# DIALECT-COPA: Extending the Standard Translations of the COPA Causal Commonsense Reasoning Dataset to South Slavic Dialects

**Nikola Ljubešić**
Jožef Stefan Institute
University of Ljubljana
nikola.ljubesic@ijs.si

**Nada Galant**
Čakavski sabor
nada.galant@gmail.com

**Sonja Benčina**
Parafraza
be.sonja@gmail.com

**Jaka Čibej**
University of Ljubljana
jaka.cibej@ff.uni-lj.si

**Stefan Milosavljević**
University of Graz
stefannmilosavljevic@gmail.com

**Peter Rupnik**
Jožef Stefan Institute
peter.rupnik@ijs.si

**Taja Kuzman**
Jožef Stefan Institute
taja.kuzman@ijs.si

## Abstract

The paper presents new causal commonsense reasoning datasets for South Slavic dialects, based on the Choice of Plausible Alternatives (COPA) dataset. The dialectal datasets are built by translating by native dialect speakers from the English original and the corresponding standard translation. Three dialects are covered – the Cerkno dialect of Slovenian, the Chakavian dialect of Croatian and the Torlak dialect of Serbian. The datasets are the first resource for evaluation of large language models on South Slavic dialects, as well as among the first commonsense reasoning datasets on dialects overall. The paper describes specific challenges met during the translation process. A comparison of the dialectal datasets with their standard language counterparts shows a varying level of character-level, word-level and lexicon-level deviation of dialectal text from the standard datasets. The observed differences are well reproduced in initial zero-shot and 10-shot experiments, where the Slovenian Cerkno dialect and the Croatian Chakavian dialect show significantly lower results than the Torlak dialect. These results show also for the dialectal datasets to be significantly more challenging than the standard datasets. Finally, in-context learning on just 10 examples shows to improve the results dramatically, especially for the dialects with the lowest results.

## 1 Introduction

Causal commonsense reasoning task has been shown to be highly useful for evaluation of the natural language understanding (NLU) capabilities of large language models (LLM) (Wang et al., 2019). It provides an insight into whether the models are able to acquire common world knowledge and, moreover, whether they are able to generalize to other languages. Among others, the Choice Of Plausible Alternatives (COPA) dataset (Roemmele et al., 2011) has been extensively used for these purposes. At the time of development of the COPA dataset, a successful application of commonsense inference to text understanding was considered to be "one of the grand challenges of natural language processing" (Gordon et al., 2012), with the most successful systems barely achieving accuracy above the random baseline. Recently, we have witnessed development of incredibly powerful language models and innovations in this field happening at an unprecedented pace. Twelve years after the introduction of the COPA dataset, the state-of-the-art pretrained language models are able to achieve accuracy higher than 99% (Chowdhery et al., 2023; Zhong et al., 2022). However, the COPA dataset was initially available only for English. When first efforts were made to develop COPA datasets also for other, less-resourced languages, the evaluations of large language models on these datasets showed that there is a large gap in their natural language understanding capabilities when applied to different languages (Ponti et al., 2020; Žagar and Robnik-Šikonja, 2022).

In this paper, we present new COPA datasets for three South Slavic dialects – the COPA dataset for Slovenian Cerkno dialect, Croatian Chakavian dialect and the Torlak dialect of Serbian.[1] We re-

[1]The Torlak dialect is a Balkan Sprachbund variety that shares features with both standard Serbian and other Balkan languages, among which most notably Macedonian and Bul-



*Proceedings of the Eleventh Workshop on NLP for Similar Languages, Varieties, and Dialects (VarDial 2024)*, pages 89–98

lease these dialectal datasets as extensions of the already existing COPA datasets in standard languages, namely Slovenian, Croatian, Serbian and Macedonian. All the datasets were translated from the English COPA dataset (Roemmele et al., 2011) following the XCOPA methodology (Ponti et al., 2020), with the difference that dialectal translations were supported both by the English original and the closest standard translation.

Recent instruction-tuned generative language models were shown to do incredibly well in this commonsense reasoning task, even in South Slavic languages, both in Latin and Cyrillic, achieving accuracy between 94% and 97%.[2] This motivated us to further evaluate the models' capabilities, by analyzing their performance on South Slavic dialects, for which there is much less texts available on the web than for standard South Slavic languages. This means that these dialects are barely present in the training data of the language models, or are not present at all. The performance of large language models on dialectal texts is a highly relevant research direction because it measures the capacity of a language model to generalize the linguistic knowledge beyond the standard languages the models have primarily been pretrained on.

The selection of dialects followed three main criteria: (1) that they are rather different from the standard, (2) that they are diverse between each other, and (3) that we can identify reliable translators into that dialect. Starting with Slovenian, several viable options are available in addition to the Cerkno region dialect (such as the Prekmurje dialect or dialects spoken by Slovenian national minorities in Italy, Austria, and Hungary), but ultimately, the decision to select the Cerkno dialect was based on the availability of a translator. For the Croatian language, given that the Slovenian standard language is rather close to the Kajkavian dialect of Croatia (Kapović, 2017), and that the Shtokavian dialect is very close to the standard language (Vidović, 2007), we chose the Chakavian dialect, again, selecting the micro-dialect of Žminj due to availability of a reliable translator. Finally, aiming at a dialect from Serbia, Macedonia, or Bulgaria, we chose the Torlak dialect which has been well researched as a distinct dialect of the Balkan Sprachbund, having relationships to Serbian, Macedonian and Bulgarian (Mišeska Tomić, 2006; Milosavljević, 2018; Živojinović, 2021; Vuković et al., 2022). Additionally, Torlak is officially listed as a vulnerable language by the UNESCO (Moseley, 2010). To go with the micro-dialect of the region near the town of Lebane (Žugić, 2005; Milosavljević, 2018), again, was based on the availability of a translator.

The reasons why we are following upon translating an existing English benchmark, rather than compiling a new one, are the following: (1) it is much cheaper, but also safer to translate an existing benchmark, proven to measure reasonably well the phenomenon of interest, especially in light of a similar culture, rather than to compile a new benchmark that would need to go through quite many tests before being reasonably safe for usage, (2) the results obtained on a translated benchmark are much more comparable to the results on the original benchmark than the results on less dependent benchmarks, which allows us to measure the comparative performance of a model in multiple languages and dialects, (3) the original and translated benchmarks can be considered also a machine translation benchmark, both between the dialect and the standard counterpart, as well as between the dialect and another language, and, finally, (4) if the benchmark was to be read to generate a spoken language understanding benchmark, aside from the new modality itself, we would also obtain benchmarks in speech to speech, but also text to speech and speech to text translation in quite many directions, the biggest novelty, again, being the dialectal feature of the benchmark.

The paper is structured as follows: firstly, in Section 2, we present the previous work on English COPA and its translations to other languages. Secondly, in Section 3, we present the developed datasets for South Slavic standard languages and dialects. We first introduce the COPA datasets for standard Slovenian, Croatian, Serbian and Macedonian languages in Subsection 3.1. Then we present the development of dialect datasets in Subsection 3.2, and mention the challenges we encountered in Subsection 3.3. We conclude this section with Subsection 3.4 where we provide an insight into the level of differences between the datasets in the standard and dialectal languages. Next, in Section 4, we apply instruction fine-tuned large language models to the South Slavic COPA standard and dialectal datasets to obtain initial insights on their capabilities on our target languages and dialects.

garian. In this specific instance, the speech of the Region of Jablanica near the town of Lebane was used, which is more similar to standard Serbian compared to the most typical Balkan Sprachbund varieties.

[2] https://github.com/clarinsi/benchich/tree/main/copa

Finally, we wrap up the paper with conclusions and suggestions for further work in Section 5.

## 2 Related Work

**English COPA** The Choice Of Plausible Alternatives (COPA) dataset (Roemmele et al., 2011) was first created in English to evaluate machine learning approaches to automated commonsense reasoning. The dataset consists of instances that contain three sentences: a premise and two possible hypotheses (alternatives), either stated as a cause or effect of the premise. Each instance has the manually-annotated label with the answer to the task of determining which of the two alternatives is more plausible than the other. The dataset was designed in such way that it necessitates the model to solve the task based on the acquired linguistic and world knowledge that is not explicitly present in the text. The dataset consists of 1000 instances of commonsense causality, split into 500 instances in training and development split (400:100) and 500 instances in the test split. The COPA dataset was first presented as an evaluation dataset in the shared task of the 6th International Workshop on Semantic Evaluation (SemEval 2012) (Gordon et al., 2012). A few years later, the usefulness of the COPA dataset was also recognized by the authors of the well-known benchmark for general-purpose natural language understanding SuperGLUE[3] (Wang et al., 2019) where COPA was selected as one of 8 included datasets. In addition to causal reasoning, supported by the COPA dataset, SuperGLUE includes question answering, textual entailment, co-reference resolution, and word sense disambiguation.

**COPA in Other Languages** The first efforts to use the COPA dataset for evaluation in other languages appeared almost 10 years after the development of the English dataset. Ponti et al. (2020) introduced the Cross-lingual Choice of Plausible Alternatives (XCOPA) dataset which includes translation of the development and test splits of the COPA dataset to 11 more languages that come from 11 distinct language families and 5 macroareas: Estonian, Haitian Creole, Indonesian, Italian, Eastern Apurímac Quechua, Kiswahili, Tamil, Thai, Turkish, Vietnamese and Mandarin Chinese. Translation of the COPA dataset was also fueled by its introduction to the SuperGLUE benchmark (Wang et al., 2019). The benchmark and the COPA dataset inside it were inter alia translated to Russian (Shavrina et al., 2020) and to Slovenian[4] (Žagar and Robnik-Šikonja, 2022). Recently, the COPA dataset was also translated to 18 Indic languages as part of the development of the natural language understanding (NLU) benchmark for Indic languages IndicXTREME (Doddapaneni et al., 2023), and to Estonian (Kuulmets et al., 2022), where low-cost alternatives to the XCOPA methodology were investigated. Namely, researchers machine-translated the dataset and then manually edited the automatic translation. In contrast, recent work by Wibowo et al. (2023) suggests a more detailed approach. Instead of translating the COPA dataset, they developed their own variant of the dataset with new instances that incorporate Indonesian local and cultural nuances, and thus provide a more natural portrayal of causal reasoning within the Indonesian culture. Interestingly, similarly to our approach, they prepare the COPA dataset both in Indonesian standard language as well as in its dialect – Jakartan Indonesian, which is a colloquial variety that is used in day-to-day conversations.

**COPA Modelling** At the first shared task that used the COPA dataset, commonsense reasoning was shown to be a very hard task for machine learning approaches (which were non-neural at the time) with the best methods achieving accuracy scores of 65.4%, only 15% higher than the random baseline (with accuracy of 50%) (Gordon et al., 2012). With the recent introduction of Transformer-based BERT-like pretrained language models, the task in English has shown to be much simpler for the models to grasp and on the SuperGLUE leaderboard, the state-of-the-art pretrained language models achieve an incredible accuracy higher than 99% (Chowdhery et al., 2023; Zhong et al., 2022). However, the introduction of the COPA datasets in other languages showed a large gap in natural language understanding capabilities between English and other languages. For Slovenian, Croatian, Indic languages and Indonesian, the best models among state-of-the-art multilingual and monolingual BERT-like pretrained language models only reach up to the accuracy between 61.8% and 65.8% (Ulčar and Robnik-Šikonja, 2021; Ljubešić and Lauc, 2021; Wibowo et al., 2023). While the BERT-like models seem not to be up to this challenging task, recently introduced instruction-tuned GPT-

[3] https://super.gluebenchmark.com/

[4] The Slovenian SuperGLUE dataset is available at http://hdl.handle.net/11356/1380

**English**: The girl found a bug in her cereal. She lost her appetite.
**Slovenian**: Dekle je v kosmičih našlo žuželko. Izgubila je apetit.
**Cerkno dialect**: Zjala je najdla hruošče u kosmičih. Zgubila je apetit.
**Croatian**: Djevojka je pronašla kukca u žitaricama. Izgubila je apetit.
**Chakavian dialect**: Mlada je našla neko blago va žitaricah. Je zgubila tiek.
**Serbian**: Девојчица је пронашла бубу у житарицама. Изгубила је апетит.
**Serbian (transliterated)**: Devojčica je pronašla bubu u žitaricama. Izgubila je apetit.
**Torlak dialect:** Девојчица нашла бубаљку међу њојне житарице. Изгубила си апетит.
**Torlak dialect (transliterated):** Devojčica našla bubaljku među njojne žitarice. Izgubila si apetit.
**Macedonian**: Девојката пронајде бубачка во нејзините житарки. Изгуби апетит.
**Macedonian (transliterated)**: Devojkata pronajde bubačka vo nejzinite žitarki. Izgubi apetit.

Figure 1: Example of a premise and a hypothesis from the COPA datasets in English, Slovenian, Cerkno dialect, Croatian, Chakavian dialect, Serbian, Torlak dialect, and Macedonian.

like models showed impressive capabilities also on non-English COPA datasets. Wibowo et al. (2023) evaluated the GPT-4 model (OpenAI, 2023), used with a 5-shot prompting strategy. Their model was reported to achieve incredible accuracy of 89.09% on standard Indonesian and 89.62% on Jakartan Indonesian.

## 3 South Slavic Standard and Dialect COPA

The newly presented COPA datasets have exactly the same content as the English COPA dataset (Roemmele et al., 2011), only the language is different. They consist of 400 training instances, 100 developmental instances and 500 test instances. Each instance consists of a premise (*The movie tickets sold out.*), a question (either *What was the cause?* or *What happened as the result?*) and two alternatives (*It was opening day for the movie.* and *The movie received poor reviews.*), where one is manually labelled to be more plausible than the other.

We first present the datasets of standard languages, namely Slovenian (Žagar et al., 2020), Croatian (Ljubešić, 2021), Serbian (Ljubešić et al., 2022b) and Macedonian (Ljubešić et al., 2022a), followed by the newly developed dialectal datasets, namely those for the Cerkno dialect, the Chakavian dialect, and the Torlak dialect (Ljubešić et al., 2024).

### 3.1 COPA in Standard South Slavic Languages

Motivated by astounding performance achieved by the large language models (LLMs) on other languages, the COPA datasets were translated for benchmarking the performance of LLMs on four standard South Slavic languages: Slovenian, Croatian, Serbian and Macedonian, resulting in the Slovenian COPA dataset as part of the SuperGLUE translation (Žagar et al., 2020), COPA-HR (Ljubešić, 2021), COPA-SR (Ljubešić et al., 2022b), and COPA-MK (Ljubešić et al., 2022a) datasets. While the Slovenian and Croatian datasets use the Latin script, Serbian and Macedonian use the Cyrillic script. Important to note here is that Serbian is a digraphic language, using the Cyrillic and the Latin script interchangeably, while Macedonian uses the Cyrillic script, but still has a transliteration technique into the Latin script that is occasionally used, especially in online communication. While translating the COPA-HR, the COPA-SR and the COPA-MK datasets, the methodology and guidelines laid out by the XCOPA authors were followed (Ponti et al., 2020), while the Slovenian version of the dataset was translated with less stringent rules. For the Croatian, Serbian and Macedonian dataset, each dataset was translated by one native speaker. Prior to the translation, the translators labelled the instances by choosing the most probable alternative for each premise. This step was not performed during the translation of the Slovenian dataset. The observed agreement of the English annotator and the Croatian translator was

perfect on the training and the validation dataset, with one different label (agreement of 99.8%) on the test dataset. In other cases – COPA-MK and COPA-SR – the translator had perfect agreement with the English gold labels. In contrast to the XCOPA, where only the test and development split were translated, the four South Slavic languages have the training split translated as well. While this necessitates more translation effort, it extends the usability of the datasets and enables research also on fine-tuning language models on the South Slavic languages, not only evaluation of their zero-shot capabilities.

### 3.2 COPA in South Slavic Dialects

In this work we extend the efforts of translating COPA to South Slavic languages by providing the first datasets that allow evaluation of the natural language understanding capabilities of large language models on South Slavic dialects. More precisely, we focus on the following dialects: the Cerkno dialect of Slovenian, spoken in the Slovenian Littoral region, specifically from the town of Idrija; the Chakavian dialect of Croatian from northern Adriatic, specifically from the town of Žminj; and the Torlak dialect from southeastern Serbia, specifically from the town of Lebane. As with the standard languages, we follow the same methodology and translation guidelines as proposed by the XCOPA dataset authors (Ponti et al., 2020). Each dialect was translated by one carefully selected translator who is a native speaker of the dialect. A novelty in this approach is that both English and standard South Slavic language were at disposal to the translator during the translation process. The training and development splits of the resulting datasets in Cerkno (COPA-SL-CER), Chakavian (COPA-HR-CKM) and Torlak (COPA-SR-TOR) dialects are made freely available,[5] while the test data are shared only upon request to prevent the contamination of large language models and the resulting invalidity of the benchmark measurements due to a possibility that the future large language models would use these data during pretraining. In Figure 1, we show an example of a premise and a hypothesis from the newly developed dialectal COPA datasets, as well as the standard language and the original English COPA (Roemmele et al., 2011) datasets. The Serbian, Torlak and Macedonian examples are, for readability purposes, represented both in the Latin and the Cyrillic script. While the Serbian (Ljubešić et al., 2022b) and Macedonian COPA datasets (Ljubešić et al., 2022a) have been published in the Cyrillic script, all three DIALECT-COPA datasets are published in the Latin script.

[5]The datasets can be downloaded from the CLARIN.SI repository: http://hdl.handle.net/11356/1766

### 3.3 Challenges with Adapting COPA to Dialects

**Spelling** When extending the COPA datasets to South Slavic dialects, we entered an uncharted territory regarding the development of benchmarks for these dialects, as they do not have a canonical spelling. Even within the dialect, some spelling variants depend on the speaker's preference (e.g., Slovenian standard word *voda* ("water") can be written in Cerkno Slovenian: *voda*, *uoda* or *woda*). Our main instruction to the translators was to translate in the manner they would consider communicating in writing with other speakers of that dialect.

**Grammar** One should note that sentence-level word order frequently differs between written standard South Slavic languages and the written dialectal text. While written language tends to follow topic-comment sequence (organizing information from known to new and emphasizing the sentence-final element), dialectal written language relies on an order closer to the spoken form, and has therefore a looser order. While translators strove to provide authentic translations in their native dialect, they mentioned that this was difficult to achieve at times, as they found many sentences in the COPA dataset to sound somewhat inauthentic and artificial and become even more so when translated to a non-standard language.

**Difference between English and South Slavic grammar** Compared to the English original, Slavic languages express grammatical gender (feminine, masculine, neuter) and number (singular, plural; and dual in the case of Slovenian). The translators strove to provide a balanced representation of all grammatical genders and numbers in examples when no such information can be gleaned from the English original.

### 3.4 Quantitative Analysis of Datasets

A first insight in the level of difference between the standard language and corresponding dialectal dataset is obtained by performing a series of character- and word-level comparisons, presented in Table 1. We first measure the average character and word similarity between each dialectal dataset

| standard | dialect | char | word | top |
|---|---|---|---|---|
| Slovenian | Cerkno | 0.647 | 0.293 | 24 |
| Croatian | Chakavian | 0.613 | 0.297 | 28 |
| Serbian | Torlak | 0.698 | 0.376 | 39 |

Table 1: Similarity between the standard and dialectal datasets calculated as average Levenshtein ratio of sentence pairs on level of characters (`char`) and the level of words (`word`), as well as the size of the intersection of the 100 most frequent words in the standard and dialectal dataset (`top`).

and its closest standard dataset via the Levenshtein ratio metric. Based on these two measurements one can see that the Torlak dialect is much more similar to Serbian, its corresponding standard language, than the Cerkno and Chakavian dialects to Slovenian and Croatian, respectively. If we compare the latter two dialects based on the level of similarity to their corresponding standard language, we see that while the Cerkno dialect is more similar to the Slovenian on the character level, the Chakavian dialect is more similar to the Croatian on the word level.

We perform a final measurement of similarity that focuses on the most frequent words, which includes most function words. We calculate the size of the intersection of sets of 100 most frequent words in the standard dataset and the dialectal dataset. The results of this measurement show again for the Torlak dialect to be the closest to its standard counterpart, but this time the Cerkno dialect being less similar to the standard than the Chakavian dialect.

The goal of these measurements is to inform the dataset users of the varying distance between the three dialectal datasets when comparing to their standard variant. We expect the research community to use these datasets in more in-depth analyses of the dialects and their corresponding standard varieties.

## 4 Baselines

In this section we present baseline results of currently best-performing open and closed instruction-tuned GPT-like large language models. For the open model (downloadable weights) we select the `Mixtral-8x7B-Instruct-v0.1`[6] model (Jiang et al., 2024), while among the closed (API access only) models we opt for the `gpt-4-0125-preview` model (OpenAI, 2023). The selection of models is based on best results obtained during preliminary experimentation across models available at the time of the writing. We use instruction-tuned models so that we can follow a uniform extraction of answers from each model via a unified prompt. The prompts used are presented in Appendix A. They were selected during preliminary experiments, showing comparable and consistent results across all models and datasets.

We perform experiments in a zero-shot and 10-shot fashion on the training portions of datasets of both the standard languages (including English) and the dialects. In both cases we use the models "off-the-shelf", without any additional fine-tuning. In the zero-shot scenario, the prompt only includes the definition of the task and the instance for which we require a label, while in the 10-shot scenario, we also provide the first ten instances from the development split with the correct answers. We opt to use the training data as our evaluation data in the baseline experiments due to the closed nature of our dialectal test data. Using test data in these experiments would significantly reduce the replicability of our results, as test data are only available upon request.

The baseline experiments showed the following. There is a significant gap between performance of models on standard languages and dialects. The Cerkno dialect proves to be by far the most challenging one, followed by the Chakavian dialect, while the Torlak dialect performs most similarly to its standardized variety. The differences in performance on dialects roughly follow the character and word similarities between the standard and the dialectal dataset, presented in Section 3.4.

The comparison of the performance of the two models shows that the closed GPT-4 model (OpenAI, 2023) is significantly more potent than the open Mixtral model (Jiang et al., 2024). Interestingly, few-shot learning significantly improves the results, especially with the hardest cases of Chakavian and Cerkno dialects and the most potent GPT-4 model, where Chakavian achieves improvement of 9 points, while Cerkno dialect achieves improvement of 14 points.

For the improvements obtained with 10-shot prompting, the main question arises whether the improvement is due to the model learning about the task itself or about the language/dialect that the model is being tested on. Additional research will be required to disentangle these two likely effects.

[6] https://huggingface.co/mistralai/Mixtral-8x7B-Instruct-v0.1

| model | n-shot | EN | SL | SL-CER | HR | HR-CKM | SR | SR-TOR | MK |
|---|---|---|---|---|---|---|---|---|---|
| Mixtral | 0 | 0.875 | 0.683 | 0.405 | 0.705 | 0.580 | 0.713 | 0.638 | 0.665 |
| Mixtral | 10 | 0.933 | 0.803 | 0.500 | 0.818 | 0.603 | 0.795 | 0.748 | 0.703 |
| GPT-4 | 0 | 0.988 | 0.960 | 0.595 | 0.963 | 0.778 | 0.968 | 0.925 | 0.945 |
| GPT-4 | 10 | **0.995** | **0.980** | **0.738** | **0.988** | **0.870** | **0.990** | **0.968** | **0.978** |

Table 2: Accuracy achieved on the training split of COPA for different models, prompting fashions (zero-shot vs 10-shot scenario), and languages and dialects. The languages and dialects presented are: English (en), Slovenian (sl), Cerkno Slovenian (sl-cer), Croatian (hr), Chakavian Croatian (hr-ckm), Serbian (sr), Serbian Torlak (sr-tor) and Macedonian (mk).

At a recent shared task based on this dataset (Chifu et al., 2024) the power of adaptation of large language models to dialects via in-context learning has been demonstrated by multiple teams, while one team has shed some light on the impact of the task semantics and the dialect semantics (Ljubešić et al., 2024), showing that both are useful, but that most improvement is coming from the side of dialect semantics.

## 5 Conclusions

This paper introduced DIALECT-COPA – a dataset for commonsense reasoning covering three South Slavic dialects, an extension of the already available translations into their respective standard varieties. The commonsense reasoning benchmark is based on the popular Choice of Plausible Alternatives (COPA) English dataset. The datasets of both dialects and standard languages were translated by native dialect speakers from the original English COPA dataset (Roemmele et al., 2011). During the translation process into each dialect, the translator also had access to the translation into the closest standard variety so that the dialectal translations exhibit a minimum of translation artifacts when compared to the standard translation.

The dialects covered are the Cerkno dialect of the Slovenian language, the Chakavian dialect of the Croatian language, and the Torlak dialect of the Serbian language. Together with the dataset, we also perform experiments on the translations of the COPA dataset into all standard South Slavic languages that are related to the evaluated dialects except Bulgarian. Such data setup enables precise measurements of the differences in performance between standard languages and dialects, but also potential transfer learning opportunities between the standard and dialect varieties.

A quantitative comparison of the dialectal datasets with their standard language counterparts shows a varying level of character-level, word-level and lexicon-level deviation of dialectal text from the standard datasets, with the observed differences rather well reproduced in baseline zero-shot and 10-shot experiments. Namely, the Slovenian Cerkno dialect and the Croatian Chakavian dialect show significantly lower results than the Torlak dialect. This suggests that the idiolect of the translator into the Torlak dialect is closer to standard Serbian, which makes the dataset simply less challenging.

Besides the difference in performance gaps between dialects, the baseline results also show, very much expectedly, that performance on standard languages is significantly better than that on dialects. The open models show also to be, similar to comparable results on other benchmarks (Gao et al., 2023; OpenAI, 2023), less capable than the closed models available only through an API.

Rather good news for large language model adaptation to dialectal texts is that in-context 10-shot learning drastically improves the performance on the worst-performing dialects, with a 14-point performance improvement on the Cerkno dialect and a 9-point improvement on the Chakavian dialect. Part of the improvement in performance can be followed back to the model in-context learning about the task itself. Further analyses are required to obtain a more detailed insight to which level this impacts the results.

There are many additional future directions we plan to follow upon. One is measuring the human performance on the presented dialects given their linguistic background. Namely, some of the presented dialects are not easy to understand by most speakers of the related standard language. Another research direction is adding a speech component to these datasets, which opens up the dataset for spoken dialectal language understanding measurements, but also dialectal speech-to-speech and speech-to-text translation and generation.

Finally, we hope that this dataset will spark interest in constructing datasets of many more dialects

of well-resourced languages. While we can consider the standard of these languages to be well-resourced, there is a wealth of linguistic diversity that has still not been well covered.

## Acknowledgements

The research presented in this paper was conducted within the research project titled Basic Research for the Development of Spoken Language Resources and Speech Technologies for the Slovenian Language (J7-4642), the research programme "Language resources and technologies for Slovene" (P6-0411), and the CLARIN.SI research infrastructure, all funded by the Slovenian Research and Innovation Agency (ARIS).

## 6 Limitations

It is important to note that the regional language variants in the dialect COPA datasets should be interpreted only as one of the possible projections of dialects into written form, not as a single canonical version. Furthermore, while we refer to these datasets as dialect translations for simplicity, we are aware that this is not in line with the view of dialectologists where dialects are purely spoken variants. It should be thus put forward that our dialect translations are just an attempt at projecting dialectal speech into a semi-canonical written form. To bridge these limitations, we are planning on creating a speech audio dataset where the native speakers would read out the COPA instances. This would provide a truer representation of dialects and also open a new front of evaluation of language models on speech COPA datasets.

## References

Adrian Chifu, Goran Glavaš, Radu Ionescu, Nikola Ljubešić, Aleksandra Miletić, Filip Miletić, Yves Scherrer, and Ivan Vulić. 2024. VarDial evaluation campaign 2024: Commonsense reasoning in dialects and multi-label similar language identification. In *Eleventh Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2024)*, Mexico City, Mexico. Association for Computational Linguistics.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

Sumanth Doddapaneni, Rahul Aralikatte, Gowtham Ramesh, Shreya Goyal, Mitesh M Khapra, Anoop Kunchukuttan, and Pratyush Kumar. 2023. Towards Leaving No Indic Language Behind: Building Monolingual Corpora, Benchmark and Models for Indic Languages. In *The 61st Annual Meeting Of The Association For Computational Linguistics*.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2023. A framework for few-shot language model evaluation.

Andrew Gordon, Zornitsa Kozareva, and Melissa Roemmele. 2012. Semeval-2012 task 7: Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In ** SEM 2012: The First Joint Conference on Lexical and Computational Semantics–Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 394–398.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Mate Kapović. 2017. The position of Kajkavian in the South Slavic dialect continuum in light of old accentual isoglosses. *Zeitschrift für Slawistik*, 62(4):606–620.

Hele-Andra Kuulmets, Andre Tattar, and Mark Fishel. 2022. Estonian Language Understanding: a Case Study on the COPA Task. *Baltic Journal of Modern Computing*, 10(3):470–480.

Nikola Ljubešić. 2021. Choice of plausible alternatives dataset in Croatian COPA-HR. Slovenian language resource repository CLARIN.SI.

Nikola Ljubešić, Boshko Koloski, Kristina Zdravkovska, and Taja Kuzman. 2022a. Choice of plausible alternatives dataset in Macedonian COPA-MK. Slovenian language resource repository CLARIN.SI.

Nikola Ljubešić, Taja Kuzman, Peter Rupnik, Stefan Milosavljević, Nada Galant, Sonja Benčina, and Jaka Čibej. 2024. "Choice of plausible alternatives" datasets in South Slavic dialects DIALECT-COPA. Slovenian language resource repository CLARIN.SI.

Nikola Ljubešić and Davor Lauc. 2021. BERTić-The Transformer Language Model for Bosnian, Croatian, Montenegrin and Serbian. In *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing*, pages 37–42.

Nikola Ljubešić, Mirjana Starović, Taja Kuzman, and Tanja Samardžić. 2022b. Choice of plausible alternatives dataset in Serbian COPA-SR. Slovenian language resource repository CLARIN.SI.

Nikola Ljubešić, Taja Kuzman, Peter Rupnik, Goran Glavaš, Fabian David Schmidt, and Ivan Vulić. 2024. JSI and WüNLP at the DIALECT-COPA Shared Task: In-Context Learning From Just a Few Dialectal Examples Gets You Quite Far . In *Eleventh Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2024)*, Mexico City, Mexico. Association for Computational Linguistics.

Stefan Milosavljević. 2018. Osobine klitičkog udvajanja u govoru jablaničkog kraja [Properties of clitic doubling in the speech of the Region of Jablanica]. In Miloš Kovačević, editor, *Savremena proučavanja jezika i književnosti: Zbornik radova sa IX naučnog skupa mladih filologa Srbije / Knjiga 1*, pages 41–52. FILUM, Kragujevac.

Olga Mišeska Tomić. 2006. *Balkan Sprachbund Morpho-Syntactic Features*. Springer, Dordrecht, The Netherlands.

Christopher Moseley. 2010. *Atlas of the World's Languages in Danger*. Unesco.

OpenAI. 2023. Gpt-4 technical report.

Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. XCOPA: A Multilingual Dataset for Causal Commonsense Reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2362–2376.

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S Gordon. 2011. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *2011 AAAI Spring Symposium Series*.

Tatiana Shavrina, Alena Fenogenova, Emelyanov Anton, Denis Shevelev, Ekaterina Artemova, Valentin Malykh, Vladislav Mikhailov, Maria Tikhonova, Andrey Chertok, and Andrey Evlampiev. 2020. RussianSuperGLUE: A Russian language understanding evaluation benchmark. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4717–4726, Online. Association for Computational Linguistics.

Matej Ulčar and Marko Robnik-Šikonja. 2021. SloBERTa: Slovene monolingual large pretrained masked language model. *Proceedings of Data Mining and Data Warehousing, SiKDD*.

Domagoj Vidović. 2007. Accentual alternations in Neo-Štokavian Ijekavian dialects of Neretvanska krajina. In *Tones and Theories: Proceedings of the International Workshop on Balto-Slavic Accentuation*, pages 199–211.

Teodora Vuković, Anastasia Escher, and Barbara Sonnenhauser. 2022. Degrees of non-standardness: Feature-based analysis of variation in a Torlak dialect corpus. *International Journal of Corpus Linguistics*, 27(2):220–247.

Jelena Živojinović. 2021. Torlak clitic doubling: A cross-linguistic comparison. In Andreas Blümel, Jovana Gajić, Ljudmila Geist, Uwe Junghanns, and Hagen Pitsch, editors, *Advances in Formal Slavic Linguistics 2018*, pages 423–441. Language Science Press, Berlin.

Radmila Žugić. 2005. Rečnik govora jablaničkog kraja. *Srpski dijalektološki zbornik*, LII:XI–XLII + 1–470.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems*, 32.

Haryo Akbarianto Wibowo, Erland Hilman Fuadi, Made Nindyatama Nityasya, Radityo Eko Prasojo, and Alham Fikri Aji. 2023. COPAL-ID: Indonesian language reasoning with local culture and nuances. *arXiv preprint arXiv:2311.01012*.

Aleš Žagar and Marko Robnik-Šikonja. 2022. Slovene SuperGLUE Benchmark: Translation and Evaluation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2058–2065.

Aleš Žagar, Marko Robnik-Šikonja, Teja Goli, and Špela Arhar Holdt. 2020. Slovene translation of SuperGLUE. Slovenian language resource repository CLARIN.SI.

Qihuang Zhong, Liang Ding, Yibing Zhan, Yu Qiao, Yonggang Wen, Li Shen, Juhua Liu, Baosheng Yu, Bo Du, Yixin Chen, et al. 2022. Toward efficient language model pretraining and downstream adaptation via self-evolution: A case study on superglue. *arXiv preprint arXiv:2212.01853*.

# A Appendix

**Zero-shot prompt** An example from the Slovenian Cerkno dataset.

*You will be given a task. The task definition is in English, but the task itself is in another language. Here is the task!*

*Given the premise "Muoje telu je metalu sinca na traua.", and that we are looking for the cause of this premise, which hypothesis is more plausible?*

*Hypothesis 1: "Sunce je šlu guor.".*

*Hypothesis 2: "Traua je bla pakuošena.".*

*Answer only with "1" or "2".*

*Answer:*

**Ten-shot prompt** An example from the Croatian Chakavian dataset.

*You will be given a task. The task definition is in English, but the task itself is in another language. You are to choose the more likely hypothesis given a premise. Take into account that we are either*

*looking for a cause or an effect of the premise. Answer only with "1" or "2". Here are some examples of the task:*

*Example 1:*

*Premise: "Muški je otpra špino."*

*Question: "effect"*

*Hypothesis 1: "Školjka ot zahoda se je napunila z oduon."*

*Hypothesis 2: "Oda je počela teć z mlaznici."*

*Answer: "2"*

*Example 2:*

*Premise: "Mlada je našla neko blago va žitaricah."*

*Question: "effect"*

*Hypothesis 1: "Nalila je mlieko va škudelico."*

*Hypothesis 2: "Je zgubila tiek."*

*Answer: "2"*

*Example 3:*

*...*

*Example 10:*

*Premise: "Šlovek je čuda popi na fešte."*

*Question: "effect"*

*Hypothesis 1: "Ta drugi dan ga je bolela glava."*

*Hypothesis 2: "Ta drugi dan mu je kapa nuos."*

*Answer: "1"*

*Now to your task!*

*Premise: "Moje tielo je hitalo hlat na travo."*

*Question: "cause"*

*Hypothesis 1: "Sunce je hodilo van."*

*Hypothesis 2: "Trava je bila pokošena."*

*Answer:*