The CLASSLA-PIQA South-Slavic Multilingual Physical Commonsense Reasoning Dataset Collection

Vuk Dinić, Sonja Benčina, Jaka Čibej, Nada Galant, Taja Kuzman Pungeršek, Mirna Potočnjak, Peter Rupnik, Biljana Stojanovska Đurić , Nikola Ljubešić

Here we describe the process followed in preparing the physical commonsense datasets for four official South Slavic languages – Slovenian (sl), Croatian (hr), Serbian (sr), Macedonian (mk) – and its two dialects – Slovenian Cerkno dialect (sl-cer) and Croatian Chakavian dialect (hr-ckm). For each language and dialect we identified one author that would construct the dataset manually by following identical instructions across the languages and dialects. The authors were first given the general instructions of the shared task:

Task:

- Create at least 100 PIQA-format examples (prompt, solution0, solution1, label) in your native language or dialect.
- Each example should describe a physical object and what can happen to it / be done with it.

Requirements:

- Physical reasoning: For each example, the solution must relate to physical properties of one or more objects.
- Common sense: For each example, an average person who speaks your language natively should know the answer.
- Culturally specific: We encourage authors to include culturally-relevant examples for their language(s). For example, some items may not be easily translatable into English or may require regional and/or cultural common sense.

The general instructions were later expanded with some practical guidance. Authors are encouraged to begin with objects around them, such as a chair, books, plants, or headphones. They may also use a dictionary to select random objects, or look to sources like Wikipedia, life hacks, or assembly manuals for inspiration. A useful strategy is to draft prompts first and leave solutions for later. To illustrate good practice, the Serbian dataset was completed in advance and made available as a reference.

Altogether, the authors generated 600 examples in four official South Slavic languages: 100 examples in Slovenian, 100 in Croatian, 150 in Serbian (in Cyrillic script), another 150 in Serbian (automatically transliterated 150 Cyrillic examples into the Latin script), and 100 in Macedonian. The authors also

prepared two dialectal datasets, both of which are already part of the DIALECT-COPA collection (https://aclanthology.org/2024.vardial-1.7/): 100 examples in the Cerkno dialect of Slovenian and 100 in the Chakavian dialect of Croatian. In this set of languages and dialects, all use the Latin script, except for Serbian, using both the Latin and Cyrillic script, and Macedonian, using the Cyrillic script only.

Once the datasets were constructed, we ran a manual cross-check of the instances and their labels. The cross-check was performed by other co-authors of this collection of datasets. For each dataset, another co-author, having a significant level of understanding of the language or dialect, has solved the task by not having gold labels at their disposal. We report the obtained accuracies in the first row in Table 1. Important to note is that the person taking the Chakavian test is just partially knowledgeable of the dialect, therefore the lower result of 76%, but also a show of how challenging the task is without a detailed understanding of a dialect. Once the cross-check was done, the authors were invited to inspect the disagreements between their labels and those given by other authors, not resulting in any changes of the labels. During that process, minor improvements in the text of the instances, based on comments of co-authors doing cross-checks, were applied as well.

Aside from the cross-check accuracy, which corresponds to human performance on the dataset, Table 1 gives basic insights into the (dis-)similarities of the 6 datasets in terms of the length distribution of instances. While the instructions and examples of good practice given to the authors were identical, the datasets still show differences in their length. The main difference is in the dialectal datasets, that are shorter on average, and lack especially long instances. The reason for this limitation is rather obvious - there are no texts available in these dialects (recipes, instruction manuals) that could have been exploited as a starting point for generating longer instances.

	Slovenian	Croatian	Serbian	Macedonian	Cerkno	Chakavian
Human performance	97%	100%	97.33%	92%	100%	76%
Median instance length	106.5	57.5	147.5	113	56	34.5
Average instance length	120.3	95	200.5	117	64.3	36.6
Minimum	13	15	13	49	8	16

instance length						
Maximum instance length	406	733	1802	236	178	71
Standard deviation	72.3	121	227.9	38	26.5	11.5

Table 1. Human performance in accuracy obtained by the manual cross-check by other authors, and statistics on the length distribution of instances (number of words).