, *Blinken*).
  - Focus on global and local events.

- **Italy:**
  - Influence of local culture and administration (e.g., *Modena*, *Carabinieri*).
  - Social and work-related topics are evident.

- **Spain:**
  - Geographic and cultural connection (e.g., *Canare*, *Picasso*).
  - Translation and integration topics highlighted.

- **UK:**
  - Economic and administrative focus (e.g., *GBP*, *HMRC*).
  - References to health services (*NHS*) and geography.

General Observations

- Distinct cultural and administrative terms reflect local adaptation.
- Media and online influence are significant across all regions.
- Romanian diaspora integrates into local issues while maintaining ties to Romania.

Regional Linguistic Insights

- **Romania:** Diverse vocabulary; focus on local and global news.
- **Germany & UK:** Practical terms related to economy and administration dominate.
- **Italy & Spain:** Balance of cultural and social integration with Romanian identity.

Overall Patterns

- Linguistic adaptation is visible through local influences in diaspora regions.
- Romanian identity remains strong across all analyzed texts.

4 Limitations

While the findings are robust, several limitations were noted:

- Dependence on word-level features, limiting contextual depth.
- Corpus diversity was constrained by the availability of textual data from specific regions.
- Advanced models such as transformers were not utilized, which could provide richer insights.

Future work will address these limitations by expanding corpus diversity and incorporating state-of-the-art NLP techniques.

5 Conclusions and Future Work

This project successfully examined the linguistic adaptations of Romanian in diaspora contexts, highlighting the socio-cultural dynamics influencing language use. Key takeaways include:

- Romanian exhibits significant adaptability, influenced by the majority language in host countries.
- Regional variations offer a window into cultural integration and identity.
- NLP pipelines are effective in quantifying and visualizing linguistic patterns.

Future directions include:

- Incorporating social media data for more dynamic analyses.
- Using transformers and contextual embeddings for deeper insights.
- Expanding the study to spoken language analysis.

References

- Susan Sanders, Tom Dotz, Tom Hoobyar (2013). "NLP: The Essential Guide"

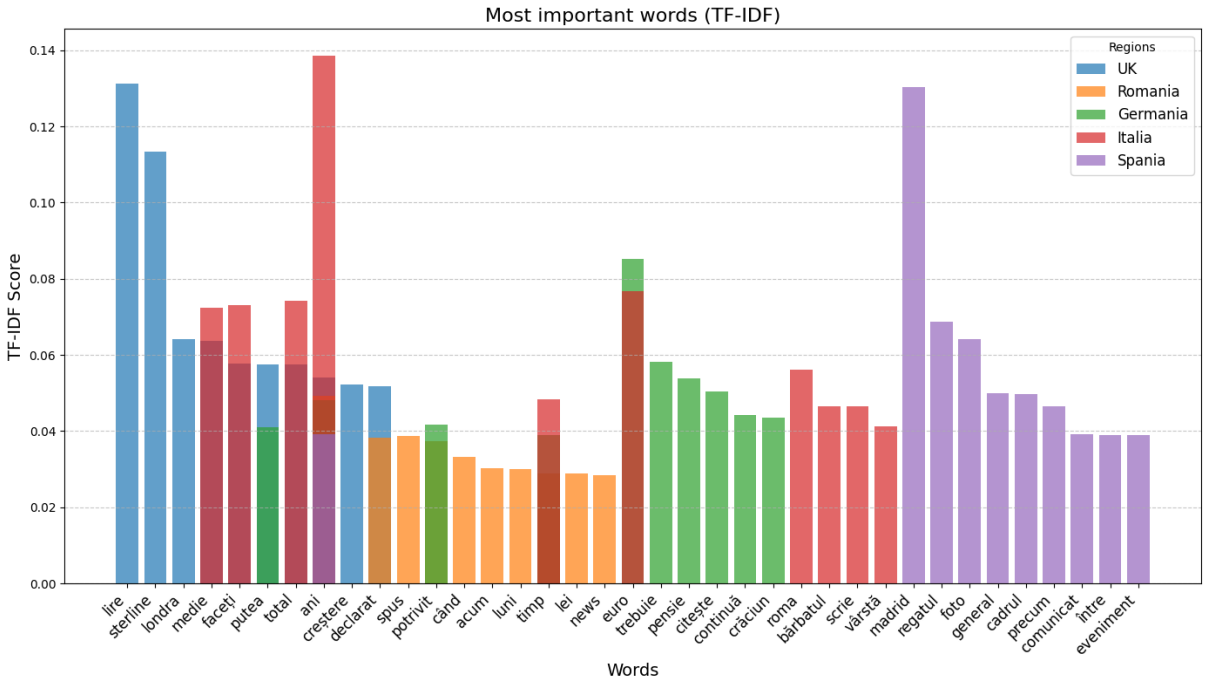

Figure 2: TF-IDF

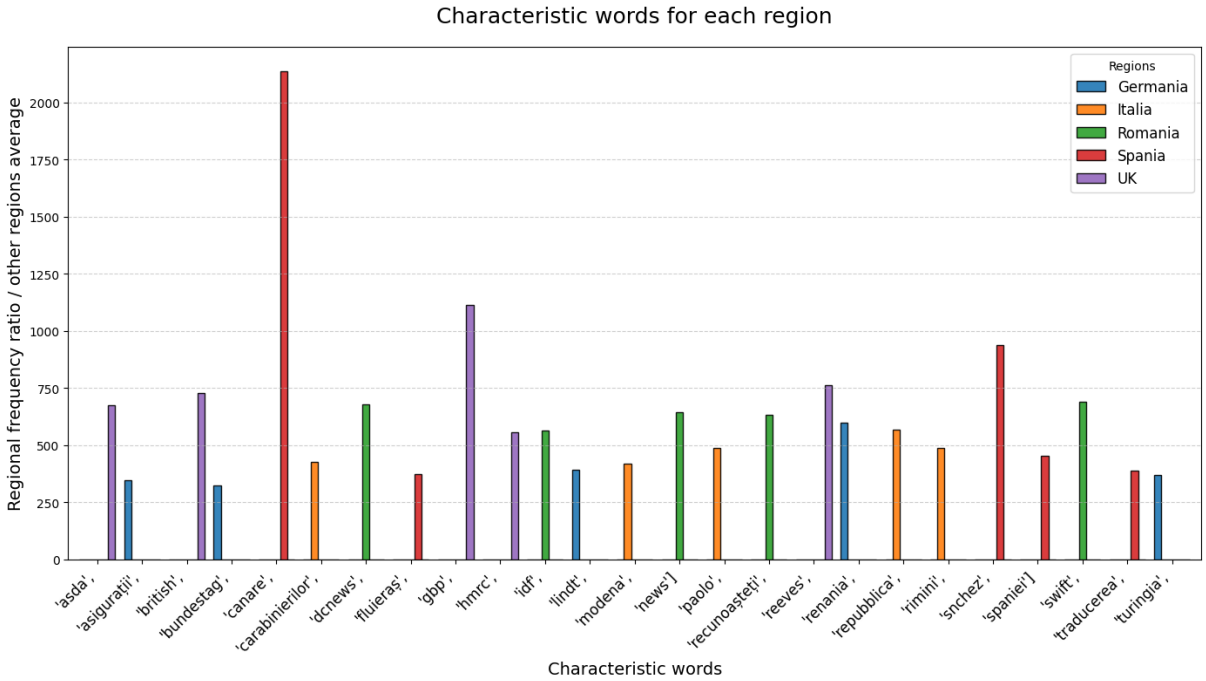

Figure 3: Characteristic Words