# PIQA-PL: Polish Physical Reasoning Dataset

**Ewa Rudnicka**
Department of Artificial Intelligence
Wroclaw Tech
Poland
ewa.rudnicka@pwr.edu.pl

**Jan Kocoń**
Department of Artificial Intelligence
Wroclaw Tech
Poland
jan.kocon@pwr.edu.pl

## Abstract

This paper describes the construction of **PIQA-PL**, a Polish dataset for physical commonsense reasoning created in response to the MRL 2025 Shared Task at EMNLP. The dataset contains 131 examples in the PIQA format, each consisting of a prompt, two candidate solutions, and a binary label. All data were authored manually by native Polish linguists, following the guidelines of the original PIQA dataset. We provide details on the collection process, content categories, perturbation types, and dataset statistics.

## 1 Introduction

Commonsense reasoning is a crucial challenge in natural language understanding. The PIQA dataset (Bisk et al., 2020) introduced a benchmark for physical commonsense reasoning in English. To enable similar research for Polish, we present PIQA-PL, a manually constructed dataset aligned with the shared task requirements.

The dataset aims to evaluate the ability of systems to reason about everyday physical situations in Polish. It follows the PIQA task format and contributes to broadening the multilingual commonsense resources.

## 2 Dataset Construction

PIQA-PL was created entirely manually by native Polish speakers with linguistic expertise. Two linguists were responsible for brainstorming and inventing examples, drawing on their knowledge of Polish history, culture, customs, and everyday habits. They first generated candidate prompts, then developed them into full examples consisting of a question and two possible solutions. In most cases, the correct solution was written first, after which one or more perturbations were introduced to form incorrect alternatives.

During this process, the linguists also assigned each example to one or more thematic **content categories** and exactly one **perturbation type**. This ensured a balanced coverage of both domains and linguistic variations.

Every constructed example was subsequently reviewed by a third, more experienced linguist who served as a supervisor. This supervisor checked the naturalness, grammaticality, and cultural plausibility of each item, often suggesting modifications or improvements. In this way, the dataset reflects a careful multi-stage process that combines creativity, linguistic accuracy, and rigorous quality control.

## 3 Dataset Statistics

PIQA-PL contains 131 examples. There are 17 distinct thematic categories and 22 perturbation types. The most frequent content categories are *practical* (49), *food* (31), and *custom* (18). Other content categories: automotive, beauty, cleaning, customs, education, equine, exercise, free time, game, gardening, health, household, hygiene, religion. The most frequent perturbations are *change of noun* (51), *change of instruction* (15), and *change of adjective* (11). Other perturbations: change of adjective and noun, change of adposition, change of adpositional phrase, change of adverb, change of adverbial, change of description, change of noun and adjective, change of noun and omission of prepositional phrase, change of noun and verb, change of noun phrase, change of nouns, change of number, change of preposition, change of prepositional attribute, change of verb, change of word order, changed word order, inversion, omission.

Table 1 summarizes basic dataset statistics, while Tables 2 and 3 show the most frequent categories.

| Statistic | Value |
|---|---|
| Total examples | 131 |
| Distinct content categories | 17 |
| Distinct perturbation types | 22 |
| Avg. prompt length | 11.97 |
| Avg. solution0 length | 11.98 |
| Avg. solution1 length | 11.69 |
| Proportion >25 (prompt+solution0) | 58.8% |
| Proportion >25 (prompt+solution1) | 55.7% |

Table 1: Summary statistics of PIQA-PL. Lengths are in words.

| Content category | Count |
|---|---|
| Practical | 49 |
| Food | 31 |
| Custom | 18 |
| Cleaning | 10 |
| Health | 7 |

Table 2: Top-5 content categories in PIQA-PL.

| Perturbation type | Count |
|---|---|
| Change of noun | 51 |
| Change of instruction | 15 |
| Change of adjective | 11 |
| Change of verb | 10 |
| Change of nouns | 7 |

Table 3: Top-5 perturbation types in PIQA-PL.

## Acknowledgments

This work was supported in part by (1) CLARIN-PL, the European Regional Development Fund, FENG programme (FENG.02.04-IP.040004/24); (2) the Polish Ministry of Science and Higher Education: the CLARIN-PL project and the program "International Projects Co-Funded"; (3) the statutory funds of the Department of Artificial Intelligence, Wroclaw Tech.

## References

Yonatan Bisk, Rowan Zellers, Jianfeng Gao, Yejin Choi, and 1 others. 2020. PIQA: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 7432–7439.