A Multilingual Physical Commonsense Dataset for Hungarian and Romanian

Andrew Hoblitzell
Purdue University, Indianapolis, USA
ahoblitz@purdue.edu

August 2025

Abstract

This data card documents a multilingual commonsense dataset for Hungarian and Romanian for the MRL 2025 Shared Task. The dataset follows PIQA and SWAG/HellaSwag formats: scenarios with two candidate completions that test physical interactions in everyday contexts. Native-language translators and writers produced and reviewed items under a defined protocol. The release contains 100 examples per language in TSV format.

Correspondence: Andrew Hoblitzell (ahoblitz@purdue.edu)

Keywords: Multilingual NLP, Commonsense Reasoning, Dataset Creation, Low-resource Languages, Physical Reasoning, Hungarian, Romanian

1 Introduction

Commonsense reasoning is a central challenge in natural language processing, requiring implicit knowledge of the physical world. Benchmarks such as PIQA [2] have advanced English evaluation, yet many languages lack resources for this capability.

The present dataset targets the MRL 2025 Shared Task on multilingual commonsense reasoning [1]. Hungarian and Romanian, two languages with distinct linguistic properties and limited NLP resources, are included to broaden evaluation beyond major languages. The dataset adopts the PIQA and SWAG/HellaSwag style [2, 3, 4], presenting binary-choice scenarios that assess physical plausibility. The data also relates to COPA-style causal plausibility and its multilingual extension XCOPA [5, 6], presenting binary-choice scenarios that assess physical plausibility.

2 Languages and Dataset Design

2.1 Language Selection

The dataset includes Hungarian to expand typological coverage given its status as a Uralic language with agglutinative morphology and relative resource scarcity. Underrepresentation in multilingual benchmarks also meant Hungarian was among the last languages to secure volunteer writers and reviewers. During curation, encoding errors affecting Hungarian orthography were corrected, notably degradation of the long double-accent characters ő (U+0151) and ű (U+0171) into plain vowels or unstable combining sequences under legacy encodings. All Hungarian text is normalized to NFC, stored as UTF-8 in the TSV release, with automated checks for stray combining marks and inconsistent punctuation.

The dataset includes Romanian to complement Hungarian with a Romance language while remaining relatively low-resource. Romanian likewise was among the last languages to recruit volunteers, reflecting under-representation. Curation resolved frequent encoding confusions between comma-below and cedilla variants by standardizing ș (U+0219) and ț (U+021B) and replacing legacy ş/ţ code points where present. As with Hungarian, all Romanian text is normalized to NFC and saved as UTF-8 TSV. The dataset is freely available for use.

The parallel development of the dataset in Hungarian and Romanian revealed interesting linguistic nuances in expressing physical concepts. For instance, scenarios involving motion and direction often required different grammatical structures due to Hungarian's rich system of case suffixes compared to Romanian's reliance on prepositions.

2.2 Task Format

Each example contains a scenario description followed by two candidate completions. The task is to choose the completion that is physically plausible. The prompt template is:

```
Scenario: [Description of initial situation]
Option A: [First possible completion]
Option B: [Second possible completion]
```

2.3 Example Items

Hungarian Example: *Scenario*: Amikor egy könnyű fémcsésze leesik a pultról...
Option A: ...összetörik, amikor földet ér.
Option B: ...felpattan vagy horpad, de nem törik össze.
Correct: A/B

Romanian Example: *Scenario*: Când o ceașcă ușoară de metal cade de pe tejghea...
Option A: ...se va sparge după ce atinge pământul.
Option B: ...va sări înapoi sau se va îndoi, dar nu se va sparge.
Correct: A/B

3 Annotation Protocol & Quality Control

3.1 Team Composition

For each language, two translators and one editor were used. Translators ensured accurate and natural language use; editors picked what they thought was the superior edit.

3.2 Authoring Guidelines

Guidelines required cultural grounding while maintaining universal physical principles. Incorrect options were written to appear superficially reasonable yet physically impossible.

3.3 Verification Process

Each item was created by a writer and then cross-validated by a second translator for linguistic accuracy and consistency with the guidelines.

3.4 Translation Notes

Notes for Hungarian: Overall, the translation process went smoothly, with most of the phrases and terminology transferring naturally into Hungarian. Structure and meaning were well preserved, and the final result should be clear and easy to follow for Hungarian speakers. For example, the phrase "user-friendly interface" translated very well into Hungarian as "felhasználóbarát felület", which is a common and accurate expression. However, the expression "cutting-edge solution" was more challenging, as the direct translation ("élvonalbeli megoldás") does not always carry the same modern and innovative tone. In such cases, the closest natural equivalent was used while keeping the original intent.

Notes for Romanian: In general, the translations respect the original meaning and manage to convey the information in a clear and natural way in Romanian. An example that was translated very well is sentence no. 3: "How do you put a stamp on a T-shirt?" → "Cum pui o ștampilă pe un tricou?". This faithfully and naturally renders the original meaning. On the other hand, an example that was not translated as well is sentence no. 1: "When a light metal cup falls off a counter, it will shatter after hitting the ground." → "Când o ceașcă ușoară de metal cade de pe tejghea, se va sfărâma după ce atinge pământul.". Here, the verb "will shatter" is not the most appropriate for a metal object (which usually deforms or bends, not shatters), which makes the translation less natural.

4 Limitations

The dataset size supports evaluation but may limit fine-tuning use. A text-only format lacks multimodal grounding. Everyday scenarios may not generalize to specialized technical domains. Future work includes expanding the number of examples and languages, adding adversarial items, and exploring multimodal variants.

5 Conclusion

This data card documents a multilingual physical commonsense dataset for Hungarian and Romanian developed for the MRL 2025 Shared Task. The contribution includes culturally grounded evaluation resources for two underrepresented languages and a documented annotation protocol. Parallel examples across typologically distinct languages can support cross-linguistic analysis of how physical reasoning is expressed.

Acknowledgments

Thanks to the annotators for their contributions. Appreciation to the MRL 2025 organizers for the shared-task framework and to the translation reviewers for ensuring linguistic quality and cultural appropriateness.

References

[1] MRL 2025 Shared Task: Multilingual Physical Reasoning Datasets. Workshop on Multilingual Representation Learning (MRL) at EMNLP 2025. https://sigtyp.github.io/st2025-mrl.html. Accessed: 30 Aug 2025.

[2] Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. PIQA: Reasoning about Physical Commonsense in Natural Language. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):7432–7439, 2020. https://arxiv.org/abs/1911.11641.

[3] Rowan Zellers, Yonatan Bisk, Roy Schwartz, and Yejin Choi. SWAG: A Large-Scale Adversarial Dataset for Grounded Commonsense Inference. In *Proceedings of EMNLP*, pages 93–104, 2018. https://arxiv.org/abs/1808.05326.

[4] Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. HellaSwag: Can a Machine Really Finish Your Sentence? In *Proceedings of ACL*, pages 4791–4800, 2019. https://arxiv.org/abs/1905.07830.

[5] Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S. Gordon. Choice of Plausible Alternatives: An Evaluation of Commonsense Causal Reasoning. In *Proceedings of the AAAI Spring Symposium on Logical Formalizations of Commonsense Reasoning*, 2011. https://ict.usc.edu/pubs/Choice%20of%20Plausible%20Alternatives-%20An%20Evaluation%20of%20Commonsense%20Causal%20Reasoning.pdf.

[6] Edoardo M. Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. XCOPA: A Multilingual Dataset for Causal Commonsense Reasoning. In *Proceedings of EMNLP*, pages 2362–2376, 2020. https://aclanthology.org/2020.emnlp-main.185.