When the Dictionary Strikes Back: A Case Study on Slovak Migration Location Term Extraction and NER via Rule-Based vs. LLM Methods

Miroslav Blšták[α] **Jaroslav Kopčan**[α] **Marek Šuppa**[β, γ] **Samuel Harvan**[δ]
Andrej Findor[β] **Martin Takáč**[β] **Marián Šimko**[α]
[α]Kempelen Institute of Intelligent Technologies, Bratislava, Slovakia
[β]Comenius University in Bratislava, Slovakia, [γ]Cisco Systems,
[δ]Central European University, Vienna
Correspondence: marek@suppa.sk

Abstract

This study explores the task of automatically extracting migration-related locations (source and destination) from media articles, focusing on the challenges posed by Slovak, a low-resource and morphologically complex language. We present the first comparative analysis of rule-based dictionary approaches (NLP4SK) versus Large Language Models (LLMs, e.g. SlovakBERT, GPT-4o) for both geographical relevance classification (Slovakia-focused migration) and specific source/target location extraction. To facilitate this research and future work, we introduce the first manually annotated Slovak dataset tailored for migration-focused locality detection. Our results show that while a fine-tuned SlovakBERT model achieves high accuracy for classification, specialized rule-based methods still have the potential to outperform LLMs for specific extraction tasks, though improved LLM performance with few-shot examples suggests future competitiveness as research in this area continues to evolve.

1 Introduction

Automated analysis of media articles on human migration has gained prominence due to ongoing global crises such as conflict, poverty, political instability, and persecution of minorities, with migration frequently occupying media coverage (Spinde, 2021). Public attitudes toward migrants are shaped by factors including the type of migration, country of origin, and gender. For example, perceptions of mothers fleeing war in neighboring countries often differ markedly from those of economic migrants from distant regions. Thus, effective migration analysis requires extracting details about migrants' origins, destinations, and whether they are transiting or settling in a given country, a problem traditionally explored in different contexts (Bonde and Dembele, 2023).

Example – Location Extraction

Input
We've put Syrian asylum applications in the Czech Republic on hold for now, which is what we typically do when a country's situation changes dramatically.

Output
Source: Syria
Destination: Czech Republic

To study migration patterns, it is essential to identify and differentiate location-related terms (such as source and destination) from other locations mentioned in text, and to determine if the migration is relevant to the target country. Term Extraction (TE) and Named Entity Recognition (NER) are core NLP tasks for this purpose: TE identifies domain-specific terms, while NER detects named entities like people, places, and organizations (Wang et al., 2023). However, standard NER methods often fall short, as they may extract irrelevant transit or unrelated locations, and may miss coreferential expressions like "this place" that refer to migration endpoints. We posit that effectively analyzing migration vectors therefore requires a dual approach: extracting migration-specific location entities while simultaneously classifying sentences to determine their relevance to the migration patterns of interest.

While robust tools exist for high-resource languages such as English (Hu et al., 2023), low-resource and morphologically complex languages like Slovak present additional challenges, including limited annotated datasets and tool availability. Consequently, our work explores and compares multiple locality detection strategies for Slovak, including rule-based and large language model (LLM) approaches, and introduces the first manu-

Proceedings of the 10th Workshop on Slavic Natural Language Processing (Slavic NLP 2025), pages 91–100
July 31, 2025 ©2025 Association for Computational Linguistics

ally annotated dataset for these tasks that is also publicly available.

Our main contributions can be summarized as follows:

- We tailor location extraction methods to migration-related term identification,
- We further experiment with TE and NER in Slovak, a low-resource, morphologically rich language,
- We provide a comparison of dictionary-based, rule-based, and LLM-driven approaches,
- Finally, we create and manually annotate of the first Slovak dataset for migration-focused locality detection.

We publicly release the dataset as well as all of the code associated with its creation and subsequent experiments at `http://github.com/MIMEDIS/bsnlp2025`.

2 Related Work

Analyzing media texts concerning migration represents a growing field of research. The proliferation of media articles necessitates automated methods for extracting key information. Beyond identifying locations (Badr et al., 2024), researchers increasingly focus on extracting attributes such as sentiment or stance towards migration (Mets et al., 2023; Hamerlik et al., 2024) and detecting related hate speech (Khatua and Nejdl, 2023). Analysis also extends to user-generated content on social media platforms like Twitter and Facebook (Chi et al., 2025). However, gaining a true understanding of migration perspectives requires considering the geographical context, as viewpoints often differ based on the migrants' country of origin.

Numerous studies provide comparative analyses of Named Entity Recognition (NER) tool performance (Hu et al., 2023), generally indicating that modern approaches leveraging deep learning or Large Language Models (LLMs) surpass traditional dictionary-based methods. This presents challenges for low-resource languages like Slovak, where the availability of robust NER tools is limited (Šeleng et al., 2025; Šuba et al., 2023). Furthermore, the task extends beyond merely identifying location names; it requires discerning the type of location and whether it refers to a specific, bounded area.

Developing a solution for this nuanced location extraction is non-trivial. To our knowledge, no prior research specifically addresses location extraction with these granular requirements for migration texts, although prior research in the adjacent domain of border security intelligence has addressed similar challenges. For instance, Atkinson et al. (2011) developed a real-time system for the EU Border Agency to extract structured information on illegal migration events from multilingual news, and Zavarella et al. (2012) specifically focused on refining news event geotagging for border security using lexico-semantic patterns. A somewhat related problem was addressed in Zhang et al. (2010), where researchers extracted data from route direction documents. However, those documents possessed a simpler structure amenable to regular expressions. In contrast, determining a location's role (e.g., source, transit, or destination) within unstructured media text typically demands a deeper contextual understanding and analysis of sentence structure, as predefined patterns are absent.

3 Datasets

For evaluation purposes, we manually curated a dataset comprising several thousand sentences on migration, sourced from Slovak media articles published in 2022 and 2024. This dataset is partitioned into two subsets tailored for our distinct tasks. While many sentences overlap between subsets, some are exclusive due to task-specific relevance. The sentences cover migration related to conflicts in Ukraine, Syria, and Gaza, supplemented by other diverse scenarios (e.g., political or economic migration) to ensure broad representation. Annotation focused on identifying source and target migration locations, excluding purely transit mentions. Near-identical sentences derived from modified press releases were deduplicated.

The first subset supports a classification task: determining if a sentence pertains to Slovakia (i.e., migration *to*, *from*, or *through* the country). Each sentence is labeled accordingly. The second subset facilitates an extraction task, with sentences annotated with identified source and target locations, where applicable.

Manual annotation was performed by three authors following the guidelines outlined in Appendix A; sentences for which there was not full agreement among annotators were excluded to ensure data quality. The annotation process navigated several complexities. Key challenges included disambiguating locations with identical names, standardizing variations in place names (e.g., ’EU’ vs. ’European Union’, ’Czechia’ vs. ’Czech Republic’), and normalizing geographic scope (e.g., ’Europe’ vs. ’Eastern Europe’). Further difficulties arose from resolving referential expressions (’our country’), linking organizations or acronyms to their associated countries (e.g., ’Slovak Catholic Charity’, ’ZSSK’), identifying locations implied by adjectives (’African refugees’), and managing mentions of the same location at different granularities within one sentence. To illustrate the dataset’s complexity, Table 1 shows the distribution of unknown (i.e. not explicitly mentioned in the text) *source/destination* localities from human annotations.

Overall, the dataset represents a comprehensive collection of human-annotated sentences related to the migration theme, derived from 2323 unique articles. Within this corpus we establish two specialized subsets: The Slovakia-relevance subset contains 2736 annotated samples. The subset for the locality extraction task comprises 1652 samples annotated by humans for the identification and extraction of geographic localities. The final dataset was partitioned using a stratified approach with a 70:20:10 ratio for train/val/test split, ensuring class consistency distribution across all splits. More detailed statistics of the dataset and sample examples can be found in Appendix C.

4 Methods and Evaluation

We evaluated several methodological approaches for comparative analysis: a rule-based dictionary method, BERT-based models, and autoregressive transformers.

4.1 Rule-based Dictionary Approach (NLP4SK)

We employed NLP4SK[1], a Slovak NLP tool, for our rule-based dictionary approach. Its strengths include an extensive database of Slovak locations (recognizing Slovak, English, and international names), a lemmatizer, and conceptual dictionaries. This allows NLP4SK to identify locations across various grammatical forms typical in Slovak (e.g., non-capitalized, non-noun forms) and covering diverse geographical features (cities, regions, mountains, etc.).

For the classification task, NLP4SK identifies sentences as Slovak-related if they contain any reference to a Slovak location or a relevant adjective (e.g., “Slovak police”).

For the extraction task, NLP4SK first identifies all location entities. It then distinguishes source and target locations using lexico-syntactic cues. Prepositions preceding an entity often indicate its role (e.g., “from” suggests a source). Grammatical case is also leveraged, as Slovak morphology can convey this information (e.g., genitive often implies source, accusative target). Additionally, adjectives indicating origin (e.g., “Ukrainian man”) are typically mapped to the source location. Results from this method are designated NLP4SK and more information about its implementation can be found in Appendix D.

4.2 Geographical Relevance Classification

To classify sentences based on their geographical relevance to Slovakia, we established a majority-class baseline (always predicting the dominant “non-Slovak” category). We also implemented the NLP4SK dictionary classifier, which labels a sentence “Slovak” if any Slovak location lexicon entry is found.

Finally, we fine-tuned SlovakBERT as a binary classifier. This model was trained to distinguish specific references to Slovakia (the country, cities, or distinctly Slovak entities) from broader mentions (e.g., Europe). After a stratified data split (train/validation/test), we fine-tuned the model for 5 epochs using the AdamW optimizer with a learning rate of 2e-5. Performance was evaluated using accuracy and macro-F1 score on the held-out test set. Results for all classifiers are reported in Table 1a.

4.3 Locality Extraction Models

For the locality extraction task, which involves identifying migration source and target locations, traditional BERT-based models may present limitations. Their rigid sequence labeling and lack of nuanced directional understanding (source vs. target) could pose a problem. We determined that autoregressive transformers, such as GPT models, are better suited. These models excel at contextual understanding, inferential reasoning, and processing even implicit information necessary to distinguish between source and target localities effectively.

We utilized the GPT-4o model, specifically - *gpt-4o-2025-03-26* version, with the temperature pa-

[1] http://arl6.library.sk/nlp4sk/

	Geo relevance	
Metric	F1	Acc
Majority class	41.95	72.26
NLP4SK	96.15	96.90
SlovakBERT	**97.75**	**98.45**

(a) Macro F1 scores for the classification task of geographical relevance to Slovak localities. Evaluated on the test set. The best performance is in bold.

	NLP4SK	GPT-4o	
		zero-shot	few-shot
Source	91.82	83.09	87.21
Destination	84.36	76.13	81.64
Combined	**88.09**	79.62	84.42

(b) Macro F1 scores for locality extraction using various approaches on human-annotated data. The *Source*, *Destination* and *Combined* refer to the respective subsets of the dataset. The best performance is boldfaced.

Figure 1: Main results obtained from our experiments.

rameter set to zero to ensure consistent, deterministic outputs.

As for the prompting strategy, incorporating best practices, we leveraged meta-prompt templates for textual output, sourced directly from OpenAI's official documentation [2]. The exact prompt can be found in Appendix B. Used meta-prompts guide the model towards a holistic understanding of the migration narrative, moving beyond simple Named Entity Recognition. The model was prompted in English and instructed to provide reasoning for its decisions along with the structured output.

For the few-shot configuration, we have randomly selected a set of five examples to provide contextual demonstrations for the model.

Results are detailed in Table 1b. Consistent evaluation criteria and subsequent human validation were applied across all approaches. The evaluation is done by string/text similarity using both exact/substring checks and token-based metrics, with configurable thresholds to determine correct matches. The same evaluation was utilized for both approaches to ensure consistency and fair comparison.

5 Results

For geo-relevance classification, Table 1a shows that both the fine-tuned SlovakBERT model and a dictionary-based classifier significantly outperformed a naive majority-class baseline. SlovakBERT achieved slightly higher scores, confirming successful fine-tuning on meaningful patterns rather than merely guessing the most frequent class. The dictionary approach, specifically designed for this task, also demonstrated strong performance, serving as a valuable benchmark.

Turning to locality extraction, results in Table 1b indicate this is a substantially more complex task. We compared a dictionary-based method against GPT-4o in zero-shot and few-shot settings using macro F1 scores. Both approaches identified source localities more effectively than destination localities. This might stem from media often explicitly stating origin countries, while destinations (especially domestic ones) might be implied. Notably, the dictionary approach surpassed both GPT-4o variants in identifying both source and destination localities, highlighting the efficacy of tailored, rule-based systems for specialized tasks in low-resource languages like Slovak. Nevertheless, GPT-4o's improved performance with few-shot examples underscores the benefit of providing contextual demonstrations to large language models.

6 Conclusion and Future Work

We explored methods for extracting localization data from migration-related texts, specifically addressing the challenges presented by the Slovak language. Our work encompassed two main tasks: classifying whether migration discussed in a text concerns Slovakia, and extracting lists of origin and destination locations for migration events.

Comparing a dictionary-based approach with GPT-4o variants revealed differing performance characteristics, particularly for the extraction task. The evaluation methodology, relying heavily on string matching, naturally favored the dictionary method's highly consistent output format. The observed lower scores for GPT-4o may partly reflect this evaluation approach; its generative capabilities can lead to syntactically varied phrasings of correctly identified locations from the text, which are penalized by strict matching despite potential semantic equivalence to the gold annotations. A summary of error types encountered in the extrac-

[2] https://platform.openai.com/docs/guides/prompt-generation

tion task is detailed in Appendix E.

Future research aims to deepen the migration analysis by automatically extracting richer information. This includes identifying who is migrating (e.g., men, women, families), their characteristics (e.g., race, nationality, religion), the reason or purpose of migration (e.g., economic, refugee status, political), the direction of migration relative to the observer (immigration, emigration, return), and the stance towards migration based on these criteria. Such detailed data extraction would enable a comprehensive analysis of how media outlets cover migration and potentially influence public opinion.

Limitations

This study's scope and generalizability are subject to several limitations. Firstly, our reliance on proprietary models accessed via paid APIs may hinder the reproducibility of certain results. Secondly, the focus on Slovak, a language with limited computational resources, means our conclusions might not directly transfer to other languages.

The dataset itself introduces constraints. Compiled from news articles dated 2022 and 2024, it predominantly covers four major migration events: the war in Ukraine, the Israel-Palestine conflict, the Syrian political situation, and migration from Africa to Europe, leading to underrepresentation of minor migration events. Migration directionality (immigration vs. emigration) is assessed from a European and Slovak perspective. Despite efforts to ensure sentence diversity and balanced country representation, the dataset inevitably overrepresents nations frequently featured in the source articles, such as Ukraine, Russia, Syria, and various African countries.

Finally, the availability of suitable Named Entity Recognition (NER) tools specifically adapted for Slovak is limited. We selected a tool best suited for our data requirements, acknowledging its constraints compared to multilingual alternatives or the performance benchmarks discussed in recent research, particularly concerning the use of language-specific cues to differentiate source and target locations.

Acknowledgments

This research was supported by the EU NextGenerationEU through the Recovery and Resilience Plan for Slovakia under the project No. 09I01-03-V04-00007, DisAI-AMPLIFIED and by grant APVV-21-0114.

References

Martin Atkinson, Jakub Piskorski, Erik Van der Goot, and Roman Yangarber. 2011. *Multilingual real-time event extraction for border security intelligence gathering*. Springer.

Hajar Badr, Zamzam Awahida, Mansour Essgaer, Asma Ajaal, and Abbas Ahessin. 2024. Named entity recognition for identifying entities related to illegal migration in libya: An analysis of twitter textual data. In *2024 IEEE 4th International Maghreb Meeting of the Conference on Sciences and Techniques of Automatic Control and Computer Engineering (MI-STA)*, pages 567–572.

Lossan Bonde and Severin Dembele. 2023. High accuracy location information extraction from social network texts using natural language processing. *arXiv preprint arXiv:2308.16615*.

Guanghua Chi, Guy J. Abel, Drew Johnston, Eugenia Giraudy, and Mike Bailey. 2025. Measuring global migration flows using online data. *Preprint*, arXiv:2504.11691.

Endre Hamerlik, Marek Šuppa, Miroslav Blšták, Jozef Kubík, Martin Takáč, Marián Šimko, and Andrej Findor. 2024. ChatGPT as your n-th annotator: Experiments in leveraging large language models for social science text annotation in Slovak language. In *Proceedings of the 4th Workshop on Computational Linguistics for the Political and Social Sciences: Long and short papers*, pages 81–89, Vienna, Austria. Association for Computational Linguistics.

Xuke Hu, Zhiyong Zhou, Hao Li, Yingjie Hu, Fuqiang Gu, Jens Kersten, Hongchao Fan, and Friederike Klan. 2023. Location reference recognition from texts: A survey and comparison. *ACM Comput. Surv.*, 56(5).

Aparup Khatua and Wolfgang Nejdl. 2023. Why do we hate migrants? a double machine learning-based approach. In *Proceedings of the 34th ACM Conference on Hypertext and Social Media*, HT '23, New York, NY, USA. Association for Computing Machinery.

Mark Mets, Andres Karjus, Indrek Ibrus, and Maximilian Schich. 2023. Automated stance detection in complex topics and small languages: the challenging case of immigration in polarizing news media. *Preprint*, arXiv:2305.13047.

Timo Spinde. 2021. An interdisciplinary approach for the automated detection and visualization of media bias in news articles. *Preprint*, arXiv:2112.13352.

Zixiang Wang, Jian Yang, Tongliang Li, Jiaheng Liu, Ying Mo, Jiaqi Bai, Longtao He, and Zhoujun Li. 2023. Multilingual entity and relation extraction from unified to language-specific training. *Preprint*, arXiv:2301.04434.

Vanni Zavarella, Jakub Piskorski, Ana Sofia Esteves, and Stefano Bucci. 2012. Refining border security news event geotagging through deployment of lexico-semantic patterns. In *2012 European Intelligence and Security Informatics Conference*, pages 334–339. IEEE.

Xiao Zhang, Prasenjit Mitra, Alexander Klippel, and Alan MacEachren. 2010. Automatic extraction of destinations, origins and route parts from human generated route directions. In *Geographic Information Science*, pages 279–294, Berlin, Heidelberg. Springer Berlin Heidelberg.

Martin Šeleng, Štefan Dlugolinský, Michal Staňo, and Ladislav Hluchý. 2025. Model for named entity extraction from short fire event-related texts. *Procedia Computer Science*, 256:557–564. CENTERIS - International Conference on ENTERprise Information Systems / ProjMAN - International Conference on Project MANagement / HCist - International Conference on Health and Social Care Information Systems and Technologies.

Dávid Šuba, Marek Šuppa, Jozef Kubík, Endre Hamerlik, and Martin Takáč. 2023. Wikigoldsk: Annotated dataset, baselines and few-shot learning experiments for slovak named entity recognition. *Preprint*, arXiv:2304.04026.

A Annotation guidelines

Locality Extraction Guidelines

Migration Vector consists of an locality origin - SOURCE and DESTINATION locality that represents the movement of people. Annotations of migration vectors should be based on explicit textual evidence, not on inference or assumption as these could be wrong. Always define localities on Slovak nominative case in the annotation.

Text Analysis Process

- **Step 1**

  Begin by carefully reading the entire text. Identify all mentioned localities and pay attention to surrounding contextual clues and linguistic markers for establishing direction of migration between them.

- **Step 2**

  After localities identification, classify each of them according to their roles in the migration vectors as SOURCE locality - if the locality functions as origin point where migration began, DESTINATION locality - if the locality functions as destination point wher migration ended. Some localities present within text might be TRANSIT localities - where migration movement did not originate or ended. Additionally there might be UNRELATED localities with no direct connection to migration patterns.

- **Step 3**

  After locality role assessment within migration patterns, establish final SOURCE-DESTINATION migration pairs that represent the migration vectors. This involves connection of origin localities with their corresponding destinations, while excluding transit or unrelated localities.

Special Considerations when identifying migration vectors from text:

- Migration within historical context require the same methodological approach as contemporary ones

- Similarly, for hypothetical migration scenarios same thorough analytical process should be done

- Annotations related to locality extraction should remain firmly anchored in the text, it is recommended to avoiding inferences about locations not explicitly mentioned or inferred from contextual clues

- If there are present multiple migration vectors within the inspected sample, treat each unique combination as a distinct migration vector

- If there is ambiguous directional information, meaning text does not clearly establish whether identified localities serves as SOURCE or DESTINATION localities, do not try to guess intended direction and annotate them as None.

Locality Relevance Guidelines
Determine whether a sentence contains content related specifically to Slovak locations.

Text Analysis Process

- Read and analyze the text for both explicit and implicit mention of Slovakia, Slovak places or direct references to Slovak people and other entities.

- ext mentioning Slovakia as a country, a specific location within Slovakia or content directly related to Slovak people, entities whether explicitly stated or implied is **considered as related to Slovak localities.**

- Text which does not mention Slovak locations or contains references to broader ranges like Europe or completely different locations is **considered as not-related to Slovak localities.**

Ambiguous cases: When encountering potentially ambiguous terms, rely on context to determine the correct reference.

B Prompting strategy

Migration Vector Extraction Prompt

Prompt Instructions
Identify migration vectors (FROM and TO localities) from a Slovak text.
Follow these instructions:
1. **Identify Localities**: Extract all localities mentioned in the text.
2. **Handle Unclear Localities**: Mark as "None" if no clear origin or destination is found and do not infer localities.
3. **Determine Migration Direction**: Establish origin (FROM) and destination (TO) localities for migration.
4. **Ignore Transit Locations**: Focus on starting point and endpoint only.
5. **Multiple Vectors**: Identify each unique FROM-TO pair if more than one vector exists.
6. **Handle Unclear Localities**: Mark as "None" if no clear origin or destination is found.
7. **Output for Each Vector**:

- **Provide Reasoning**: Detail the identified localities with references from the text.

- **Conclude**: State the appropriate migration vector locality pairs on the final line.

- **Confidence Level**: Specify as High, Medium, or Low.

- **Format**: "FROM: [locality], TO: [locality]" ensuring localities are listed without qualifying adjectives.

8. **Language**: Use Slovak (nominative case).
Ensure this applies to historical or hypothetical scenarios as well.

Steps
1. Analyze text for specific mentions of localities. 2. Interpret context clues for implicit localities. 3. Determine the full migration vector by excluding mere transit points. 4. Document findings and reasoning.

Output Format
1. Analysis of localities mentioned 2. Reasoning for migration vector identification 3. Final answer with locality pairs, or "None" if not identifiable (naming in Slovak nominative case)

Example Output
FROM: Sýria, TO: Nemecko

C Dataset Samples and Statistics

C.1 Samples

Below are examples demonstrating scenarios in which migration vectors contain undetermined origin or destination points.

Example – Source Locality Unknown

Input
In 2018, during a visit to a migrant facility in Texas, she wore a jacket with the slogan 'I Really Don't Care, Do U?'

Output
Source: None
Destination: Texas

Example – Destination Locality Unknown

Input
"We're determined to do whatever we can to stop Syria from falling apart, prevent masses of people fleeing from Syria, and naturally, to curb the spread of terrorism and extremism," according to the minister, as reported by AFP news agency.

Output
Source: Syria
Destination: None

Example – Both Localities Unknown

Input
The Defense Minister also highlighted how Smer's longstanding positions on the Ukraine conflict and migration issues are proving prescient. He pointed out that events are increasingly validating what the party has maintained all along.

Output
Source: None
Destination: None

C.2 Statistics

The Figures below depict various statistics of the dataset, such as its character (Figure 2) and token (Figure 3) length distributions, label distributions (Figure 4) and locality distribution (Table 1).

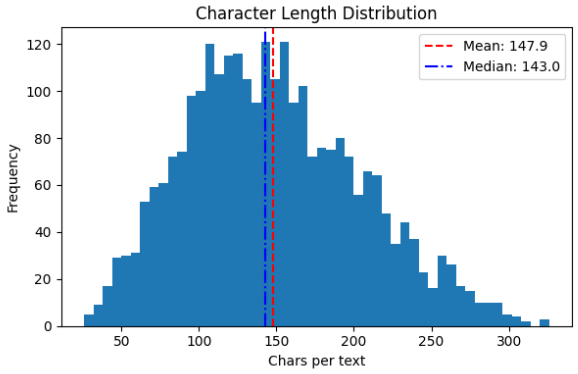

Figure 2: Distribution of the dataset: character length in the final dataset

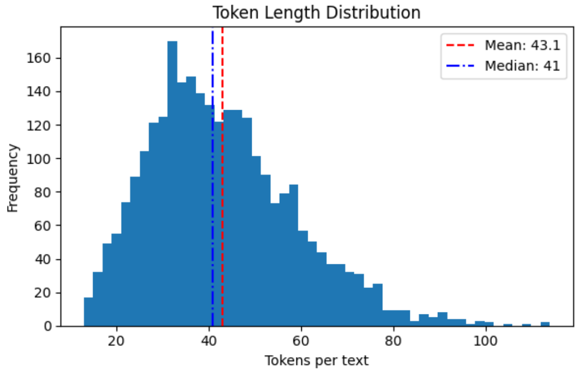

Figure 3: Distribution of the dataset: token length in the final dataset. The tokens originate from the SlovakBERT tokenizer.

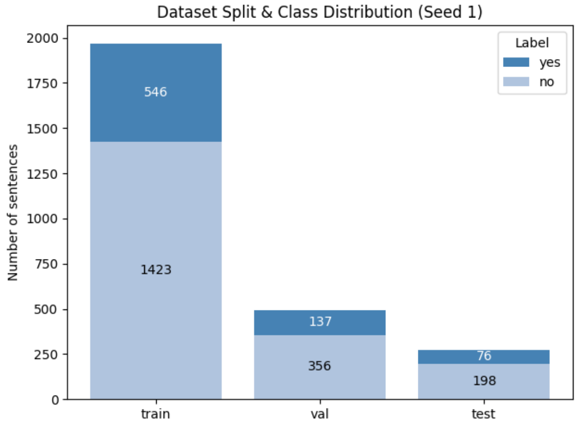

Figure 4: Final relevance dataset distribution across train, validation, and test splits with consistent class ratios.

Locality	Unknown count	Percentage
Source	717	43.40
Destination	441	26.69
Combined	0	0.00

Table 1: The distribution of ”Unknown” localities (i.e. when either the Source or Destination field is not filled in for a specific sample) in the various subsets of the final dataset. Note that the result in the *Combined* row shows that either Source or Destination field are always filled in in the final dataset.

D Rule-Based System Implementation

The dictionary rule-based approach contains several modules that are useful, especially for our problem, and utilizes a huge database of locations in the Slovak language.

- Location Database - an extensive database of geographical entities in the Slovak language context
- Morphological Flexibility - allowing locality matching beyond the standard nominative case, in a wide variety of grammatical forms. Utilization of lexico-syntactic information about prepositions, which is useful to distinguish between source and target locations.
- Semantic Labeling - labels obtained from conceptual dictionaries which allows us to extract location not only from typical noun form, but also in other different forms what is very often in high flexible language as Slovak language is (e.g. words mentioned as adjectives: “ukrainian man” or entities which do not start with capital letter).
- Migration Specific text detection based on keywords.

Georelevance Rule

- If a text token is in the dictionary of Slovak locations (containing all Slovak regions, districts, counties, cities, municipalities), then mark the text as geographically relevant.

  Example: "Migranti dočasne bývajú v tábore pre utečencov v Kútoch." (Migrants are temporarily staying in refugee camp in Kúty.) → Geographically relevant: Kúty is in the Slovak locations dictionary

Location Extraction Rules

1. **Source rule**: When a Slovak location name appears after the preposition "z" (from), identify and extract this as the place something or someone comes from.

   Example: "Utečenci z Bratislavy hľadajú nové ubytovanie." (Refugees from Bratislava are looking for new accommodation.) → Extracted Source: Bratislava

2. **Destination rule**: When a Slovak location name appears after the preposition "do" (to), identify and extract this as the place something or someone is going to.

   Example: "Migranti cestujú do Košíc za prácou." (Migrants are traveling to Košice for work.) → Extracted Destination: Košice

3. **Adjective Transformation Rule**: When a nationality or location-based adjective describes a noun, convert this relationship to show the noun originates from that location.

   Example: "Ukrajinskí utečenci potrebujú dlhodobú pomoc." (Ukrainian refugees need long-term assistance.) → Extracted Source: Ukrajina

   Example: "Sýrske rodiny cestujúce do Bratislavy." (are traveling to Bratislava.) → Extracted Source: Sýria, Extracted Destination: Bratislava

E Error Analysis for Locality Extraction Task

model	error type	count
GPT-4o	*label error*	4
	model error	38
NLP4SK	*label error*	6
	model error	28

Table 3: Error analysis summary o then locality extraction test set for both models. We note that the full test set is comprised of 166 samples.

The Table 3 shows error analysis results for locality extraction on samples that were mismatched by GPT model.

- **Label error** represent instances where the human annotators incorrectly labeled the localities in the text.

LLM	Method	EN prompt	SVK prompt
gpt-4o	FewShot	87.67	86.75
	ZeroShot	85.42	**86.54**
gemini-2.5-flash-preview-05-20	FewShot	**87.95**	86.75
	ZeroShot	77.11	88.55
llama-3.3-70B	FewShot	85.54	**87.35**
	ZeroShot	82.53	83.73
deepseek-chat-v3-0324	FewShot	87.39	87.27
	ZeroShot	86.74	**88.23**

Table 2: Macro F1 scores for the location extraction task with combined results for both FROM and TO comparing different models across EN CoT and SVK CoT prompt versions.

- GPT **model error** represent instances where LLM failed to correctly identify or extract locality information that was present in the text.
- NLP4SK **model error** represent instances where Dictionary failed to correctly identify or extract locality information that was present in the text

Table 2 presents the performance results for both evaluation methods across a range of language models. To better understand model limitations, we conducted a detailed error analysis of the model outputs, identifying systematic error patterns and the most significant challenges of the locality extraction task.

Systematic Error Analysis

A recurring pattern of errors was observed across all models on the 161-item test set. The most common errors, averaged per model, were ranked by frequency:

- **Incorrect Destination Extraction:** 15 cases
- **Destination Hallucination:** 12 cases
- **Source Hallucination:** 6 cases
- **Incorrect Source Extraction:** 4 cases
- **Omitted Destination:** 3 cases

Key Challenges in Locality Extraction

Destination Extraction. The analysis reveals that identifying destination localities is the most error-prone aspect of the task for large language models. All models exhibited a strong tendency to hallucinate destinations, even when none were present in the source text. These findings indicate that destination extraction is significantly more challenging than source extraction.

Linguistic Challenges. Several linguistic phenomena proved difficult for the models to handle correctly:

Geographic Specificity: Models frequently confused broad regions with specific countries (e.g., substituting “Eastern Europe” for a specific nation) or conflated locations into larger areas (e.g., mapping “Africa/Libya” to “Northern Africa”).

Prepositional Ambiguity: Models struggled to interpret Slovak directional prepositions (*do*, *z*, *v*), often incorrectly inferring movement from statements that merely described a location outside of a migration context.

Contextual Disambiguation: A common failure was the inability to distinguish between locations relevant to migration (i.e., source or destination) and those that were part of the narrative setting, especially within complex refugee or immigration accounts.

Entity Role Identification: Uncertainty in identifying the role of mentioned individuals (e.g., as a migrant, an observer, or an aid worker) negatively impacted the accuracy of the locality extraction process.