VarDial 2024

VarDial 2024 - The Eleventh Workshop on NLP for Similar Languages, Varieties and Dialects

Proceedings of the Workshop

June 20, 2024

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
317 Sidney Baker St. S
Suite 400 - 134
Kerrville, TX 78028
USA
Tel: +1-855-225-1962
acl@aclweb.org

ISBN 979-8-89176-104-9

Preface

These proceedings include the 22 papers presented at the Eleventh Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2024), co-located with the 2024 annual conference of the North American Chapter of the Association for Computational Linguistics (NAACL). VarDial was held in Mexico City, Mexico, in a hybrid format, allowing participants to attend on-site or remotely.

Now at its eleventh edition, we are pleased to see that VarDial continues to serve the community as the main venue for researchers interested in the computational processing of diatopic language variation. The papers accepted this year address a wide range of topics, such as language variety identification, corpus creation, and machine translation. These proceedings once again illustrate the great linguistic diversity that VarDial embodies. They include work on Germanic (historical Dutch, Limburgish, Norwegian, Swiss German), Romance (Portuguese, Spanish, Occitan), and Slavic languages (South Slavic, Slovak), as well as Arabic and Nahuatl.

As in previous editions, VarDial 2024 features an evaluation campaign with two shared tasks: The DIALECT-COPA shared task on dialectal causal commonsense reasoning, and the DSL-ML shared task on multi-label classification of similar languages. Both tasks were organized for the first time this year, although DSL-ML relies on datasets built for earlier tasks. This volume includes the system description papers prepared by the participating teams, as well as a report written by the task organizers summarizing the results and findings of the evaluation campaign.

Finally, we would like to take this opportunity to thank all the shared task organizers and the participants for their hard work. We further thank the VarDial program committee members, and particularly the 14 PC members who newly joined this year, for being an important part of the workshop's success.

The VarDial workshop organizers:

Yves Scherrer, Tommi Jauhiainen, Nikola Ljubešić, Preslav Nakov, Jörg Tiedemann, and Marcos Zampieri

https://sites.google.com/view/vardial-2024

Organizing Committee

Organizers

Tommi Jauhiainen, University of Helsinki
Nikola Ljubešić, Jožef Stefan Institute
Preslav Nakov, Mohamed bin Zayed University of Artificial Intelligence
Yves Scherrer, University of Oslo
Jörg Tiedemann, University of Helsinki
Marcos Zampieri, George Mason University

Program Committee

Program Committee

Noëmi Aepli, University of Zurich
César Aguilar, Universidad Veracruzana
Sina Ahmadi, University of Zurich
Laura Alonso Alemany, Universidad Nacional de Cordoba
Delphine Bernhard, Lilpa, Université de Strasbourg
Gabriel Bernier-Colborne, National Research Council Canada
Verena Blaschke, LMU Munich
Aoife Cahill, Dataminr
David Chiang, University of Notre Dame
Adrian-Gabriel Chifu, Aix-Marseille Universite, Universite de Toulon, CNRS, LIS, Marseille, France
Steven Coats, University of Oulu
Jon Dehdari, Fidelity Investments
Stefanie Dipper, Ruhr University Bochum
Mark Dras, Macquarie University
Jonathan Dunn, University of Illinois Urbana-Champaign
Pablo Gamallo, CITIUS, University of Santiago de Compostela
Cyril Goutte, National Research Council Canada
Nizar Habash, New York University Abu Dhabi
Radu Tudor Ionescu, University of Bucharest
Anjali Kantharuban, University of California, Berkeley
Ekaterina Lapshinova-Koltunski, University of Hildesheim
Lung-Hao Lee, National Yang Ming Chiao Tung University
John P. McCrae, Insight Center for Data Analytics, National University of Ireland Galway
Aleksandra Miletić, Department of Digital Humanities, University of Helsinki
Filip Miletić, University of Stuttgart
Maciej Ogrodniczuk, Institute of Computer Science, Polish Academy of Sciences
Petya Osenova, Sofia University St. Kl. Ohridskiand IICT-BAS
Jelena Prokic, Leiden University
Christoph Purschke, University of Luxembourg
Francisco Manuel Rangel Pardo, Universitat Politècnica de València
Reinhard Rapp, University of Mainz
Tanja Samardžić, University of Zurich
Serge Sharoff, University of Leeds
Milena Slavcheva, Institute of Information and Communication Technologies, Bulgarian Academy of Sciences
Aarohi Srivastava, University of Notre Dame
Joel Tetreault, Dataminr
Rob Van Der Goot, IT University of Copenhagen
Pidong Wang, Google
Taro Watanabe, Nara Institute of Science and Technology
Çağrı Çöltekin, University of Tübingen

Table of Contents

VarDial Evaluation Campaign 2024: Commonsense Reasoning in Dialects and Multi-Label Similar Language Identification

Adrian-Gabriel Chifu[1], Goran Glavaš[2], Radu Tudor Ionescu[3], Nikola Ljubešić[4,5], Aleksandra Miletić[6], Filip Miletić[7], Yves Scherrer[6,8], Ivan Vulić[9]

[1]Aix-Marseille University, [2]University of Würzburg, [3]University of Bucharest, [4]Jožef Stefan Institute, [5]University of Ljubljana, [6]University of Helsinki, [7]University of Stuttgart, [8]University of Oslo, [9]University of Cambridge

Abstract

This report presents the results of the shared tasks organized as part of the VarDial Evaluation Campaign 2024. The campaign is part of the eleventh workshop on Natural Language Processing (NLP) for Similar Languages, Varieties and Dialects (VarDial), co-located with NAACL 2024. Two shared tasks were included this year: dialectal causal commonsense reasoning (DIALECT-COPA), and Multi-label classification of similar languages (DSL-ML). Both tasks were organized for the first time this year, but DSL-ML partially overlaps with the DSL-TL task organized in 2023.

1 Introduction

The workshop series on *NLP for Similar Languages, Varieties and Dialects* (VarDial), traditionally co-located with international conferences, has reached its eleventh edition. Since the first edition, VarDial has hosted shared tasks on various topics such as language and dialect identification, morphosyntactic tagging, question answering, and cross-lingual dependency parsing. The shared tasks have featured many languages and dialects from different families and data from various sources, genres, and domains (Aepli et al., 2023, 2022; Chakravarthi et al., 2021; Gaman et al., 2020; Zampieri et al., 2019, 2018, 2017; Malmasi et al., 2016; Zampieri et al., 2015, 2014).

As part of the VarDial Evaluation Campaign 2024, we offered two shared tasks which we present in this paper:

- **DIALECT-COPA:** Dialectal causal commonsense reasoning[1]
- **DSL-ML:** Multi-label classification of similar languages[2]

DSL-ML continues the long line of language and dialect identification (Jauhiainen et al., 2019) shared tasks at VarDial, whereas DIALECT-COPA features a task novel to the evaluation campaigns.

The evaluation campaign took place in January – March 2024. The call for participation and the training data sets for the shared tasks were published in the second half of January, and the results were due to be submitted on March 11th.[3]

In the following sections, the two tasks are discussed in detail, focusing on the data, the participants' approaches, and the obtained results. Section 2 is dedicated to DIALECT-COPA and Section 3 to DSL-ML.

2 The DIALECT-COPA Task on Causal Commonsense Reasoning

2.1 Motivation

The causal commonsense reasoning (CCR) task has been established as an important task in evaluation of natural language understanding (NLU) capabilities of pretrained language models, including the latest family of the so-called Large Language Models (LLMs). The original English dataset, Choice Of Plausible Alternatives (COPA) (Roemmele et al., 2011) has been used as the standard evaluation benchmark for the English CCR task since its release, and it is also included in the English SuperGLUE benchmark (Wang et al., 2019).

Language-specific variants of COPA have also been created, where the bulk of the data is covered in the multilingual XCOPA dataset (Ponti et al., 2020). The original XCOPA covers 11 standard language varieties from 11 language families, including some lower-resource languages such as Haitian Creole, Tamil, and Southern Quechua. It has been included into the established XTREME-R benchmark (Ruder et al., 2021) for the evalua-

[1]Task organizers: Nikola Ljubešić, Ivan Vulić, Goran Glavaš.

[2]Task organizers: Adrian Chifu, Radu Ionescu, Aleksandra Miletić, Filip Miletić, Yves Scherrer.

[3]https://sites.google.com/view/vardial-2024/shared-tasks

Proceedings of the Eleventh Workshop on NLP for Similar Languages, Varieties, and Dialects (VarDial 2024), pages 1–15
June 20, 2024

tion of cross-lingual transfer, and has consequently been used as a de facto evaluation benchmark for CCR in cross-lingual and multilingual scenarios. Besides XCOPA, there also exist single-language translations of adaptations of COPA into other languages such as Slovenian (Žagar and Robnik-Šikonja, 2022), Russian (Shavrina et al., 2020), and Catalan,[4] among others.

While COPA and XCOPA were considered challenging benchmarks for previous encoder-style models such as BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019) and XLM-R (Conneau et al., 2020), current state-of-the-art LLMs now provide impressive performance on these datasets (Chowdhery et al., 2023; Zhong et al., 2022; Shi et al., 2023): they are able to reach $\geq 90\%$ accuracy for diverse languages such as Thai, Estonian, Indonesian, Tamil, Vietnamese or Turkish (Shi et al., 2023). Whereas LLMs have been proven to perform extremely well on high-resource and even moderately resourced standard languages, their ability to conduct CCR for truly low-resource languages (Senel et al., 2024) and *especially dialects* (Joshi et al., 2024) has been much less investigated and empirically measured. For instance, lower performance on the standard lower-resource languages of the XCOPA dataset (e.g., Haitian Creole, Quechua, Swahili) already indicates additional difficulty for and reduced capability of current LLMs.

All COPA datasets to date comprise the same set of instances covering the same or similar set of topics. The only core difference between different datasets is the actual, target language variety of a particular dataset. Another property of COPA and its derivatives is its simple and easy-to-evaluate data format. In a nutshell, each data instance consists of three sentences: a statement (*premise*) and two possible *effects* or *causes* (termed *alternatives*) for the premise. Given an English example, a premise *'The man turned on the faucet.'* is combined with two alternatives *'The toilet filled with water.'* and *'Water flowed from the spout'*. The task is then to select the alternative that more plausibly has a causal relation with the premise, where each instance is manually annotated with a correct answer. The standard evaluation measure is accuracy, where the random baseline is therefore at 50% accuracy, and errors made by the systems could be due to subtle details related to understanding causality relationships.

The above background related to CCR in general and COPA-style datasets in particular has motivated us to create a first shared task on CCR for *dialectal data*, DIALECT-COPA, which we discuss next. In summary, the selection of the task has been guided by the following observations and criteria:

- CCR is an established and important NLU task for the evaluation of language models in monolingual, multilingual, and cross-lingual setups;
- CCR has never been in focus of VarDial evaluation campaigns and, vice versa, there have been no attempts to date to extend the CCR task and the corresponding COPA-style data to non-standard language varieties and dialects;
- CCR based on the standard COPA data format offers an excellent balance between the structural simplicity and semantic complexity of the task, with clear and straightforward evaluation protocols and measures.
- The standardized COPA format and the multi-parallel nature of COPA-based datasets in different standard language varieties combined with newly created dialectal COPA variants offer ample opportunity for cross-linguistic and cross-dialectal analyses and studies of model behavior and performance, as part of the shared task as well as for future research.
- For dialects chosen for DIALECT-COPA, obtaining large quantities of raw text is typically not possible, which renders good out-of-the-box performance of LLMs for them difficult and unlikely; this calls for new and creative approaches in order to mitigate the current gaps of LLMs when faced with CCR on dialectal data.

2.2 Data

The focus of the first DIALECT-COPA shared task has been on *micro-dialects* of several South-Slavic languages. This choice has been partially motivated by the recent creation of COPA datasets for standard language varieties of several, moderately resourced in NLP terms, South-Slavic languages: Slovenian COPA-SL (Žagar and Robnik-Šikonja, 2022), Croatian COPA-HR (Ljubešić, 2021), Serbian COPA-SR (Ljubešić et al., 2022b) and Macedonian COPA-MK (Ljubešić et al., 2022a). All the datasets were translated by human translators,

[4] https://huggingface.co/datasets/projecte-aina/COPA-ca

native speakers of the target languages, from the English COPA dataset (Roemmele et al., 2011), with all the datasets, except for COPA-SL, following the XCOPA translation and adaptation methodology (Ponti et al., 2020). COPA-SL was translated without any additional adaptation as part of the Slovenian SuperGLUE benchmark (Žagar and Robnik-Šikonja, 2022). Serbian and Macedonian datasets are written in Cyrillic, while the other data are in the Latin script.

For the shared task, the COPA-* datasets in the standard South-Slavic languages were then extended to three micro-dialects that are spoken in narrow micro-geographical areas (Ljubešić et al., 2024a): **1)** the Cerkno dialect of Slovenian (COPA-SL-CER), spoken in the Slovenian Littoral region, specifically from the town of Idrija; **2)** the Chakavian dialect of Croatian from northern Adriatic (COPA-HR-CKM), specifically from the town of Žminj, and **3)** the Torlak dialect from southeastern Serbia (COPA-SR-TOR), specifically from the town of Lebane in Serbia.

The three dialectal datasets featuring in the DIALECT-COPA task were again created following the established translation and adaptation methodology of XCOPA. All data instances were translated and adapted from the closest standard language COPA (e.g., COPA-HR was used to derive COPA-HR-CKM), allowing the human translators to also consult the original English COPA as the additional source. Following the original COPA data split, COPA-SL-CER and COPA-SR-TOR contain 400 instances for training, 100 for development and 500 test instances. COPA-HR-CKM was treated as a surprise dialect, and it comprises only the 500 translated and adapted test instances. We allowed the use of any external data except the 500 test instances in any language for which a COPA dataset variant exists,[5] given the multi-parallel nature of the COPA datasets.

While the contamination of todays' LLMs with the English COPA dataset is very likely, we are rather sure that there is a minimum danger of the results of this shared task to be contaminated, and this is for the following reasons: (1) the dialectal datasets were not published before this shared task, (2) inspections of performance of various recent LLMs has shown not-perfect results on the English dataset, and (3) comparable results to the English ones were achieved on the non-English datasets, that are available for a short period of time. Finally, to ensure future validity of the measurements on this shared task's data, the test data of the DIALECT-COPA dataset are not published publicly, but are available only upon request of fellow researchers.

The evaluation metric regularly used in the COPA datasets, as well as inside this shared task, is accuracy, which puts the random baseline, given the binary nature of the task, at 50%. Ljubešić et al. (2024a) propose already competitive baselines, with Mixtral 8x7B Instruct (Jiang et al., 2024) zero-shotting achieving results around 70% accuracy on standard South Slavic datasets, but random to 63% accuracy on the dialectal datasets. Similarly, with zero-shotting the GPT-4 model (OpenAI et al., 2024), results of around 95% accuracy are reported for the standard South Slavic datasets, while the dialectal datasets achieve results between 60% and 93%. The significantly lower results on dialectal datasets, regardless of the model applied, show for the DIALECT-COPA dataset to be a very much open challenge and therefore a great fit for this evaluation campaign.

2.3 Participants

gmu-nlp. The team from the George Mason University submitted 10 runs, which is the maximum number of allowed runs in the shared task. Their approach (Faisal and Anastasopoulos, 2024) primarily focused on adaptation to dialects through various techniques of data augmentation: namely transforming *cause* instances into *effect* instances (and vice versa) by switching the place of the premise and the correct hypothesis, generating the non-available Chakavian training data by translating the standard data into the dialect via the the Claude 3 (Anthropic, 2024) and GPT-4 (OpenAI et al., 2024) models prompted with dialectal translation examples and rules, and fine-tuning a model on a combination of training data from specific languages and dialects. They inspected two models: the smaller Electra-style BERTić model (Ljubešić and Lauc, 2021), and the mT5-based aya-101 model (Üstün et al., 2024). The authors also used the 'trick' of independently fine-tuning a *cause* and an *effect* model.

JSI. The team from the Jožef Stefan Insitute submitted six runs, all based on zero- and few-shotting the Mixtral 8×7B Instruct model (Jiang et al., 2024)

[5]This of course refers to all the other 'COPA languages' beyond the South Slavic languages, e.g., all the XCOPA languages, Russian, and Catalan

team	run	name	API-only	adapt	sl-cer	hr-ckm	sr-tor	mean
gmu-nlp	1	orgl_hr_ckm_test	N	FT	0.700	0.750	0.824	0.758
gmu-nlp	2	aya	N	FS	0.694	0.756	0.84	0.763
gmu-nlp	3	orglc_omix_mk_hr_ckm_test	N	FT	0.690	0.756	0.836	0.761
gmu-nlp	4	orgl_sl_cer_test	N	FT	0.686	0.718	0.836	0.747
gmu-nlp	5	orgl_test	N	FT	0.682	0.760	0.824	0.755
gmu-nlp	6	orgl_mk_hr_ckm_test	N	FT	0.660	0.742	0.848	0.750
gmu-nlp	7	orgl_mk_hr_ckm	N	FT	0.582	0.634	0.682	0.633
gmu-nlp	8	all_train_rev_genx_omixmatch_select	N	FT	0.576	0.622	0.692	0.630
gmu-nlp	9	orgl_mk_hr_ckm_10	N	FT	0.572	0.626	0.722	0.640
gmu-nlp	10	orgl_10	N	FT	0.540	0.622	0.700	0.621
JSI	1	gpt4-zero	Y	ZS	0.594	0.754	0.908	0.752
JSI	2	gpt4-task	Y	FS	0.734	0.890	0.974	0.866
JSI	3	gpt4-list	Y	FS	0.696	0.846	0.946	0.829
JSI	4	mixtral-zero	N	ZS	0.518	0.576	0.706	0.600
JSI	5	mixtral-task	N	FS	0.542	0.640	0.724	0.635
JSI	6	mixtral-list	N	FS	0.578	0.618	0.722	0.639
WueNLP	1	MixtralLoRA-en-last	N	FT	0.562	0.626	0.714	0.634
WueNLP	2	MixtralLoRA-en-val	N	FT	0.574	0.620	0.706	0.633
WueNLP	3	MixtralLoRA-x-last	N	FT	0.556	0.606	0.738	0.633
WueNLP	4	MixtralLoRA-x-val	N	FT	0.550	0.608	0.738	0.632
UNIRI	1	RAG_simple_1	Y	ZS	0.688	0.760	-	-
UNIRI	2	simple_1	Y	ZS	0.664	0.774	0.894	0.777
UNIRI	3	RAG_with_reasoning_1	Y	ZS	0.708	0.764	-	-
UNIRI	4	with_reasoning_1	Y	ZS	0.608	0.664	0.806	0.693

Table 1: Official results on the DIALECT-COPA shared task. The evaluation metric is accuracy, with a random baseline of 0.5. The *API-only* column encodes whether the system is based on a closed model, available only through API calls or not. The *adapt* column categorizes the system adaptations whether they are based on fine-tuning (FT), few-shot (FS) or zero-shot (ZS) approaches.

and the GPT-4 model (OpenAI et al., 2024), the few-shotting approach exploiting their finding that correct answers are not crucial for the in-context learning of the dialect, and that the first N test instances, where correct answers are not given, can easily be exploited for that task, with great enhancements in results (Ljubešić et al., 2024b). The team also investigated a plethora of other models, the two selected models being by far the best performing in the group of open-source models (Mixtral 8x7B) and closed-source models (GPT-4).

WueNLP. The team from the University of Würzburg submitted four runs, all being focused on LoRA-fine-tuning the Mixtral 8×7B Instruct model (Jiang et al., 2024) either on English or on standard language data, following upon the logic that dialectal data might not be available for fine-tuning the model (Ljubešić et al., 2024b). The team regularly fine-tuned the model on the training subset only, keeping the development data for selecting the checkpoint with the best results.

UNIRI. The team from the University of Rijeka submitted four runs, all exploiting the GPT-4 model, the basic zero-shot approach being extended with a step-by-step-reasoning prompt and a retrieval-augmented-generation-based use of dialectal lexicons (Perak et al., 2024). The dialectal lexicons, available for two out of the three dialects in question, have previously been extended with examples generated by GPT-4.

2.4 Results

The official results of the four teams that have submitted their system descriptions are given in Table 1. The first observation to be made is that all of the runs on all of the systems have beaten the random baseline of 50% accuracy.

Starting with the *gmu-nlp* team, their results show an expected improvement in results when the aya-101 model is employed (runs 1-6) in comparison to the smaller BERTić model (runs 7-10). While the team provides very interesting approaches to data augmentation, the second run, based only on few-shotting the aya model, achieves very competitive results to the remaining runs employing the same model, but relying on LoRA-fine-tuning on various combinations and enhancements of the training data. Important to note is that the *gmu-nlp* team provided the best results overall when an open-source backbone LLM is used.

Moving on to the *JSI* team, they have reached the best results overall, but with the API-only closed-source GPT-4 model. They propose a simple zero-shot prompt, and two improvements of that prompt, both exploiting the first 10 instances from the test set. While the *list* prompt only gives exemplary sentences of the target dialect, the *task* prompt contains the structure and the goal of the task, but without an answer given. Both 10-shot prompts improve the zero-shot approach significantly, the *list* prompt being inferior to the *task* prompt, showing that, while learning about the dialect in-context is the biggest source of improvement, learning about the task itself does help further.

The *WueNLP* team, exploiting LoRA-based fine-tuning of Mixtral, obtained very similar results to those few-shot results of the *JSI* team. This shows that fine-tuning an LLM on 400 training instances on the specific task, either on English data (runs 1 and 2), or on the standard language data closest to the target dialect (runs 3 and 4), is equivalent to in-context learning from 10 instances in the target dialect (*JSI* team runs 5 and 6), even if the task itself (*JSI* team run 6), or an answer (*JSI* team run 5), are not provided. Interestingly, there is no difference in the results regardless of whether the English or the standard-variety training data are used for fine-tuning, showing that fine-tuning successfully informs the model of the task (the results are three points better than *JSI* team run 4 - Mixtral zero-shot results), but not of the final dialect.

Finally, the *UNIRI* team exploits, similarly to the *JSI* team, the GPT-4 model, but obtains better results on simple zero-shotting (*UNIRI* team run 2 vs. *JSI* team run 1), quite likely due to a better stated prompt, starting with *This is a reasoning task*. Where *UNIRI* do not improve is with the step-by-step-reasoning prompt, which lowers all their results (run 4). Interestingly enough, the step-by-step-reasoning prompt improves their results on standard languages (reported in their paper), showing that even GPT-4 is challenged by reasoning in a dialect to a level where the step-by-step-reasoning requirement hurts the performance. Interestingly, the retrieval-augmented-generation approach of *UNIRI* does help on the Slovenian Cerkno dialect, but slightly hurts the performance on the Chakavian dialect. A potential reason is that the overall performance on the Cerkno dialect is lower: therefore, the additional lexical information is more helpful than in the case with the Chakavian dialect.

2.5 Conclusions

The overall conclusions that can be drawn from the results of the DIALECT-COPA task are the following. First, there is a large dialectal gap present, given the difference between the results reported on the standard datasets and the dialectal datasets. Second, open-source models do not perform as well as the closed API-based models; however, few-shot or fine-tuned open models achieve the level of performance of zero-shot closed models. Third, data augmentation or retrieval-augmented-generation through dialectal lexicons seems to be as efficient as simply in-context learning from a few dialectal examples. Finally, the highly-efficient in-context learning seems to benefit mostly from the additional information on the dialect to be processed, rather than on the task itself.

3 The DSL-ML Task on Multi-Label Similar Language Identification

3.1 Motivation

VarDial has run shared tasks on the topic of discriminating between similar languages and varieties since its first edition. The DSL shared tasks organized from 2014 to 2017 focused on languages with several varieties like English, Spanish, Portuguese, and BCMS (Bosnian, Croatian, Montenegrin, Serbian) (Zampieri et al., 2017; Malmasi et al., 2016; Zampieri et al., 2015, 2014). These tasks were based on the DSL Corpus Collection (DSLCC Tan et al., 2014),[6] a collection of journalistic texts compiled assuming that each instance's variety label is determined by where the text is retrieved from. Previous research (e.g. Goutte et al., 2016) has shown the limitations of this problem formulation, as some texts (especially short texts such as single sentences) may not contain any linguistic marker that would allow systems, or even native speakers, to discriminate between two similar language varieties. In the past years, several proposals were made to address this issue:

- The DSL-TL dataset (Zampieri et al., 2023), introduced in conjunction with a shared task

[6] http://ttg.uni-saarland.de/resources/DSLCC/

	English	Portuguese	Spanish	French	BCMS
Number of varieties	2 (UK, US)	2 (PT, BR)	2 (AR, ES)	4 (BE, CA, CH, FR)	4 (BS, HR, ME, SR)
Annotation	Human	Human	Human	Automatic	Human
Train labeling	Multi-label	Multi-label	Multi-label	Multi-label	Single-label
Dev labeling	Multi-label	Multi-label	Multi-label	Multi-label	Multi-label
Test labeling	Multi-label	Multi-label	Multi-label	Single-label	Multi-label
Named entities	Present	Present	Present	Masked	Present
Avg. tokens/instance in *train*	33	38	52	64	5548
Training instances	2097	3467	3467	340,363	368
Multi-label instances in *dev*	13%	14%	32%	0.7%	20%

Table 2: Key properties of the datasets used in the DSL-ML task.

at VarDial 2023 (Aepli et al., 2023), contains Spanish, Portuguese and English sentences that were manually annotated using crowd-sourcing. The annotation setup is restricted to two varieties per language (e.g. Peninsular and Argentinian Spanish), but allows a third option "Both or neither" if the instance does not provide sufficient grounds for reliable classification.

- Bernier-Colborne et al. (2023) argue that language variety identification is best framed as a multi-label classification problem. They analyze the FreCDo corpus (Găman et al., 2023) used in the VarDial 2022 FDI shared task (Aepli et al., 2022) and find substantial amounts of near-duplicate sentences associated with different labels in FreCDo. This near-duplicate analysis allows them to automatically derive a variant of FreCDo where ambiguous instances are annotated with multiple labels.

- Keleg and Magdy (2023) analyze different datasets used for Arabic dialect identification and find that many of the analyzed samples are valid in multiple dialects. As a result, the performance of dialect identification models is underestimated, as about two thirds of false positives are actually not true errors. Like Bernier-Colborne et al. (2023), they recommend multi-label annotations as a solution for future dialect identification tasks.

- Miletić and Miletić (2024) propose a reannotation of a single-annotator, single-label dataset for BCMS based on Twitter data (Rupnik et al., 2023). They explicitly introduce multi-label annotation based on labels produced by multiple annotators from all target regions. A re-evaluation of a previously proposed DSL system (Rupnik et al., 2023) against the multi-label annotation shows an improvement of the accuracy assessment (+4.1 points), indicating that some of the model predictions that were considered as wrong in the single-label setting are not necessarily errors. These results further support the multi-label annotation for the DSL task.

3.2 Data

The DSL-ML task is based on three data sources from five different languages. The choice of languages was mainly motivated by the availability of existing multi-label-annotated datasets. The five datasets have rather distinct properties in terms of size, instance lengths, genre, annotation and preprocessing. Table 2 summarizes these differences across the datasets (detailed statistics are provided in Table 3 in the appendix). For this reason, we provide distinct datasets for the five languages and evaluate the participants' submissions separately on each of them.

English, Portuguese, Spanish. For these languages, we re-use the DSL-TL dataset with the same split as in the VarDial 2023 task. We merely transform the "neither/both" labels to a comma-separated list of variant annotations. For example, the generic label `ES` becomes `ES-ES,ES-AR`.

French. The French training and development sets are obtained by combining the FreCDo (Găman et al., 2023) and DSLCC v4 (Tan et al., 2014) datasets, which comprise French (`FR-FR`), Swiss (`FR-CH`), Belgian (`FR-BE`), and Canadian (`FR-CA`) samples of text collected from the news domain. The topics used to collect most of the training and development data are available in the FreCDo paper. For the test data, we choose a new set of

Language	Label	Training # Samples	Training # Tokens	Development # Samples	Development # Tokens	Test # Samples	Test # Tokens
English	EN-GB	755	21,011	211	5,767	114	3068
	EN-GB, EN-US	273	8,686	76	2,409	30	978
	EN-US	1,069	49,761	312	12,380	156	6352
	Total	2097	79,458	599	20,556	300	10,398
	Multi-label	13.0%		12.7%		10.0%	
Spanish	ES-AR	851	49,009	227	12,725	133	8,034
	ES-AR, ES-ES	1,131	61,559	318	17,421	156	8,528
	ES-ES	1,485	93,584	444	28,021	206	13,290
	Total	3,467	204,152	989	58,167	495	29,852
	Multi-label	32.6%		32.2%		31.5%	
Portuguese	PT-BR	2,136	98,061	588	26,848	299	13,605
	PT-BR, PT-PT	420	17,684	134	5,562	59	2,232
	PT-PT	911	38,524	269	11,379	137	5,887
	Total	3,467	154,269	991	43,789	495	21,724
	Multi-label	12.1%		13.5%		11.9%	
French	FR-BE	120,653	8,147,415	7,444	508,853	3,000	333,001
	FR-BE, FR-CA			2	108		
	FR-BE, FR-CH	603	44,991	31	1,920		
	FR-BE, FR-CH, FR-FR	61	2,681				
	FR-BE, FR-FR	1,052	81,602	82	5,295		
	FR-CA	19,041	557,468	2,167	148,669	3,000	334,755
	FR-CA, FR-FR			2	161		
	FR-CH	115,664	7,530,080	1,021	70,245	3,000	317,727
	FR-CH, FR-FR	162	12,218	3	186		
	FR-FR	83,127	5,280,740	6,338	432,269	3,000	323,485
	Total	339,537	21,657,195	17,090	1,167,706	12,000	1,308,959
	Multi-label	0.6%		0.7%		0.0%	
BCMS	BS	45	257,856	7	66,186	10	65,660
	BS, HR			4	29,596	3	9,661
	BS, HR, ME					1	1,634
	BS, HR, ME, SR			1	7,294		
	BS, ME			5	24,791	4	42,262
	BS, ME, SR					2	26,958
	BS, SR			4	23,398	1	2,015
	HR	53	385,385	16	128,760	16	131,821
	HR, SR			6	25,496	2	10,247
	ME	34	242,084	4	20,385	8	66,157
	ME, SR			5	45,738	3	17,340
	SR	236	1,489,997	70	434,136	73	479,606
	Total	368	2,375,322	122	805,780	123	853,361
	Multi-label	0.0%		13.0%		13.0%	

Table 3: Distribution of samples and tokens in the DSL-ML datasets.

topics, namely "inflation" (En.: "inflation"), "jeux olympiques" (En.: "olympic games"), and "reine d'angleterre" (En.: "queen of england"). Each topic was used to query two sources per country. We underline that the training and test topics and sources are disjoint, which generates a cross-domain evaluation setting. Multi-label annotations are inferred using the approach of Bernier-Colborne et al. (2023), which converts near duplicates into multi-label samples. After applying this data cleaning procedure, the training set remains with 340,363 samples, while the development and test sets consist of 17,090 and 12,000 samples, respectively. The training and development data are multi-label, meaning that samples may belong to more than one class, while the testing samples are single-label.[7] In contrast to the datasets of the other languages, named entities are

[7]Running the code of Bernier-Colborne et al. (2023) on the test data did not result in finding near duplicates.

replaced with the NE tag to prevent systems from learning named-entity-related shortcuts. The complete dataset contains approximately 370K samples and 33M tokens.

BCMS. The training set is the same as the BENCHIC-langTwitter training set (Rupnik et al., 2023) (except that retweets were removed from the data for the shared task) and thus only contains single-label annotations. The development and test sets come from the same collection, but were manually reannotated with multiple labels (Miletić and Miletić, 2024). The instances in this dataset cover the entire tweet production of a user and are thus much longer than the single-sentence instances of the other datasets.

Table 3 shows the number of samples and tokens per label and split for all DSL-ML languages, as well as the corresponding percentages of multi-label samples.

3.3 Baseline

The baseline proposed by the shared task organizers is based on an SVM classifier applied on a combination of TF-IDF-weighted character and word n-grams.[8] The classifier follows a multi-class (but not multi-label) setup where label combinations are added as distinct atomic labels. For example, the English task would have three distinct labels: the two single-variety labels `EN-GB` and `EN-US` as well as the multi-variety-label `EN-GB,EN-US`. This setup is equivalent to the one used in DSL-TL, except that the `EN` label is renamed to `EN-GB,EN-US`.

3.4 Participants

Brandeis. The Brandeis team (Sälevä and Palen-Michel, 2024) submitted 3 runs for each of the five languages. Their first run is based on a simple classifier applied to bag-of-n-gram features, where the n-grams are considered at both word and character levels. Aside from count n-gram-based statistics, they also employ the TF-IDF scheme as an alternative representation. For the classification, they alternatively consider logistic regression models, linear-kernel SVMs and random forest models.

For their second run, Sälevä and Palen-Michel (2024) employ a pre-trained multi-lingual BERT (mBERT) (Devlin et al., 2019) and independently fine-tune it on each subset of languages. To address the multi-label classification task, the authors attach a linear classification layer with a sigmoid activation for each unit, and use a threshold of 0.5 for the label to be included in the set of predicted labels. However, if there is no label surpassing the initial threshold, they gradually lower the threshold to 0.25 and 0.05, respectively.

The third run submitted by Brandeis is a variation of the second run, where the fine-tuning of mBERT is jointly performed on all languages (from all sub-tasks) at once.

Jelly. The Jelly team (Gillin, 2024) submitted 3 runs for English, Spanish and Portuguese and 1 run for French; they did not participate in the BCMS subtask. All submitted runs except one are based on one-shot prompting a large language model (LLM). The authors choose the open-source Mistral-7B model (Jiang et al., 2023). For each test sample, the authors provide a prompt containing one training example per language variety and expect the model to produce the multi-label prediction for the given test sample. The different runs differ in the postprocessing of the model output and the back-off strategy chosen if the model output did not contain any valid label.

For the English sub-task, run 2 refers to a variant of in-context learning where the prompt also contains instructions for the labeling task, and run 3 is an ensemble of runs 1 and 2. This team also submitted the raw outputs of Mistral-7B without postprocessing and backoff for comparison - these runs are marked as *open*.

VLP. The VLP team (Ngo et al., 2024) submitted one or two runs for each language. Their first run is based on a bidirectional long short-term memory network (BiLSTM) (Graves et al., 2013). It comprises an embedding layer, several BiLSTM layers and two dense layers, where the last one performs the classification of samples via softmax.

The second run employs the same architecture, but the input is based on ConceptNet embeddings (Speer et al., 2017). More specifically, the authors use ConceptNet Numberbatch semantic vectors, which provide a representation of word meanings extracted from ConceptNet. The ConceptNet embeddings are not available for BCMS, therefore only run 1 is submitted for that subtask. The VLP submissions consider all target labels as atomic, in the same way as the baseline.

[8] The code for the baseline system is available at `https://github.com/yvesscherrer/DSL-ML-2024/tree/main/baseline`. The system described here corresponds to the *atomic* option in the provided script.

English				
Rank	Team	Run	Macro-F1	Multi-label EM
1	Brandeis	3	0.855	0.267
2	Brandeis	2	0.853	0.267
3	Brandeis	1	0.806	0.267
4	VLP	2	0.770	0.167
5	VLP	1	0.759	0.267
6	Jelly	2	0.755	0.133
7	Jelly	2-open	0.752	0.367
8	*Baseline*		0.751	0.100
9	Jelly	1	0.751	0.300
10	Jelly	3	0.750	0.367
11	Jelly	1-open	0.717	0.233

Spanish				
Rank	Team	Run	Macro-F1	Multi-label EM
1	Brandeis	2	0.823	0.500
2	Brandeis	3	0.821	0.551
3	*Baseline*		0.770	0.391
4	VLP	1	0.754	0.455
5	Brandeis	1	0.746	0.455
6	VLP	2	0.741	0.423
7	Jelly	1	0.663	0.333
8	Jelly	2	0.655	0.289
9	Jelly	3	0.649	0.289
10	Jelly	1-open	0.601	0.199

Portuguese				
Rank	Team	Run	Macro-F1	Multi-label EM
1	Brandeis	3	0.752	0.424
2	Brandeis	1	0.724	0.220
3	Brandeis	2	0.714	0.136
4	*Baseline*		0.683	0.068
5	VLP	1	0.664	0.136
6	Jelly	1	0.629	0.356
7	Jelly	2	0.593	0.136
8	Jelly	3	0.586	0.136
9	VLP	2	0.566	0.000
10	Jelly	1-open	0.388	0.034

French			
Rank	Team	Run	Macro-F1
1	Brandeis	3	0.385
2	*Baseline*		0.372
3	Jelly	1	0.313
4	Brandeis	1	0.270
5	Brandeis	2	0.265
6	VLP	2	0.260
7	VLP	1	0.257

BCMS					
Rank	Team	Run	Macro-F1	Weighted F1	Multi-label EM
1	Brandeis	1	0.762	0.843	0.000
2	Brandeis	2	0.719	0.756	0.125
3	*Baseline*		0.606	0.737	0.000
4	VLP	1	0.272	0.370	0.000
5	Brandeis	3	0.199	0.453	0.000

Table 4: Results of the DSL-ML shared task. The official metric is macro F1 score. We do not report weighted F1 score for English, Spanish, Portuguese and French since their test sets are (relatively) balanced and produce the same ranking. For BCMS, we report both macro-averaged and weighted F1-scores. *Multi-label exact match (EM)* refers to the proportion of correctly predicted instances with multiple labels. The French test set does not have multiple labels.

3.5 Results

We evaluate each subtask separately, using macro-averaged F1-score as the main metric. We additionally report weighted-average F1-score for the BCMS task since the class distribution in the test set is much less balanced than in the other tasks.

Furthermore, we measure the models' ability to perform multi-label classification by measuring *multi-label exact match*, i.e., the proportion of gold instances containing two or more labels for which the same set of labels was predicted. The results are presented per language in Table 4.

In general, we see that *Brandeis* is the only team that consistently beats the baseline on all subtasks. While their traditional machine learning submission (run 1) obtained first rank for BCMS, the BERT-based submissions (runs 2 and 3) are ranked highest on the other subtasks. *VLP* beats the baseline for English, is slightly below the baseline for Spanish and Portuguese, and considerably lower for French and BCMS. Their two runs perform roughly on par. Finally, *Jelly* narrowly outperforms the baseline for English, but remains several points below it for the other subtasks.

It can also be seen that the baseline does a comparatively poor job in correctly predicting the multi-labeled instances. While all three participating teams outperform the baseline in terms of multi-label exact match, team *Brandeis* again shows the most consistent performance.

Multi-Label Classification of the DSL-TL Data. Among all languages of this subtask, the overall results are the most encouraging for English. Seven out of ten submitted runs scored above the baseline based on the macro-F1 score (all three runs from *Brandeis*, runs 1 and 2 from *VLP*, and runs 2 and 2-open from *Jelly*), with the top-ranked system achieving a 10% improvement over the baseline. All systems also outperform the baseline on the multi-label exact match score. However, the multi-label exact match score remains relatively low, with the best score at 36.67%, achieved by runs 2-open and 3 submitted by *Jelly*, which are based on the Mistral-7B model. These runs ranked 7th and 10th, respectively.

For Spanish, only runs 2 and 3 by *Brandeis* score above the baseline, with the best VLP system scoring 3% below the baseline, and the *Jelly* runs lagging by 10 or more points. On this language, highly ranked systems also achieve solid results on the multi-label exact match score compared to other languages. In particular, run 3 from *Brandeis* reaches 55.13%.

For Portuguese, the three runs from *Brandeis* are the only systems that outperform the baseline on the macro-F1 score. Overall, the results on the multi-label exact match score are lower for this dataset than for other languages except BCMS. However, the top-ranked system does achieve 42.37%, and the second-best system on this metric is run 1 from *Jelly*, with 35.59%. This is another example of a system that lags behind the baseline based on the macro-F1 score (in this case, by 6 points), but which has a solid performance compared to other systems when it comes to labelling multi-label instances.

Multi-Label French Dialect Identification. For French, two models, one proposed by the Brandeis team and the other by the organizers, stand out from the rest. The top scoring model is based on jointly fine-tuning the mBERT model on all languages. Interestingly, this model is significantly better than the mBERT version fine-tuned on French data (run 2 of Brandeis team), indicating a large benefit from training on multiple languages.

The baseline is a shallow approach (linear SVM) based on basic features, which generalizes fairly well to the cross-domain setup of the French subtask. It is able to compete with the deep model based on multi-lingual fine-tuning submitted by the Brandeis team, being only 1.3% behind.

The third best model, submitted by the Jelly team, uses the Mistral-7B LLM based on in-context learning. Although in-context learning seems to work fairly well, the approach is clearly below the system based on multi-lingual fine-tuning proposed by Brandeis. The Jelly team (Gillin, 2024) obtained much better results on the English sub-task, likely because Mistral-7B is mostly trained on English text. Therefore, in the future, it would be interesting to explore approaches that combine fine-tuning and in-context learning.

The other models submitted by the participants are barely able to surpass the random chance baseline (with an F1 score of 0.25). The last three models are based on deep architectures, and their poor results are likely to be attributed to overfitting. In summary, we conclude that the French sub-task proposed for the 2024 edition of VarDial is very challenging, particularly because of the domain-shift between training and test data, as well as the generally short text samples which may not always contain dialectal patterns.

Multi-Label BCMS Variety Identification. Only Brandeis and VLP submitted runs on the BCMS data. Runs 1 and 3 by the Brandeis team score above the baseline, whereas the remaining submissions score significantly lower on both reported F1 scores. The top two systems achieve solid F1 results, on par with the ones they achieve on Portuguese, although lagging somewhat behind the top scores on English and Spanish.

As noted above, the Brandeis run 1, based on traditional machine learning approaches, achieves the best overall scores on BCMS. However, it is notable that only the Brandeis run 2, based on mBERT, scores above zero on the Multi-label Exact Match score. In other words, this is the only system that manages to correctly label any multi-label instances in the test set. Overall, the multi-label EM is the lowest on BCMS out of all of the languages of this subtask. These results indicate that, while the general task of distinguishing between the varieties of BCMS may be less difficult than it is for French, correctly labelling multi-label instances remains very challenging.

3.6 Conclusions

For the first time at VarDial, we proposed a language and dialect identification task that accepts multi-label scenarios with any number of classes. It includes three two-country settings (with three possible labels, for English, Spanish and Portuguese) as well as two four-country settings (with up to fifteen possible labels, for French and BCMS).

Among the five languages, French turned out to be the most challenging one in terms of obtained macro F1-scores. There are several possible explanations for this. The French data distinguishes itself from the other datasets by a domain shift between training and test data, by its reliance on automatic labeling (both for the initial single-label annotations and the inference of multi-label annotations), and by the masking of named entities. The relative impact of these properties is hard to quantify at the moment and will require additional experiments.

The BCMS task has also been found difficult, especially in terms of multi-label exact match. Eleven labels (country combinations) occur in the test set, but only four of them were observed in the training data, and nine of them in the development set. In such scenarios, it is crucial to use specific multi-label classifiers that can produce combinations of labels unseen at training time.

In terms of methods, both traditional classifiers and embedding-based models were proposed, but none of the two approaches clearly outperforms the other across languages. The *Jelly* submission introduces few-shot prompting as a potentially appealing training-free approach, but the results are not competitive yet with task-specific models. The used large language model often fails to provide the output labels in the correct format, and therefore heavy post-processing is required.

The five datasets used in the DSL-ML task differ widely in size and annotation procedures, and it can be seen that the different submissions are sensitive to different aspects of multi-label classification of similar varieties. We hope to have paved the way for further tasks that embrace the multi-label scenario.

4 Conclusion

This paper presented an overview of the two shared tasks organized as part of the VarDial Evaluation Campaign 2024: Dialectal causal commonsense reasoning (DIALECT-COPA) and Multi-label classification of similar languages (DSL-ML).

Among all the conclusions from the results on the DIALECT-COPA shared task presented in Section 2.5, the most interesting one is that in-context learning on dialectal examples seems to be a highly potent method of adapting an LLM to dialectal tasks. The intuition we have developed through this shared task is that it is all about managing expectations of LLMs, and that letting the LLM simply know about the modified language variant it will be tested on improves its performance significantly.

When it comes to the DSL-ML task, the observations stemming from this iteration further justify the multi-label approach to this task. This is supported both by the proportion of multi-label instances found in the data and by the multi-label exact match scores, which point to the difficulty of the task. We also noted that there were no clear winners in terms of methods between traditional classifiers and embedding-based models. However, as indicated above, the level of disparity between the five datasets used in this year's shared task makes it challenging to identify the impact of different factors on model performance. One possible way forward for this task would consist in creating a homogeneous dataset, taking advantage of best practices from the existing datasets.

Both tasks were shown to be rather challenging, opening up opportunities for future evaluation campaigns.

Acknowledgements

We thank all the participants for their interest in the shared tasks.

The work related to the DIALECT-COPA shared task has received partial funding from the Slovenian Research and Innovation Agency (J7-4642, P6-0411).

The work related to the DSL-ML shared task has received partial funding from the Academy of Finland (funding decision no. 341798).

References

Noëmi Aepli, Antonios Anastasopoulos, Adrian-Gabriel Chifu, William Domingues, Fahim Faisal, Mihaela Gaman, Radu Tudor Ionescu, and Yves Scherrer. 2022. Findings of the VarDial evaluation campaign 2022. In *Proceedings of the Ninth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 1–13, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Noëmi Aepli, Çağrı Çöltekin, Rob Van Der Goot, Tommi Jauhiainen, Mourhaf Kazzaz, Nikola Ljubešić, Kai North, Barbara Plank, Yves Scherrer, and Marcos Zampieri. 2023. Findings of the VarDial evaluation campaign 2023. In *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*, pages 251–261, Dubrovnik, Croatia. Association for Computational Linguistics.

Anthropic. 2024. The claude 3 model family: Opus, sonnet, haiku.

Gabriel Bernier-Colborne, Cyril Goutte, and Serge Leger. 2023. Dialect and variant identification as a multi-label classification task: A proposal based on near-duplicate analysis. In *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*, pages 142–151, Dubrovnik, Croatia. Association for Computational Linguistics.

Bharathi Raja Chakravarthi, Gaman Mihaela, Radu Tudor Ionescu, Heidi Jauhiainen, Tommi Jauhiainen, Krister Lindén, Nikola Ljubešić, Niko Partanen, Ruba Priyadharshini, Christoph Purschke, Eswari Rajagopal, Yves Scherrer, and Marcos Zampieri. 2021. Findings of the VarDial evaluation campaign 2021. In *Proceedings of the Eighth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 1–11, Kiyv, Ukraine. 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.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Fahim Faisal and Antonios Anastasopoulos. 2024. Data-Augmentation based Dialectal Adaptation for LLMs. In *Eleventh Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2024)*, Mexico City, Mexico. Association for Computational Linguistics.

Mihaela Găman, Adrian-Gabriel Chifu, William Domingues, and Radu Tudor Ionescu. 2023. FreCDo: A large corpus for French cross-domain dialect identification. *Procedia Computer Science*, 225:366–373.

Mihaela Gaman, Dirk Hovy, Radu Tudor Ionescu, Heidi Jauhiainen, Tommi Jauhiainen, Krister Lindén, Nikola Ljubešić, Niko Partanen, Christoph Purschke, Yves Scherrer, and Marcos Zampieri. 2020. A report on the VarDial evaluation campaign 2020. In *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 1–14, Barcelona, Spain (Online). International Committee on Computational Linguistics (ICCL).

Nat Gillin. 2024. One-shot Prompt for Language Variety Identification. In *Eleventh Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2024)*, Mexico City, Mexico. Association for Computational Linguistics.

Cyril Goutte, Serge Léger, Shervin Malmasi, and Marcos Zampieri. 2016. Discriminating similar languages: Evaluations and explorations. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1800–1807, Portorož, Slovenia. European Language Resources Association (ELRA).

Alex Graves, Abdel-Rahman Mohamed, and Geoffrey Hinton. 2013. Speech recognition with deep recurrent neural networks. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 6645–6649. IEEE.

Tommi Jauhiainen, Marco Lui, Marcos Zampieri, Timothy Baldwin, and Krister Lindén. 2019. Automatic Language Identification in Texts: A Survey. *Journal of Artificial Intelligence Research*, 65:675–782.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B. *arXiv preprint arXiv:2310.06825*.

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*.

Aditya Joshi, Raj Dabre, Diptesh Kanojia, Zhuang Li, Haolan Zhan, Gholamreza Haffari, and Doris Dippold. 2024. Natural language processing for dialects of a language: A survey. *arXiv preprint arXiv:2401.05632*.

Amr Keleg and Walid Magdy. 2023. Arabic dialect identification under scrutiny: Limitations of single-label classification. In *Proceedings of ArabicNLP 2023*, pages 385–398, Singapore (Hybrid). Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

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 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, Kiyv, Ukraine. Association for Computational Linguistics.

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, Nada Galant, Sonja Benčina, Jaka Čibej, Stefan Milosavljević, Peter Rupnik, and Taja Kuzman. 2024a. DIALECT-COPA: Extending the standard translations of the COPA causal commonsense reasoning dataset to South Slavic dialects. In *Eleventh Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2024)*, Mexico City, Mexico. Association for Computational Linguistics.

Nikola Ljubešić, Taja Kuzman, Peter Rupnik, Goran Glavaš, Fabian David Schmidt, and Ivan Vulić. 2024b. 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.

Shervin Malmasi, Marcos Zampieri, Nikola Ljubešić, Preslav Nakov, Ahmed Ali, and Jörg Tiedemann. 2016. Discriminating between similar languages and Arabic dialect identification: A report on the third DSL shared task. In *Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial3)*, pages 1–14, Osaka, Japan. The COLING 2016 Organizing Committee.

Aleksandra Miletić and Filip Miletić. 2024. A gold standard with silver linings: Scaling up annotation for distinguishing Bosnian, Croatian, Montenegrin and Serbian. In *Proceedings of the 4th Workshop on Human Evaluation of NLP Systems*, Turin, Italy. European Language Resources Association.

The Quyen Ngo, Thi Anh Phuong Nguyen, My Linh Ha, Thi Minh Huyen Nguyen, and Phuong Le-Hong. 2024. Improving Multi-label Classification of Similar Languages by Semantics-Aware Word Embeddings. In *Eleventh Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2024)*, Mexico City, Mexico. Association for Computational Linguistics.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report.

Benedikt Perak, Slobodan Beliga, and Ana Meštrović. 2024. Incorporating Dialect Understanding into LLM Using RAG and Prompt Engineering Techniques for Causal Commonsense Reasoning. In *Eleventh Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2024)*, Mexico City, Mexico. Association for Computational Linguistics.

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*.

Sebastian Ruder, Noah Constant, Jan Botha, Aditya Siddhant, Orhan Firat, Jinlan Fu, Pengfei Liu, Junjie Hu, Dan Garrette, Graham Neubig, and Melvin Johnson. 2021. XTREME-R: Towards more challenging and nuanced multilingual evaluation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10215–10245, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Peter Rupnik, Taja Kuzman, and Nikola Ljubešić. 2023. BENCHić-lang: A benchmark for discriminating between Bosnian, Croatian, Montenegrin and Serbian. In *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*, pages 113–120, Dubrovnik, Croatia. Association for Computational Linguistics.

Jonne Sälevä and Chester Palen-Michel. 2024. Brandeis at VarDial 2024 DSL-ML Shared Task: Multilingual models, simple baselines and data augmentation. In *Eleventh Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2024)*, Mexico City, Mexico. Association for Computational Linguistics.

Lütfi Kerem Senel, Benedikt Ebing, Konul Baghirova, Hinrich Schuetze, and Goran Glavaš. 2024. Kardeş-NLU: Transfer to low-resource languages with the help of a high-resource cousin – a benchmark and evaluation for Turkic languages. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1672–1688, St. Julian's, Malta. Association for Computational Linguistics.

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.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. 2023. Language models are multilingual chain-of-thought reasoners. In *The Eleventh*

International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023. OpenReview.net.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. ConceptNet 5.5: an open multilingual graph of general knowledge. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, page 4444–4451. AAAI Press.

Liling Tan, Marcos Zampieri, Nikola Ljubešic, and Jörg Tiedemann. 2014. Merging comparable data sources for the discrimination of similar languages: The dsl corpus collection. In *Proceedings of the 7th Workshop on Building and Using Comparable Corpora (BUCC)*, pages 11–15, Reykjavik, Iceland.

Ahmet Üstün, Viraat Aryabumi, Zheng-Xin Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, et al. 2024. Aya model: An instruction finetuned open-access multilingual language model. *arXiv preprint arXiv:2402.07827*.

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.

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.

Marcos Zampieri, Shervin Malmasi, Nikola Ljubešić, Preslav Nakov, Ahmed Ali, Jörg Tiedemann, Yves Scherrer, and Noëmi Aepli. 2017. Findings of the VarDial evaluation campaign 2017. In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, pages 1–15, Valencia, Spain. Association for Computational Linguistics.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Ahmed Ali, Suwon Shon, James Glass, Yves Scherrer, Tanja Samardžić, Nikola Ljubešić, Jörg Tiedemann, Chris van der Lee, Stefan Grondelaers, Nelleke Oostdijk, Dirk Speelman, Antal van den Bosch, Ritesh Kumar, Bornini Lahiri, and Mayank Jain. 2018. Language identification and morphosyntactic tagging: The second VarDial evaluation campaign. In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018)*, pages 1–17, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Marcos Zampieri, Shervin Malmasi, Yves Scherrer, Tanja Samardžić, Francis Tyers, Miikka Silfverberg, Natalia Klyueva, Tung-Le Pan, Chu-Ren Huang, Radu Tudor Ionescu, Andrei M. Butnaru, and Tommi Jauhiainen. 2019. A report on the third VarDial evaluation campaign. In *Proceedings of the Sixth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 1–16, Ann Arbor, Michigan. Association for Computational Linguistics.

Marcos Zampieri, Kai North, Tommi Jauhiainen, Mariano Felice, Neha Kumari, Nishant Nair, and Yash Bangera. 2023. Language variety identification with true labels. *arXiv preprint arXiv:2303.01490*.

Marcos Zampieri, Liling Tan, Nikola Ljubešić, and Jörg Tiedemann. 2014. A report on the DSL shared task 2014. In *Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects*, pages 58–67, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.

Marcos Zampieri, Liling Tan, Nikola Ljubešić, Jörg Tiedemann, and Preslav Nakov. 2015. Overview of the DSL shared task 2015. In *Proceedings of the Joint Workshop on Language Technology for Closely Related Languages, Varieties and Dialects*, pages 1–9, Hissar, Bulgaria. Association for Computational Linguistics.

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*.

What Drives Performance in Multilingual Language Models?

Sina Bagheri Nezhad, Ameeta Agrawal
Portland State University
{sina.bagherinezhad, ameeta}@pdx.edu

Abstract

This study investigates the factors influencing the performance of multilingual large language models (MLLMs) across diverse languages. We study 6 MLLMs, including masked language models, autoregressive models, and instruction-tuned LLMs, on the SIB-200 dataset, a topic classification dataset encompassing 204 languages. Our analysis considers three scenarios: ALL languages, SEEN languages (present in the model's pretraining data), and UNSEEN languages (not present or documented in the model's pretraining data in any meaningful way). We examine the impact of factors such as pretraining data size, general resource availability, language family, and script type on model performance. Decision tree analysis reveals that pretraining data size is the most influential factor for SEEN languages. However, interestingly, script type and language family are crucial for UNSEEN languages, highlighting the importance of cross-lingual transfer learning. Notably, model size and architecture do not significantly alter the most important features identified. Our findings provide valuable insights into the strengths and limitations of current MLLMs and hope to guide the development of more effective and equitable multilingual NLP systems.[1]

1 Introduction

Multilingual large language models (MLLMs) have revolutionized natural language processing by enabling applications like machine translation and sentiment analysis across numerous languages (Barbieri et al., 2022; Yang et al., 2023). Understanding how these models perform across languages with diverse linguistic properties is crucial for further development (Devlin et al., 2019; Wu and Dredze, 2020; Scao et al., 2022; Lai et al., 2023; Ahuja et al., 2023). Despite significant progress, linguistic disparities persist in NLP, highlighting the need for models that perform effectively and safely across a wider range of languages (Joshi et al., 2020; Ranathunga and de Silva, 2022; Agrawal et al., 2023; Wang et al., 2023).

The factors contributing to the effectiveness of MLLMs, however, remain unclear. While several studies suggest the amount of language-specific pretraining data as a key factor (Wu and Dredze, 2020; Scao et al., 2022; Shliazhko et al., 2022; Ahuja et al., 2023), most investigations are limited in scope, focusing on a small set of languages, specific tasks, or training paradigms like masked language modeling (MLM) or autoregressive models. Crucially, prior work often overlooks the distinction between languages encountered during pretraining (SEEN), languages entirely new to the model (UNSEEN), and the complete set of languages available in the evaluation dataset (ALL). The question remains – *what factors are important in the case of unseen languages where language-specific pretraining data is not one of the relevant factors?* This distinction is essential for understanding how MLLMs generalize to languages with varying levels of familiarity.

Our work takes a deeper look at the various factors under several experimental settings. Our key contributions are as follows:

- We conduct a comprehensive evaluation of 6 MLLMs, including MLM, autoregressive, and instruction-tuned LLMs, on a text classification task spanning a wide range of languages. This diverse set of models includes `mBERT` (Devlin et al., 2019), `XLM-R` (Conneau et al., 2020), `GPT-3.5` (Brown et al., 2020), `Bloom` (Scao et al., 2022) in 5 sizes, `Bloomz` (Muennighoff et al., 2023) in 5 sizes, and `XGLM` (Lin et al., 2022) in 4 sizes. Additionally, we consider three training scenarios: zero-shot, 2-shot, and fully supervised.
- We consider four key factors in our analysis: pre-

[1] https://github.com/PortNLP/MLLMs_performance

Proceedings of the Eleventh Workshop on NLP for Similar Languages, Varieties, and Dialects (VarDial 2024), pages 16–27
June 20, 2024

Reference	Factors	Task	Languages
Wu and Dredze (2020)	Pretraining data size, Task-specific data size, Vocabulary size	NER	99
Scao et al. (2022)	Pretraining data size, Task-specific data size, Language family, Language script	Probing	17
Shliazhko et al. (2022)	Pretraining data size, Language script, Model size	Perplexity	61
Ahuja et al. (2023)	Pretraining data size, Tokenizer fertility	Classification, QA, Sequence Labeling, NLG	2-48
Ours	Pretraining data size, Language family, Language script, General resource availability	Text classification	204

Table 1: Factors considered in related works and this work.

training data size, general resource availability levels, language family, and script type. This allows for a more nuanced understanding of the factors influencing MLLM performance.

- We leverage the recently introduced SIB-200 dataset (Adelani et al., 2023), which includes 204 languages, enabling us to investigate MLLM performance across a diverse and extensive linguistic landscape. Between the languages pertaining to the models and the dataset, we are able to further distinguish them along the dimensions of SEEN, UNSEEN, or ALL, depending on whether the languages were seen during pretraining, or unseen during pretraining, or the set of all languages available in the evaluation dataset, respectively.

By analyzing these factors across different models and training setups, we aim to provide deeper insights into the development of effective and equitable MLLMs for a truly multilingual NLP landscape.

2 Related Work

Multilingual NLP research has flourished in recent years, with the development and evaluation of numerous multilingual language models trained on diverse and extensive language datasets. Notable examples include `mBERT` (Devlin et al., 2019), `XLM-R` (Conneau et al., 2020), `mBART` (Liu et al., 2020), `mT5` (Xue et al., 2021), `BLOOM` (Scao et al., 2022), `GPT-3` (Brown et al., 2020), `GPT-4` (OpenAI, 2023), `LLaMA` (Touvron et al., 2023), `PaLM` (Chowdhery et al., 2022), and `PaLM 2` (Anil et al., 2023).

Researchers are increasingly interested in investigating the factors influencing MLLM performance. Wu and Dredze (2020) examined the impact of pretraining data size, task-specific data size, and vocabulary size on named entity recognition performance. Scao et al. (2022) explored the correlation between probing performance and factors like language family, task-specific dataset size, and pretraining dataset size for the BLOOM model. Shliazhko et al. (2022) assessed the impact of language script, pretraining corpus size, and model size on language modeling performance, while Ahuja et al. (2023) investigated the influence of tokenizer fertility and pretraining data on MLLM performance.

While these studies provide valuable insights, they often focus on a limited set of languages, primarily due to the historical scarcity of annotated multilingual datasets. Additionally, research by Blasi et al. (2022) highlights the significant inequalities in the development and performance of language technologies across the world's languages, with a strong bias towards resource-rich languages like English and other Western European languages. Further exacerbating this issue is the lack of representation for dialects, varieties, and closely-related languages within existing datasets. As noted by Faisal et al. (2024), this absence hinders the development of NLP systems capable of effectively handling the nuances of linguistic diversity. However, the recent emergence of comprehensive multilingual datasets like SIB-200 (Adelani et al., 2023), and GLOT500 (ImaniGooghari et al., 2023) offers exciting opportunities for more extensive and nuanced analyses. Table 1 summarizes the factors considered in related works and our study. For a more comprehensive overview of contributing factors to cross-lingual transfer in multilingual language models, readers are encouraged to refer to the review by Philippy et al. (2023).

3 Methodology

Several factors can influence the performance of multilingual models. In this section, we briefly describe the distinct factors related to typology and data, the dataset of more than 200 languages used for evaluation, and the models we consider in this study.

3.1 Typology and Data Factors

We consider various factors to understand their impact on model performance including:

- **Pretraining Data Size:** This refers to the percentage of language-specific data used during the pretraining of each model[2].
- **General Resource Availability (Res Level):** Beyond model-specific resources such as pretraining data size, we also consider a more general notion of resource availability, as per the linguistic diversity taxonomy which categorizes languages into six resource levels (Joshi et al., 2020), where level 0 corresponds to low-resource and level 5 corresponds to high-resource level languages. This classification helps us understand the influence of more general resource availability on model performance, and may serve as a proxy when model-specific statistics may not be available (such as in the case of proprietary models). Language resource levels generally correlate positively with models pretraining data sizes, with varying degrees of alignment across different models: mBERT (0.52) and XLM-R (0.48) exhibit relatively stronger correlations, while GPT-3 (0.18), BLOOM (0.37), and XGLM (0.31) show comparatively weaker associations.
- **Language Family (Lang Family):** The language families that the languages belong to capture some of their linguistic relationships. The information was sourced from the Ethnologue[3] (Ethnologue, 2022).
- **Script:** The script of a language refers to the writing system it employs. This information was sourced from ScriptSource[4].

[2] We obtained the train dataset distribution values for `mBERT` from `https://github.com/mayhewsw/multilingual-data-stats` and for `GPT-3.5` we use proxy statistics from `https://github.com/openai/gpt-3/blob/master/dataset_statistics/languages_by_word_count.csv`. Distribution of train dataset for `XLM-R`, `BLOOM`, `BLOOMZ` and `XGLM` were obtained from their respective papers.

[3] `https://www.ethnologue.com`

[4] `https://www.scriptsource.org`

3.2 Data

We systematically study the multilingual models under an important NLP task – text classification (Chang and Bergen, 2023). The SIB-200 dataset (Adelani et al., 2023) offers a valuable resource for evaluating MLLM performance in a large-scale text classification task, enabling simultaneous analysis of approximately 200 languages, with text samples categorized into one of seven classes. F1 score is used as the metric for this task.

Exploratory analysis of the dataset reveals several interesting insights:

- As shown in Figure 1, most languages in SIB-200 are classified as resource level 1, indicating a deliberate focus on low-resource languages. This allows us to assess how MLLMs perform on languages with limited linguistic resources available.
- Figure 4 in Appendix B illustrates the distribution of language families within the SIB-200 dataset. Notably, the dataset encompasses 23 different language families, providing a rich linguistic landscape for our analysis. Indo-European languages constitute a significant portion (approximately 36%) of SIB-200, reflecting their status as the most widely spoken language family globally (Ethnologue, 2022). However, Niger-Congo, Afro-Asiatic, and Austronesian languages also have considerable representation in the dataset. This diverse language family distribution enables us to analyze MLLM performance across different linguistic groups.
- The SIB-200 dataset encompasses text samples written in 29 different script types, offering a diverse range of writing systems for our analysis. As shown in Figure 5 in Appendix B, the Latin script, used by nearly 70% of the global population (Vaughan, 2020), is the most prevalent writing system in the dataset, followed by Arabic and Cyrillic scripts. This distribution allows us to investigate the impact of script type on MLLM performance.

For all evaluations, we use the default train and test splits recommended by the SIB-200 authors. This ensures consistency and comparability across different models and training settings.

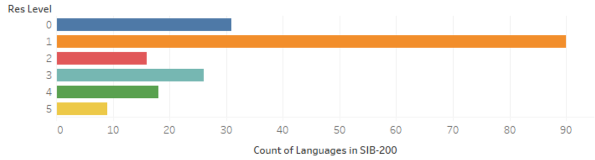


Figure 1: Distribution of resource levels in SIB-200.

3.3 Models

We study the following 6 multilingual language models spanning various architectures and sizes:

- Masked Language Models (MLMs):
  - `mBERT` (bert-base-multilingual-cased) (Devlin et al., 2019)
  - `XLM-R` (xlm-roberta-base) (Conneau et al., 2020)
- Autoregressive Language Models
  - `GPT-3.5` (text-davinci-003) (Brown et al., 2020)
  - `Bloom` (Scao et al., 2022) in 5 sizes (560m, 1.1b, 1.7b, 3b, and 7.1b parameters)
  - `XGLM` (Lin et al., 2022) in 4 sizes (564m, 1.7b, 2.9b, and 7.5b parameters)
- Instruction-tuned LLMs:
  - `Bloomz` (Muennighoff et al., 2023) in 5 sizes (560m, 1.1b, 1.7b, 3b, and 7.1b parameters)

These models were chosen for several key reasons:

1. These models provide broad language coverage, allowing us to analyze performance across a diverse set of languages and maximize the linguistic diversity in our study.

2. By including MLMs, autoregressive models, and instruction-tuned LLMs, we can investigate how different model architectures influence performance.

3. The inclusion of models with varying parameter sizes allows us to investigate the interplay between model scale and the factors influencing performance.

4. `mBERT` and `XLM-R`, despite being relatively smaller models, have demonstrated competitive performance even compared to larger models like `ChatGPT` after fine-tuning (Lai et al., 2023; Zhu et al., 2023).

5. The inclusion of both `Bloom` and `XGLM`, both autoregressive models, allows us to investigate the impact of pretraining data composition. `Bloom` focuses more on low-resource languages during pretraining, whereas `XGLM` emphasizes high-resource languages. This deliberate selection enables us to analyze how the distribution of languages in the pretraining data affects performance across different resource levels.

Note that we primarily focus on models that are open-source or have made the list of pretraining languages and data composition available.

Additionally, we consider the following training and inference scenarios:

- Zero-shot: `GPT-3.5`, `Bloom`, `Bloomz`, and `XGLM` were evaluated directly on the test set without any specific fine-tuning. This assesses the model's ability to generalize to unseen tasks and languages based on its pretrained knowledge.

- Two-shot In-Context Learning (ICL): `Bloom`, `Bloomz`, and `XGLM` were also evaluated in two-shot ICL setting where the models were provided with two labeled examples for each class from the train set. This allows us to particularly investigate effective factors for improving performance of unseen languages. We opted for two demonstrations in ICL to keep the input length shorter than the context length of our models across all languages.

- Full-shot: `mBERT` and `XLM-R` were fine-tuned on the SIB-200 training set and evaluated on the test set.

For full-shot training of mBERT and XLM-R, we adhered to the hyperparameters recommended by the SIB-200 paper authors to ensure consistency with the original dataset benchmarks. For Bloom, Bloomz, and XGLM in both zero-shot and two-shot ICL settings, as well as for GPT-3.5 in zero-shot setting, we use prompts to frame the text classification task, which are detailed in Appendix A.

4 Results and Analysis

Now we discuss the results of our comprehensive experiments. We focus on analyzing the performance of models across three distinct scenarios: ALL, SEEN, and UNSEEN. The ALL

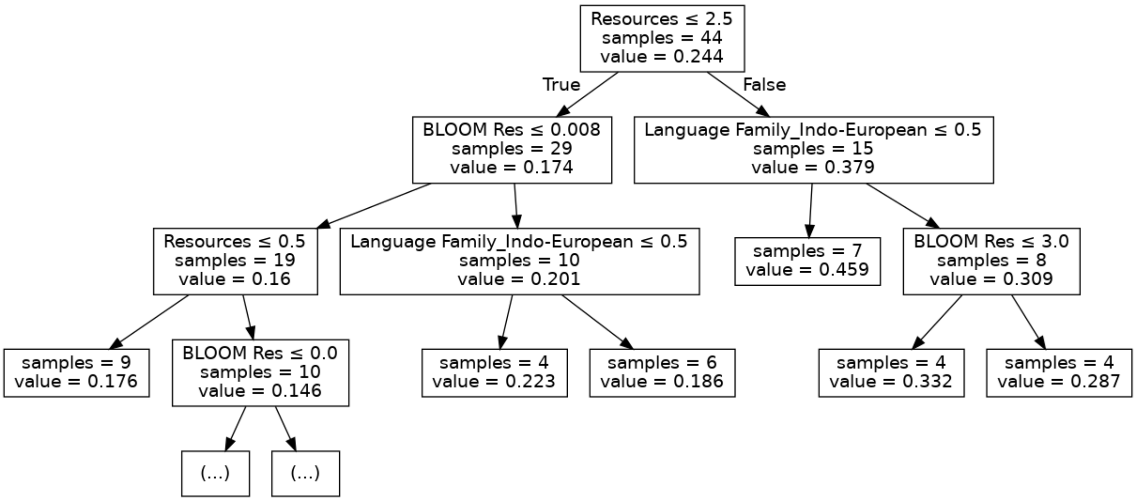


Figure 2: Decision tree for Bloom-560m (zero-shot, SEEN languages). "General resource level" emerges as the most important feature, with a significant performance difference between languages above and below the 2.5 threshold ($p < 0.001$ as per Mann-Whitney U test).

scenario considers all languages in the SIB-200 dataset for which resource level information is available[5]. The SEEN scenario focuses on languages included in the pretraining data of the respective MLLMs, while the UNSEEN scenario examines performance on languages not present in the pretraining data.

In total, results are obtained from 93 distinct experimental settings (models of different sizes, training scenarios, and language categories of seen/unseen/all).

To understand the complex interplay of multiple factors influencing MLLM performance, we employ decision tree analysis for statistical inference. This approach is well-suited for handling factors of different types, including categorical, ordinal, and numeric data. Decision trees are trained to predict the F1 score of models based on language features. By analyzing the resulting tree structure, we can gain insights into the relative importance of different features and their interactions.

As decision trees were trained on the entirety of our data, traditional methods for testing their performance were not applicable. Instead, we employed the Mann-Whitney U test (Mann and Whitney, 1947), to ensure that the features appearing at the root of the decision trees were indeed relevant and contributed significantly to the differentiation between the language splits. This approach allowed us to validate the significance of the features identified by the decision tree in delineating distinct language groups without relying solely on the performance metrics of the decision tree models themselves.

Figure 2 presents the decision tree analysis for the `Bloom-560m` model on SEEN languages, revealing *general resource level* as the most influential feature. Specifically, the tree distinguishes between languages with resource levels below 2.5 (levels 0,1,2) and those above 2.5 (levels 3,4,5). Among the 44 SEEN languages, the 29 languages with resource levels below 2.5 exhibit a mean F1 score of 0.174, while the 15 languages with higher resource levels achieve a significantly higher mean F1 score of 0.379. A Mann-Whitney U test confirms a statistically significant difference in performance between these two groups ($p < 0.001$). This suggests that for the `Bloom-560m` model on SEEN languages, the general resource level of a language plays a crucial role in determining its performance, with higher resource levels leading to better performance. By employing this combined approach of decision tree analysis and statistical testing, we can effectively disentangle the complex relationships between various factors and their impact on MLLM performance.

The summarized results[6] of all 93 decision tree analyses are presented in Table 2. We observe distinct patterns in feature importance across the three scenarios:

[5]This information is available for 190 languages.

[6]Detailed decision trees for all models and setups are available in our repository: `https://github.com/PortNLP/MLLMs_performance`

	Zero-shot		
Model	ALL	SEEN	UNSEEN
Bloom-560m	Pretrain data (<=0.125%)	Resource level (<=2.5)	Script (Latin or not)
Bloom-1b1	Pretrain data (<=0.125%)	Resource level (<=2.5)	Script (Devanagari or not)
Bloom-1b7	Pretrain data (<=0.175%)	Resource level (<=2.5)	Script (Latin or not)
Bloom-3b	Pretrain data (<=0.175%)	Resource level (<=2.5)	Script (Latin or not)
Bloom-7b1	Pretrain data (<=0.125%)	Resource level (<=2.5)	Script (Devanagari or not)
Bloomz-560m	Script (Latin or not)	Pretrain data (<=0.03%)	Script (Latin or not)
Bloomz-1b1	Pretrain data (<=0.008%)	Pretrain data (<=0.03%)	Script (Latin or not)
Bloomz-1b7	Pretrain data (<=0.008%)	Pretrain data (<=0.03%)	Script (Latin or not)
Bloomz-3b	Pretrain data (<=0.002%)	Pretrain data (<=0.013%)	Script (Latin or not)
Bloomz-7b1	Pretrain data (<=0%)	Pretrain data (<=0.9%)	Script (Latin or not)
XGLM-564m	Pretrain data (<=0.003%)	Resource level (<=2)	Lang. family (Austronesian or not)
XGLM-1.7b	Pretrain data (<=0.006%)	Pretrain data (<=1.487%)	Script (Devanagari or not)
XGLM-2.9b	Pretrain data (<=0.003%)	Script (Latin or not)	Script (Devanagari or not)
XGLM-7.5b	Pretrain data (<=0%)	Pretrain data (<=1.122%)	Script (Devanagari or not)
GPT-3.5	Resource level (<= 2.5)	Pretrain data (<=0.003%)	Lang. family (Indo-European or not)
	Two-shot ICL		
Model	ALL	SEEN	UNSEEN
Bloom-560m	Pretrain data (<=0.045%)	Pretrain data (<=0.045%)	Lang. family (Indo-European or not)
Bloom-1b1	Pretrain data (<=0.095%)	Pretrain data (<=0.095%)	Script (Latin or not)
Bloom-1b7	Pretrain data (<=0.175%)	Pretrain data (<=0.175%)	Script (Latin or not)
Bloom-3b	Pretrain data (<=0.008%)	Pretrain data (<=0.008%)	Script (Latin or not)
Bloom-7b1	Pretrain data (<=0.008%)	Pretrain data (<=0.008%)	Script (Latin or not)
Bloomz-560m	Pretrain data (<=0.03%)	Pretrain data (<0.03%)	Script (Devanagari or not)
Bloomz-1b1	Pretrain data (<=0.008%)	Pretrain data (<=0.013%)	Script (Latin or not)
Bloomz-1b7	Pretrain data (<=0.005%)	Pretrain data (<=0.013%)	Script (Cyrillic or not)
Bloomz-3b	Pretrain data (<=0%)	Pretrain data (<=0.9%)	Script (Latin or not)
Bloomz-7b1	Pretrain data (<=0%)	Pretrain data (<=0.013%)	Script (Latin or not)
XGLM-564m	Pretrain data (<=0.003%)	Pretrain data (<=0.095%)	Lang. family (Niger-Congo or not)
XGLM-1.7b	Pretrain data (<=0.003%)	Resource level (<=2)	Script (Devanagari or not)
XGLM-2.9b	Pretrain data (<=0.003%)	Script (Latin or not)	Lang. family (Indo-European or not)
XGLM-7.5b	Pretrain data (<=0.003%)	Pretrain data (<=0.15%)	Lang. family (Indo-European or not)
	Full-shot		
Model	ALL	SEEN	UNSEEN
mBERT	Pretrain data (<=0.032%)	Pretrain data (<=0.073%)	Lang. family (Indo-European or not)
XLM-R	Pretrain data (<=0.005%)	Pretrain data (<=0.031%)	Lang. family (Indo-European or not)

Table 2: Top features identified by decision tree analysis for each model and scenario. For SEEN languages, pretraining data size and resource level dominate (except for XGLM-2.9b, where script type is most influential). For UNSEEN languages, linguistic characteristics (script type and language family) take precedence. All features exhibit statistically significant differences in performance ($p < 0.001$).

ALL Languages:

- For the ALL languages scenario, decision trees clearly reveal that pretraining data is the most influential factor in 29 out of 31 cases. Because ALL includes languages SEEN and UNSEEN, notably, our deeper look at the decision tree analyses indicates that this factor in most cases boils down to *whether the language was part of the training set or not, rather than the amount of language-specific data*, as indicated by the values

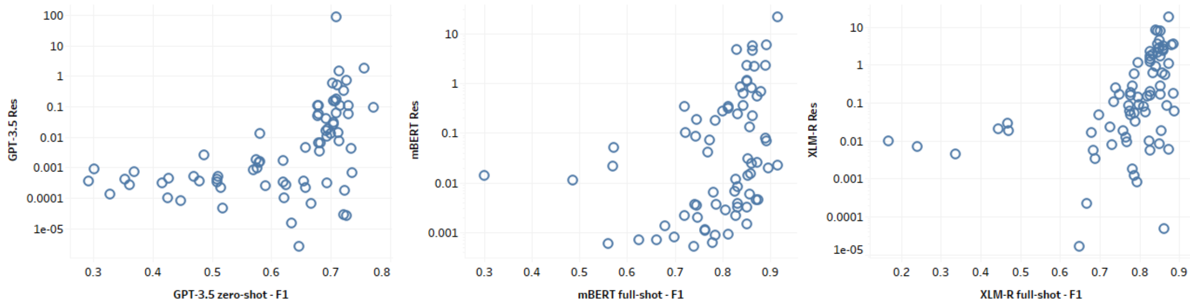


Figure 3: F1 Score vs. model-specific pretraining data (percentage) for GPT-3.5, mBERT and XLM-R models.

of the pretraining data percentages which range from 0% to at most 0.175%. GPT-3.5 model draws the distinction along general resource levels whether a language is low resource (0, 1, or 2) or level 3 and higher.

SEEN Languages:

- For SEEN languages, model-specific pretraining data continues to remain the most influential factor in 22 out of 31 model and scenario combinations. However, this time because there are no unseen languages in the mix, the model performance seems to be impacted by the amount of pretraining data, as indicated by the slightly higher percentage values as compared to the ALL languages scenario.

- Interestingly, general resource availability based on linguistic diversity taxonomy (Joshi et al., 2020) appears to be the most important factor for `Bloom` models in the zero-shot setup, as well as for `xglm-564m` (zero-shot) and `xglm-1.7b` (two-shot). For `Bloom` models, the distinction is along resource levels 0/1/2 or higher, whereas for `xglm` models, it is along 0/1 and higher. Additionally, `xglm-2.9b` in both zero-shot and two-shot scenarios shows a stronger influence of script type (Latin or not). These cases indicate that factors beyond pretraining data size can also play a significant role for specific models and settings.

- Furthermore, Figure 3 plots the performance of `mBERT`, `XLM-R`, and `GPT-3.5` models in relation to model-specific pretraining data amounts. The figure demonstrates a clear trend: as the model-specific language data increases, so does the model's performance. This observation aligns with the finding that pretraining data size is a crucial factor for SEEN languages.

UNSEEN Languages:

- In contrast to SEEN languages, UNSEEN languages show quite a different pattern. Naturally, because UNSEEN languages do not have pretraining data as one of their relevant factors, it is absent from this column. However, out of 31 models, 23 are most impacted by script type, and 8 are most influenced by language family. This shift in importance towards linguistic features suggests that when models encounter unfamiliar languages, they rely more heavily on similarities in writing systems to generalize from their existing knowledge.

- Within the scripts and language families, there are nuanced differences. For instance, while generally the models make the distinction along the lines of whether the script is Latin or not, occassionally Devanagari script also seems important, particularly for `XGLM` models. Similarly, while Indo-European is the most common influential language family, we also observe an instance each of Austronesian and Niger-Congo. Additionally, models of different sizes from the same family may prefer not just a different script or a different language family when moving from zero-shot to two-shot setting, they may prefer an entirely different factor (e.g., `Bloom-560m` in zero-shot vs. two-shot settings), further complicating the matters.

5 Discussion

Our comprehensive analysis of 6 multilingual models on the SIB-200 dataset reveals valuable insights into the factors influencing their performance across a diverse range of languages.

Our key findings can be summarized as follows:

- Pretraining data size consistently emerges as a crucial factor, but the distinction is less along the quantity of data but rather whether the languages have been encountered during training or not.

- For UNSEEN languages, script type and language family are influential, suggesting that MLLMs rely on cross-lingual transfer learning to generalize to unfamiliar languages.
- General resource availability plays a less prominent role overall but appears to be important for one specific model under one setting (`Bloom` in zero-shot for seen languages).
- Interestingly, the performance of `Bloomz`, an instruction-tuned model, is more influenced by the distribution of languages in its pretraining corpus than the fine-tuned dataset used for instruction tuning. This suggests that the initial pretraining stage plays a crucial role in shaping the model's capabilities, even after further fine-tuning for specific tasks.
- Finally, our analysis also indicates that while model size and architecture may influence overall performance, they do not significantly alter the most important features identified by the decision trees. The distribution of languages in the pretraining data and the linguistic characteristics of the target languages consistently emerge as the dominant factors regardless of the specific model architecture or scale.

Several future directions remain to be explored. We observed that script type can be more influential for specific models and settings. Further investigation is needed to understand the reasons behind these preferences and how they can be leveraged to achieve more consistent performance across languages. It is also not clear why models lean towards different factors under different settings (for instance, resource level is important in `Bloom-560m` zero-shot setting but pretraining data is important in its two-shot ICL setting).

6 Conclusion

This study analyzed 6 multilingual language models on the SIB-200 dataset, revealing key insights into their performance across around 200 languages. We found that the size of the pretraining data significantly affects performance. For unseen languages, script type and language family become more crucial, highlighting the importance of cross-lingual transfer learning. While general resource availability plays a less prominent role overall, it can be significant for specific models and settings. Interestingly, model size and architecture do not significantly change the most important features identified in our analysis. Our work contributes to a deeper understanding of MLLMs and hopes to guide the development of more effective and equitable multilingual NLP systems.

Limitations

This study provides insights into multilingual language model performance, but it is important to acknowledge certain limitations. The SIB-200 dataset, while extensive, may contain biases in language representation and genre distribution, potentially affecting the generalizability of our findings. Additionally, our analysis focuses on the text classification task, and the findings may not directly generalize to other NLP tasks. While we analyzed a diverse set of models, our findings may not be fully representative of the entire MLLM landscape. Finally, our analysis is based on the current state of MLLMs, and the relative importance of different factors may change as these models continue to evolve. Future research should address these limitations by expanding to more diverse datasets, investigating different NLP tasks, evaluating a broader range of models, and conducting longitudinal studies.

Ethics Statement

The experimental setup and code implementation ensured adherence to ethical guidelines, data usage agreements, and compliance with the terms of service of the respective language models and data sources. The research team also recognized the importance of inclusivity and fairness by considering a diverse set of languages and language families in the evaluation, thus avoiding biases and promoting balanced representation.

Acknowledgements

We are grateful to the anonymous reviewers whose feedback and thought-provoking questions enhanced this paper. The engaging discussions and collaborative spirit within the PortNLP research group were instrumental in shaping this research. We acknowledge the National Science Foundation for their financial support through grants (CRII:RI

2246174 and SAI-P 2228783), which made this work possible.

References

David Ifeoluwa Adelani, Hannah Liu, Xiaoyu Shen, Nikita Vassilyev, Jesujoba O Alabi, Yanke Mao, Haonan Gao, and Annie En-Shiun Lee. 2023. Sib-200: A simple, inclusive, and big evaluation dataset for topic classification in 200+ languages and dialects. *arXiv preprint arXiv:2309.07445*.

Ameeta Agrawal, Lisa Singh, Elizabeth Jacobs, Yaguang Liu, Gwyneth Dunlevy, Rhitabrat Pokharel, and Varun Uppala. 2023. All translation tools are not equal: Investigating the quality of language translation for forced migration. In *2023 IEEE 10th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 1–10.

Kabir Ahuja, Rishav Hada, Millicent Ochieng, Prachi Jain, Harshita Diddee, Samuel Maina, Tanuja Ganu, Sameer Segal, Maxamed Axmed, Kalika Bali, et al. 2023. Mega: Multilingual evaluation of generative ai. *arXiv preprint arXiv:2303.12528*.

Rohan Anil, Andrew M. Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, Eric Chu, Jonathan H. Clark, Laurent El Shafey, Yanping Huang, Kathy Meier-Hellstern, Gaurav Mishra, Erica Moreira, Mark Omernick, Kevin Robinson, Sebastian Ruder, Yi Tay, Kefan Xiao, Yuanzhong Xu, Yujing Zhang, Gustavo Hernandez Abrego, Junwhan Ahn, Jacob Austin, Paul Barham, Jan Botha, James Bradbury, Siddhartha Brahma, Kevin Brooks, Michele Catasta, Yong Cheng, Colin Cherry, Christopher A. Choquette-Choo, Aakanksha Chowdhery, Clément Crepy, Shachi Dave, Mostafa Dehghani, Sunipa Dev, Jacob Devlin, Mark Díaz, Nan Du, Ethan Dyer, Vlad Feinberg, Fangxiaoyu Feng, Vlad Fienber, Markus Freitag, Xavier Garcia, Sebastian Gehrmann, Lucas Gonzalez, Guy Gur-Ari, Steven Hand, Hadi Hashemi, Le Hou, Joshua Howland, Andrea Hu, Jeffrey Hui, Jeremy Hurwitz, Michael Isard, Abe Ittycheriah, Matthew Jagielski, Wenhao Jia, Kathleen Kenealy, Maxim Krikun, Sneha Kudugunta, Chang Lan, Katherine Lee, Benjamin Lee, Eric Li, Music Li, Wei Li, YaGuang Li, Jian Li, Hyeontaek Lim, Hanzhao Lin, Zhongtao Liu, Frederick Liu, Marcello Maggioni, Aroma Mahendru, Joshua Maynez, Vedant Misra, Maysam Moussalem, Zachary Nado, John Nham, Eric Ni, Andrew Nystrom, Alicia Parrish, Marie Pellat, Martin Polacek, Alex Polozov, Reiner Pope, Siyuan Qiao, Emily Reif, Bryan Richter, Parker Riley, Alex Castro Ros, Aurko Roy, Brennan Saeta, Rajkumar Samuel, Renee Shelby, Ambrose Slone, Daniel Smilkov, David R. So, Daniel Sohn, Simon Tokumine, Dasha Valter, Vijay Vasudevan, Kiran Vodrahalli, Xuezhi Wang, Pidong Wang, Zirui Wang, Tao Wang, John Wieting, Yuhuai Wu, Kelvin Xu, Yunhan Xu, Linting Xue, Pengcheng Yin, Jiahui Yu, Qiao Zhang, Steven Zheng, Ce Zheng, Weikang Zhou, Denny Zhou, Slav Petrov, and Yonghui Wu. 2023. Palm 2 technical report.

Francesco Barbieri, Luis Espinosa Anke, and Jose Camacho-Collados. 2022. XLM-T: Multilingual language models in Twitter for sentiment analysis and beyond. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 258–266, Marseille, France. European Language Resources Association.

Damian Blasi, Antonios Anastasopoulos, and Graham Neubig. 2022. Systematic inequalities in language technology performance across the world's languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5486–5505, Dublin, Ireland. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Tyler A Chang and Benjamin K Bergen. 2023. Language model behavior: A comprehensive survey. *arXiv preprint arXiv:2303.11504*.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. Palm: Scaling language modeling with pathways.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Ethnologue. 2022. What are the largest language families?

Fahim Faisal, Orevaoghene Ahia, Aarohi Srivastava, Kabir Ahuja, David Chiang, Yulia Tsvetkov, and Antonios Anastasopoulos. 2024. Dialectbench: A nlp benchmark for dialects, varieties, and closely-related languages.

Ayyoob ImaniGooghari, Peiqin Lin, Amir Hossein Kargaran, Silvia Severini, Masoud Jalili Sabet, Nora Kassner, Chunlan Ma, Helmut Schmid, André Martins, François Yvon, and Hinrich Schütze. 2023. Glot500: Scaling multilingual corpora and language models to 500 languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1082–1117, Toronto, Canada. Association for Computational Linguistics.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Viet Dac Lai, Nghia Trung Ngo, Amir Pouran Ben Veyseh, Hieu Man, Franck Dernoncourt, Trung Bui, and Thien Huu Nguyen. 2023. Chatgpt beyond english: Towards a comprehensive evaluation of large language models in multilingual learning. *arXiv preprint arXiv:2304.05613*.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona Diab, Veselin Stoyanov, and Xian Li. 2022. Few-shot learning with multilingual language models.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual Denoising Pre-training for Neural Machine Translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Henry B Mann and Donald R Whitney. 1947. On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, pages 50–60.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2023. Crosslingual generalization through multitask finetuning.

OpenAI. 2023. Gpt-4 technical report.

Fred Philippy, Siwen Guo, and Shohreh Haddadan. 2023. Towards a common understanding of contributing factors for cross-lingual transfer in multilingual language models: A review. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5877–5891, Toronto, Canada. Association for Computational Linguistics.

Surangika Ranathunga and Nisansa de Silva. 2022. Some languages are more equal than others: Probing deeper into the linguistic disparity in the nlp world. *arXiv preprint arXiv:2210.08523*.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Oleh Shliazhko, Alena Fenogenova, Maria Tikhonova, Vladislav Mikhailov, Anastasia Kozlova, and Tatiana Shavrina. 2022. mgpt: Few-shot learners go multilingual. *arXiv preprint arXiv:2204.07580*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models.

Don Vaughan. 2020. The world's 5 most commonly used writing systems.

Wenxuan Wang, Zhaopeng Tu, Chang Chen, Youliang Yuan, Jen-tse Huang, Wenxiang Jiao, and Michael R Lyu. 2023. All languages matter: On the multilingual safety of large language models. *arXiv preprint arXiv:2310.00905*.

Shijie Wu and Mark Dredze. 2020. Are all languages created equal in multilingual bert? *arXiv preprint arXiv:2005.09093*.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Wen Yang, Chong Li, Jiajun Zhang, and Chengqing Zong. 2023. Bigtrans: Augmenting large language models with multilingual translation capability over 100 languages. *arXiv preprint arXiv:2305.18098*.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Lingpeng Kong, Jiajun Chen, Lei Li, and Shujian Huang. 2023. Multilingual machine translation with large language models: Empirical results and analysis. *arXiv preprint arXiv:2304.04675*.

A Appendix: Prompts

This appendix provides the specific prompts used for evaluating `Bloom`, `Bloomz`, `XGLM`, and `GPT-3.5` in the zero-shot and two-shot in-context learning (ICL) settings on the SIB-200 text classification task.

Zero-shot Prompt (Bloom, Bloomz, XGLM):

```
SENTENCE: “{input sentence}”
Is this SENTENCE science, travel, politics,
sports, health, entertainment, geography?
OPTIONS:
-science
-travel
-politics
-sports
-health
-entertainment
-geography
ANSWER:
```

Two-shot ICL Prompt (Bloom, Bloomz, XGLM):

```
What category does SENTENCE belong to?

SENTENCE: “{sentence1}”
LABEL: {label1}
SENTENCE: “{sentence2}”
LABEL: {label2}
...
SENTENCE: “{sentence14}”
LABEL: {label14}
SENTENCE: “{input sentence}”
OPTIONS:
-science
-travel
-politics
-sports
-health
-entertainment
-geography

LABEL:
```

Zero-shot Prompt (GPT-3.5):

```
You will be provided with a text, and
your task is to classify its category
as science, travel, politics, sports,
health, entertainment, geography.
{input sentence}

Category:
```

B Appendix: Supplemental plots

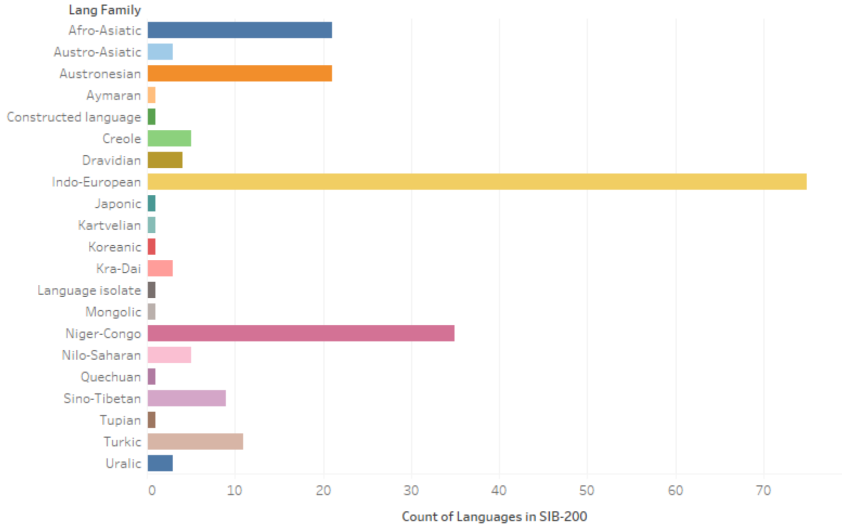


Figure 4: Distribution of language family in SIB-200.

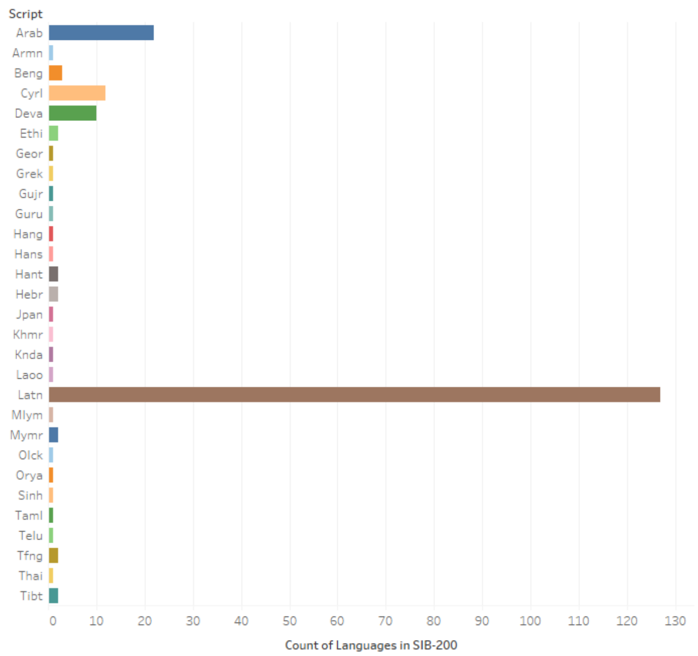


Figure 5: Distribution of scripts in SIB-200.

Does Whisper Understand Swiss German? An Automatic, Qualitative and Human Evaluation

Eyal Liron Dolev
Linguistics Center Zurich
German Department
University of Zurich
eyalliron.dolev@uzh.ch

Clemens Fidel Lutz
Department of Computational Linguistics
Phonetics & Speech Sciences
University of Zurich
clemensfidel.lutz@uzh.ch

Noëmi Aepli
Department of Computational Linguistics
University of Zurich
noemi.aepli@uzh.ch

Abstract

Whisper is a state-of-the-art automatic speech recognition (ASR) model (Radford et al., 2022). Although Swiss German dialects are allegedly not part of Whisper's training data, preliminary experiments showed that Whisper can transcribe Swiss German quite well, with the output being a speech translation into Standard German. To gain a better understanding of Whisper's performance on Swiss German, we systematically evaluate it using automatic, qualitative, and human evaluation. We test its performance on three existing test sets: SwissDial (Dogan-Schönberger et al., 2021), STT4SG-350 (Plüss et al., 2023), and Swiss Parliaments Corpus (Plüss et al., 2021). In addition, we create a new test set for this work, based on short mock clinical interviews.

For automatic evaluation, we used word error rate (WER) and BLEU. In the qualitative analysis, we discuss Whisper's strengths and weaknesses and analyze some output examples. For the human evaluation, we conducted a survey with 28 participants who were asked to evaluate Whisper's performance.

All of our evaluations suggest that Whisper is a viable ASR system for Swiss German, so long as the Standard German output is desired.

1 Introduction

Swiss German is the name of a group of Alemannic (High German) dialects spoken in German-speaking Switzerland by around 5.5 million people.[1] German-speaking Switzerland displays a state of diglossia (Ferguson, 1959; Rash, 1998), more specifically a medial diglossia: spoken contexts evoke Swiss German, written contexts evoke Standard German (Kolde, 1983; Haas, 2004). According to this principle, Swiss German is used as a spoken language in almost all settings, with the exception of some restricted, specific formal settings in which Standard German is spoken, e.g., on the news or at school, as well as a *lingua franca* with non Swiss German speakers (Hogg et al., 1984).

Swiss German has no spoken standard variety and no written variety, and therefore no orthographic norms. In writing, Standard German is used. Thus, whenever spoken language (Swiss German) has to be written down, e.g., subtitles to a TV program or minutes of a meeting, Standard German is used. If Swiss German is written, it happens in situations that are conceptually spoken ("*konzeptionell mündlich;*" Koch and Oesterreicher, 1994) and which are situated on the immediacy-end of Koch and Oesterreichs's communication model (*Nähe-Distanz-Modell*, cf. Koch and Oesterreicher, 1985), e.g., ads or chat messages, cf. Ueberwasser and Stark (2017), which created a corpus of text messages written in Swiss German.

To summarize, in German-speaking Switzerland there is a state of medial diglossia: the spoken language is Swiss German (a group of dialects with no standard variety); the written language is a Swiss vareity of Standard German. Swiss German and Standard German are, although genetically and systematically very close, two different languages, whereas only Standard German has a codified written form (Berthele, 2004). The task of putting down Swiss German speech to written form is, therefore, not a transcription task, but rather a translation task, translating Swiss German to Standard German. This spoken–written juxtaposition of Swiss German and Standard German explains why almost all the automatic speech recognition

[1] Bundesamt für Statistik: Hauptsprachen seit 1910, accessed on 23.04.2024

Proceedings of the Eleventh Workshop on NLP for Similar Languages, Varieties, and Dialects (VarDial 2024), pages 28–40

efforts for Swiss German until now have dealt with Swiss German speech to Standard German text (see Section 2).

Whisper is a state-of-the-art multilingual model for automatic speech recognition (ASR) (Radford et al., 2022). Although Swiss German is not officially part of Whisper's training data[2], in preliminary trials, we observed that Whisper could recognize Swiss German quite well, with the output produced being Standard German. According to Ruder (2024) most large language models (LLMs) have likely encountered some data for most languages available on the web, which is probably the case here too.

We intentionally refrain from attempting to fine-tune Whisper. Not only did Sicard et al. (2023)'s fine-tuning attempts of Whisper on Swiss German data worsen the model's performance; we find Whisper's zero-shot performance on Swiss German, at this stage, already impressive and applicable. Before any costly GPU hours are spent in an attempt to improve Whisper, we think it should first be scrutinized and analyzed in its current state.

In this work, we evaluate Whisper's performance on Swiss German audio in different settings and modes. We automatically evaluated Whisper on three large corpora, namely SPC (Plüss et al., 2021), STT4SG-350 (Plüss et al., 2023), and SwissDial (Dogan-Schönberger et al., 2021), measuring word error rate (WER) and BLEU.

To test Whisper on real-life spoken language, we created a new test set for which we translated into Standard German mock clinical interviews held in Swiss German. The total length of the interviews is approx. 30 minutes. To test Whisper's performance on this test set, we offer a qualitative analysis of Whisper's output and a human evaluation based on a survey ($n = 28$).

2 Previous Work

ASR for Swiss German is an ambiguous term. While the audio input to the system is always Swiss German, the text output can be: (a) dialectal writing – loosely phonemic representation of Swiss German; (b) normalized writing – transcriptions resembling standard German that are relatively consistent but distant from the acoustic signal (Nigmatulina et al., 2020); (c) Standard German translation.

In recent years, Swiss German has enjoyed a proper boom in the field of speech corpora, ASR and speech generation. The first major corpus with Swiss German audio was ArchiMob, which includes dialectal as well as normalized writing (Samardžić et al., 2016; Scherrer et al., 2019). Nigmatulina et al. (2020) used the ArchiMob corpus to compare systems producing dialectal and normalized writing and concluded that performance is better with standardized writing. Dogan-Schönberger et al. (2021) created SwissDial, a large corpus containing Standard German as well as Swiss German transcriptions in eight dialects.

Some work concentrated on ASR with Standard German speech translation and leveraged existing Transformer and XLS-R ASR models, fine-tuning them with Swiss German data. Plüss et al. (2021) published the "Swiss Parliaments Corpus", and experimented further with ASR models for Swiss German with Standard German output. Plüss et al. (2022) presented SDS-200, a corpus of Swiss German dialectal speech with Standard German text translations containing 200 hours of speech. They also experimented with training Transformer models and fine-tuning Wav2Vec2 XLS-R models on their data. Their best model (XLS-R) reached a WER of 21.6 and a BLEU score of 64.0. Recently, Plüss et al. (2023) presented the as-of-today largest corpus of Swiss German dialectal speech with Standard German text, containing 343 hours of speech. They fine-tuned a Wav2Vec2 XLS-R model on the corpus and reached a WER of 14.0 and a BLEU score of 74.7 on their test set.

Most recently, Sicard et al. (2023) turned to Whisper and tested it in a zero-shot setting on select Swiss German/Standard German test sets (SwissDial, SDS-200, SPC). Reportedly, their fine-tuning experiments on Whisper (medium version) worsened performance, leading the model to suffer from catastrophic forgetting.

3 Test Sets

To evaluate Whisper's performance on Swiss German, we test it using WER and BLEU on three test sets: SwissDial (Dogan-Schönberger et al., 2021), Swiss Parliaments Corpus (Plüss et al., 2021), STT4SG-350 (Plüss et al., 2023). We additionally created a new test set based on short Swiss German mock clinical interviews, which we additionally evaluate using a qualitative analysis and a human survey.

[2]https://github.com/openai/whisper, accessed on 23.04.2024

3.1 Mock Clinical Interviews

This work serves as a preparation step towards a large longitudinal study in the field of suicide prevention.[3] During this study, patients from a Zurich-based psychiatric clinic will be interviewed several times. We test how reliable and viable Whisper is for transcribing/translating these interviews.

To this end, i.e., to test Whisper in a naturalistic and applied setting containing spontaneous speech, we used mock clinical interviews that were held in Swiss German and recorded for instructional and training purposes in a total length of approx. 30 minutes. The interviews were recorded with three women interviewees using a lapel microphone[4] and simple convertible laptops. We, the authors of this work, then translated these interviews into Standard German according to some basic translation guidelines we created to maintain consistency. We call this ad-hoc test set "Mock Clinical Interviews".

This test set will be automatically evaluated using WER and BLEU as well as using a qualitative analysis to discuss Whisper's strengths and weaknesses in Swiss German, and a human evaluation, for which we conducted a survey ($n = 28$).

3.2 SwissDial

For the creation of SwissDial, eight speakers, speaking eight different dialects[5] were asked to translate Standard German prompts to their own dialects and then record the translations. The prompts were made of sentences crawled from the internet, encompassing different text genres: news stories, Wikipedia articles, weather reports and short stories (Dogan-Schönberger et al., 2021). Because the prompts were translated into Swiss German by each of the speakers, sometimes greater departures from the Standard German source occur. See Figure 1, containing the first three dialect entries from the first entry in the corpus, for an example.

As can be seen in Figure 1, the German word *derzeit* "currently" was translated by the different dialect speakers as *zur Ziit*, *momentan* and *derziit*, respectively. One cannot, however, expect that Whisper translates all of these different Swiss German words back to the Standard German original, especially considering that the Swiss German words each have a closer Standard German equivalent (*zur Zeit*, *momentan* and *derzeit*, respectively).

To circumvent this problem and include prompts

```
{
  "id": 0,
  "de": "Derzeit␣ist␣er␣in␣
      ↪ \"Parasite\",␣dem␣
      ↪ Siegerfilm␣von␣Cannes,␣zu␣
      ↪ sehen.",
  "ch_sg": "Zur␣Ziit␣isch␣er␣in␣
      ↪ \"Parasite\",␣en␣
      ↪ Siegerfilm␣vo␣Cannes,␣
      ↪ zgseh.",
  "ch_be": "Momentan␣ischer␣in␣
      ↪ \"Parasite\"␣z␣gseh,␣em␣
      ↪ Siegerfium␣vo␣Cannes.",
  "ch_gr": "Derziit␣isch␣er␣in␣
      ↪ \"Parasite\",␣am␣
      ↪ Siegerfilm␣vu␣Cannes,␣z␣
      ↪ gseh.",
    ...
}
```

Figure 1: The first three dialectal translations of the first entry in the SwissDial corpus. The first word in the Standard German source ("de"), *derzeit*, is translated differently in each dialect: *zur ziit*, *momentan*, *derziit*.

that are less likely to contain major departures from the source, which might unfairly fail Whisper when the produced output is compared to the original prompt, we created an ad-hoc test set: We calculated for each Standard German prompt and its respective dialectal translations the chrF score (Popović, 2015) using SacreBLEU's implementation (Post, 2018). We then evaluated Whisper's performance on the 500 prompts with the best chrF scores for each dialect.

3.3 Swiss Parliaments Corpus

The Swiss Parliament Corpus (SPC) is a dataset containing sentences taken from speeches held at the Grand Council (Grosser Rat) of the Canton of Bern (Plüss et al., 2021).[6] Almost all speakers hold their speeches in Bernese German. For the creation of the corpus, Plüss et al. (2021) split the audio into segments, so-called sentences, whereas segments shorter than one second and longer than 15 seconds were discarded. The corpus creators also made sure that the speech segments were unique within the set. The speech segments were then force-aligned to the Standard German minutes (i.e., translations), which were created by the Canton of Bern. The result is parliament speeches split into segments (sentences) with their corresponding Standard German transla-

[3] MULTICAST

[4] RØDE smartLav+

[5] The dialects of Zurich, Bern, Basel, Aargau, Grisons, St. Gallen, Lucerne, and the Walser

[6] The name of the corpus is thus a misnomer – it is not a corpus representing the whole diversity of Swiss German.

tions from the minutes. We tested Whisper on the test set part of the corpus[7].

3.4 STT4SG-350

Like SPC, STT4SG-350[8] is a corpus containing single sentences of Swiss German speech with Standard German translations (Plüss et al., 2023). Unlike the former, STT4SG-350 includes an almost even split between seven different dialect regions.[9] The sentences produced by speakers were taken from Swiss newspapers and proceedings of two Swiss Parliaments. Participants, who were recruited either via a crowdsourcing platform or academic or personal channels as well as news ads, self-reported their dialect region, age group, gender, and where they grew up and/or went to school. The whole corpus consists of 343 hours of speech. We tested Whisper on the test set part which contains 34 hours of speech in approx. 25k sentences.

4 Evaluation

4.1 Automatic Evaluation

Usually, word error rate (WER) is used as a metric to automatically evaluate ASR systems. However, the type of ASR for Swiss German that we evaluate in this work is Swiss German audio input with Standard German text output – a speech translation task. This means, as is generally the case in translation, that it is not uncommon for a sentence to have several possible translations. Standard German translations of Swiss German are, in that sense, no different, although in many cases, there are clear one-to-one correspondences in vocabulary and grammatical structures between Standard and Swiss German. But when correspondences are ambiguous, the translator has to make a conscious decision on how to translate vocabulary or grammatical constructions. For example, Swiss German only has one tense referring to past events – the perfect. Standard German has, at least formally, two past tenses – the perfect and the preterite. The translator thus has to choose, according to context, how to translate the Swiss German perfect.

This ambiguity in translation, a typical problem in evaluating machine translation systems, makes the usual metric used for ASR systems – word error rate (WER) – not unproblematic. We thus additionally use BLEU (Papineni et al., 2002), a typical metric used to evaluate machine translation systems. This will also help compare the performance of Whisper to previous Swiss German ASR models, as previous work also reports WER and BLEU.

To compute WER, we used JiWER's[10] implementation. For BLEU we used SacreBLEU's implementation (Post, 2018).

Mode	WER	BLEU
Continuous recordings	**0.33**	**52.03**
Segmented clips	0.37	44.19

Table 1: Whisper's performance on our *Mock Clinical Interviews* test set, comparing continous recordings vs. segmented clips. Best results in bold.

4.2 Qualitative & Human Evaluation

In addition to testing Whisper's performance on several datasets and evaluating its performance automatically, we offer a qualitative and human evaluation of our Mock Clinical Interviews (see Section 3.1). In the qualitative evaluation, we will show examples of Whisper's output, analyze errors, and shed some light on the strengths and weaknesses of Whisper's performance.

Our human evaluation, in which we recruited 28 people – university students, colleagues, and acquaintances – via personal channels to evaluate Whisper's output, offers more informative feedback about how humans perceive Whisper's output.

5 Results: Automatic Evaluation

5.1 Mock Clinical Interviews

We tested Whisper's large-v3 model on our test set ("Mock Clinical Intverviews", see Section 3.1). We compared Whisper's performance on continuous recordings versus short clips containing single speech segments. Given segmented clips, WER and BLEU scores were 0.37 and 44.19, respectively. With the continuous recordings, WER and BLEU scores were 0.33 and 52.03, respectively, see Table 1. We conclude that Whisper performs better on longer, continuous recordings than on short clips.

This comes, however, at a slight risk of hallucinations: Four out of sixteen transcriptions/translations generated by Whisper included one sentence

[7] 6 hours, 3332 segments

[8] Standing for "Speech-to-text for Swiss German"

[9] These seven regions are Basel, Bern, Grisons, Central Switzerland, East Switzerland, Valais and Zurich.

[10] https://github.com/jitsi/jiwer, accessed on 23.04.2024

that was not uttered in the original audio, see Section 6.3 for more details.

5.2 SPC, STT4SG & SwissDial

We further tested Whisper's large-v3 model on the three other test sets: SPC, STT4SG-350, and SwissDial (see Section 3). The results, compared to results reported by other works, can be seen in Table 2. We always picked the best result reported in each of the other works.

Whisper's latest large model, version 3, outperforms Whisper's previous model, as reported by Sicard et al. (2023). However, fine-tuned Wav2Vec2 models on the SPC and the STT4SG-350 training sets outperform Whisper on the respective test sets, as reported by Plüss et al. (2023) and Schraner et al. (2022). Whisper does come close to the Conformer model pre-trained by Plüss et al. (2021) with a difference of only 1.7 *p.p.* and 1.6 *p.p.* in WER and BLEU, respectively.

For SPC, STT4SG-350, and SwissDial, we also computed WER and BLEU for each sentence separately and then computed the mean and standard deviation (so-called micro average). As can be seen in Table 3, the standard deviations for WER and BLEU are quite big, ranging at 0.24–0.25 for WER and 27.95–32.24 for BLEU. This shows that Whisper's performance measured in WER and BLEU fluctuates considerably. For some sentences in STT4SG-350 for example, BLEU scores went up to 100. See also Figures 2 and 3 in Appendix A.4.

It should be noted that the SPC corpus contains some considerable deviations between audio and reference translations, which were taken from the parliament's proceedings (see Section 3.3). For instance, in one clip[11], the heard audio is *und das isch schlächt*. The reference translation is "Das ist schlecht", excluding the coordinating conjunction *und* "and". Whisper perfectly transcribed this as "Und das ist schlecht", but this is penalized with a WER score of 0.33. It is not inconceivable, that the models trained by Plüss et al. (2021) learned these deviating translations, which might explain their better performance on the SPC test set. As the case may be, comparing WER and BLEU scores for SPC between Whisper's performance and Plüss et al. (2021) may raise concerns, and its meaningfulness can and should be questioned. For more examples of perfect output by Whisper penalized by diverging reference translations, see Table 10 in Appendix A.

For SwissDial, we also evaluated Whisper's performance on the different dialects. As can be seen in Table 4, the Grisons dialect has the best WER and BLEU scores; the Walser dialect has the worst scores.[12] Why Whisper performs differently on different dialects and which phonetic, phonological or grammatical traits affect Whisper's performance should be more closely examined in future work.

To conclude, we consider Whisper's results impressive, especially considering that it operates in a zero-shot setting. Its output is without doubt meaningful and useful.

6 Qulatitative Analysis

6.1 General Impression

In general, we were genuinely impressed with Whisper's performance. The Standard German output corresponds in meaning and style to the Swiss German audio to almost the full extent. Whisper generated entire error-free passages that are fluent, consistent in style, retain the original word order and correspond fully to the original (see Table 6 in Appendix A.1 for examples).

However, some things are not always consistent. For example, the Swiss German perfect tense is translated sometimes as the Standard German perfect tense and sometimes as the preterite. The output switches inconsistently between the two past forms within the same passage. See Table 7 in Appendix A.1 for examples.

We noticed that in certain cases, words are changed when translated to Standard German, even when the Swiss German word has an identical corresponding word in Standard German. One example of this is the Swiss German word *lässig* which is translated to Standard German *toll*, both meaning "cool, nice". In this case, this is desired since in Standard German, *lässig* means rather "casual, easy-going" – Swiss German *lässig* and Standard German *lässig* are false friends. Another example is the translation of Swiss German *Sache* to Standard German *Dinge*, both meaning "things", however, *Dinge* is used mostly for tangible things and in the given contexts *Sachen* would have been a better translation.

[11] `82495971-6523-4f96-be13-753b8bb564cf.flac`

[12] The Walser dialect is also considered in Switzerland the most difficult to understand.

Test Set	Model	Mode	WER	BLEU	
Mock Interviews	Whisper large-v3	zero-shot	0.372	44.3	This work
SPC	Conformer	pre-trained	0.278	58.6	Plüss et al. (2021)
	Wav2Vec2	**fine-tuned**	**0.237**	**60.7**	Schraner et al. (2022)
	Whisper large	zero-shot	0.332	55.6	Sicard et al. (2023)
	Whisper large-v3	zero-shot	0.295	57.0	This work
STT4SG-350	XLS-R	fine-tuned	0.153	72.2	Schraner et al. (2022)
	Wav2Vec2	**fine-tuned**	**0.140**	**74.7**	Plüss et al. (2023)
	Whisper large-v3	zero-shot	0.230	63.1	This work
SwissDial	Whisper large	zero-shot	0.294	56.2	Sicard et al. (2023)
	Whisper large-v3	**zero-shot**	**0.230**	**61.0**	This work

Table 2: WER (lower is better) and BLEU (higher is better) scores for our corpora, compared to results reported in previous works.

Test Set	WER	BLEU
SPC	0.30 (0.24)	54.01 (27.95)
STT4SG-350	0.24 (0.25)	60.61 (32.24)
SwissDial	0.25 (0.24)	57.23 (31.51)

Table 3: Mean and standard deviation WER and BLEU for the corpora when computed for each sentence separately.

Dialect	WER	BLEU
Aargau	0.272	55.40
Bern	0.210	64.95
Basel	0.209	63.24
Grisons	0.169	69.99
Lucerne	0.276	55.06
St. Gallen	0.209	64.03
Walser	0.297	53.46
Zurich	0.229	60.67

Table 4: WER and BLEU scores for each dialect in the SwissDial corpus.

6.2 Concise Style

We notice that Whisper's translations are of a style that is more concise than the original. This is especially noticeable in the removal of modal particles and conjunctions: Modal particles with little semantic content but with an information structural function like *halt* or *einfach* might disappear from the output. Conjunctions like *dann* "then" or *und* "and" are not always included. In one case, however, conjunctions and particles were hallucinated by Whisper. Whisper deals then inconsistently with particles and conjunctions, mostly removing them but rarely also adding them by hallucination.

It is a known phenomenon that during translation, the explicitness of cohesive markers, such as the particles and conjunctions mentioned above, can shift (Blum-Kulka, 1986). Leaving out such markers, as evidenced in Whisper's output, can be seen as a case of implicitation, cf. Lapshinova-Koltunski et al. (2022) (which refers to them as "discourse connectives"). If we assume that the target side of the training data was more concise and less explicit than the spoken Swiss German, then this would explain Whisper's behavior.

It should, however, be noted that such modal particles usually serve an information structural function (Musan, 2010). Thus, they do not necessarily affect the truth value of an utterance and, therefore, have little influence on the overall meaning (Krifka, 2007). For examples of removed particles, see Table 8 in Appendix A.1.

6.3 Hallucinations

Four out of sixteen transcriptions of whole conversations contained hallucinations – a sentence that was generated by Whisper without a corresponding utterance in the source audio.

In one conversation, in which the interviewee recounted the death of her mother, the following sentence was hallucinated:

> *Sie blieb nicht mehr in unserem pegen... Meine Frau, die ich so sehr liebte.* ("She didn't remain in our GIBBERISH... My wife, whom I loved so much.")

In another conversation, a sentence was continued by a hallucination (marked in bold):

> *Ähm ... Ja, jetzt bin ich immer noch etwas groggy, aber es geht etwas.* ***Ich bin ganz müde. Äh ...***

> ***Okay, ich kann ... Äh ... Zuerst schon.***, ("Ehm ... Yes, now I'm still somewhat groggy, but I'm managing. **I am really tired. Eh ... Okay, I can ... Eh ... Firstly.**")

In a different case, a sentence was preceded by a hallucination (in bold):

> ***Und ... äh ... Das hat mich sehr angestrengt. Äh ...*** *Das hat mich sehr viel aufgewühlt.* ("**And ... eh ... That really strained me. Eh ...** That really upset me.")

At the end of one conversation, *Untertitel von S G*[13] ("Subtitles by...") was added.[14]

We couldn't identify a pattern as to when and why hallucinations happen, but they seem to be a generally known problem with Whisper and are not specific to Swiss German audio.[15] Therefore, users should be aware that there is a possibility of hallucinations being added and in doubt re-check the audio.

7 Human Evaluation

7.1 Motivation

Performance of ASR systems is usually reported in WER , cf. Radford et al. (2022); Baevski et al. (2020). However, it is less meaningful for evaluating ASR for Swiss German with Standard German output since several outputs can be considered correct (see also Sections 1 and 4.1). Therefore, BLEU established itself as a second metric reported in works on ASR for Swiss German (Plüss et al., 2022, 2023; Schraner et al., 2022).

BLEU is meaningful mostly as a relative metric, comparing several systems; as an absolute score, it is less meaningful. It has been the object of criticism since Callison-Burch et al. (2006). Even its significance as a relative metric use has been harshly criticized, with Kocmi et al. (2021) complaining that "the sole use of BLEU impeded the development of improved models leading to bad deployment decisions." If we acknowledge that language technology is made for human beings, then its most important evaluation should be what humans think about it. We therefore conducted a short survey to evaluate how human beings perceive Whisper's performance.

[13]Whisper's output included a real person's name, which we anonymize here for privacy reasons.

[14]Obviously due to subtitles being part of the training data.

[15]A DuckDuckGo search for "openai whisper hallucination" returns many web pages discussing the issue.

7.2 Survey

For the survey, we randomly picked three of the conversations recorded as Mock Clinical Interviews (see Section 3.1) and extracted 119 sentence pairs consisting of our reference translation (sentence A) and Whisper's output (sentence B).

In the evaluation task, participants were asked, on a scale of 1 to 5, to rate:

1. To what extent is the meaning of sentence A retained in sentence B?
2. To what extent is sentence B fluent and natural?

with 1 being the worst and 5 being the best score. To assist the participants, each grade on the scale was given a verbal description (see Table 9 in Appendix A.2 for details). The participants were instructed to rate the fluency of sentence B (Whisper's output) independently from sentence A (reference) and to ignore punctuation marks.

Twenty-eight university students, colleagues, and acquaintances, who were recruited via personal channels, participated in the survey, all of them native speakers of German or Swiss German. The mean scores for meaning and fluency among all raters were 4.358 $\pm$ 0.046 (*SD* 0.239) and 4.39 $\pm$ 0.074 (*SD* 0.387), respectively, out of a maximum of 5 points. These scores suggest a very high human satisfaction with Whisper's performance.

7.3 Worst Rated Sentences

In an attempt to uncover some of Whisper's weaknesses, we picked the six sentences with the lowest mean score across all raters, see Table 5.

In sentence 1, the output includes the word *Riesiges* "huge" instead of the original *Kleines* "small", which is the exact opposite. In sentence 2, the subject of the sentence changes from the original *ich* "I" to *sie* "they", and the verb changes from *genommen* "took" to *liess* "let", causing the output to diverge greatly in meaning from the reference. Also, the preposition changes from *zur* in the reference to *an der* in Whisper's output. In sentence 3, the name of a train line in Zurich (*Forchbahn*) is "misheard" as *Furchtbahn* "fright train". Sentence 4 diverges greatly from the reference, with the use of the 3rd person accusative pronoun *ihn* without first introducing its referent, resulting in a genuine *non sequitur*. In sentence 5, the time mentioned in the original (*viertel nach sechs* "quarter past six") was changed to *4.15*. Finally, in sentence 6, the word *Schlummer-Taste* "snooze button" was

	Reference	Whisper	Mean
1	also meistens etwas **Kleines**, weil ich am Abend nicht so hunger habe	Also meistens etwas **Riesiges**, weil ich am Abend nicht so Hunger habe.	2.31
2	und habe meine Sachen genommen und dann **bin ich** auf den Bus gelaufen also **zur** Bushaltestelle	Ich nahm meine Sachen und **liess sie** auf den Bus. Also **an der** Bushaltestelle.	2.50
3	Auf das Tram gegangen, auf die **Forchbahn** und dann hierher gekommen.	Auf die **Furchtbahn** gingen. Dann kam ich hierher.	2.54
4	und dann bin ich bis zum **Stadelhofen** gefahren und habe ich noch die Tram genommen und dann bin ich hierher gekommen	Ich fuhr bis zum **Stadelhof** und nahm **ihn** aus der Rampe. Dann kam ich hierher.	2.58
5	Ja also ich bin um **viertel nach sechs** aufgestanden	Ich bin um **4.15** Uhr aufgestanden.	2.65
6	Da habe ich aber zuerst noch die **Schlummer-Taste** gedrückt zweimal.	Ich habe zuerst die **Schlamasseltasche** gedrückt, zweimal.	2.69

Table 5: The six sentences rated the worst in the human evaluation.

misheard as *Schlmasseltasche*, a gibberish word meaning “bad luck bag”.

There is no recurring pattern in these sentences. It seems, however, that the transcription of named entities (*Forchbahn*, *Stadelhofen*, cf. sentences 3 and 4) and numbers (cf. sentence 5) might result in errors.

8 Conclusion

We evaluated Whisper’s performance on Swiss German audio using automatic evaluation (WER and BLEU), a qualitative analysis and a human survey. All three evaluation types are evidential of very high performance: WER and BLEU are on par or slightly below other systems (cf. Table 2). The qualitative analysis revealed very high quality, retaining almost always the original meaning with only slight changes in style and some removal of cohesion markers such as particles and connectors. The human evaluation showed high human satisfaction (mean: 4.36/5.00, n=28).

We are therefore convinced that Whisper can be used, as is and out-of-the-box, without any further adaptations, for transcribing Swiss German, providing that the desired output is Standard German and that some loss of cohesion markers is acceptable.

However, as with any AI-based tool, Whisper should be used with caution. The qualitative analysis revealed some cases of changes in meaning, especially of numbers, as well as some hallucinations, though these were rare (one sentence in four out of sixteen 2-minute clips). In case of doubt, users should always refer to the original audio. Nevertheless, for the task of transcribing large portions of Swiss German audio or as a first step in a pipeline with other tasks in mind, such as keyword extraction or sentiment analysis, we think Whisper is a helpful, useful, and viable ASR tool.

Limitations

In this work, we evaluated Whisper’s performance on Swiss German using automatic evaluation (WER and BLEU). We restricted ourselves to these metrics, since these are the metrics that are reported in previous works on ASR for Swiss German. Granted, other potentially better-suited metrics also come to mind, e.g., chrF (Popović, 2015) and BERTScore (Zhang et al., 2020). However, since models from previous works are not publicly available, we could not test them using different metrics besides WER and BLEU and had to rely on the scores reported in the respective works. Previous models not being publicly available also explains why we could not test the performance of previous models on our own test set (Mock Clinical Interviews), which would have been desirable.

Acknowledgements

Eyal Liron Dolev is a doctoral student at the MULTICAST project, Swiss National Science Foundation project no. 205913. He would like to express his gratitude to his doctoral supervisor, Prof. Guido Seiler, for allowing him the freedom to conduct this work. This work was also supported by the Swiss National Science Foundation, project no. 191934. We thank the FHNW Institute for Data Science for making the SPC corpus available, as well as SwissNLP for making the STT4SG-350 dataset available. We also kindly thank the reviewers for their valuable feedback.

References

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations.

Raphael Berthele. 2004. Vor lauter Linguisten die Sprache nicht mehr sehen – Diglossie und Ideologie in der deutschsprachigen Schweiz. In Helen Christen, editor, *Dialekt, Regiolekt und Standardsprache im Sozialen und Zeitlichen Raum*, pages 111–136. De Gruyter, Vienna.

Shoshana Blum-Kulka. 1986. Shifts of cohesion and coherence in translation. *Interlingual and intercultural communication: Discourse and cognition in translation and second language acquisition studies*, 272:17.

Chris Callison-Burch, Miles Osborne, and Philipp Koehn. 2006. Re-evaluating the role of Bleu in machine translation research. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 249–256, Trento, Italy. Association for Computational Linguistics.

Pelin Dogan-Schönberger, Julian Mäder, and Thomas Hofmann. 2021. Swissdial: Parallel multidialectal corpus of spoken Swiss German.

Charles A. Ferguson. 1959. Diglossia. *Word*, 15(2):325–340.

Walter Haas. 2004. Die sprachsituation der deutschen Schweiz und das Konzept der Diglossie. In Helen Christen, editor, *Dialekt, Regiolekt und Standardsprache im Sozialen und Zeitlichen Raum*, pages 81–110. De Gruyter, Vienna.

Michael A. Hogg, Nicholas Joyce, and Dominic Abrams. 1984. Diglossia in Switzerland? A Social Identity Analysis of Speaker Evaluations. *Journal of Language and Social Psychology*, 3(3):185–196.

Peter Koch and Wulf Oesterreicher. 1985. Sprache der Nähe — Sprache der Distanz. Mündlichkeit und Schriftlichkeit im Spannungsfeld von Sprachtheorie und Sprachgeschichte. *Romanistisches Jahrbuch*, 36(1):15–43.

Peter Koch and Wulf Oesterreicher. 1994. Schriftlichkeit und Sprache.

Tom Kocmi, Christian Federmann, Roman Grundkiewicz, Marcin Junczys-Dowmunt, Hitokazu Matsushita, and Arul Menezes. 2021. To ship or not to ship: An extensive evaluation of automatic metrics for machine translation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 478–494, Online. Association for Computational Linguistics.

Gottfried Kolde. 1983. Sprachkontakte in gemischtsprachigen Städten: vergleichende Untersuchungen über Voraussetzungen und Formen sprachlicher Interaktion verschiedensprachiger Jugendlicher in den Schweizer Städten Biel/Bienne und Fribourg/Freiburg i. Ue. volume 37, Wiesbaden.

Manfred Krifka. 2007. Basic Notions of Information Structure. pages 13–56.

Ekaterina Lapshinova-Koltunski, Christina Pollkläsener, and Heike Przybyl. 2022. Exploring Explicitation and Implicitation in Parallel Interpreting and Translation Corpora. *The Prague Bulletin of Mathematical Linguistics*, 119:5–22.

Renate Musan. 2010. *Informationsstruktur*, volume 9 of *Kurze Einführung in Die Germanistische Linguistik*. Universitätsverlag Winter, Heidelberg.

Iuliia Nigmatulina, Tannon Kew, and Tanja Samardžić. 2020. ASR for non-standardised languages with dialectal variation: the case of Swiss German. In *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 15–24, Barcelona, Spain (Online). International Committee on Computational Linguistics (ICCL).

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Michel Plüss, Jan Deriu, Yanick Schraner, Claudio Paonessa, Julia Hartmann, Larissa Schmidt, Christian Scheller, Manuela Hürlimann, Tanja Samardžić, Manfred Vogel, and Mark Cieliebak. 2023. STT4SG-350: A speech corpus for all Swiss German dialect regions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1763–1772, Toronto, Canada. Association for Computational Linguistics.

Michel Plüss, Manuela Hürlimann, Marc Cuny, Alla Stöckli, Nikolaos Kapotis, Julia Hartmann, Malgorzata Anna Ulasik, Christian Scheller, Yanick Schraner, Amit Jain, Jan Deriu, Mark Cieliebak, and Manfred Vogel. 2022. SDS-200: A Swiss German speech to Standard German text corpus. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3250–3256, Marseille, France. European Language Resources Association.

Michel Plüss, Lukas Neukom, Christian Scheller, and Manfred Vogel. 2021. Swiss parliaments corpus, an automatically aligned swiss german speech to standard german text corpus.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision.

Felicity J. Rash. 1998. *The German Language in Switzerland: Multilingualism, Diglossia and Variation*. German linguistic and cultural studies. P. Lang.

Sebastian Ruder. 2024. True Zero-shot MT. `https://newsletter.ruder.io/p/true-zero-shot-mt`. Accessed: 2024-03-10.

Tanja Samardžić, Yves Scherrer, and Elvira Glaser. 2016. ArchiMob - a corpus of spoken Swiss German. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4061–4066, Portorož, Slovenia. European Language Resources Association (ELRA).

Yves Scherrer, Tanja Samardžić, and Elvira Glaser. 2019. Digitising Swiss German: How to process and study a polycentric spoken language. *Language Resources and Evaluation*, 53(4):735–769.

Yanick Schraner, Christian Scheller, Michel Plüss, and Manfred Vogel. 2022. Swiss German speech to text system evaluation.

Clément Sicard, Victor Gillioz, and Kajetan Pyszkowski. 2023. Spaiche: Extending state-of-the-art ASR models to Swiss German dialects. In *Proceedings of the 8th edition of the Swiss Text Analytics Conference*, pages 76–83, Neuchatel, Switzerland. Association for Computational Linguistics.

Simone Ueberwasser and Elisabeth Stark. 2017. What's up, Switzerland? A corpus-based research project in a multilingual country. *Linguistik Online*, 84(5).

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

A Appendix

A.1 Examples of Whisper's Performance

Table 6 offers two excerpts from Whisper's output for our *Mock Clinical Interviews* test set (see Section 3.1). The excerpts exemplify Whisper's consistent high-quality performance over a longer passage of spontaneous and continuous speech.

Table 7 offers a speech excerpt from an interview in which the interviewee describes a past narrative (morning routine). It is an example of how Whisper inconsistently translates the Swiss German perfect sometimes as the Standard German preterite and sometimes as the Standard German perfect, cf. Section 6.

A.2 Human Evaluation Guidelines

Table 9 shows the rating guidelines for the raters in the human evaluation survey, cf. Section 7.

A.3 SPC Examples

Table 10 shows some examples of diverging reference translations that unjustly penalize Whisper's performance, cf. Section 5.2.

A.4 Plots

Figures 2 and 3 contain boxplots of the distribution of WER and BLEU scores of Whisper's performance on the three test sets: SPC, STT4SG and SwissDial. They show that Whisper's performance measured in WER and BLEU fluctuates considerably; for some sentences in STT4SG-350 for example, BLEU scores went up to 100.

Reference translation	Whisper
und es ist einfach so mega heimelig irgendwie also so	Es ist einfach so mega heimelig irgendwie.
es gibt mir so ein warmes Gefühl	Es gibt mir so ein warmes Gefühl.
und mir gefällt auch wie sie sich anziehen dort also,	Und mir gefällt auch, wie sie sich dort anziehen.
also ich finde es auch so schön es zu sehen,	Ich finde es auch so schön, es zu sehen.
nicht nur weil es lustig ist und beruhigend sondern auch, weil es schön ist zum anschauen	Nicht nur, weil es lustig und beruhigend ist, sondern auch, weil es schön ist, es anzuschauen.
also man hat Einzeltherapie zum Beispiel, oder Gruppentherapie mit anderen zusammen	also man hat Einzeltherapie zum Beispiel oder Gruppentherapie mit anderen zusammen
oder Musiktherapie, Maltherapie, oder auch so Entspannungsgruppen	oder Musiktherapie, Maltherapie oder auch so Entspannungsgruppen.
teilweise kann man auch selber Sport machen wenn man das will, also nicht in der Gruppe sondern alleine	Teilweise kann man auch selber Sport machen, wenn man das will, also nicht in der Gruppe, sondern alleine,
oder irgendwie so Walkinggruppen, oder so Achtsamkeitsgruppen, wo man in die Natur geht.	oder irgendwie so Walking-Gruppen oder so Achtsamkeitsgruppen, wo man in die Natur geht.

Table 6: Excerpts of Whisper's performance on continuous speech from our "Mock Clinical Interviews" test set (the segmentation into sentences is only for the sake of readability). These excerpts are evidential of Whisper high-quality performance.

Swiss German	Whisper
Denn **bin** i richtig **ufgstandfe**	Dann **bin** ich richtig **aufgestanden**.
Dänn **bin** i go dusche **ggange**	Ich **ging** duschen.
Dänn **han** i mi **aazoge**	Dann **zog** ich mich an.
Dänn mine chatz no fuetter **gää**, will si di ganzi ziit am maue **gsi isch** und unbedingt **het welle** esse	Ich **gab** meinen Katzen Futter, weil sie die ganze Zeit am Mauen **waren** und essen **wollten**.
Dänn **bin** i mit ire id stube abe ggange	Dann **ging** ich mit ihr in den Wohnzimmer.
Dänn **han** *ich* öppis **ggässe** dänn **het** *sii* öppis **ggässe**	Ich **habe** etwas **gegessen**, dann **hat** sie etwas **gegessen**.
Und dänn **bin** i wider uffe go zäh putze	Ich **ging** wieder hoch, um die Zähne zu putzen.

Table 7: Excerpt from a conversation in which the interviewee recounted what they did the same morning. It exemplifies how Whisper switches between perfect and preterite in Standard German. The input is always in the perfect tense. Perfect/preterites are marked in bold.

Reference	Hypothesis
“weil ich **dann halt** wieder auf mich gestellt bin.”	“weil ich wieder auf mich gestellt bin.”
“und darum ist es **ein bisschen** beides.”	“Darum ist es beides.”
“**ja und** ich find’s **einfach nur** spannend”"	“Ich finde es spannend”
“**Halt irgend so** eine Einschlafmeditation von einer Person...”	“Eine Einschlafmeditation von einer Person...”
“und er bekommt 50’000 Franken”	“und **dann** bekommt man **irgendwie noch** 50’000 Franken”"

Table 8: Examples for the removal of particles and conjunctions in Whisper’s output. Words in bold are particles/-conjunctions missing in the reference/hypothesis.

	Sinn – Ist der originale Sinn beibehalten? Entspricht Satz B Satz A?
5	Entspricht sinngemäss voll und ganz dem Original
4	Etwas ist verloren gegangen, die Bedeutung ist aber im grossen und ganzen gleich
3	Stimmt teilweise, aber nicht in allen Teilen
2	Entspricht kaum noch dem originalen Sinn
1	Gar nicht
	Flüssigkeit. Bezogen auf Satz B – ist das gutes Deutsch?
5	Ja, voll und ganz. Natürlich und einwandfrei.
4	Relativ flüssig
3	Nicht ganz flüssig, etwas merkwürdig
2	Kaum akzeptabel
1	Inakzeptabel

Table 9: Rating guidelines for the raters participating in the survey of human evaluation.

Swiss German Audio	SPC Reference	Whisper	WER
...nachhinei muss me döt de iibürgerigswillige sägge, er het scho...	So muss den Einbürgerungswilligen im Nachhinein gesagt werden:	Nachhinein muss man den Einbürgerungswilligen sagen, er hat schon	0.75
Dir wüssed scho vo de römerziite her	Aus Römerzeiten wissen Sie schon:	Ihr wisst schon von den Römerzeiten her,	1.4
...und das isch schlächt	Das ist schlecht.	Und das ist schlecht.	0.33
Während acht jahr isch s in betriib gsi	Während acht Jahren wurde es betrieben.	Während acht Jahren war es in Betrieb.	0.5
...u es het halt i Gotts name oo mitem finanzielle z tüe...	Und es hat halt auch wirklich mit dem finanziellen Aspekt zu tun.	Und es hat halt in Gottes Namen auch mit dem Finanziellen zu tun.	0.42
Töu vo euch erinnere sech müglicherwiis aa experiment ir physik oder chemie	Manche von Ihnen erinnern sich möglicherweise an missglückte Experimente in Physik oder Chemie.	Ein Teil von euch erinnert sich möglicherweise an Experimente in Physik oder Chemie.	0.38

Table 10: Examples for perfect performance of Whisper penalized by strongly divergent reference translations in the SPC corpus.

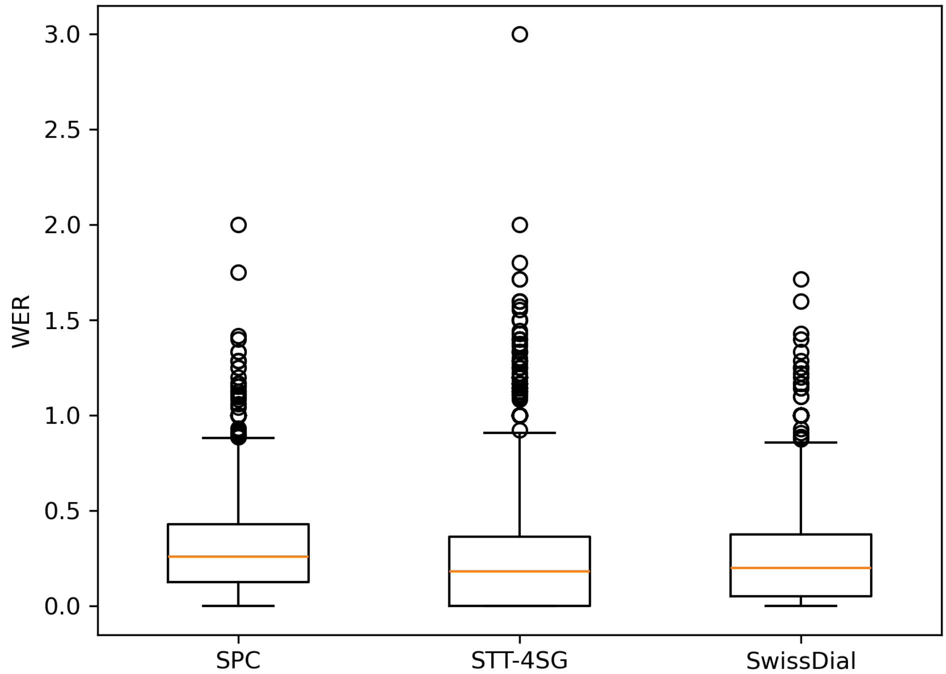


Figure 2: Distribution of WER scores for each corpus.

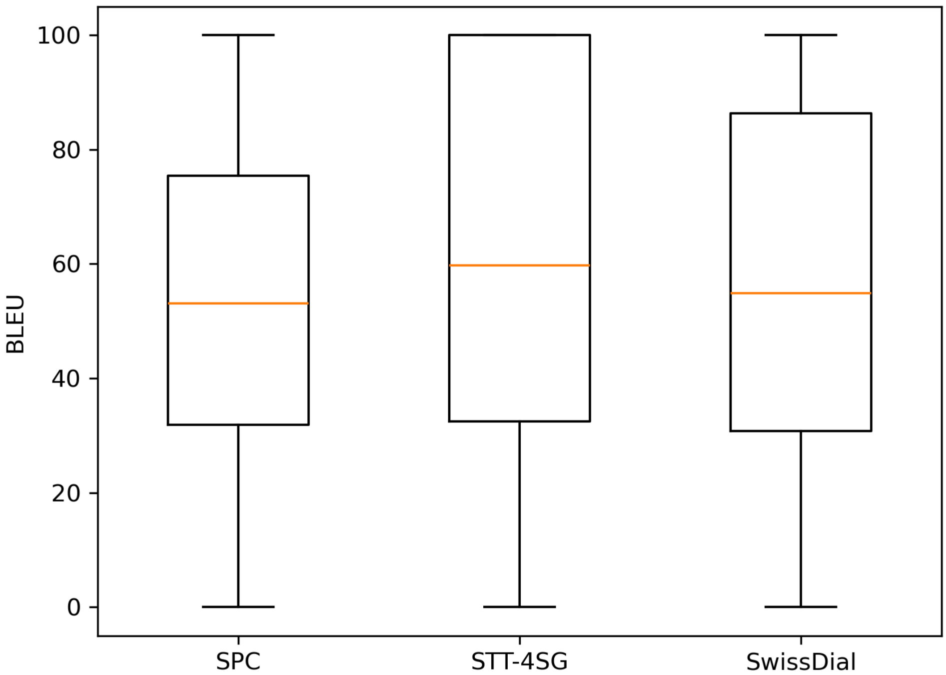


Figure 3: Distribution of BLEU scores for each corpus.

How Well Do Tweets Represent Sub-Dialects of Egyptian Arabic?

Mai Mohamed Eida[1], Mayar Nassar[2], Jonathan Dunn[1],
[1]University of Illinois Urbana-Champaign, [2]Ain Shams University
Correspondence: maimm2@illinois.edu

Abstract

How well does naturally-occurring digital text, such as tweets, represent sub-dialects of Egyptian Arabic (EA)? This paper focuses on two EA sub-dialects: Cairene Egyptian Arabic (CEA) and Sa'idi Egyptian Arabic (SEA). We use morphological markers from ground-truth dialect surveys as a distance measure across four geo-referenced datasets. Results show that CEA markers are prevalent as expected in CEA geo-referenced tweets, while SEA markers are limited across SEA geo-referenced tweets. SEA tweets instead show a prevalence of CEA markers and higher usage of Modern Standard Arabic. We conclude that corpora intended to represent sub-dialects of EA do not accurately represent sub-dialects outside of the Cairene variety. This finding calls into question the validity of relying on geo-referenced tweets alone to represent dialectal differences.

1 Egyptian Arabic Sub-Dialects

Existing work on Egyptian Arabic (EA) sub-dialects primarily uses geo-referenced data to represent specific varieties. The question here is whether existing EA corpora adequately represent the intended sub-dialects: do existing written corpora of EA equally represent both majority varieties (e.g., Cairene Egyptian Arabic: CEA) and minority varieties (e.g., Sa'idi Egyptian Arabic: SEA)? This is an important question for two reasons: first, representation within the training data (upstream) influences representation within language technology (downstream). This means that dialect adaptation for less prestigious varieties like SEA depends on these dialects being adequately represented in training corpora (Biber, 1993; Dunn, 2020). Second, spoken and written register variation in Arabic can impact dialect representation. For example, results in this paper suggest that speakers of CEA freely use their dialect in tweets but speakers of SEA revert to Modern Standard Arabic (MSA). This implies that the relationship between dialect and register is not predictable across sub-dialects.

Current work on Dialectal Arabic (DA) resources and applications does not take into account DA variation beyond the country level (Abdul-Mageed et al., 2020a, 2020b, 2021; Bouamor et al., 2018; Tachicart et al., 2022). Further, this work has not considered spoken and written register variation across sub-dialects. Therefore, this paper addresses two specific questions. First, which Egyptian Arabic sub-dialects are represented within existing digital written corpora, specifically tweets? To find out, we compare these corpora with ground-truth dialect surveys (Behnstedt and Woidich, 1985; Khalafallah, 1969). Second, is EA in a digital written register, specifically tweets, equally representative of spoken sub-dialects? To find out, we compare the relative usage of DA vs. MSA features in tweets across two sub-dialects of Egyptian Arabic.

If current datasets are representative of EA sub-dialects, and if register variation across written and spoken EA is limited, then NLP tasks like Arabic micro-dialect identification, machine-translation, and morphological parsing can be adapted for dialectal varieties using Tweet-based corpora. In other words, this would mean that digital written data, as a register, remains representative of EA sub-dialects. However, if the current datasets are not representative of EA sub-dialects, this means that sub-dialects (beyond CEA) are low-resource and that more data collection is needed to represent all EA sub-dialects. Further, the possibility that digital written registers are not equally valid for all sub-dialects means that other sources of EA sub-dialect data, such as speech, should be explored. In other words, if speakers of less prestigious dialects like SEA revert to standardized forms in written registers then spoken data must also be used for dialect adaptation.

The primary contribution of this paper is to measure how well written registers represent different

Proceedings of the Eleventh Workshop on NLP for Similar Languages, Varieties, and Dialects (VarDial 2024), pages 41–55

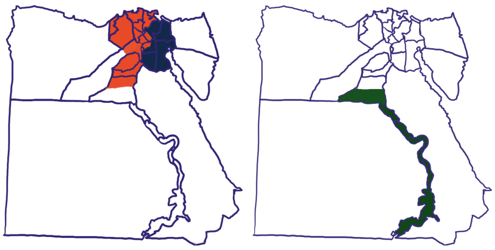


Figure 1: Map of Egyptian Arabic Sub-dialects (Woidich, 1996): CEA and other rural dialects (left), SEA dialect (right)

sub-dialects of Egyptian Arabic as well as the validity of specific datasets designed to capture these sub-dialects. We find that the more prestigious CEA is well-represented but that resources intending to represent the less prestigious SEA fail to do so.

The outline of this paper is as follows: Section 2 provides an overview of Egypt's sub-dialects, addresses register variation in Arabic, and previous work on sub-dialect data collection and dialect identification. In Section 3, we provide an overview of specific features of two Egyptian sub-dialects: Cairene Egyptian Arabic (CEA) and Sa'idi Egyptian Arabic (SEA). These features are drawn from ground-truth dialect surveys. Section 4 discusses a baseline corpus, a large reference corpus to which other datasets are compared, and three sub-dialect corpora. The remaining sections present results and discuss the significance of representation and register variation across CEA and SEA.

2 Related Work

2.1 Egyptian Arabic Sub-dialects

Arabic is a diglossic language (Ferguson, 1959), with Modern Standard Arabic (MSA) considered the High-variety, and Dialectal Arabic (DA) the Low-variety. While MSA is the official language of Egypt, Egyptian Arabic (EA) is the dialectal variety spoken among Egyptians. EA sub-dialects are classified by geographical location, and can be grouped into 4-5 sub-dialects with variation across phonology, morphology, syntax, semantics, and lexicon (Behnstedt and Woidich, 1985; Badawi, 1973). The most prestigious dialect is Cairene Egyptian Arabic (CEA), the sub-dialect spoken by approximately 40% of Egyptians, specifically middle class Egyptians in Cairo and urban cities (Gadalla, 2000; Hanna, 1962; Harrell, 1957; Hospers, 1973; Norlin, 1987; Leddy-Cecere and Schroepfer, 2019). On the other hand, Sa'idi Egyptian Arabic (SEA) is the "the most ridiculed, stigmatised and stereotyped" sub-dialect of EA (Bassiouney, 2017), yet is also the second most spoken EA sub-dialect by 40% of Egyptians. Thus, these sub-dialects are equal in usage but unequal in prestige. EA is the most thoroughly researched DA, yet work on sub-dialects other than CEA is extremely limited. With the exception of Behnstedt and Woidich (1985), and Khalafallah (1969) there exists no recent dialectal surveys on linguistic features of Sa'idi Egyptian Arabic.

2.2 Register Variation

The only standardized and codified writing system of Arabic is MSA (Brustad, 2017; Håland, 2017; Høigilt and Mejdell, 2017). In the past century, EA was "rarely written, and ha[d] little prestige among the people" (Harrell, 1957, p.1). Therefore, there remains no codified written system for EA. It was not until the spread of Social Networking Sites across the past three decades generated a wealth of content written using EA, despite the lack of EA codification (Kindt and Kebede, 2017). Written EA output contains inconsistencies in orthographic representations due to a mixture between using codified MSA as well as developing new orthographic representations for linguistic features exclusive to EA. These features can be dialectal markers; however, the defaulting to MSA in orthographic representation despite different dialectal phonetic representations is exceedingly common. This is a result of influences of standard language ideology and emphasis on 'correctness' in language use (Bassiouney, 2014).

An example is the phonetic representation of the lexical item 'camel'. It is orthographically represented as 'جمل' [dʒamal] in MSA, pronounced as [damal] in SEA, and [gamal] in CEA. Speakers of both dialects orthographically represent this word using the codified MSA form, when SEA could represent it to be phonologically reflective of one variation in their sub-dialect 'دمل'. Using the codified MSA form is common, making it difficult to detect dialectal markers across Arabic sub-dialects. To our knowledge, there has been no empirical corpus analysis of EA written orthographic patterns across sub-dialects. However, there has been a large effort to identify orthographic patterns in DA written data for the purpose of facilitating and enhancing computational parsing of DA inconsistent orthographic patterns (Altantawy et al., 2010; Habash et al., 2005; Habash, 2007; Habash et al., 2012; Fashwan and Alansary, 2021). The complexity of Arabic orthography, lack of DA codification, prevalence of MSA as the medium of writing, and lack of empirical research across DA written/spoken registers all motivate the validation of collected DA written text before using this data to represent EA sub-dialects.

2.3 Resources and Tasks

A survey of EA corpora from ***The Linguistic Data Consortium*** (LDC), ***MASADER***[1], and ***InfoGuistics***[2] show that existing Egyptian Arabic corpora primarily feature CEA. MADAR (Bouamor et al., 2018) is a multi-dialect corpus across 25 Arabic cities, one of which is a SEA city. However, MADAR is translated from English and French and not a naturally-occurring corpus, and thus excluded for the purpose of this paper.

DA sub-dialects have been an Arabic NLP focus mostly for dialect identification. A number of other efforts to identify DA sub-dialects on the city level include NADI2020 (Abdul-Mageed et al., 2020a) and NADI2021 (Abdul-Mageed et al., 2021), two series of dialect identification shared tasks. These tasks target micro-dialect identification through matching each Tweet to its corresponding city, with approximately 56 Egyptian cities represented in the datasets. Teams mainly used transformer-based methods for this challenge.

The question is whether the corpora assumed to represent sub-dialects actually do so. The NADI2020 & 2021 sub-dialect shared task's difficulty is reflected in the low F1 scores achieved, 6.39% in NADI2020 and a slight improvement to 8.6% in NADI2021. One reason could be that not all cities have distinct sub-dialects, with some spanning across many cities with minimal distinctions (Behnstedt and Woidich, 1985). Therefore, predicting a specific city is too specific a task when the underlying dialectal features are specific to all cities in the same area. It is also possible that the geo-referenced tweets are not representative of the intended sub-dialects because speakers avoid using less prestigious varieties in certain settings, instead reverting to MSA. This is further explored in this paper.

Abdul-Mageed et al. (2020b) present another contribution towards micro-dialect identification by fine-tuning BiGRU and mBERT models to distinguish sub-dialects in around 21 Arabic countries and 319 cities. They report human annotation at the city level was deemed nearly impossible, as they employed annotators from various Arabic countries to identify sub-dialects outside their native country and dialect. This task would likely be difficult but feasible within a single country. For instance, while a Moroccan might struggle to identify Egyptian sub-dialects across Egyptian cities, an Egyptian might have the linguistic experience necessary to make such distinctions. Despite such annotation efforts, including adjustments for diglossia and code-switching within the data, the system's peak performance was an F1 score of 20.11% and accuracy of 19.88%. The system performed better when utilizing dialectal Arabic alone without inclusion of MSA data. Performance was higher when fewer cities were included.

3 Sub-Dialect Distance Measures

3.1 Dialectal Features

This paper relies on dialectal features from ground-truth dialect surveys to measure the distance between sub-dialects of EA and their expected patterns. Starting with morphological and grammatical features of each sub-dialect, we focus on demonstratives, interrogatives, prepositions, adverbs, and negation particles, as reported for CEA and SEA in existing dialectal surveys (Behnstedt and Woidich, 1985; Khalafallah, 1969; Leddy-Cecere and Schroepfer, 2019). Our motivation is to select features where there is a distinction between SEA and CEA in orthography, yet are essential to the syntax of SEA and CEA to maximize the likelihood of their presence in the text. Based on the ground-truth surveys, we the believe selected features are sufficient to indicate how well a corpus represents each sub-dialect, although discrepancies in the orthographic representation of these features can vary. For this reason, we rate each feature for markedness; sample features are illustrated in Table 1, and a full list of features in the appendix.

We exclude possible overlap corresponding to MSA features when possible. We use regular expressions to further account for spelling mistakes, such as usage of 'ي، ى' or 'آ ، إ ، أ ، ا' interchangeably, and different orthographic representations 'برضه' vs 'برضو' or 'بردو', and allomorphs of selected features 'ما' vs 'مـ'. We tested features

[1] https://arbml.github.io/masader/
[2] aucegypt.edu/infoguistics/directory/Corpus-Linguistics

in isolation to ensure validity and reliability.

A quantitative analysis of feature validity considers the likelihood of capturing false positives. For example, Ad4 in Figure 2 (left) is able to capture different orthographic representations with less than 5% false positive results. For features with false positives higher than 5%, we analyse them qualitatively. For example, in Figure 2 (right), the SEA negation particle coded Neg4 occurs by adding the suffix 'شي' at the end of a verb. The regex captures this representation, along with any part of speech ending in 'شي', resulting in a large number of false positives, which are then checked manually.

Feature	MSA	CEA	SEA
`Interrogative***`	أيضاً 'also'	برضه 'also'	برضك 'also'
`Adverb***`	الآن 'now'	دلوقتي 'now'	دلوق 'now'
`Particle*`	ليس 'not'	مش 'not'	مش 'not'
`Preposition*`	في 'in'	في 'in'	ف 'in'
`Demonstrative***`	اولائك 'these'	دول 'these'	داكهما 'these'

Table 1: Sample grammatical features distinctions between MSA, CEA, and SEA. *** indicate most marked features, and * the least marked.

We also elected to exclude features which do not have a clear orthographic distinctions between SEA and CEA from quantitative results. Some features differ between SEA and CEA in phonetic, morphological, or semantic distinctions, however these distinctions are not indicated in the orthographic form. For example, the free negation particle, Neg1, is less likely to be followed by perfective or imperfective verbs in CEA, but this is common within SEA. Accordingly, we elected to examine the data qualitatively with non-orthographically distinct features in order to capture some dialectal features.

3.2 MSA and DA

We measure the usage of MSA in tweets across sub-dialects in one dataset. The more MSA is used in the dataset, the less DA is used and, therefore, the less the sub-dialect is actually represented. To determine usage of MSA, we identify MSA morphological features and isolate tweets, then manually annotate for correctness. Two annotators, native speakers of EA, manually annotate the Micro-Dialect dataset (Abdul-Mageed et al., 2020b) for MSA, DA, and code-switching of both using annotation guidelines for Arabic dialectness by Habash et al. (2008). We group annotation guidelines of 1 & 2 as MSA, 3 as code-switching, and 4 & 5 as DA. Inter-annotator reliability across a sample 1000 tweets measured 86%.

4 Datasets

We use this section to first discuss the baseline or reference corpus which is used to validate the expected features, to determine whether our extraction method does in fact capture the variants which we intend to use to explore sub-dialects of EA. We then describe the corpora used to test whether georeferenced tweets from specific cities contain the dialectal variants expected given the ground-truth dialect surveys.

4.1 Cairo Baseline Corpus

For the baseline corpus, we use Cairo georeferenced tweets from Dunn (2020), shown in Table 2. The purpose of this corpus is to ensure the validity of our feature extraction method. Therefore, this corpus is used to measure prevalence of CEA features in Cairo tweets. Tweets include both DA and MSA, and have been pre-processed to only include the Arabic text in the Tweet. With exception of prepositions, CEA features do not overlap with MSA features, therefore, the results should reflect CEA usage in tweets. Due to the size of this corpus, that it is extracted from Cairo, we expect to find high representation of the selected CEA features. Due to migration from SEA cities to Cairo, we also expect to find some SEA features represented by SEA users who might have migrated to Cairo, though much less than its CEA counterparts. Therefore, this corpus is a baseline to ensure the validity and reliability of the script in capturing features by georgraphical location.

Dataset	Tweets	Tokens	MSA/DA
`Baseline`	808,312	12,233,632	Both

Table 2: Baseline Corpus (Dunn, 2020) across Cairo. Tokens by\ s.

بَرْدو	بَرْضو	بَرْضُه	كُوتْشي	شافْشي	نِمْشي
b<a>r.du:	b<a>r.dˤu:	b<a>r.dˤ<u>h	k<u>t.ʃi:	ʃa:f.-ʃi:	n<i>m.ʃi:
also.ADV	also.ADV	also.ADV	sneakers.NOM	see-3.S.M.PERF-NEG	walk-1.P.IPFV
also	also	also	sneakers	he did not see	we walk

Figure 2: Examples of distinct orthographic representations resulting in false positives (right) vs. alternating orthographic representations of the same word with no false positives (left)

4.2 Sub-Dialect Datasets

We examine three datasets of tweets geo-tagged by city from Arabic micro-dialect identification shared tasks. Datasets include MicroDialect Identification (Abdul-Mageed et al., 2020b), NADI 2020 (Abdul-Mageed et al., 2020a), and NADI 2021 datasets (Abdul-Mageed et al., 2021) across eleven cities. Tweets were collected in 2019 over 10 months, from users who exclusively tweeted from the same location.

Dataset	Tweets	Tokens	MSA/DA
`MicroDialect`	6,056	77,173	Both
`NADI2020`	1,021	13,288	Both
`NADI2021`	798	7,324	DA
`Total`	7,875	97,785	

Table 3: CEA Datasets: Tweets span across Cairo, New Cairo City, Suez, PortSaid and Ismailia (Abdul-Mageed et al., 2020b; Abdul-Mageed et al., 2020a; Abdul-Mageed et al., 2021). Tokens by\ s.

SEA and CEA cities were determined based on reported dialectal surveys (Behnstedt and Woidich, 1985). Except for NADI2021 (Abdul-Mageed et al., 2021), all tweets include both MSA and DA. All datasets were pre-processed for punctuation, replies, other embedded foreign tokens, hashtags, or indicators for cross-posting on other platforms except for MicroDialect datasets. We elected to not pre-process this corpus to further examine the results on both pre-processed and unprocessed datasets. Some of the original datasets included 10M tweets but cannot be obtained due to API limitations at the time of this paper; therefore, we examine the limited data released within the training and development datasets.

SEADataset	Tweets	Tokens	MSA/DA
`MicroDialect`	3,076	39,292	Both
`NADI2020`	1,862	24,693	Both
`NADI2021`	1,863	16,507	DA
`Total`	6,801	80,492	

Table 4: SEA Datasets: Tweets span across Qena, Asyut, Aswan, Luxor, Sohag, and BeniSeuf (Abdul-Mageed et al., 2020b; Abdul-Mageed et al., 2020a; Abdul-Mageed et al., 2021). Tokens by\ s.

5 Results

5.1 Does the Baseline Cairo Corpus Contain CEA Features?

To test the validity of sub-dialect morphological CEA and SEA features reported in Behnstedt and Woidich (1985), Khalafallah (1969), and Leddy-Cecere and Schroepfer (2019), we measure the distance between spoken CEA features reported and their prevalence in the written Cairo Baseline corpus. As illustrated in Figure 3, CEA features are overwhelmingly prevalent in the Cairo corpus, while SEA features are not. Each feature is an alternation (i.e., the CEA vs SEA variant). This figure shows the percentage of CEA features used in the baseline Cairo corpus. Feature names correspond to the feature list in the appendix.

This high usage of CEA variants and low usage of SEA variants in the baseline corpus confirms the validity of using these features to measure the distance between dialects. Therefore, we conclude that Cairo is representative of the sub-dialect reported in the dialectal surveys: Cairene Egyptian Arabic. In the next section, we measure the SEA datasets for its representation of SEA sub-dialectal features.

5.2 Do SEA Corpora Contain SEA Features?

The first question is whether we see a greater share of expected SEA features in corpora used to rep-

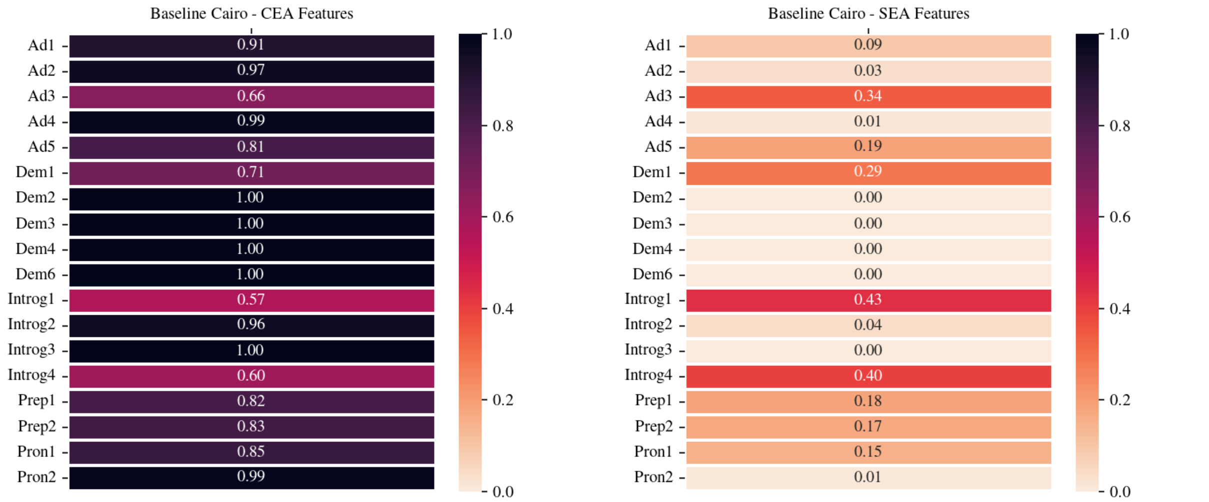


Figure 3: Share of Expected CEA (left) and SEA (right) Variants for each Alternation for Cairo Baseline Corpus. MSA and DA Corpus. Features are complementary.

resent SEA. We take a feature-by-feature look in Figure 4, here using the share of SEA variants for each of the alternations discussed above from the dialect survey. These are complementary features, so that if the share of SEA usage is 25%, then the share of CEA usage must be 75%. Each row is a feature, corresponding with the feature descriptions found in the appendix. The first column represents the baseline corpus of Cairo tweets. The second column represents the CEA cities from NADI2020, NADI 2021, and MicroDialect Corpus and the third column the SEA cities from the same datasets. We would expect, then, that there would be a much higher share of SEA usage in the final column.

First, many features remain unobserved (hence a 0.00 value), even though the annotation methods discussed above accurately identify these variants and some are observed in the Cairo Baseline corpus. This means that the features are simply not observed in these relatively small corpora.

Second, we see that only a few of the overall alternations show the pattern expected from the dialect surveys: Ad3, Dem1, Introg1, Introg4, Prep2, and Pron1 are all markedly more common in SEA as expected. The other features show either no difference at all or the opposite pattern as the dialect surveys. However, what is significant across these specific features is their distinction from their CEA counterparts in either the shortening or elongation of existing vowels or the loss of voiceless final consonants. For example, SEA dialectal surveys report the lack of [h] at the end of Introg4 in SEA features written as ‘لي’ [le:], while CEA surveys report its presence in CEA features written as ‘ليه’ [le:h]. This distinction is not as marked as the distinct realization of Dem6, where SEA features add a stop /k/ at the end of the word written as ‘برضك’ [bard^{ʕ}ak], a phoneme more marked than /h/.

Other highly marked SEA features, such as negation particles, are not observed in SEA datasets yet do occur in the Cairo Baseline corpus. For example, Neg4, Neg3 (Table 5) SEA features are marked with either dropping the CEA negative prefix ‘ما’, and adding a long vowel ‘ي’ to the CEA negation suffix ‘ش’. Qualitative analysis of Neg 3, 4 shows these features are not observed within any SEA dataset, yet are observed in the Cairo Baseline corpus across 50 instances such as ‘ينفعشي’ meaning ‘not possible’, ‘شافشي’ meaning ‘did not see’. This finding extends to other SEA features observed only in the Cairo Baseline corpus, except for Dem 2-6.

Third, overall there is little difference between the CEA and SEA corpora in usage of SEA features. One reason could be language change which has taken place since the dialect surveys. Another reason could be the impacts of internal migration from rural to urban areas (Miller, 2005). If this were the case, then we would expect that some users in CEA cities would maintain clear SEA fea-

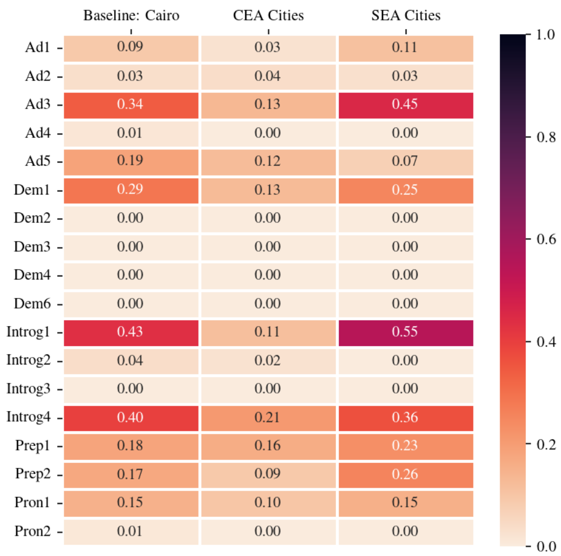


Figure 4: Share of Expected SEA Features for each Alternation across SEA cities and CEA cities in NADI2020, NADI2021, and MicroDialect Corpus compared with baseline Cairo corpus.

tures. This is the goal of the analysis in the next section, where we look at individual cities within each dialect area.

5.3 Are Cities Consistent Within Regions?

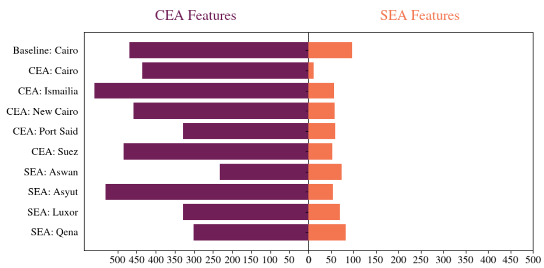


Figure 5: Prevalence of CEA and SEA features by city, using frequency per 10k words. MicroDialect Corpus.

The next question is whether the features expected from the ground-truth dialect surveys appear in the tweets representing different cities within SEA and CEA. While the overall aggregated usage might be unexpected, perhaps some cities have changed (i.e., from SEA to CEA), thus disguising usage in the core SEA cities. This is shown for the mixed MSA and DA Micro-Dialect corpus in Figure 5 & 6, where each city is a bar. The purple values on the left represent the overall frequency per 10k words of CEA features (the prestige dialect), and the pink values on the right represent the same quantity for SEA features (the non-prestige dialect).

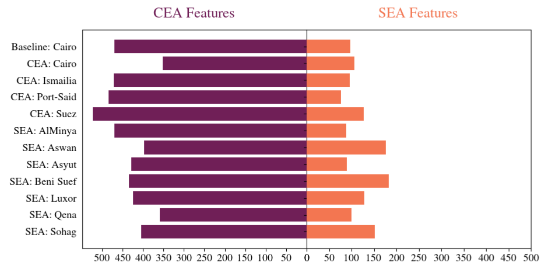


Figure 6: Prevalence of CEA and SEA features by city, using frequency per 10k words. NADI 2020 Corpus.

What we see, first, is that CEA features are overall much more common than SEA features, across both dialect areas. There is a high prevalence of CEA features even in cities expected to represent SEA, such as Asyut, although most SEA cities have a lower rate of usage. Second, we see that there is a relatively equal usage of SEA features across cities, even central CEA locations like Cairo. Because this data represents a small number of users, the figure includes a baseline corpora of other tweets from Cairo, a much larger corpus as described in the Data section. We contrast this raw unprocessed corpus with the cleaned version in Figure 7, here using the NADI 2021 shared-task corpus. Because MSA samples have been removed here, the overall rate is much higher. However, while the density of dialectal features is higher, there is still no sharp distinction in the usage of CEA features in CEA and SEA locations. On the other hand, SEA features are slightly more frequent in SEA locations. Further, qualitative analysis of Neg1 and Neg2 reflect the results of the quantitative analysis. NADI2021 shows twice Neg1 SEA usage than its CEA counter

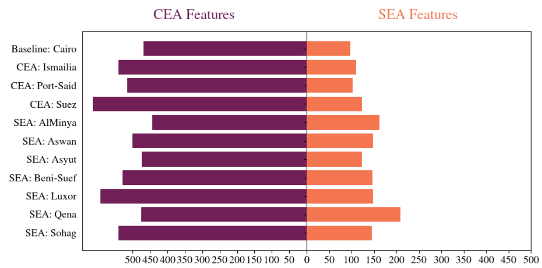


Figure 7: Prevalence of CEA and SEA features by city, using frequency per 10k words. NADI 2021 Corpus.

part, observing instances of the particle followed by imperfective verbs such as 'مش تستني' meaning 'do not wait' and 'مش فيه' meaning 'there is none'. In CEA, both instances are negated using the Neg2 feature, represented as 'متستنيش' and 'مفيش', respectively. The same analysis for the NADI 2020 corpus is shown in Figure 6. Again the CEA variants dominate across all cities, although to a lesser degree in SEA cities.

5.4 Are Users Consistently Writing in SEA?

The basic finding here is that the corpora representing SEA dialect areas do not contain a substantial usage of the expected SEA features. Instead, CEA features are found across all cities. Why? One possibility is that language change has taken place since the dialect survey was undertaken, although this would be an unusually fast process of change. Another possibility is that older or less connected speakers retain the SEA features but are not represented on social media. A third possibility is that SEA speakers do not produce SEA variants in this digital written setting. We will consider these possibilities further in the discussion.

For now, it is possible that individuals from SEA and CEA cities have changed locations. Thus, we might expect users to consistently use one or the other sub-dialect but to be located in unexpected cities. This is the purpose of the analysis in the next section.

To find out if there are users of each dialect who are out of place, perhaps because of internal migration within Egypt, we visualize the distances between MicroDialect user-specific corpora in Figure 8. Here each point is a corpus representing a single user; the style of each point refers to the dialect area it is supposed to represent. Points are then positioned within a two-dimensional space by using PCA to reduce the usage of all dialectal features into two main components. Taken together, these two components explain 96% of the variance across features; thus, we take this as a reliable visualization of the dialectal relations between user-specific corpora.

First, it is clear that individuals taken to represent both CEA (circles) and SEA (x's) are inter-mingled. This would indicate that the previous overlap in feature usage across CEA and SEA is not because some individuals retain expected usage and others do not. Rather, the usage patterns of individuals are not organized around the expected dialects. In other words, the disconnect between SEA corpora and expected SEA features is not a result of individual differences across users.

Second, since each SEA user is closely patterned with at least one CEA user, this indicates that the core expected SEA speakers are not actually producing that dialect. One possibility is that these users are instead producing either more standard dialectal features (CEA) or are simply reverting to non-dialectal production (MSA). This is explored in the final section.

5.5 Who is Reverting to MSA?

To explore whether SEA users are producing CEA features or resorting to non-dialect production, we annotated the largest SEA corpus, MicroDialect corpus, for MSA, DA, and code-switching. As illustrated in Figure 9, SEA users seem to be using MSA approximately as much their usage of DA. However, CEA users are using DA significantly more than MSA. There were no significant differences in code-switching among both groups.

This tells us two things. First, SEA tweets include a high number of MSA tweets, therefore, chances of SEA feature representation in SEA dataset has lowered by 50% of the overall dataset. Second, when using DA, SEA users do not use highly marked SEA features, but rather resort to either CEA features or SEA features which carry closer resemblance to their CEA counterparts. Regardless, SEA datasets are not representative of the targeted sub-dialect SEA. It is worth noting that the data released is across 11 SEA and 11 CEA users, thereby, limiting any generalizations about SEA data in general. However, insofar as these corpora are taken to represent SEA production, this results show that the non-prestige sub-dialect is inadequately represented compared to the prestige sub-dialect.

6 Discussion

As highlighted by the results, geo-referenced SEA datasets are not representative of SEA sub-dialects. Results are consistent across DA datasets, MSA and DA datasets, and processed and unprocessed datasets. One possibility could be that language change has taken place since the dialect survey was undertaken. However, there is prevalent evidence of SEA features in current SEA speech. Therefore, this cannot be attributed to SEA drastic language change. Another possibility is that older or less

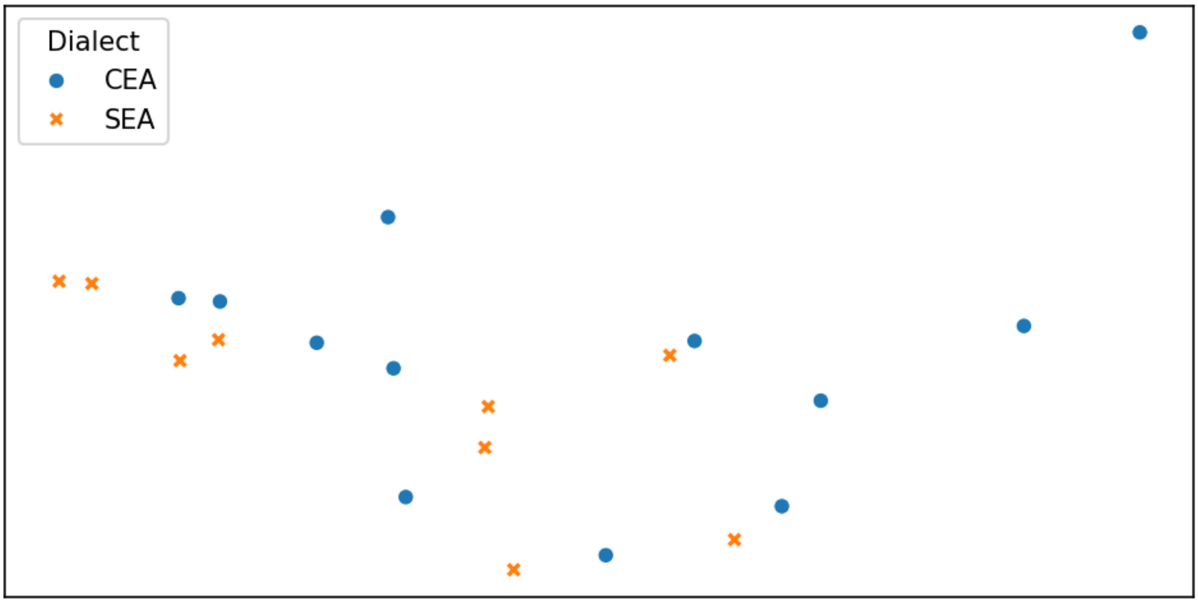


Figure 8: User-by-user plots of feature usage, visualized using PCA for dimension reduction. The original vectors undergoing PCA are the relative frequency of each dialectal feature.

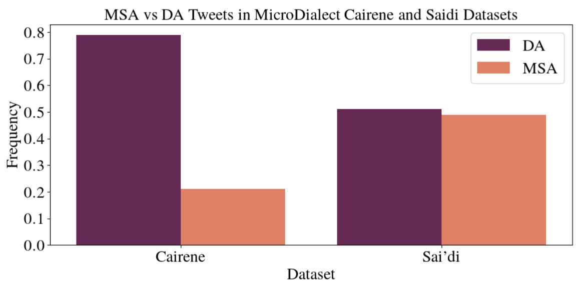


Figure 9: MSA vs. DA Tweet Frequencies in CEA and SEA Datasets.

connected speakers retain the SEA features but are not represented on social media. Kindt and Kebede (2017) report Cairene Egyptians prefer using written MSA vs. DA based on education, age, gender, and platform. Egyptians between the ages of 13-34 use DA significantly more frequently than Egyptians over 50, and women are more likely to write in MSA than men. Given the limited user demographic information beyond consistently tweeting from the same location for over 10 months, we cannot conclude if some, or any, demographic variables contribute to the lack of SEA features or DA use in SEA datasets. A third possibility is that SEA speakers do not produce SEA variants in this digital written setting. While there is a lack of recent SEA dialectal surveys, there is evidence on low attitudes and stigmatized perceptions of the SEA dialect (Bassiouney, 2014; Bassiouney, 2017). SEA users could be avoiding SEA markers in an attempt to position themselves differently across digital platforms. A larger geo-referenced written digital corpus is needed to explore these possibilities further. Regardless, the examined SEA datasets are not representative of the SEA sub-dialect, and register variation is significant across SEA spoken and written registers.

7 Conclusion

This paper finds that EA sub-dialects (except CEA) are low-resourced, and existing Tweet datasets are not representative of EA sub-dialects. Further, register variation between SEA speech and naturally-occurring digital written tweets is significant, therefore, these results call into question the validity of relying on geo-referenced tweets alone to represent dialectal differences. This paper further highlights the need for more representation across DA resources to include DA sub-dialects (Tachicart et al., 2022), and more empirical research on register variation across Dialectal Arabic written sub-dialects and their orthographic patterns in digital spaces.

8 Limitations

Given the inconsistencies across Arabic written DA orthography, the selected morphological markers' orthographic representation is not the ground-truth, but rather the most frequent patterns observed by the authors in EA sub-dialects and DA digital written contexts, in alignment with the dialectal surveys. This could explain the 0.00 consistent results for some features, although we do account for this through experimenting with all possible orthographic representations. These features could be restricted to speech, or have fallen out of use among the demographic of SEA users online.

Another limitation includes our choice to restrict dialectal markers in quantitative analysis to ones captured with minimal false positives after several iterations of analysis, limiting our quantitative analysis to the most explicit features. We expect some features might be underrepresented or overrepresented in the Cairene Baseline Corpus due its large size, especially if they overlap with similar MSA patterns. We also recognize that some SEA and CEA features are observed in other rural and urban sub-dialects, such as Alexandrian Egyptian Arabic or Shara'wi Egyptian Arabic. However, given the geo-referenced nature of the datasets, we limit our analysis to cities that use CEA and SEA only.

There could be more significant evidence of SEA lexical markers in SEA datasets, however, we do not examine lexical choices between SEA and CEA tweets in this paper. Further, we recognize that one of the most popular Egyptian TV and cinema genres have focused on Sai'di settings (Bassiouney, 2017), and the datasets explored could possibly include quotes or references from such works, and accordingly impact SEA results.

Acknowledgments

We thank Dr. Nizar Habash for the fruitful discussions on representation of Egyptian Arabic sub-dialects. We are grateful for Dr. Muhammad Abdul-Mageed and his group for compiling and sharing the Micro-Dialect Egyptian dataset used in this paper. Special thanks to Dr. Lane Schwartz for his guidance and valuable input. We also thank the anonymous reviewers for their constructive feedback.

References

Muhammad Abdul-Mageed, Chiyu Zhang, Houda Bouamor, and Nizar Habash. 2020a. NADI 2020: The first nuanced Arabic dialect identification shared task. In *Proceedings of the Fifth Arabic Natural Language Processing Workshop*, pages 97–110, Barcelona, Spain (Online). Association for Computational Linguistics.

Muhammad Abdul-Mageed, Chiyu Zhang, AbdelRahim Elmadany, Houda Bouamor, and Nizar Habash. 2021. Nadi 2021: The second nuanced arabic dialect identification shared task. *arXiv preprint arXiv:2103.08466*.

Muhammad Abdul-Mageed, Chiyu Zhang, AbdelRahim Elmadany, and Lyle Ungar. 2020b. Toward micro-dialect identification in diaglossic and code-switched environments. *arXiv preprint arXiv:2010.04900*.

Mohamed Altantawy, Nizar Habash, Owen Rambow, and Ibrahim Saleh. 2010. Morphological analysis and generation of arabic nouns: A morphemic functional approach. In *Language Resources and Evaluation Conference*.

El-Said Badawi. 1973. *Mustawayat al-Arabiyah al-muasirah fī Misr : bahth fi alaqat al-lughah bi-al-hadarah*. Dār al-Mārif, Cairo.

Reem Bassiouney. 2014. *Language and identity in modern Egypt*. Edinburgh University Press.

Reem Bassiouney. 2017. *Identity and dialect performance: A study of communities and dialects*. Routledge.

Peter Behnstedt and Manfred Woidich. 1985. Die ägyptisch-arabischen dialekte. *Tübinger Atlas des Vorderen Orients/Beihefte/B*, 50.

Douglas Biber. 1993. Representativeness in corpus design. *Literary and linguistic computing*, 8(4):243–257.

Houda Bouamor, Nizar Habash, Mohammad Salameh, Wajdi Zaghouani, Owen Rambow, Dana Abdulrahim, Ossama Obeid, Salam Khalifa, Fadhl Eryani, Alexander Erdmann, et al. 2018. The madar arabic dialect corpus and lexicon. In *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*.

Kristen Brustad. 2017. Diglossia as ideology. In *The politics of written language in the Arab world*, pages 41–67. Brill.

Jonathan Dunn. 2020. Mapping languages: The corpus of global language use. *Language Resources and Evaluation*, 54(4):999–1018.

Amany Fashwan and Sameh Alansary. 2021. A morphologically annotated corpus and a morphological analyzer for egyptian arabic. *Procedia Computer Science*, 189:203–210.

Charles A Ferguson. 1959. Diglossia. *Word*, 15(2):325–340.

Hassan AH Gadalla. 2000. *Comparative morphology of standard and Egyptian Arabic*, volume 5. Lincom Europa Munich.

Nizar Habash. 2007. On arabic transliteration. in van den bosch, a. and soudi, a., editors, arabic computational morphology: Knowledge-based and empirical methods.

Nizar Habash, Mona Diab, and Owen Rambow. 2012. Conventional orthography for dialectal Arabic. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 711–718, Istanbul, Turkey. European Language Resources Association (ELRA).

Nizar Habash, Owen Rambow, Mona Diab, and Reem Kanjawi-Faraj. 2008. Guidelines for annotation of arabic dialectness. In *Proceedings of the LREC Workshop on HLT & NLP within the Arabic world*, pages 49–53.

Nizar Habash, Owen Rambow, and George Anton Kiraz. 2005. Morphological analysis and generation for arabic dialects. In *Proceedings of the ACL Workshop on Computational Approaches to Semitic Languages*, pages 17–24.

Eva Marie Håland. 2017. Adab sākhir (satirical literature) and the use of egyptian vernacular. In *The politics of written language in the Arab world*, pages 142–165. Brill.

H Morcos Hanna. 1962. *The phrase structure of Egyptian colloquial Arabic*, volume 35. Walter de Gruyter GmbH & Co KG.

Richard Slade Harrell. 1957. *The phonology of colloquial Egyptian Arabic*. American Council of Learned Societies.

Jacob Høigilt and Gunvor Mejdell. 2017. *The politics of written language in the Arab world: Writing change*. Brill.

Johannes Hendrik Hospers. 1973. *A Basic Bibliography for the Study of the Semitic Languages: Volume I*, volume 1. Brill Archive.

Abdelghany A Khalafallah. 1969. *A descriptive grammar of saidi Egyptian colloquial Arabic*, volume 32. Walter de Gruyter GmbH & Co KG.

Kristian Takvam Kindt and Tewodros Aragie Kebede. 2017. A language for the people?: Quantitative indicators of written darija and ammiyya in cairo and rabat. In *The politics of written language in the Arab World*, pages 18–40. Brill.

Thomas Leddy-Cecere and Jason Schroepfer. 2019. *Egyptian Arabic*, 2 edition, pages 433–457. Routledge.

Catherine Miller. 2005. Between accomodation and resistance: Upper egyptian migrants in cairo.

Kjell Norlin. 1987. *A phonetic study of emphasis and vowels in Egyptian Arabic*, volume 30. Lund University.

Ridouane Tachicart, Karim Bouzoubaa, Salima Harrat, and Kamel Smaïli. 2022. Morphological Analyzers of Arabic Dialects: A survey. *Studies in Computational Intelligence*, 1061.

Manfred Woidich. 1996. Rural dialect of egyptian arabic: an overview. *Egypte/Monde Arabe*, (27-28):325–354.

A Appendix - Supplementary Materials

Feature	Gloss	Code	SEA	CEA	MSA
Adverbs	Now	Ad1**	دلوقت	دلوقتي	الآن
Adverbs	Very	Ad2*	واصل، خالص	خالص	كثيراً
Adverbs	Outside	Ad3**	برا	برة، بره	الخارج
Adverbs	Also	Ad4***	برضك، برض	برضه، برضو	أيضاً
Adverbs	Very	Ad5***	قوي	أوي	جداً
Prepositions	On	Prep1*	ع، على	ع، على	على
Prepositions	In	Prep2*	ف، في	في	في
Interrogative	Why	Introg1**	لي	ليه	لماذا
Interrogative	Where	Introg2***	وين	فين	أين
Interrogative	When	Introg3***	ميتى	امتى، امتا	متى
Interrogative	How	Introg4**	كيف	ازاي	كيف
Particles	Negation - free	Neg1*	مش +	مش +	ما، ليس، لا
Particles	Negation- bound	Neg2*	ما+ش	ما+ش	-
Particles	Negation	Neg3**	ما+شي	-	-
Particles	Negation	Neg4***	شي+	-	-
Pronouns	I (am)	Pron1**	آني	أنا	أنا
Pronouns	We (are)	Pron2**	نحنا	احنا	نحن
Demonstratives	This (m.)	Dem1*	دا	ده	هذا
Demonstratives	That (m.)	Dem3**	داك	دا	ذلك، ذاك
	This (f.)	Dem5**	داكهي، دي	دي، ديه	هذه
Demonstratives	That (f.)	Dem6**	داكي	دي، ديه	تلك
Demonstratives	These	Dem7**	داكهما	دول	هؤلاء
Demonstratives	Those	Dem8**	ديكهما	دول	أولائك

Table 5: Sample grammatical features distinctions between SEA, CEA, and MSA. *** indicate most marked features, and * the least marked.

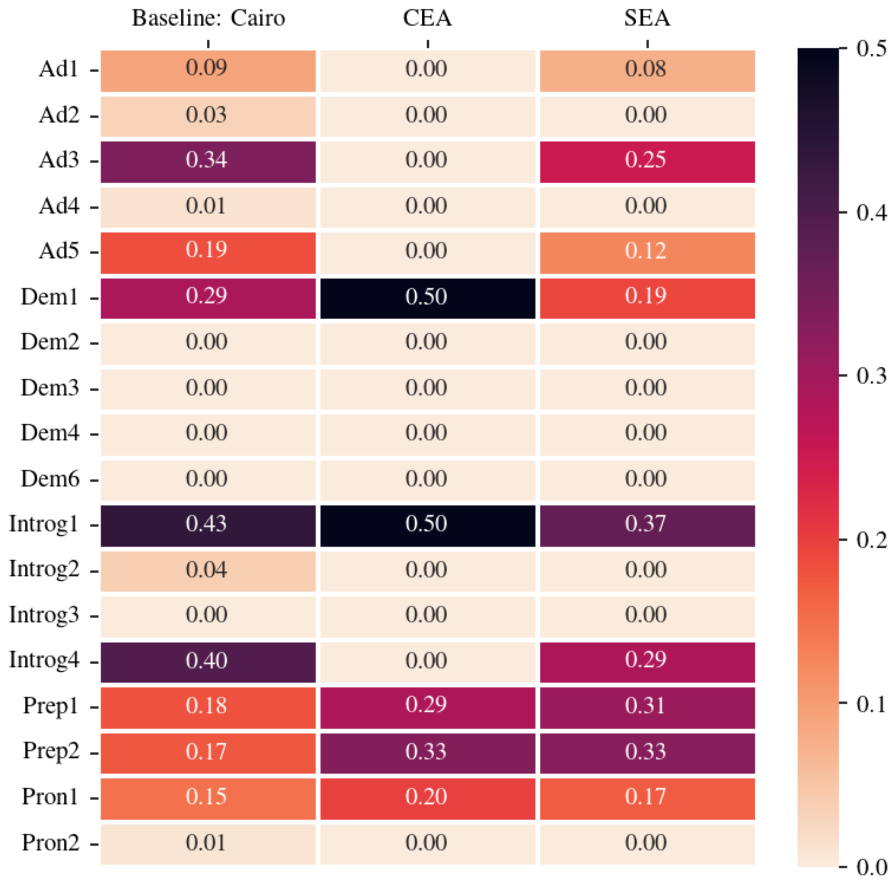


Figure 10: Share of SEA Variants for each Alternation. NADI2020 Corpus compared with Baseline Cairo corpus.

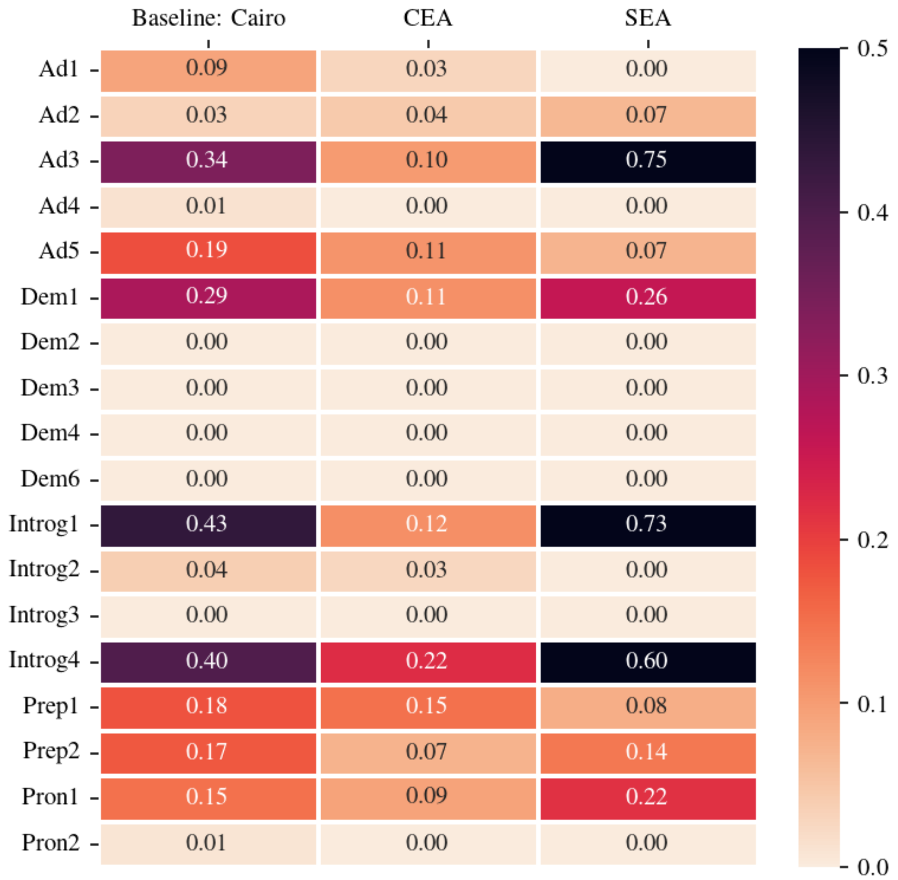


Figure 11: Share of SEA Variants for each Alternation. MicroDialect Corpus compared with Baseline Cairo corpus.

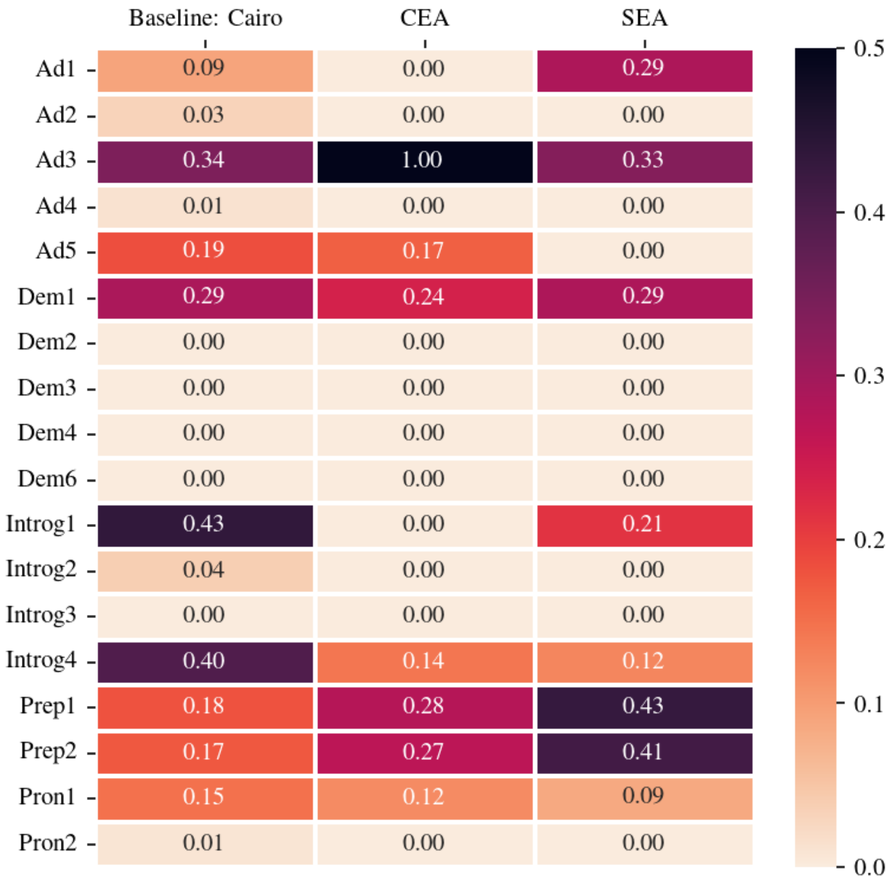


Figure 12: Share of SEA Variants for each Alternation. NADI2021 Corpus compared with Baseline Cairo corpus.

When Elote, Choclo and Mazorca are *not* the Same. Isomorphism-Based Perspective to the Spanish Varieties Divergences

Cristina España-Bonet[1] and **Ankur Bhatt**[1,2,3] and
Koel Dutta Chowdhury[2] and **Alberto Barrón-Cedeño**[4]
[1]DFKI GmbH, [2]Saarland University, Saarland Informatics Campus, Germany
[3]Rheinland-Pfälzische Technische Universität Kaiserslautern-Landau, Germany
[4]Universitá di Bologna, Forlì, Italy
{cristinae, ankur.bhatt}@dfki.de,
koel.duttachowdhury@uni-saarland.de, a.barron@unibo.it

Abstract

Spanish is an official language in 20 countries; in 19 of them, it arrived by means of overseas colonisation. Its close contact with several coexisting languages and the rich regional and cultural diversity has produced varieties that divert from each other. We study these divergences with a data-based approach and according to their qualitative and quantitative effects in word embeddings. We generate embeddings for Spanish in 24 countries and examine the topology of the spaces. Due to the similarities between varieties —in contrast to what happens to different languages in bilingual topological studies— we first scrutinise the behaviour of three isomorphism measures in (quasi-)isomorphic settings: relational similarity, Eigenvalue similarity, and Gromov-Hausdorff distance. We then use the most trustworthy measure to quantify the divergences among varieties. Finally, we use the departures from isomorphism to build relational trees for the Spanish varieties by hierarchical clustering, and observe that *voseo* is the phenomenon that leaves the strongest imprint in the embeddings.

1 Introduction

Language is a reflection of the needs and behaviours of the community that uses and continually transforms it. One language spoken by diverse communities and/or in various regions can exhibit different characteristics. Spanish is a prototypical example: it lies only behind Chinese in terms of number of native speakers (Eberhard et al., 2023) and, different from it, it is a global overseas language (Ammon, 2010) spoken across c. 11.7 M km^2 by people with diverse cultures and needs.

Originating in the Iberian peninsula as a dialect of Latin, Spanish spread throughout America as a consequence of colonisation. The contact with the indigenous languages present in America in the 16th century, subsequent immigration fluxes, diverse language policies, and societal differences have created a wide variety of *Spanishes*. Figure 1 shows the linguistic zones. Some of these factors operate at the country level (e.g., language policies), but most of them operate at the regional level, where a region may be part of a single country or span across several countries. Consequently, political borders do not uniquely define the varieties.

We study the Spanish varieties using data-based approaches. Since large amounts of textual data for Spanish are only available with, at best, country of origin identifiers, one of our goals is to investigate whether natural language processing (NLP) techniques allow to derive relations among countries and varieties from them. Although the varieties are intrinsically different, divergences among them are less prominent than divergences among languages (e.g., Spanish from Mexico and Spanish from Spain are more similar than Spanish and Portuguese). Because of this, methods in NLP that are adequate and work well in multilingual settings might not properly work for language varieties.

For the study, we create per-country word embeddings and examine the topology of the embedding spaces and their relations using isomorphism metrics, which measure distances between embedding spaces and, in our case, between language varieties. We question whether these measures, used mostly in bilingual scenarios, could be adequate in monolingual settings. We widely explore their performance in controlled quasi-isomorphic scenarios (being our conclusions also relevant for bilingual scenarios) and then use the most reliable configurations to measure distances among our embedding spaces and to derive relational trees. Finally, we interpret the Spanish data-based tree in terms of linguistic characteristics. The work aims at two interrelated goals: (*i*) stressing and evaluating isomorphism measures when applied to language variation and (*ii*) studying Spanish varieties in a new data-based approach to gain linguistic insights. Data and models are available on the project website.[1]

[1]https://cereal-es.github.io/CEREAL/

Proceedings of the Eleventh Workshop on NLP for Similar Languages, Varieties, and Dialects (VarDial 2024), pages 56–77

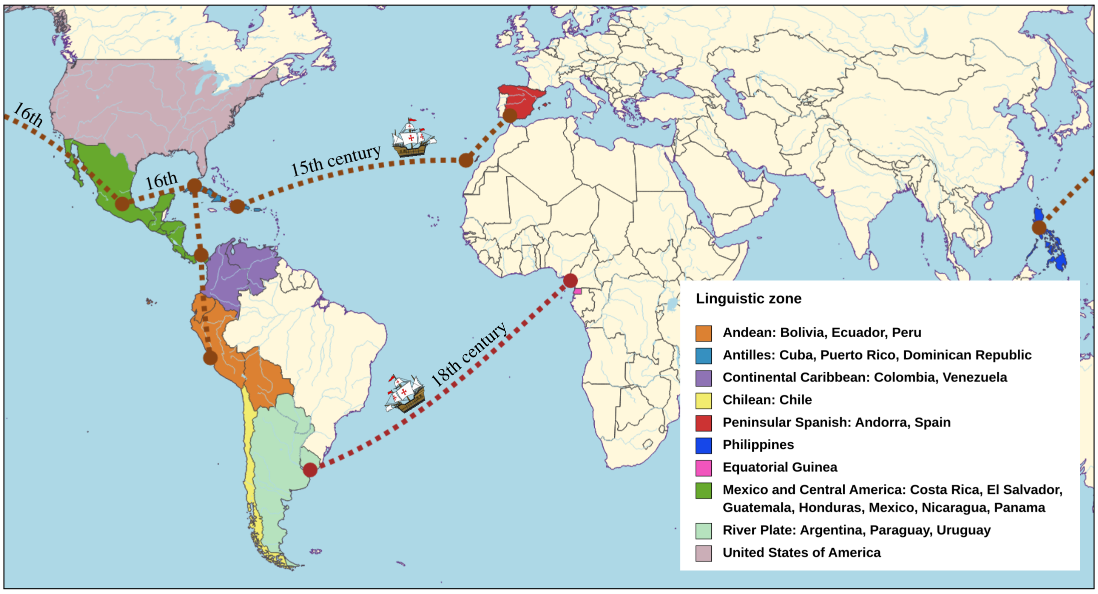


Figure 1: Common geographic Spanish linguistic zones as described by the *Real Academia Española*. The mapping between these linguistic zones and the Spanish varieties might not be one-to-one (Lipski, 2012).

2 The Origins of the Spanish Varieties

Spanish was first derived from Latin in contact with pre-Roman languages in the Iberian peninsula. Different aspects caused proto-Spanish to move away from other Romance languages to be (e.g., Catalan, Portuguese). Two come from long-term rulings during which no language imposition existed. During the Visigothic ruling, multiple words of Germanic origin were introduced, such as `guerra` (war), `riqueza` (wealth) and `yelmo` (helmet). The Arab–Berber control of (up to) two-thirds of the Iberian peninsula, from the 8th until the 15th century, imported novel knowledge and new artifacts, resulting in the introduction of more than $4k$ Arabisms to the lexicon (Alatorre, 1989, p. 74, 80).

The fall of the last remains of Arabic and Mozarabic language varieties came from the 13th until the end of the 15th century, a period during which the kingdom of Castille influenced the speech of neighbouring kingdoms, such as Leon and Aragon (Penny, 2002, p. 19). Internally, two Spanish norms contended: the one of Toledo (later on Madrid) and the one of Seville.

The path of overseas Spanish expansion started with the conquest of the Canary Islands, from which Columbus departed to arrive in the Caribbean in the late 15th century, unfolding in the conquest of Hispaniola (current Dominican R./Haiti) and Cuba. It is from Cuba that the conquest of both Mexico and Peru was launched, as depicted in Figure 1. The norm exported to the new territories was guided by the origin of the migrant population (e.g., priests, soldiers, settlers). Since almost 50% left from Andalusia and Extremadura (López Morales, 1998), these regions hold Spanish varieties derived from the Seville norm. As a result, they share phenomena such as *seseo*,[2] aspirate `/h/`, and an absence of formal/informal differentiation for the second person plural: only `ustedes` is used in America while both `ustedes` (formal) and `vosotros` (informal) are used in Spain (Penny, 2002, pp. 22–23).

It is assumed that the well-established connections across some regions kept the varieties of Spanish in Mexico and Peru closer to those from Spain. A weaker influence on more remote or difficult to reach regions (e.g., Argentina, Paraguay, Uruguay, Central America) allowed for the organic development of farther varieties. The most accepted theory (Penny, 2002, p. 25) is that the influence was driven by centres of power and strength of communication. Consequently, most of Mexico, Peru, Bolivia, and Andean Ecuador share the retention of syllable-final `/s/`, influenced by central Peninsular settlers, whereas other areas miss it, influenced by Southern Peninsular and Canarian Spanish. The pronoun `tú` (you), as used in Spain, is predominant in Mexico, the Caribbean, most of Bolivia and Peru, and part of Venezuela, whereas `vos` competes with `tú` in more remote areas (e.g., Chile, Ecuador,

[2]Phenomenon in which c and s share the phoneme `/s/`.

Colombia) and is predominant in others (e.g., Argentina, Uruguay, Paraguay). The phenomenon is called *voseo* and implies a change in the verbal forms used after the pronoun (Benavides, 2003).

Another relevant factor in the development of the varieties is the long-term influence of coexisting languages. During the whole period of Spanish rule, the native language of most of the population in America was prehispanic, be it of American or of African origin (López Morales, 1998, p. 3). As a result, Spanish in different regions incorporated large vocabularies from other languages. Also, phonetics was affected. One of the most distinctive features of (South) Argentinian and Uruguayan speakers is *zheísmo* (Staggs, 2019), the pronunciation of both `y` and `ll` as [ʒ] due to the influence of the local Amerindian languages. Nowadays, the sound has drifted in some regions to [ʃ] being called *sheísmo*.

Since independence, Spanish varieties in American countries have been significantly influenced by imported languages. During the 19th and 20th centuries, heavy immigration from Italy to Argentina (Cuadrado Rey, 2020) introduced lexical borrowings from Italian (e.g., `gamba` to refer to leg), Neapolitan and others (Bihan, 2011). Due to geographical vicinity and global influence, two foreign languages were the most influential in this same period: French in Spain and English elsewhere (e.g., `computadora` from English is used in most of Latin America vs `ordenador`, from the French `ordinateur`, in Spain).

Due to these intricacies, there is no straight line to draw on a map to separate the Spanish varieties. The varieties might form a continuum and scholars suggest different categorisations. The closest to our purposes is based on the lexicon; Henríquez Ureña (1921) distinguishes the varieties according to the indigenous language substrate: Nahuatl, Caribbean languages, Quechua, Mapudungun/Araucanian, and Guarani. Lipski (2012) defines 21 varieties (11 for Spain, 10 for America), and Soler Montes (2015) defines 8 (3 for Spain, 5 for America). In the former, the classification is based mainly upon phonetic, lexical, and morphosyntactic features; in the latter on geolinguistics. But these are only two examples. Sippola (2021) summarises 4 classifications taking into account geographical variations mostly including phonetic features. Still, the geographical variations are not aligned with geopolitical borders and this has an impact on data-based approaches. Even though few resources exist for the linguistically motivated varieties with city of origin indications (Robelo, 1904; Prieto and Roseano, 2013; Albelda Marco and Estellés, sd), large textual corpora in Spanish are, at best, tagged only with country of origin (Gonçalves and Sánchez, 2014; Tellez et al., 2023; RAE, 2024; España-Bonet and Barrón-Cedeño, 2024).

3 Isomorphism in NLP

Early empirical results using bilingual dictionaries (Youn et al., 2016) and vector embeddings calculated on textual corpora (Mikolov et al., 2013) show that concepts in natural language are structured in a similar way across languages. Vector embeddings in different languages appear to be isomorphic —or at least geometrically similar (Marchisio, 2023). However, other studies show that isomorphism does not always hold, and the more distant a pair of languages or the domain is, the weaker the isomorphism (Søgaard et al., 2018; Patra et al., 2019; Marchisio et al., 2020). But language and domain are not the only factors, differences in training corpus size, training time or the algorithm used to compute the embeddings have a significant effect too (Vulić et al., 2020; Marchisio et al., 2020).

Isomorphism metrics have been introduced in the context of bilingual lexical induction (BLI) where most of the previous conclusions have been drawn. In this context, metrics are used to quantify the similarity (or distance) between embedding spaces of different languages and to observe how they correlate with BLI accuracy.

Several metrics deal with word embeddings from different points of view. Isospectral metrics treat embedding spaces as graphs in the context of spectral graph theory: with respect to the spectral characteristics (e.g., eigenvalues and eigenvectors) of the matrix structures (e.g., adjacency and Laplacian matrices) that represent an embedding space. The Eigenvector similarity distance (Søgaard et al., 2018), the effective condition number (Dubossarsky et al., 2020) and the Spectral Graph-based Matching distance (Dutta Chowdhury et al., 2021) are examples. Isometric measures treat word embeddings as coordinates in a metric space. Earth Mover's distance is a measure of the closeness between the distribution of two sets of words (Zhang et al., 2017) and Relational Similarity is the Pearson's correlation between their cosine similarities (Vulić et al., 2020). The Gromov-Hausdorff distance scores the largest distance between a word from one space and the nearest neighbours from the other space after an isometric transformation between the spaces (Patra et al., 2019).

The mathematical definition of isomorphism, in which two structures are either isomorphic or not, is an approximation in NLP. In NLP, one deals with *degrees of isomorphism* between representative substructures instead. Due to the large vocabularies, and the richness and nuances of natural language, embedding spaces are usually represented by a subgraph/subset formed by up to 5–10 k words. The number of words and which words are used is an ad-hoc decision.

Going beyond the correlation between the isomorphism scores and BLI, the previous metrics have been used to quantify the isomorphism between embedding spaces. When multiple metrics are used, it becomes evident that they do not correlate with each other (Dubossarsky et al., 2020; Dutta Chowdhury et al., 2021; España-Bonet and Barrón-Cedeño, 2022; Marchisio et al., 2022).

In this work, we analyse relations among 24 varieties of Spanish using isomorphism metrics. We expect differences across varieties of the same language to be much smaller than across different languages. Therefore, we first calibrate the isomorphism measures in isomorphic settings —same language, same training data, same embedding algorithm, and hyperparameters (Section 6). As these metrics do not correlate, this allows us to determine the best metric and configuration (number and selection of words) to perform the fine-grained analysis among varieties (Section 7).

4 Isomorphism Measures

We select three measures that capture the isomorphic/isometric degree between two embedding spaces E_1 and E_2 represented by nearest-neighbour graphs G_1 and G_2 and sets of points S_1 and S_2. We assume that the embeddings E_1 and E_2 are mean-centred and length-normalised.

Relational similarity (RS) (Vulić et al., 2020). One can presume that the similarity between words is distributed similarly in different spaces and, so, the cosine similarity of aligned words should be similar in both spaces. RS uses a list with k words from E_1 aligned to k words from E_2 (a dictionary) and calculates the cosine similarities between all the pairs of words in E_1 and E_2 independently:

$$\begin{aligned} &\mathrm{sim}_{E_1}(S^p, S^r) \;\; \forall S^p, S^r, p \neq r \in \mathrm{list}(E_1) \\ &\mathrm{sim}_{E_2}(S^p, S^r) \;\; \forall S^p, S^r, p \neq r \in \mathrm{list}(E_2) \end{aligned} \quad (1)$$

RS is the Pearson correlation ρ between the sorted lists of similarities resulting from the spaces:

$$\mathrm{RS} = \rho\left(\mathrm{sim}_{E_1}^{\mathrm{sorted}}, \mathrm{sim}_{E_2}^{\mathrm{sorted}}\right). \quad (2)$$

Eigenvector similarity (EV) distance (Søgaard et al., 2018). A total of k words in E_i are used to construct n-nearest neighbour unweighted graphs G_i. The nearest neighbours are extracted by computing the cosine similarity between the k words in E_i and all the words in E_j. Given G_i, EV estimates the degree of isomorphism from the eigenvalues of the Laplacian of G_1 and G_2. Let the Laplacian be

$$L_i = D_i - A_i(G_i), \quad (3)$$

where A_i is the adjacency matrix of G_i, and D_i is the diagonal matrix of degrees. After computing the Laplacian eigenvalues, following Søgaard et al. (2018), one finds the smallest m such that the sum of the m largest Laplacian eigenvalues is <90% of the total. Using the smallest m of E_1 and E_2, EV is defined as

$$\mathrm{EV} = \sum_{j=1}^{m} (\lambda_{1j} - \lambda_{2j})^2, \quad (4)$$

where λ_{ij} are the top j eigenvalues of L_i.

Gromov-Hausdorff (GH) and Bottleneck distances (Patra et al., 2019). GH is an isometric measure that treats word embeddings as coordinates in a metric space. It gives the worst-case distance (E_1 vs E_2) of nearest neighbours in a shared embedding space after an optimal isometric transformation.

For each word x in S_i, one finds its nearest neighbour y in S_j (NN_j). The Hausdorff distance H is the largest of the two distances:

$$\mathrm{H} = \max\left(\mathrm{dist}(x_1, \mathrm{NN}_2), \mathrm{dist}(x_2, \mathrm{NN}_1)\right). \quad (5)$$

The Gromov-Hausdorff distance is the infimum of the Hausdorff distances under all possible orthogonal transformations. Since computing GH is an NP-hard problem, the Bottleneck distance B, bounded by GH, is used as an approximation (Chazal et al., 2009). B is the shortest distance for which there exists a perfect matching between the points p and r of the persistent diagrams[3] built from S_1 and S_2:

$$\mathrm{B} = \inf_{matches} \max_{(p,r)} ||p - r||_\infty. \quad (6)$$

[3] A persistent diagram is a set of points in $\mathbb{R}^2$ in the half-plane above the diagonal.

Country & Code		CEREAL Segments	CEREAL Words	CEREAL Vocab.	Twitter Vocab.
Andorra	ad	13,023	543,047	2,671	–
Argentina	ar	20,950,705	986,413,066	284,191	673,424
Bolivia	bo	975,429	49,518,821	53,799	47,012
Chile	cl	12,079,476	548,257,312	199,493	282,737
Colombia	co	8,323,794	375,326,751	163,212	324,635
Costa Rica	cr	825,513	37,760,657	45,893	103,086
Cuba	cu	1,919,998	93,368,177	82,275	18,682
Dominican R.	do	1,183,336	48,726,587	52,409	108,655
Ecuador	ec	1,624,269	66,662,454	64,312	147,560
Spain	es	20,950,705	880,495,659	596,842	571,196
Eq. Guinea	gq	4,050	329,469	1,698	1,167
Guatemala	gt	561,714	23,421,191	35,860	95,252
Honduras	hn	656,212	24,971,660	35,707	60,580
Mexico	mx	20,875,244	912,645,564	250,313	438,136
Nicaragua	ni	405,935	18,921,537	31,345	68,605
Panama	pa	448,974	18,431,387	31,268	111,635
Peru	pe	5,066,369	213,937,404	122,884	178,113
Philippines	ph	1,382	75,761	405	–
Puerto Rico	pr	128,103	5,619,179	15,062	23,062
Paraguay	py	775,101	33,771,401	46,513	124,162
El Salvador	sv	401,348	17,068,212	29,433	73,833
USA	us	376,839	21,335,770	34,368	292,465
Uruguay	uy	1,804,329	85,809,183	75,491	200,032
Venezuela	ve	1,201,624	55,514,289	59,334	271,924

Table 1: Number of segments and words used to compute the variety-specific word embeddings.

Data Points (Word) Selection In all the measures above, we characterise each embedding space by k words ($k \in \{100, 500, 1000, 2500, 5000\}$) following 5 criteria:

• **Most frequent words (Frequent, MFW).** We use the top-k words in an embedding space ranked by frequency. This is the standard choice for EV and GH in previous work.

• **Random words (Random).** We randomly select k words within the top half of the frequency-ranked embeddings.

• **Aligned random words (Random BiDict).** As in Random, but only words that appear simultaneously in the two spaces are considered. This is equivalent to using a bilingual lexicon in the general case, which is the standard choice for RS.

• **Numbers.** Random k numbers appearing simultaneously in the two spaces.

• **Named Entities (NEs).** Random k NEs appearing simultaneously in the two spaces. The list of NEs contains 3,416 single words extracted from the CoNLL-2002 shared task (Tjong Kim Sang, 2002).

We adapt the implementation for RS, EV, and GH in Vulić et al. (2020)[4] to consider our lists.

[4] https://github.com/cambridgeltl/iso-study

5 Variety-Specific Embedding Spaces

We use the CEREAL corpus (España-Bonet and Barrón-Cedeño, 2024) to obtain embedding spaces for 24 varieties of Spanish. CEREAL contains documents in Spanish extracted from OSCAR (Open Super-large Crawled Aggregated coRpus version 22.01 (Ortiz Suárez et al., 2019; Abadji et al., 2021) and annotated with the country of origin. We use the documents in CEREAL where the country of publication is codified in the URL and discard documents whose country was inferred automatically (CEREALex). To compute the embeddings, we eliminate sentences having only punctuation and numbers, as well as those with at least one Arabic, Chinese, Cyrillic or Greek character. We then normalise and tokenise the texts using Moses' scripts (Koehn et al., 2007) and lowercase. Table 1 shows the statistics of the final dataset together with the code we use to identify each variety. We estimate fastText (Bojanowski et al., 2017) embeddings using the default skip-gram configuration to train 300-dimensional embeddings for tokens appearing at least 20 times.

The amount of text in Spanish from Spain in CEREAL is significantly larger than for the other varieties (70.5 M segments for es vs 20.9 M for ar, the second largest). For comparability reasons, we use a subset of 20 M segments to train es embeddings. With this, ar, es, and mx have a similar amount of training data, whereas ad, gq, ph, pr, sv and us have less than 0.5 M segments and are discarded for our high-resourced experiments.

Our calibration experiments (Section 6) are done with Spanish from Spain embeddings. We create 10 models from CEREAL: 5 models using 5 seeds for fastText on a fixed subset of 20 M segments (model perturbation) and 5 models using 5 random extractions of 20 M segments over the whole 70.5 M segments (data perturbation).

Our exploration experiments (Section 7) consider embeddings for the 24 varieties. We generate 3 embedding models per variety with different seeds and show the mean in our results. In this case, we also use existing Twitter embeddings (currently X) for 22 varieties (Tellez et al., 2023).[5] Since the training corpus is not available, we use their pretrained embeddings with a single run. Otherwise, our setting with CEREAL is comparable to theirs except for the minimum frequency of in-vocabulary

[5] https://ingeotec.github.io/regional-spanish-models

tokens (the default being 5 in their case) and the fact that they remove diacritics from the data.

6 Isomorphism Measures Calibration

If isomorphism metrics are a good measure to account for distances among languages, they should drop to zero when computing the distance between embeddings of a single language —or approach 1 when they imply correlations. As described in Section 3, differences in size and domain of the training data and in the algorithm used to train the embeddings affect their performance. In this section, we isolate all these factors and evaluate the metrics in an isomorphic setting: same variety (es), same corpus (CEREAL) and same algorithm (skip-gram). We perturb the basic setting by (*i*) applying several initialisations for training the embeddings while keeping skip-gram and all its parameters constant (model perturbation) and (*ii*) subsampling the training data from a larger dataset (data perturbation). With these variations, we aim to study the robustness of the metrics to minor changes and at determining the best configuration for each of them. This study provides insights on the feasibility of using one or more isomorphism metrics to explore relations between language varieties in Section 7.

	Model Perturbation			**Data Perturbation**		
	RS↑	**EV↓**	**GH↓**	**RS↑**	**EV↓**	**GH↓**
Frequent						
100	0.989±0.000	2±1	0.02±0.00	0.860±0.056	2±1	0.03±0.01
500	0.982±0.001	2±1	0.02±0.00	0.314±0.032	3±1	0.02±0.00
1000	0.979±0.002	3±1	0.02±0.00	0.131±0.014	2±1	0.02±0.00
2500	0.976±0.002	3±1	0.01±0.00	0.038±0.000	4±1	0.01±0.00
5000	0.974±0.003	5±2	0.01±0.00	0.015±0.001	5±4	0.01±0.00
Random						
100	0.000±0.008	3±1	0.17±0.12	0.002±0.015	5±1	0.15±0.06
500	0.000±0.002	5±3	0.15±0.06	0.000±0.001	5±2	0.19±0.10
1000	0.000±0.000	6±2	0.16±0.07	0.000±0.000	5±1	0.13±0.05
2500	0.000±0.000	10±3	0.11±0.03	0.000±0.000	7±1	0.13±0.03
5000	0.000±0.000	14±7	0.07±0.01	0.000±0.000	14±5	0.12±0.04
Random BiDict						
100	0.959±0.002	1±1	0.03±0.01	0.884±0.008	3±2	0.05±0.02
500	0.959±0.002	3±1	0.02±0.00	0.882±0.004	4±1	0.03±0.01
1000	0.959±0.002	4±1	0.02±0.00	0.883±0.002	6±4	0.03±0.01
2500	0.960±0.002	7±3	0.02±0.00	0.883±0.001	7±2	0.03±0.01
5000	0.960±0.002	5±2	0.01±0.00	0.883±0.000	8±4	0.02±0.00
Numbers						
100	0.997±0.000	3±1	0.05±0.05	0.604±0.087	3±1	0.06±0.04
500	0.994±0.000	4±1	0.02±0.00	0.116±0.012	5±2	0.05±0.00
1000	0.993±0.001	5±2	0.02±0.00	0.061±0.008	9±6	0.02±0.00
2500	0.988±0.001	9±5	0.02±0.00	0.037±0.007	12±4	0.02±0.00
5000	0.985±0.001	7±1	0.02±0.00	0.022±0.003	11±3	0.05±0.01
NEs						
100	-0.003±0.013	3±1	0.10±0.03	-0.001±0.018	2±1	0.08±0.02
500	0.002±0.007	7±3	0.07±0.03	-0.003±0.007	3±1	0.07±0.03
1000	0.002±0.006	6±3	0.09±0.03	-0.002±0.004	7±2	0.07±0.02
2500	0.000±0.003	7±3	0.03±0.00	-0.002±0.002	7±2	0.05±0.02
5000	–	–	–	–	–	–

Table 2: Mean and standard deviation ($\mu \pm \sigma$) score for the three isomorphism metrics used in this study. Perfect isomorphism implies RS 1, and EV and GH 0.

Model Perturbation We use 5 embedding models for es trained with different seeds on the same data (i.e. the vocabulary is the same for all 5). We calculate RS, EV and GH for 10 combinations of embeddings with the 25 possible configurations of Section 4. Detailed results for the pairwise combinations are in Appendix A and the mean over the 10 combinations is in Table 2.

For all three metrics, there are definite trends when the mean is considered, but the trends are less evident when looking at individual embedding pairs. The variations, due only to different runs, are significant. Frequent and Random BiDict perform the best; i.e. distances EV and GH are the smallest and correlation RS the highest. In this setting, the most frequent words in both es spaces are the same and therefore behave similarly to a dictionary —this does not need to happen in the data perturbation setting and even less in the general multilingual setting. As expected, random words unrelated across spaces perform the worst. Also, numbers and NEs do not perform well (except for RS with numbers). This might be related to the fact that they cluster in a specific region of the space and cannot represent the topology of the whole. In terms of the number of datapoints, EV performs best with few, GH with a large set and RS in this setting does not seem to be sensitive to the volume.

Data Perturbation We use 5 embedding models for es trained with different random subsets of the same corpus (i.e. the vocabulary of the models is different). As before, we calculate RS, EV, and GH for a total of the 10 combinations of embeddings, and use 5 different types and a number of datapoints.

Contrary to what one could expect, the perturbation of the dataset —within the same corpus— does not bring more variability on the results of the metrics than the perturbation of the model as measured by the standard deviations (Table 2). The trends with respect to the number and types of points are also similar to the previous case; the global scores are slightly worse but compatible within the 1σ CIs for GH and EV; differences are larger for RS. Ideally, a good metric would score a distance of 0 (EV and GH) and correlation 1 (RS); EV achieves this at 2σ level, especially when using the most frequent words. We consider this configuration, EV (MFW

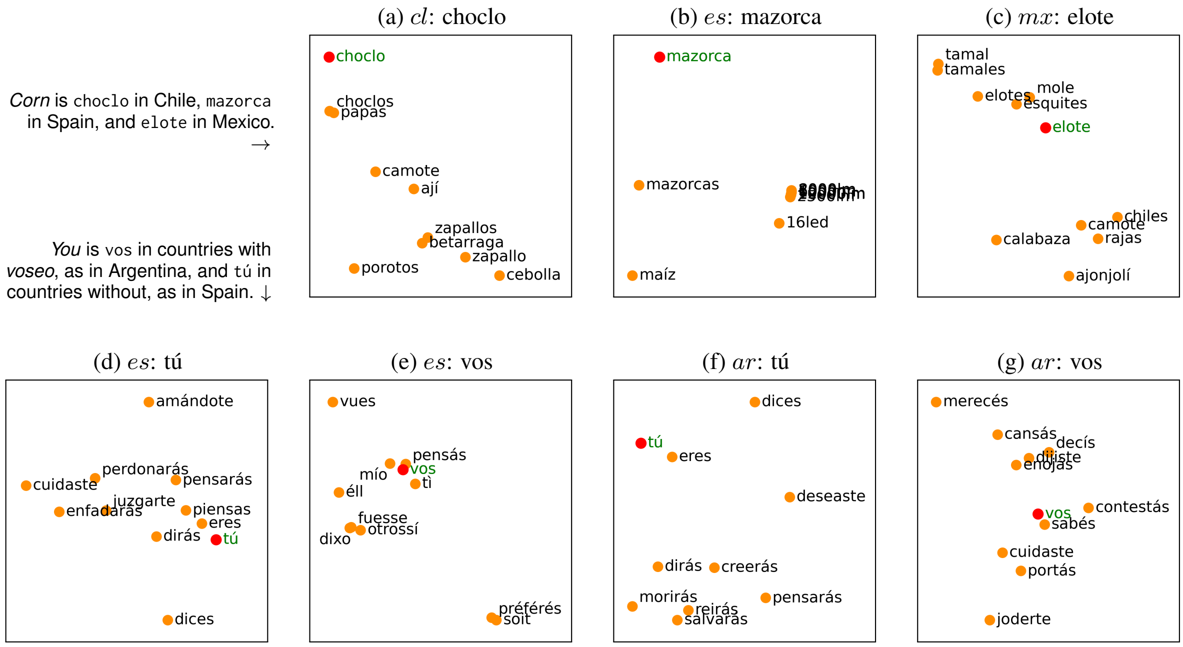


Figure 2: t-SNE projections (van der Maaten and Hinton, 2008) of the neighbourhood for the Spanish words equivalent to *corn* (top plots) and *you* (bottom plots).

100), the best metric to measure isomorphism.

This setting is close to our case of study: language varieties. We extract the data for training the embeddings from the same corpus and use skip-gram with the same configuration. Larger/smaller scores and standard deviations for the isomorphism metrics than the ones we see here should be attributed to language differences and to the quality of the embeddings given by the amount of training data per variety.

7 Spanish Varieties Relations

Qualitative Behaviour Different lexicons and cultural-dependent (near) synonyms change the topology of the embedding spaces. *Corn* in English translates into `elote` (from the Náhuatl `elotitutl`) in most of Mesoamerica, `choclo` (from the Quechua `chuqllu`) in South America, `mazorca` (from the Arabic `masúrqa`) in Colombia, Cuba and Spain and `jojoto` in Venezuela. The importance of this cereal in Spain is irrelevant in comparison to Mesoamerica, where it is so essential that it goes beyond staple food, and that changes the usage of the word.

This is reflected in Figure 2 (top plots), which shows the 10-nearest neighbours for `choclo` in cl, `mazorca` in es, and `elote` in mx. For cl and mx, `choclo` and `elote` are surrounded by other food-related words, but the intersection is almost null. For es, `mazorca` is surrounded mostly by words related to another sense of the word (a kind of light bulb). Appendix B, shows the same three words as located in all three embedding spaces —the three synonyms never appear in the same region of the space and `elote` does not even appear in cl. The surrounding words also vary from being local food when the word is shown in its native embedding space to foreign food when the word is in the embedding space corresponding to another country. Similarly, there are differences across varieties in the verbal forms usage and other grammatical issues, such as *voseo*, which also distort the embedding spaces. As Figure 2 (bottom plots) shows, for countries without *voseo*, such as Spain, the word `vos` is surrounded mostly by non-Spanish words (since it is a nearly-deprecated pronoun for this variety). In Argentina, we observe *verbal voseo*, that is, the usage of the modified 5th inflexion and the 7th verbal inflection instead of the 2nd inflection (e.g., `sabés` or `decís` rather than `sabes` or `dices`).

Isomorphism Following the results of Section 6, we select EV (MFW 100) for the main analysis and include the top-2 performing configurations per metric in the Appendix C as they give more insights on the behaviour of the metrics.

Figure 3 shows the results for EV. The heatmap combines the results with our `CEREAL` embeddings (top–right triangle) and the publicly available Twit-

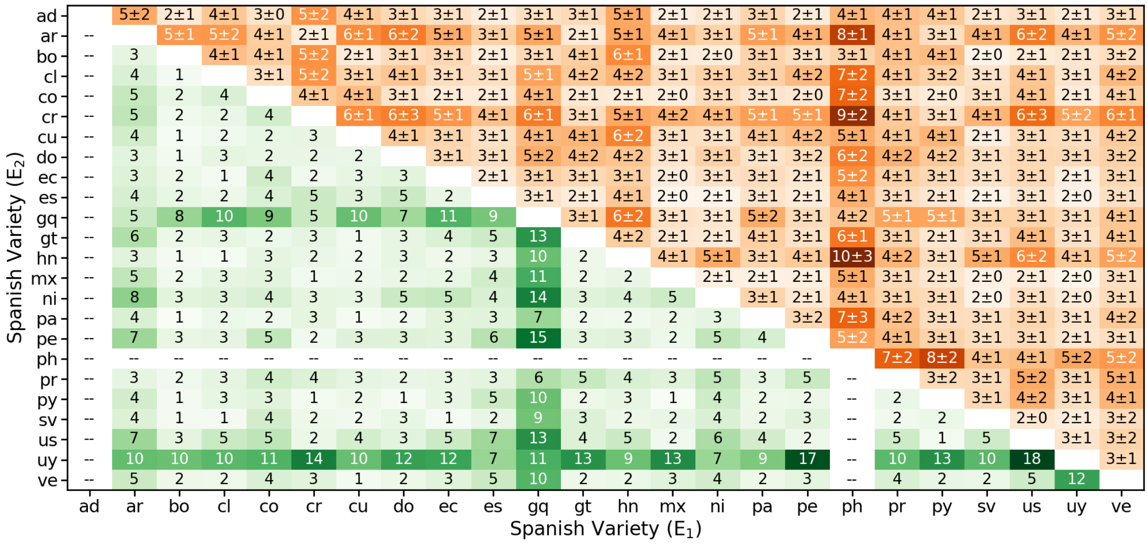


Figure 3: EV with 100 most frequent words for the 24 Spanish varieties. Top–right triangles (orange) correspond to the mean results with the `CEREAL` corpus and bottom–left triangle (green) to the `Twitter` corpus.

ter embeddings (bottom–left triangle). With our in-house embeddings, we calculate mean and standard deviation over 9 combinations per language pair, as we have three runs per variety. This is not possible with Twitter, so we expect more robust results with `CEREAL`.

Comparing the heatmap of EV with those of RS and GH, one sees that both but especially GH are sensitive to the size of the training data. The fact that GH needs to use more datapoints (5,000 MFW in contrast to the 100 for EV/RS) might make the effect of the data size stronger on GH. Varieties *ad*, *gq* and *ph* for `CEREAL` and *gq* for Twitter include less than 15 k training sentences. Pairs involving these varieties have statistically significant higher GH and lower RS values (and to a lesser extent also higher EV values) systematically for all the pairs. Vulić et al. (2020) showed in their correlation analysis between the isomorphism metrics and BLI accuracy that one needs at least 500 k training sentences to convergence in BLI accuracy. Therefore isomorphism scores with embeddings trained with less data might be suboptimal.

The results with `CEREAL` and `Twitter` are significantly different both in the absolute magnitude of the scores for pairs of varieties and in the relations between the varieties. This could be a consequence of the different volumes of training data but also of the differences in the register used in both genres. As Lipski (2012) notes, social factors are also relevant in the variation and both genres might be representative of different population profiles. The standard deviations in Figure 3 are of the same order of magnitude as in our calibration experiments (Section 6) and do not depend on the quality of the embeddings as measured by the data size. Therefore, differences in the scores across language pairs, that is, different departures from isomorphism, are representative of the distances (relations) among varieties. Next, we use hierarchical clustering to have a clearer overview of these relations.

Phylogenetic (Relational) Trees Clustering these results over varieties allows for building a Spanish phylogenetic tree. Notice that, strictly speaking, we do not construct a phylogenetic tree but a relational tree. Spanish was acquired in most American countries almost simultaneously, and varieties have been evolving in parallel since then. Also, word embeddings are static (in time) and they do not provide evolutionary relationships but an average snapshot of the language relationships. Following Dutta Chowdhury et al. (2020), we use agglomerative clustering with variance minimisation (Ward Jr, 1963) for this purpose.

We show the results for all the varieties and the best configuration option for RS and GH in Figure 4 as the comparison among metrics is especially relevant. Appendix D includes the remaining configurations. This representation makes more evident the fact that GH groups the varieties according to the amount of training data —and therefore the quality of the embeddings. On the left-hand side of the GH dendrogram are the least resourced varieties: *ph*, *ad* and *gq*. On the right-hand side are

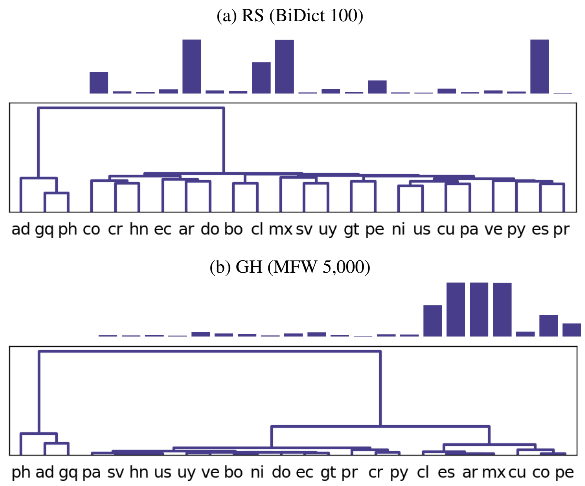


Figure 4: Relational trees derived from the CEREAL embeddings. The distribution of data is shown on top of the dendrograms as an illustration.

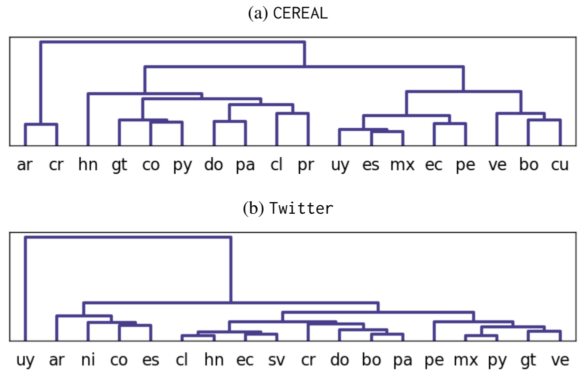


Figure 5: Relational trees for the subset of the highest resourced varieties with EV (MFW 100).

the highest resourced ones: *cl*, *es*, *ar*, *mx*, *cu*, *co* and *pe*. RS also clearly clusters *ad*, *gq* and *ph* and puts all the other varieties at a similar level. The Spearman rank correlation between the number of segments used to train the embeddings and the flattened version of the hierarchical clustering output is 0.8 for GH, 0.2 for RS and -0.1 for EV. The limitations with GH and RS are in agreement with the observations of the previous sections.

In Figure 5, we analyse in detail EV for CEREAL and Twitter for the highest resourced varieties. A visual representation on a map is in Appendix D. None of these trees groups the varieties according to their geographical position or the linguistic zones described by RAE (cf. Figure 1). It is worth pointing out that phonetic differences are in principle not observable with word embeddings on textual data but might leave traces on Twitter embeddings as a result of misspellings. This translates into Argentina (*ar*) and Uruguay (*uy*) —countries where *zheísmo* is present— lying apart in the CEREAL dendrogram, but not in the Twitter one.[6] Trends related to grammar are more evident. The right-hand side of the plots group together varieties without *voseo*: in the case of CEREAL, the exceptions are Uruguay (*uy*) and Dominican Republic (*do*) which should be swapped according to this characteristic; in the case of Twitter, Spain (*es*) sneaks in the region with *voseo*.

Different substrates are in general not observed. Contrary to *voseo* and the grammatical differences it implies, different substrates or neologisms are not global: Quechuan languages were/are spoken in what is today Argentina, Bolivia, Chile, Colombia, Ecuador and Peru; but in Bolivia there are 36 other languages such as Quechua but also Aimara, Chiquitano, etc.

[6]Both countries also share an Italian substrate.

8 Conclusions

Spanish is not a monolithic language. Five centuries of distinct but related evolution across territories have created a rich set of varieties. We study these varieties from a data-based perspective, building specific embeddings with textual data for 24 countries. We then relate the similarities and differences among embedding spaces to the divergences among varieties.

Divergences are subtle in comparison to divergences among languages. Because of this, we explore three common isomorphism metrics in quasi-isomorphic settings. Our results show that EV is the best performing metric in the controlled scenario (data perturbation). GH does not perform far, but subsequent experiments with the variety of embeddings show that it is the metric that depends the most on the amount of training data. RS rapidly degrades when we depart from the controlled experiments and it is less sensitive to the variations.

Lots of characteristics of the language coexist in written documents. The indigenous language substrate and other borrowings, grammatical characteristics such as *voseo*, and verbal tense changes are manifested in word embeddings. *Voseo* showed to be the strongest feature and its imprint is clearly seen in the relational trees we build from the departures from isomorphism obtained with EV. Informal (and sometimes incorrect) text used to create Twitter embeddings also reflects distinctive phonetic traits such as *zheísmo*.

Limitations

We have done an exhaustive exploration of the behaviour of the isomorphism metrics when the same language (Spanish from Spain) is used. The effect of the training domain and data size has been explored before in bilingual settings (Vulić et al., 2020). In this work, we do not systematically quantify the effect that the different sizes in the training data per variety imply, further than removing the varieties with less data according to the conclusions in Vulić et al. (2020). Differences in the amount of data can also imply differences in the domain (especially when few data are available) and these variations have to be taken into account when drawing conclusions.

Acknowledgements

This work has been supported by the German Research Foundation (Deutsche Forschungsgemeinschaft) under grant SFB 1102: Information Density and Linguistic Encoding and by the LT-Bridge Project (GA 952194).

References

Julien Abadji, Pedro Javier Ortiz Suárez, Laurent Romary, and Benoît Sagot. 2021. Ungoliant: An optimized pipeline for the generation of a very large-scale multilingual web corpus. Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-9) 2021. Limerick, 12 July 2021 (Online-Event), pages 1–9, Mannheim. Leibniz-Institut für Deutsche Sprache.

Antonio Alatorre. 1989. *Los 1001 años de la lengua española*. El Colegio de México / Fondo de Cultura Económica, Mexico City, Mexico.

Marta Albelda Marco and María Estellés. sd. Corpus ameresco. [online] Last accessed on 13.03.2024.

Ulrich Ammon. 2010. World languages: Trends and futures. In *The Handbook of Language and Globalization*, chapter 4, pages 101–122. John Wiley & Sons, Ltd.

Carlos Benavides. 2003. La distribución del voseo en hispanoamérica. *Hispania*, 86(3):612–623.

Ulysse Le Bihan. 2011. *Italianismos en el habla de la Argentina: herencia de la inmigración italiana*. Ph.D. thesis, Unniversiteter i Oslo.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Frédéric Chazal, David Cohen-Steiner, Leonidas J Guibas, Facundo Mémoli, and Steve Y Oudot. 2009. Gromov-Hausdorff stable signatures for shapes using persistence. In *Computer Graphics Forum*, volume 28, pages 1393–1403. Wiley Online Library.

Analía Cuadrado Rey. 2020. *El italiano en la fraseología actual del español hablado en Argentina*. Edizioni Ca' Foscari.

Haim Dubossarsky, Ivan Vulić, Roi Reichart, and Anna Korhonen. 2020. The secret is in the spectra: Predicting cross-lingual task performance with spectral similarity measures. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2377–2390, Online. Association for Computational Linguistics.

Koel Dutta Chowdhury, Cristina España-Bonet, and Josef van Genabith. 2020. Understanding translationese in multi-view embedding spaces. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6056–6062, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Koel Dutta Chowdhury, Cristina España-Bonet, and Josef van Genabith. 2021. Tracing source language interference in translation with graph-isomorphism measures. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 375–385, Held Online. INCOMA Ltd.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig, editors. 2023. *Ethnologue: Languages of the World*. SIL International, Dallas, TX.

Cristina España-Bonet and Alberto Barrón-Cedeño. 2022. The (undesired) attenuation of human biases by multilinguality. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2056–2077, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Cristina España-Bonet and Alberto Barrón-Cedeño. 2024. Elote, Choclo and Mazorca: on the Varieties of Spanish. In *Proceedings of the 2024 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, Mexico City, Mexico. Association for Computational Linguistics.

Bruno Gonçalves and David Sánchez. 2014. Crowdsourcing dialect characterization through Twitter. *PloS one*, 9(11):e112074.

Pedro Henríquez Ureña. 1921. Observaciones sobre el español en América. *Revista de filología española*, 8:357.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

John M. Lipski. 2012. *Geographical and Social Varieties of Spanish: An Overview*, chapter 1. John Wiley & Sons, Ltd.

Humberto López Morales. 1998. La aventura del español en América.

Kelly Marchisio. 2023. *Multilinguality from Static Embedding Spaces: Algorithmic, Geometric, and Data Considerations*. Ph.D. thesis, Johns Hopkins University.

Kelly Marchisio, Kevin Duh, and Philipp Koehn. 2020. When does unsupervised machine translation work? In *Proceedings of the Fifth Conference on Machine Translation*, pages 571–583, Online. Association for Computational Linguistics.

Kelly Marchisio, Ali Saad-Eldin, Kevin Duh, Carey Priebe, and Philipp Koehn. 2022. Bilingual lexicon induction for low-resource languages using graph matching via optimal transport. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2545–2561, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Tomas Mikolov, Quoc V Le, and Ilya Sutskever. 2013. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*.

Pedro Javier Ortiz Suárez, Benoît Sagot, and Laurent Romary. 2019. Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures. Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-7) 2019. Cardiff, 22nd July 2019, pages 9–16, Mannheim. Leibniz-Institut für Deutsche Sprache.

Barun Patra, Joel Ruben Antony Moniz, Sarthak Garg, Matthew R. Gormley, and Graham Neubig. 2019. Bilingual lexicon induction with semi-supervision in non-isometric embedding spaces. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 184–193, Florence, Italy. Association for Computational Linguistics.

Ralph Penny. 2002. *A History of the Spanish Language*. Cambridge University Press.

Pilar Prieto and Paolo Roseano. 2013. Atlas interactivo de la entonación del español. [online] Last accessed on 13.03.2024.

Real Academia Española RAE. 2024. Corpus del Español del Siglo XXI (CORPES). [online] Last accessed on 13.03.2024.

Cecilio A. Robelo. 1904. *Diccionario de aztequismos*.

Eeva Sippola. 2021. *Morphosyntactic Variation in Spanish: Global and American Perspectives*, pages 209–232. Cambridge University Press, United Kingdom.

Anders Søgaard, Sebastian Ruder, and Ivan Vulić. 2018. On the limitations of unsupervised bilingual dictionary induction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 778–788, Melbourne, Australia. Association for Computational Linguistics.

Carlos Soler Montes. 2015. El modelo de lengua en el aula de ELE: Adecuación de la variedad lingüística desde un punto de vista pluricéntrico. In *La enseñanza de ELE centrada en el alumno*, pages 1237–1244. Universidad Carlos III de Madrid/ Asociación para la Enseñanza de Español como Lengua Extranjera.

Cecelia Staggs. 2019. A perception study of rioplatense spanish. *McNair Scholars Research Journal*, 14(1):11.

Eric S. Tellez, Daniela Moctezuma, Sabino Miranda, Mario Graff, and Guillermo Ruiz. 2023. Regionalized models for Spanish language variations based on Twitter. *Language Resources and Evaluation*, pages 1–31.

Erik F. Tjong Kim Sang. 2002. Introduction to the CoNLL-2002 shared task: Language-independent named entity recognition. In *COLING-02: The 6th Conference on Natural Language Learning 2002 (CoNLL-2002)*.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing Data using t-SNE . *Journal of Machine Learning Research*, 9:2579–2605.

Ivan Vulić, Sebastian Ruder, and Anders Søgaard. 2020. Are all good word vector spaces isomorphic? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3178–3192, Online. Association for Computational Linguistics.

Joe H Ward Jr. 1963. Hierarchical grouping to optimize an objective function. *Journal of the American statistical association*, 58(301):236–244.

Hyejin Youn, Logan Sutton, Eric Smith, Cristopher Moore, Jon F Wilkins, Ian Maddieson, William Croft, and Tanmoy Bhattacharya. 2016. On the universal structure of human lexical semantics. *Proceedings of the National Academy of Sciences*, 113(7):1766–1771.

Meng Zhang, Yang Liu, Huanbo Luan, and Maosong Sun. 2017. Earth mover's distance minimization for unsupervised bilingual lexicon induction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1934–1945, Copenhagen, Denmark. Association for Computational Linguistics.

A Calibration of the Isomorphism Metrics

Tables 3 to 8 show the detailed results for the experiments in Section 6. In all cases, tables report the results for a given isomorphism metric (either RS, EV or GH) using the 5 types of data points and 5 different number of points defined in Section 4. Metrics are evaluated on pairs of embeddings spaces $\{E_i, E_j\}$ all of them belonging to Spanish from Spain under the two training conditions of Section 6: model perturbation (Tables 3, 4 and 5) and data perturbation (Tables 6, 7 and 8).

		$\mu \pm \sigma$	E1 E2	E1 E3	E1 E4	E1 E5	E2 E3	E2 E4	E2 E5	E3 E4	E3 E5	E4 E5
Frequent	100	0.989 ± 0.000	0.989	0.988	0.988	0.989	0.990	0.989	0.990	0.989	0.990	0.989
	500	0.982 ± 0.001	0.981	0.981	0.981	0.981	0.984	0.981	0.985	0.981	0.984	0.981
	1000	0.979 ± 0.002	0.978	0.977	0.978	0.978	0.982	0.978	0.983	0.978	0.982	0.977
	2500	0.976 ± 0.002	0.974	0.974	0.975	0.974	0.979	0.974	0.981	0.975	0.980	0.974
	5000	0.974 ± 0.003	0.973	0.972	0.974	0.972	0.979	0.973	0.981	0.973	0.980	0.972
Random	100	0.000 ± 0.008	0.014	-0.005	0.002	0.000	0.009	0.016	0.002	0.000	0.001	-0.013
	500	0.000 ± 0.002	0.000	-0.004	0.002	-0.004	0.002	0.003	0.000	-0.001	0.000	0.001
	1000	0.000 ± 0.000	-0.001	-0.001	0.000	-0.001	0.000	0.001	0.000	-0.001	0.000	0.000
	2500	0.000 ± 0.000	-0.001	-0.001	0.000	0.000	0.000	0.000	0.000	0.001	0.000	0.000
	5000	0.000 ± 0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Random BiDict	100	0.959 ± 0.002	0.960	0.957	0.955	0.957	0.963	0.958	0.964	0.957	0.963	0.958
	500	0.959 ± 0.002	0.959	0.958	0.956	0.958	0.963	0.959	0.965	0.959	0.963	0.958
	1000	0.959 ± 0.002	0.959	0.958	0.956	0.958	0.963	0.959	0.965	0.958	0.963	0.958
	2500	0.960 ± 0.002	0.960	0.958	0.957	0.959	0.963	0.959	0.965	0.958	0.963	0.958
	5000	0.960 ± 0.002	0.960	0.959	0.957	0.959	0.964	0.959	0.965	0.959	0.964	0.959
Numbers	100	0.997 ± 0.000	0.997	0.997	0.998	0.997	0.998	0.997	0.998	0.997	0.998	0.997
	500	0.994 ± 0.000	0.995	0.994	0.994	0.995	0.996	0.994	0.996	0.996	0.994	0.994
	1000	0.993 ± 0.001	0.993	0.993	0.992	0.993	0.994	0.992	0.995	0.993	0.995	0.993
	2500	0.988 ± 0.001	0.988	0.988	0.987	0.988	0.990	0.988	0.990	0.988	0.990	0.988
	5000	0.985 ± 0.001	0.986	0.985	0.984	0.985	0.987	0.985	0.987	0.985	0.987	0.985
NEs	100	-0.003 ± 0.013	-0.018	0.013	0.000	0.001	0.013	0.006	-0.011	-0.017	0.007	-0.027
	500	0.002 ± 0.007	0.002	0.002	-0.003	-0.010	0.011	0.015	0.007	0.006	0.007	-0.008
	1000	0.002 ± 0.006	0.000	0.004	-0.005	-0.010	0.002	0.015	-0.002	0.007	0.005	0.004
	2500	0.000 ± 0.003	0.001	0.000	-0.003	0.000	-0.001	0.004	-0.005	0.007	0.000	-0.002
	5000	–	–	–	–	–	–	–	–	–	–	–

Table 3: Complete results for the RS metric with combinations of 5 embedding spaces build from the same partition of the corpus but trained using different seeds.

		$\mu \pm \sigma$	E1 E2	E1 E3	E1 E4	E1 E5	E2 E3	E2 E4	E2 E5	E3 E4	E3E5	E4E5
Frequent	100	2 ± 1	2.3	2.8	2.6	3.3	1.8	2.2	1.6	0.9	1.0	1.5
	500	2 ± 1	4.5	2.8	4.3	3.8	2.4	1.5	1.7	1.6	1.5	2.0
	1000	3 ± 1	3.7	2.8	3.6	5.8	3.5	3.5	2.5	2.0	5.9	4.4
	2500	3 ± 1	3.0	3.6	3.2	4.9	2.3	2.6	2.6	2.5	3.3	5.9
	5000	5 ± 2	6.8	7.9	7.8	9.4	6.7	1.7	1.6	8.6	1.7	6.9
Random	100	3 ± 1	3.5	3.1	4.9	3.2	2.3	5.1	4.3	2.5	1.7	2.2
	500	5 ± 3	3.4	3.5	4.8	5.7	1.9	3.1	8.7	3.9	10.0	11.4
	1000	6 ± 2	10.1	4.7	3.4	4.6	11.4	5.3	9.6	5.2	3.9	3.8
	2500	10 ± 3	10.5	8.1	8.5	5.8	6.4	18.1	10.6	11.0	9.0	16.7
	5000	14 ± 7	6.6	7.4	23.0	9.6	13.7	22.8	4.8	27.6	15.1	17.4
Random BiDict	100	1 ± 1	2.6	1.4	0.9	1.1	2.4	2.1	1.7	1.6	1.1	0.7
	500	3 ± 1	3.7	3.0	1.5	5.7	3.5	4.0	4.1	4.0	4.3	4.4
	1000	4 ± 1	3.1	3.8	6.3	5.2	4.5	3.5	4.1	8.1	6.3	3.9
	2500	7 ± 3	9.8	8.1	4.3	9.8	2.6	10.9	2.6	8.6	4.0	11.1
	5000	5 ± 2	5.6	10.5	5.5	7.8	5.2	5.1	4.4	6.5	5.6	2.3
Numbers	100	3 ± 1	5.2	1.4	1.9	5.7	5.7	3.6	2.6	2.0	5.7	3.0
	500	4 ± 1	3.4	1.3	3.1	4.6	4.0	7.7	3.6	2.3	3.9	6.9
	1000	5 ± 2	9.6	4.3	9.1	6.2	5.2	4.7	3.9	5.7	3.0	2.7
	2500	9 ± 5	11.8	13.4	13.4	20.9	1.9	4.6	7.2	6.2	8.4	8.3
	5000	7 ± 1	6.4	4.3	8.6	8.5	8.9	9.3	8.4	6.9	8.3	3.2
NEs	100	3 ± 1	2.1	4.3	3.7	2.1	3.9	4.3	2.7	8.0	4.9	3.0
	500	7 ± 3	7.5	10.8	11.3	8.9	7.7	9.7	6.9	3.7	2.4	2.5
	1000	6 ± 3	2.0	6.5	5.7	7.3	8.2	6.4	6.4	10.3	13.5	2.7
	2500	7 ± 3	12.7	7.3	4.9	5.5	14.8	7.4	7.8	6.3	6.8	2.7
	5000	–	–	–	–	–	–	–	–	–	–	–

Table 4: Complete results for the EV metric with combinations of 5 embedding spaces build from the same partition of the corpus but trained using different seeds.

		$\mu \pm \sigma$	E1 E2	E1 E3	E1 E4	E1 E5	E2 E3	E2 E4	E2 E5	E3 E4	E3 E5	E4 E5
Frequent	100	0.02 ± 0.00	0.027	0.031	0.038	0.033	0.025	0.037	0.025	0.020	0.017	0.036
	500	0.02 ± 0.00	0.025	0.021	0.021	0.017	0.025	0.023	0.030	0.020	0.022	0.018
	1000	0.02 ± 0.00	0.025	0.026	0.020	0.018	0.025	0.018	0.026	0.017	0.022	0.015
	2500	0.01 ± 0.00	0.018	0.020	0.016	0.015	0.021	0.024	0.023	0.019	0.018	0.013
	5000	0.01 ± 0.00	0.017	0.019	0.014	0.015	0.021	0.018	0.010	0.015	0.012	0.013
Random	100	0.17 ± 0.12	0.086	0.073	0.244	0.127	0.072	0.317	0.054	0.318	0.053	0.371
	500	0.15 ± 0.06	0.100	0.256	0.160	0.273	0.171	0.128	0.189	0.107	0.029	0.118
	1000	0.16 ± 0.07	0.084	0.245	0.080	0.270	0.160	0.100	0.185	0.202	0.052	0.226
	2500	0.11 ± 0.03	0.123	0.136	0.125	0.068	0.079	0.075	0.125	0.127	0.205	0.077
	5000	0.07 ± 0.01	0.044	0.076	0.046	0.095	0.052	0.064	0.089	0.086	0.086	0.077
Random BiDict	100	0.03 ± 0.01	0.018	0.021	0.039	0.021	0.030	0.042	0.029	0.061	0.026	0.035
	500	0.02 ± 0.00	0.021	0.032	0.017	0.022	0.016	0.019	0.020	0.029	0.034	0.035
	1000	0.02 ± 0.00	0.031	0.032	0.029	0.027	0.020	0.036	0.017	0.039	0.030	0.038
	2500	0.02 ± 0.00	0.022	0.017	0.016	0.027	0.020	0.021	0.019	0.014	0.022	0.031
	5000	0.01 ± 0.00	0.028	0.026	0.018	0.018	0.013	0.019	0.015	0.015	0.012	0.014
Numbers	100	0.05 ± 0.05	0.030	0.043	0.055	0.038	0.015	0.025	0.027	0.039	0.208	0.035
	500	0.02 ± 0.00	0.023	0.027	0.025	0.027	0.031	0.032	0.021	0.025	0.019	0.022
	1000	0.02 ± 0.00	0.020	0.025	0.028	0.019	0.014	0.026	0.019	0.040	0.024	0.020
	2500	0.02 ± 0.00	0.031	0.025	0.019	0.019	0.022	0.028	0.023	0.025	0.020	0.028
	5000	0.02 ± 0.00	0.019	0.022	0.024	0.024	0.020	0.017	0.021	0.032	0.032	0.026
NEs	100	0.10 ± 0.03	0.151	0.080	0.093	0.062	0.125	0.170	0.135	0.121	0.105	0.041
	500	0.07 ± 0.03	0.049	0.073	0.072	0.080	0.061	0.032	0.107	0.079	0.154	0.076
	1000	0.09 ± 0.03	0.120	0.053	0.068	0.134	0.094	0.132	0.031	0.063	0.102	0.146
	2500	0.03 ± 0.00	0.029	0.045	0.033	0.029	0.039	0.027	0.017	0.040	0.035	0.025
	5000	–	–	–	–	–	–	–	–	–	–	–

Table 5: Complete results for the GH metric with combinations of 5 embedding spaces build from the same partition of the corpus but trained using different seeds.

		$\mu \pm \sigma$	E1 E2	E1 E3	E1 E4	E1 E5	E2 E3	E2 E4	E2 E5	E3 E4	E3 E5	E4 E5
Frequent	100	0.860 ± 0.056	0.839	0.837	0.768	0.882	0.901	0.871	0.943	0.787	0.945	0.828
	500	0.314 ± 0.032	0.333	0.312	0.264	0.331	0.329	0.329	0.347	0.277	0.357	0.263
	1000	0.131 ± 0.014	0.151	0.123	0.119	0.147	0.132	0.129	0.138	0.109	0.152	0.114
	2500	0.038 ± 0.000	0.040	0.039	0.033	0.045	0.042	0.037	0.041	0.030	0.043	0.031
	5000	0.015 ± 0.001	0.017	0.015	0.014	0.017	0.016	0.013	0.015	0.013	0.018	0.014
Random	100	0.002 ± 0.015	0.014	-0.010	0.009	0.026	0.004	-0.007	0.003	-0.033	0.013	0.008
	500	0.000 ± 0.001	-0.001	0.000	-0.002	-0.004	0.001	0.000	0.001	0.000	-0.003	0.003
	1000	0.000 ± 0.000	0.000	0.000	0.000	0.000	0.001	-0.001	0.000	0.000	0.000	0.000
	2500	0.000 ± 0.000	-0.001	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	5000	0.000 ± 0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Random BiDict	100	0.884 ± 0.008	0.870	0.894	0.894	0.882	0.889	0.892	0.879	0.889	0.883	0.869
	500	0.882 ± 0.004	0.880	0.881	0.878	0.886	0.881	0.885	0.878	0.891	0.889	0.879
	1000	0.883 ± 0.002	0.885	0.885	0.883	0.883	0.883	0.883	0.878	0.887	0.886	0.881
	2500	0.883 ± 0.001	0.884	0.883	0.882	0.886	0.884	0.883	0.880	0.885	0.885	0.881
	5000	0.883 ± 0.000	0.882	0.884	0.885	0.885	0.883	0.884	0.883	0.885	0.884	0.884
Numbers	100	0.604 ± 0.087	0.809	0.687	0.631	0.533	0.639	0.516	0.539	0.549	0.534	0.604
	500	0.116 ± 0.012	0.133	0.119	0.134	0.104	0.131	0.098	0.112	0.113	0.100	0.120
	1000	0.061 ± 0.008	0.049	0.068	0.073	0.071	0.067	0.057	0.058	0.049	0.060	0.067
	2500	0.037 ± 0.007	0.042	0.045	0.036	0.055	0.036	0.040	0.034	0.028	0.037	0.026
	5000	0.022 ± 0.003	0.021	0.028	0.022	0.026	0.021	0.029	0.017	0.019	0.021	0.021
NEs	100	-0.001 ± 0.018	0.029	0.008	-0.005	-0.007	-0.005	0.003	-0.029	-0.011	-0.027	0.030
	500	-0.003 ± 0.007	-0.002	0.006	0.008	-0.008	-0.015	-0.008	-0.008	-0.002	-0.011	0.002
	1000	-0.002 ± 0.004	-0.008	0.004	0.0	-0.008	-0.006	-0.005	0.000	-0.003	-0.005	0.003
	2500	-0.002 ± 0.002	0.000	0.003	-0.004	-0.004	0.000	-0.003	0.001	-0.002	-0.006	-0.005
	5000	–	–	–	–	–	–	–	–	–	–	–

Table 6: Complete results for the RS metric with combinations of 5 embedding spaces build from 5 different random partitions of the CEREAL corpus.

		$\mu \pm \sigma$	E1 E2	E1 E3	E1 E4	E1 E5	E2 E3	E2 E4	E2 E5	E3 E4	E3 E5	E4 E5
Frequent	100	2 ± 1	2.7	3.0	3.4	2.9	0.9	1.2	1.9	0.7	1.5	2.0
	500	3 ± 1	6.7	4.0	5.5	1.8	2.8	4.4	4.6	3.0	2.2	3.7
	1000	2 ± 1	3.6	3.5	3.8	5.8	1.8	1.6	2.6	2.2	2.6	1.6
	2500	4 ± 1	3.8	4.4	1.5	5.3	1.6	5.1	7.6	5.9	6.1	6.6
	5000	5 ± 4	1.1	3.3	2.8	11.8	2.1	2.3	10.9	3.8	11.1	9.0
Random	100	5 ± 1	6.1	3.5	5.6	2.8	9.7	5.9	5.0	5.6	3.9	6.8
	500	5 ± 2	8.0	5.4	4.7	4.8	3.9	10.3	3.1	4.6	1.5	5.6
	1000	5 ± 1	3.7	6.4	3.8	5.4	5.4	4.2	5.5	6.4	8.8	5.1
	2500	7 ± 1	3.3	6.4	8.1	6.4	6.9	7.4	6.6	11.2	9.6	6.6
	5000	14 ± 5	19.1	13.1	6.6	8.6	10.3	16.1	25.3	6.9	21.1	14.1
Random BiDict	100	3 ± 2	2.9	2.2	8.0	2.8	7.6	1.7	2.8	1.3	1.7	5.5
	500	4 ± 1	2.0	5.6	4.1	3.3	6.6	2.6	3.3	2.2	4.7	6.1
	1000	6 ± 4	10.9	15.0	7.5	1.6	1.8	9.6	3.9	4.0	5.1	10.3
	2500	7 ± 2	5.2	7.5	13.3	9.5	9.1	6.8	8.3	9.1	4.6	5.0
	5000	8 ± 4	6.0	9.1	4.5	4.9	5.1	21.1	6.2	9.5	10.4	6.0
Numbers	100	3 ± 1	3.2	7.1	3.5	1.6	2.6	2.0	3.1	2.0	7.4	4.3
	500	5 ± 2	1.8	4.6	8.7	5.0	5.3	8.6	5.7	8.8	3.0	5.3
	1000	9 ± 6	3.6	3.2	20.1	8.0	2.8	15.3	4.6	17.0	6.9	16.3
	2500	12 ± 4	12.3	17.3	15.2	19.0	14.2	14.1	6.6	5.1	11.6	14.4
	5000	11 ± 3	16.2	6.8	9.3	15.8	12.6	10.1	6.1	13.7	11.3	15.3
NEs	100	2 ± 1	4.2	1.1	1.8	1.1	3.9	5.5	3.5	1.7	2.0	3.3
	500	3 ± 1	3.5	4.0	3.1	2.8	7.3	5.3	5.6	3.4	1.9	1.9
	1000	7 ± 2	4.0	11.1	5.9	11.0	8.0	7.9	6.8	5.7	4.9	8.5
	2500	7 ± 2	8.5	10.1	14.6	7.4	4.8	9.6	5.3	5.3	4.5	7.2
	5000	–	–	–	–	–	–	–	–	–	–	–

Table 7: Complete results for the EV metric with combinations of 5 embedding spaces build from 5 different random partitions of the CEREAL corpus.

		$\mu \pm \sigma$	E1 E2	E1 E3	E1 E4	E1 E5	E2 E3	E2 E4	E2 E5	E3 E4	E3 E5	E4 E5
Frequent	100	0.03 ± 0.01	0.026	0.058	0.023	0.024	0.038	0.028	0.034	0.035	0.062	0.027
	500	0.02 ± 0.00	0.029	0.022	0.028	0.020	0.022	0.025	0.023	0.021	0.026	0.022
	1000	0.02 ± 0.00	0.019	0.015	0.028	0.020	0.018	0.021	0.022	0.020	0.026	0.022
	2500	0.01 ± 0.00	0.013	0.015	0.019	0.020	0.015	0.015	0.020	0.021	0.021	0.021
	5000	0.01 ± 0.00	0.018	0.018	0.014	0.020	0.016	0.016	0.017	0.018	0.021	0.014
Random	100	0.15 ± 0.06	0.037	0.072	0.209	0.134	0.086	0.215	0.133	0.228	0.206	0.198
	500	0.19 ± 0.10	0.373	0.299	0.188	0.332	0.107	0.184	0.075	0.133	0.074	0.155
	1000	0.13 ± 0.05	0.065	0.153	0.176	0.073	0.129	0.123	0.070	0.253	0.146	0.132
	2500	0.13 ± 0.03	0.143	0.091	0.188	0.160	0.129	0.116	0.067	0.186	0.146	0.103
	5000	0.12 ± 0.04	0.090	0.068	0.163	0.116	0.085	0.160	0.096	0.166	0.102	0.195
Random BiDict	100	0.05 ± 0.02	0.036	0.031	0.045	0.106	0.033	0.088	0.029	0.040	0.051	0.057
	500	0.03 ± 0.01	0.044	0.039	0.028	0.030	0.027	0.080	0.031	0.052	0.025	0.028
	1000	0.03 ± 0.01	0.069	0.023	0.034	0.036	0.031	0.025	0.031	0.037	0.035	0.036
	2500	0.03 ± 0.01	0.036	0.032	0.043	0.022	0.019	0.065	0.031	0.024	0.023	0.037
	5000	0.02 ± 0.00	0.016	0.019	0.024	0.028	0.025	0.027	0.030	0.023	0.022	0.024
Numbers	100	0.06 ± 0.04	0.031	0.040	0.125	0.030	0.028	0.124	0.022	0.120	0.028	0.114
	500	0.05 ± 0.00	0.050	0.040	0.054	0.043	0.060	0.043	0.047	0.052	0.055	0.071
	1000	0.02 ± 0.00	0.024	0.030	0.031	0.033	0.033	0.033	0.020	0.026	0.038	0.023
	2500	0.02 ± 0.00	0.024	0.026	0.022	0.034	0.027	0.033	0.030	0.023	0.031	0.036
	5000	0.05 ± 0.01	0.044	0.021	0.070	0.073	0.040	0.067	0.069	0.065	0.059	0.041
NEs	100	0.08 ± 0.02	0.063	0.123	0.053	0.057	0.129	0.055	0.121	0.098	0.101	0.066
	500	0.07 ± 0.03	0.106	0.113	0.081	0.045	0.023	0.069	0.107	0.075	0.114	0.052
	1000	0.07 ± 0.02	0.102	0.058	0.123	0.066	0.046	0.077	0.058	0.096	0.040	0.076
	2500	0.05 ± 0.02	0.067	0.027	0.030	0.043	0.084	0.064	0.052	0.041	0.054	0.041
	5000	–	–	–	–	–	–	–	–	–	–	–

Table 8: Complete results for the GH metric with combinations of 5 embedding spaces build from 5 different random partitions of the CEREAL corpus.

B Qualitative Behaviour of the Embedding Spaces

Figure 6 shows the 10-top nearest neighbours for three different varieties of Spanish words that would translate into *corn*: `elote`, `choclo`, and `mazorca` (cf. Section 7). The results with respect to three embedding spaces —cl, es and mx— show the differences associated to the three concepts. For instance, `elote` goes from inexistence in cl to all the way into a neighbourhood of regional ingredients and dishes in mx, passing through a concept more associated to foreign cuisine in es.

Figure 6: t-SNE projections (van der Maaten and Hinton, 2008) for the neighbouring spaces for the word `corn`, used as `choclo` in Chile (cl), `mazorca` in Spain (es) and `elote` in Mexico (mx).

C Extended Isomorphism Results on the Variety-Specific Spanish Embeddings

Figures 7, 8 and 9 show the extended results for the experiments in Section 7. Here, in addition to MFW 100 for EV (best configuration reported in the main text), we show the results for the top-2 best configurations for the three isomorphism metrics, RS, EV and GH: RS on MFW 100 and random BiDict 100, EV on BiDict 100, and GH on the MFW 5,000 and random BiDict 5,000 (in cases where 5,000 points are not available we use the maximum number of available points). In all cases, figures represent the scores for a given isomorphism metric using the embeddings computed on `CEREAL` and on `Twitter` data.

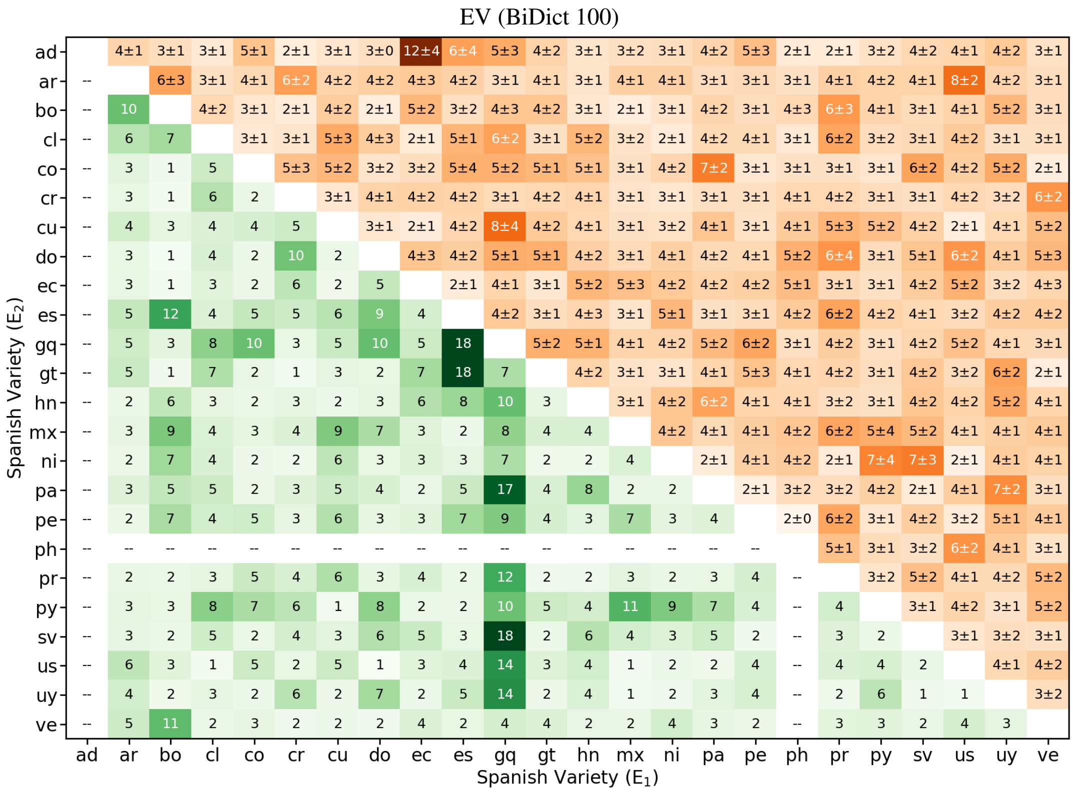


Figure 7: EV with random BiDict 100 words for the 24 Spanish varieties. Top–right triangles (orange) correspond to the results with the `CEREAL` corpus and bottom–left triangle (green) to the `Twitter` corpus.

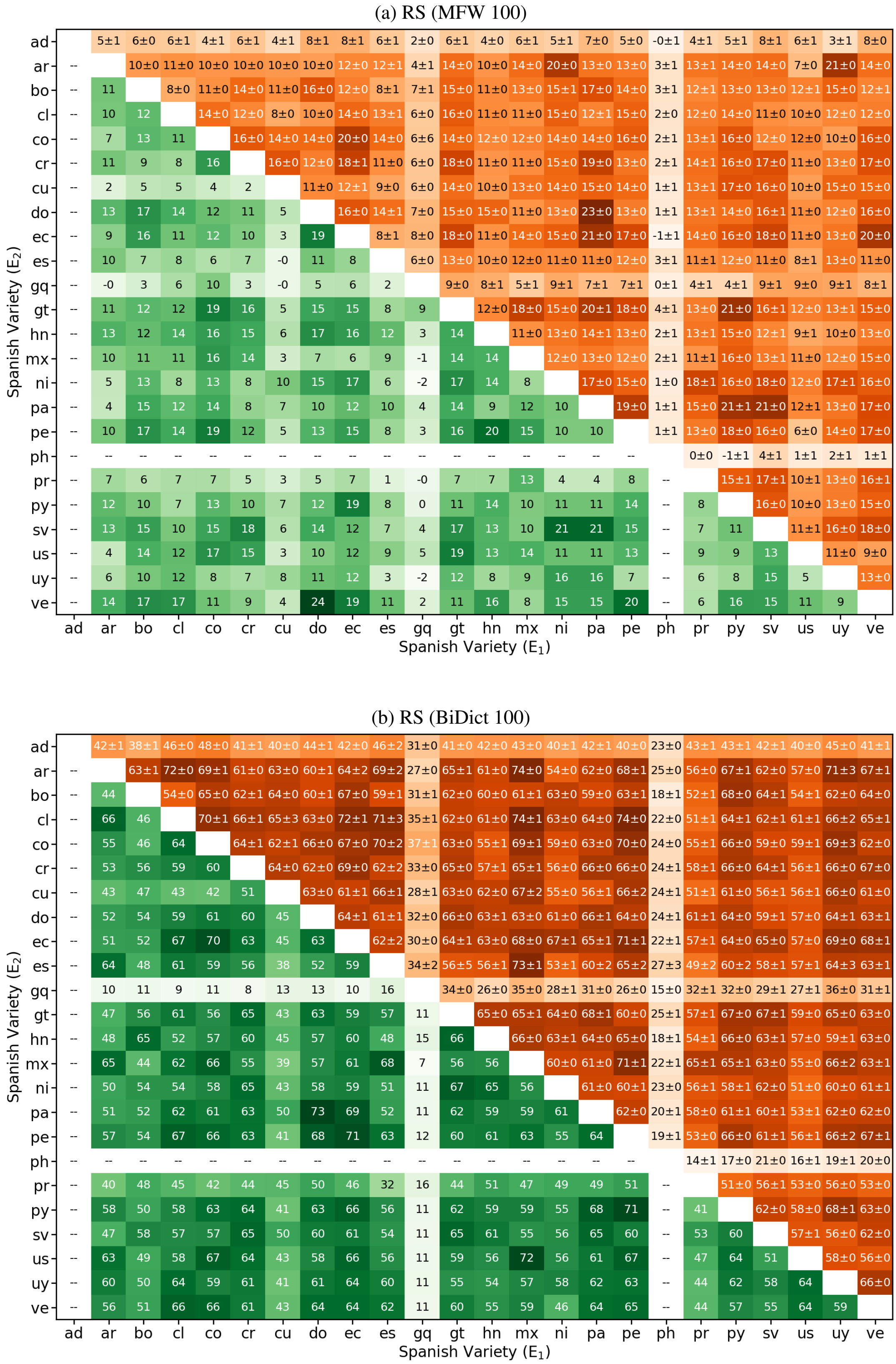


Figure 8: RS with (a) 100 most frequent words (MFW) and (b) BiDict entries multiplied by 100 for better readability. Top–right triangles (orange) correspond to the mean results with the `CEREAL` corpus and bottom–left triangle (green) to the `Twitter` corpus.

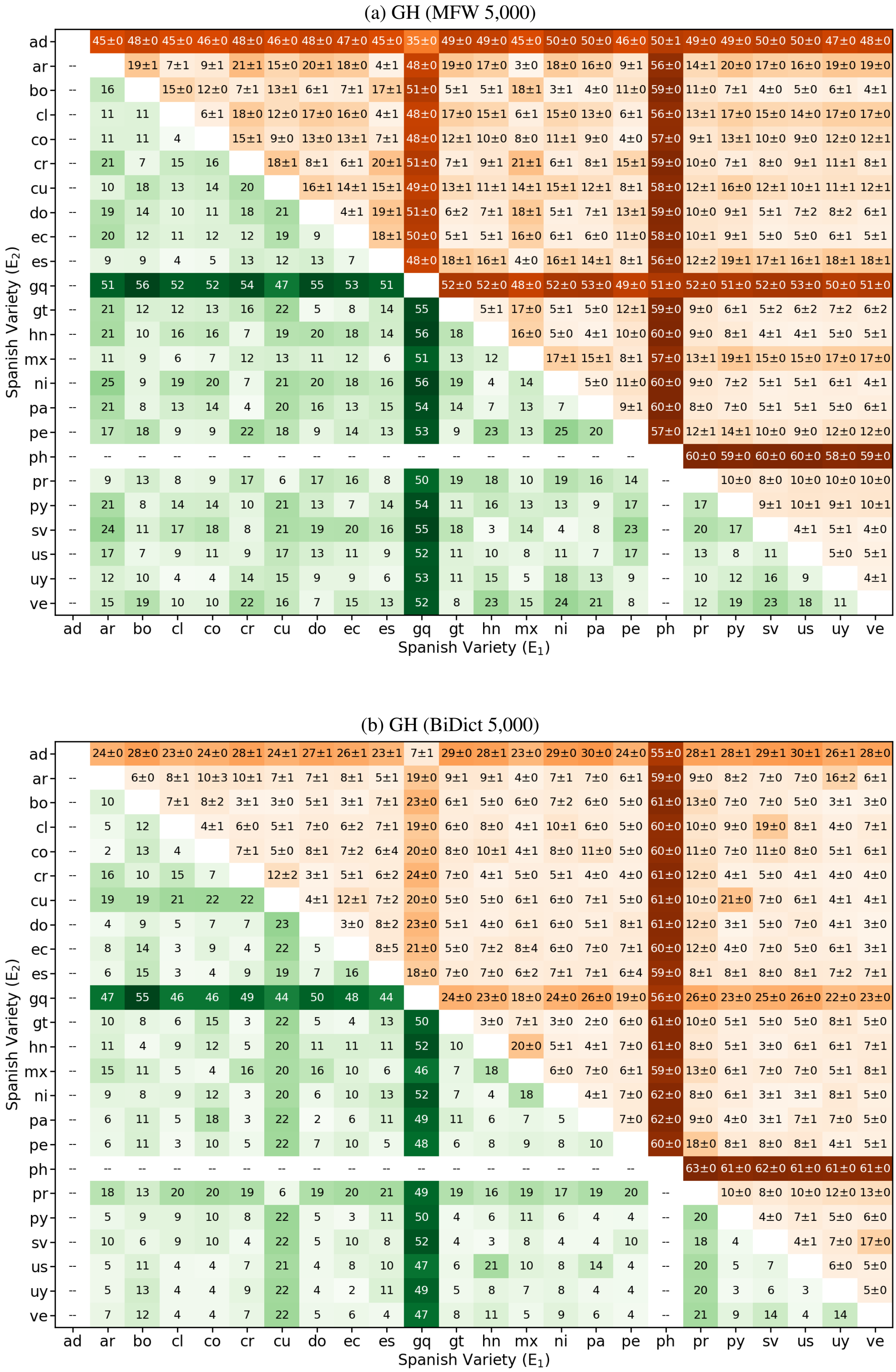


Figure 9: (GH with (a) 5,000 most frequent words (MFW) and (b) GH with 5,000 random BiDict words for the 24 Spanish varieties. Results are multiplied by 100 for better readability. Top–right triangles (orange) correspond to the mean results with the `CEREAL` corpus and bottom–left triangle (green) to the `Twitter` corpus.

D Extended Analysis on Phylogenetics

D.1 Visual Representation

Figure 10 focuses on EV and represents the regions that can be drawn on the basis of the resulting clusters. A comparison against the Spanish linguistic zones as defined by RAE (see Figure 1) reveals some divergences. Among them, Central America not necessarily being tied to Mexico as well as Colombia and Venezuela, which here appear differentiated.

(a) `CEREAL` (b) `Twitter`

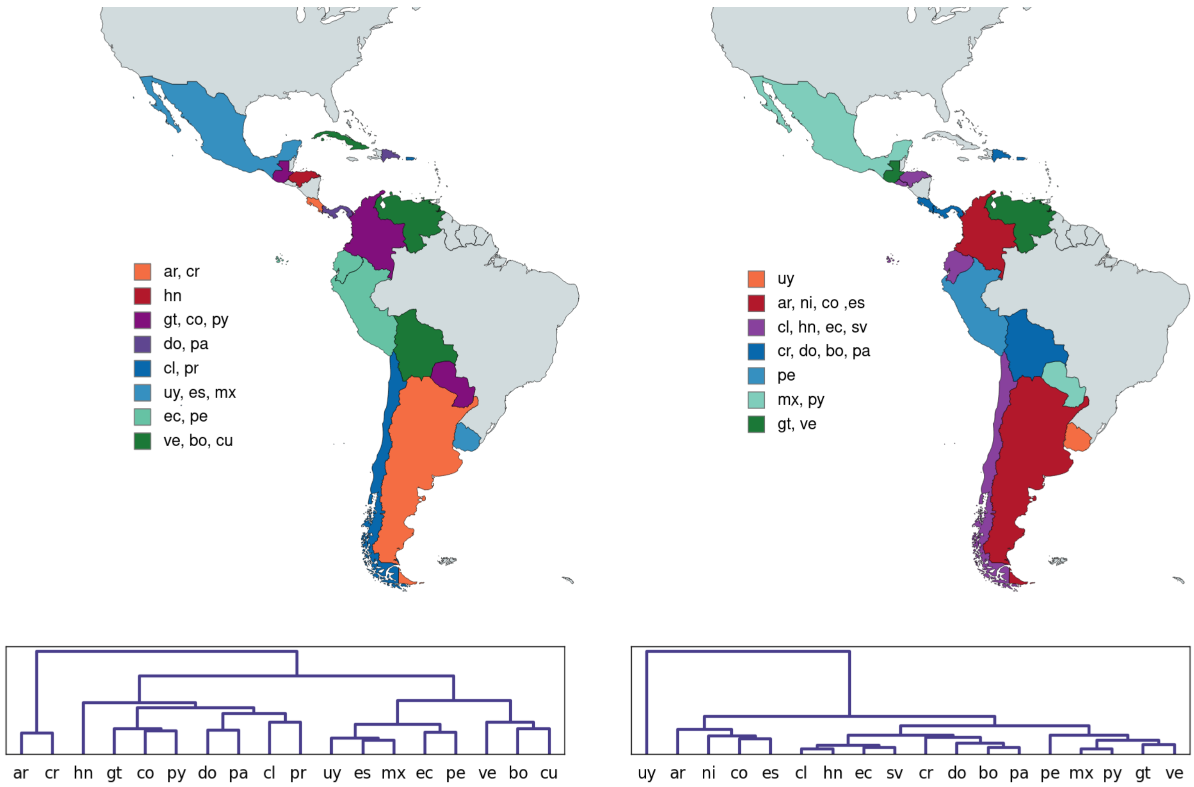


Figure 10: Geographical representation of the Spanish varieties clustered according the EV (MFW 100) score; es is omitted from the plot for visibility reasons, but it is included in the legend together with the family it groups with. Plots are done with MapChart (`https://www.mapchart.net`).

D.2 Extended Results on the Hierarchical Clustering Experiments

As in the previous sections, we show, for completeness, results for the top-2 best configurations for the 3 isomorphism metrics: RS, EV and GH. Figure 11 depicts the phylogenetic (relational) trees obtained from scores on the embeddings built with `CEREAL` for the 2nd best performing configurations (1st one is in the main text): RS on random BiDict 100, EV on BiDict 100, and GH on random BiDict 5,000. We compare the trees for 24 varieties and the subset of the 17 highest resourced varieties. Figure 12 shows the top-2 configurations for the scores derived from the embeddings computed on `Twitter` data.

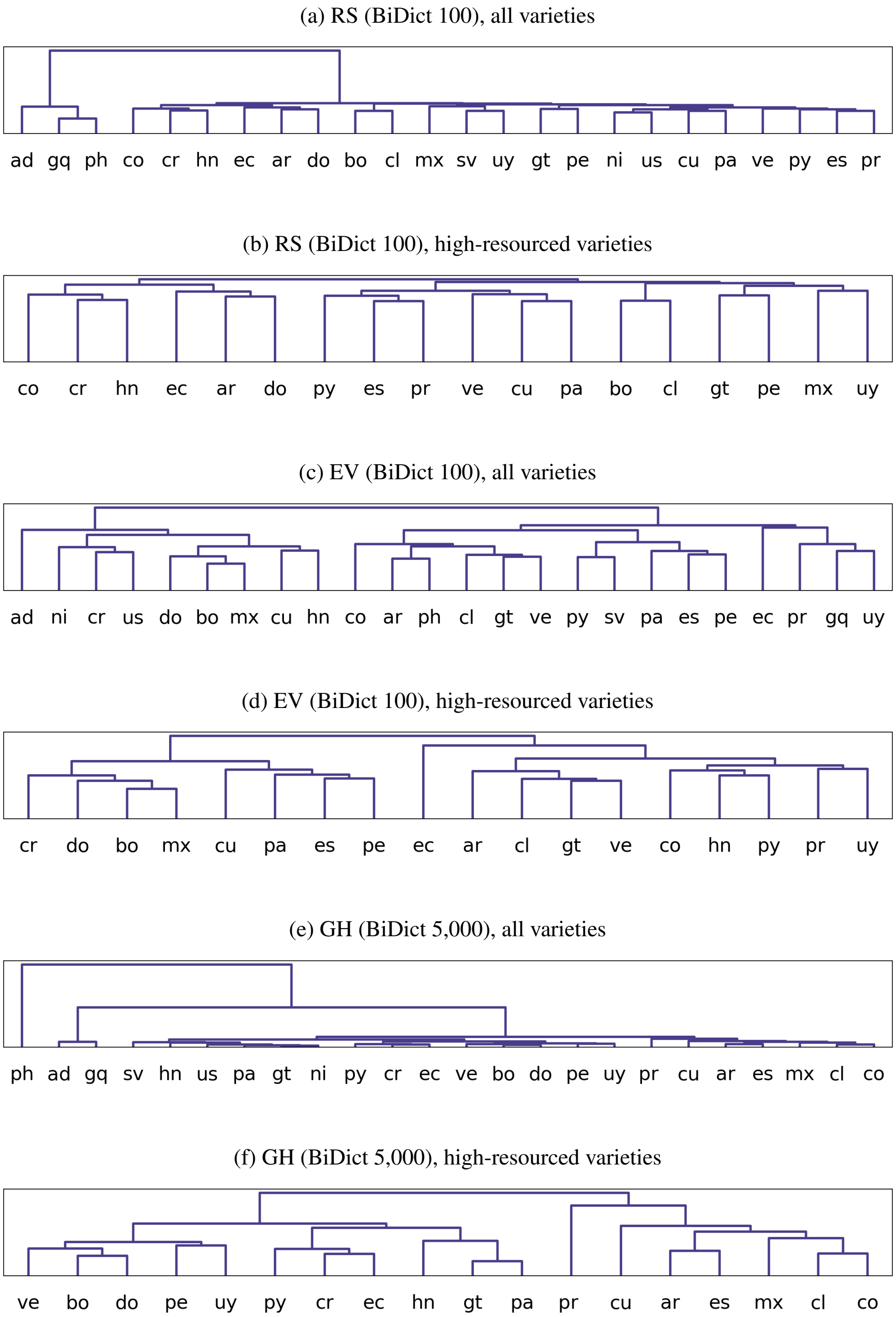


Figure 11: Hierarchical clustering on the outputs of the isomorphism measures obtained in Section 7 for the embeddings computed using the CEREAL corpus.

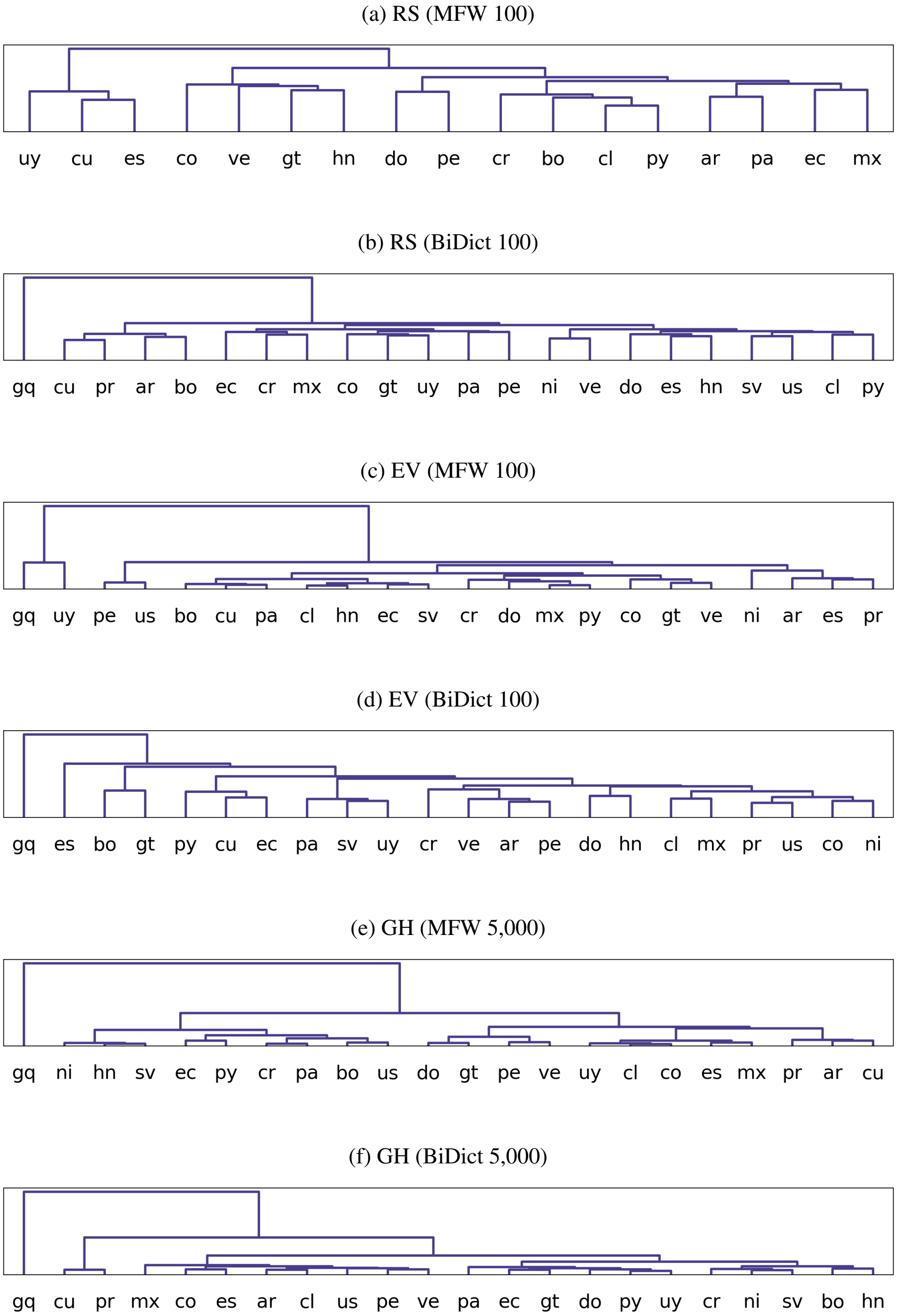


Figure 12: Hierarchical clustering on the outputs of the isomorphism measures obtained in Section 7 with Twitter embeddings.

Modeling Orthographic Variation in Occitan's Dialects

Zachary William Hopton
Language and Space Lab
University of Zurich
zacharywilliam.hopton@uzh.ch

Noëmi Aepli
Department of Computational Linguistics
University of Zurich
naepli@cl.uzh.ch

Abstract

Effectively normalizing textual data poses a considerable challenge, especially for low-resource languages lacking standardized writing systems. In this study, we fine-tuned a multilingual model with data from several Occitan dialects and conducted a series of experiments to assess the model's representations of these dialects. For evaluation purposes, we compiled a parallel lexicon encompassing four Occitan dialects. Intrinsic evaluations of the model's embeddings revealed that surface similarity between the dialects strengthened representations. When the model was further fine-tuned for part-of-speech tagging and Universal Dependency parsing, its performance was robust to dialectical variation, even when trained solely on part-of-speech data from a single dialect. Our findings suggest that large multilingual models minimize the need for spelling normalization during pre-processing.

1 Introduction

Traditionally, natural language processing pipelines have been designed to reduce noise in the data during pre-processing, either by removing it entirely (i.e., as one may do with URLs) or by normalizing noisy forms. Normalization can either improve users' understanding of a text or serve as a system-internal process that is meant to reduce noise, allowing a model to better learn from the vocabulary presented during training (Costa Bertaglia and Volpe Nunes, 2016).

However, with many of the recent successes in text normalization coming from neural networks, such as sequence-to-sequence (seq2seq) models that map orthographic variants to canonical forms, normalization can become computationally costly (Lusetti et al., 2018). Furthermore, such supervised methods are generally impractical for low-resource languages, as these languages often lack labeled datasets with standardized word forms. In the case of non-standardized languages, there is also the issue of not having canonical forms to standardize to. Consequently, recent research has shifted its focus to determining the necessity of orthographic normalization and identifying how noisy data might prove advantageous (Srivastava and Chiang, 2023; Aepli and Sennrich, 2022; Al Sharou et al., 2021).

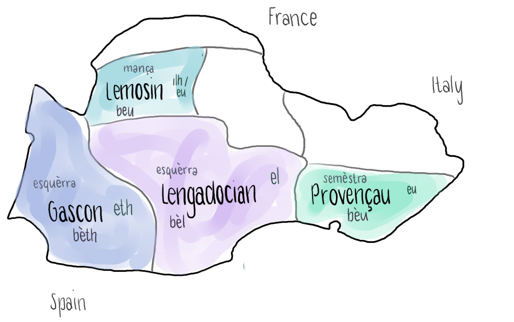


Figure 1: Dialect map of Occitan. The four dialects included in this study are highlighted, along with examples of lexical (i.e.,"mança" and "senèstra") and spelling (i.e., "bèu" and "beu") variation between the dialects.

In the present study, we fine-tune a multilingual, large language model using data from various dialects of Occitan, a Western Romance language (see Figure 1). We perform experiments to assess the model's ability to accurately represent the low-resource test language even without reducing the dialectal variation in the fine-tuning data, which manifests as both lexical and orthographic variation between dialects. Specifically, we carry out experiments on analogy computation and lexicon induction to test the fine-tuned model's intrinsic representations of Occitan's dialects, as well as Universal Dependency parsing and part-of-speech tagging experiments to study the usefulness of these fine-tuned representations in down-stream tasks. In doing so, we investigate the extent to which low-resource NLP systems that rely on transfer learning are robust to dialectal variation in fine-tuning data.

Proceedings of the Eleventh Workshop on NLP for Similar Languages, Varieties, and Dialects (VarDial 2024), pages 78–88
June 20, 2024

This enables us to avoid costly and infeasible normalization during pre-processing.

2 Linguistic Context

Occitan is a Western Romance language spoken in southern France, the Val d'Aran in Spain, and Italy's Piedmont region. Occitan coexists in a diglossic relationship with French, Spanish, Catalan, and Italian and lacks official status outside of the Val d'Aran. The six dialects that are typically delineated are Auvernhàs, Gascon, Lemosin, Lengadocian, Provençau, and Vivaroaupenc (Bec, 1995). Occitan is not standardized, and indeed, there is a great deal of geographical variation from many perspectives, including speakers' phonetic and lexical inventories, orthography, and syntax (Miletic et al., 2020b). At the level of phonetics, for example, Gascon stands out from other Occitan dialects and Western Romance languages with its use of the phone [h] (Corral et al., 2020). Occitan dialects also differ in their realization of /v/, with Lengadocian and Gascon speakers tending toward the phone [b] and Lemosin and Provençau speakers typically pronouncing [v] (Arbousset et al., 2003).

As a whole, Occitan dialects share some morphological and syntactic features with each other, many of which are more similar to Catalan and Spanish than French. For example, unlike French, Occitan is a pro-drop language (i.e., subject pronouns are not necessary) and continues the use of the past preterit and imperfect subjunctive inflections outside of writing (Bernhard et al., 2021). Patterns of verbal inflection vary across the dialects, as do the augmentative and diminutive suffixes (Miletic et al., 2020a; Arbousset et al., 2003). At the syntactic level, Gascon stands out with the use of enunciative particles to mark clauses' types (Morin, 2008; Vergez-Couret and Urieli, 2014).

A classical orthography inherited from medieval literature dominates Occitan writing in most dialects, though in Provençau, the "Mistralian norm" is more commonly used (Blanchet, 2004). Some local writing systems have been standardized for purposes such as teaching (Bernhard et al., 2021). However, individuals vary in their conventions, often in ways heavily influenced by French orthography. Figure 1 highlights some of the variations in Occitan dialects' lexicons and spelling conventions. Besides the lack of a single orthographic convention, phonetic differences between the dialects lead to different spellings for words with the same meaning. For instance, in Provençau the word for *"bedroom"* tends to be written as *"cambra,"* and the initial /c/ is pronounced as velar [k] (Arbousset et al., 2003). However, there is generally a palatalization of the consonant [k] among Occitan's Northern dialects such as Lemosin (Buckley, 2009). This is often reflected in written forms, such as the Lemosin word for *"bedroom," "chambra."*

Beyond just differences in spelling, dialects of Occitan vary at the lexical level. For instance, *"achaptar"* (*"to buy"*) is used in Lemosin while *"crompar"* or *"comprar"* are used in dialects to the south. Or, while speakers of Provençau tend to use the phrase *"aver fam"* (*"to be hungry"*), speakers in other dialects might say *"aver talent."*

These lexical and spelling variations between dialects of Occitan, along with a relative lack of data, pose challenges for NLP research. Nonetheless, there is a body of work on Occitan language technology, such as text-to-speech systems, part-of-speech taggers, universal dependency parsers, and lemmatizers (Corral et al., 2020; Vergez-Couret and Urieli, 2014; Miletic et al., 2020a; Miletić and Siewert, 2023). There is also a body of theoretical work about Occitan, such as experiments with continuous numerical representations of Occitan via cross-lingual word embeddings with related languages (Woller et al., 2021). Most NLP research focusing on Occitan has been with four out of the six dialects that are generally delineated: Lengadocian, Lemosin, Provençau, and Gascon.

3 Related Work

At its simplest, normalization using rule-based word edits can collapse variants into standard forms (Reffle, 2011). However, this requires language knowledge and becomes infeasible in cases of ambiguity. Thus, more context-sensitive approaches to text normalization, such as statistical string transduction and seq2seq neural networks, have been developed (Rios and Castro Mamani, 2014; Lusetti et al., 2018). Bawden et al. (2022) note that in both their statistical and neural machine translation approaches to normalizing Early Modern French, adding a rule-based post-processing step that constrains output to words in a contemporary French lexicon is particularly helpful. Moreover, Lusetti et al. (2018) improved downstream machine translation scores for Swiss German following orthographic normalization with a character-level encoder-decoder model accompanied by a word-level language model. Recent work in normalizing both user-generated and multi-dialect data seems to confirm the effectiveness of working at the character and byte-level during normalization (Kuparinen et al., 2023; van der Goot et al., 2021a). However, framing normalization as a machine translation task requires large amounts of supervised data and is therefore not feasible in the case of many other low-resource languages, such as Occitan.

Faced with an inability to remove orthographic noise to improve performance on downstream tasks, some have attempted to learn from the character level rather than the word or subword level. For instance, machine translation with character-level encoding can outperform subword-level encodings for morphologically rich languages, but requires deeper architectures, longer sequences, and—in models with no word or subword representations—becomes more difficult to interpret (Tang et al., 2020). Despite these hurdles, another study has found that character representations result in better downstream machine translation performance than rule-based normalization of Swiss German training data (Honnet et al., 2018). Thus, character-level modeling seems to have potential as a substitute for spelling normalization during pre-processing.

When modeling subword tokens, one way to better encode data with orthographic variation is by adding dropout to the byte-pair encoding (BPE) subwords (Sennrich et al., 2016) during subtokenization. BPE-dropout (Provilkov et al., 2020) randomly removes a certain percentage of merges while applying BPE models to subtokenize a corpus. Provilkov et al. (2020) found that applying BPE with dropout led to better machine translation of text with artificial misspellings. While exploring their models' embeddings, they find that subtokens in models trained with dropout tend to be similarly represented when they share sequences of characters. Thus, for training data with a non-standardized orthography, BPE dropout allows for more robust representations of spelling variants.

Cross-lingual transfer learning is another potential means of alleviating the sparsity of data induced by orthographic variation. Recent work has shown that for Occitan, the quality of word embeddings can indeed be improved upon if jointly trained with more data from related languages, such as Catalan and Spanish (Woller et al., 2021). Other work has sought to find normalization strategies for low-resource languages by pre-training classifiers with source languages different than the intended target language (van der Goot, 2021). This resulted in improvements over the baseline used, though interestingly, language pairs that performed the best were not always related languages. In recent work by Aepli and Sennrich (2022), the authors show that besides the relatedness of two languages, their surface similarity is also a critical factor for effective transfer learning. They find that by augmenting encoders' pre-training languages with random character noise, the models become more robust to spelling variation, and transfer learning is more effective based on performance on downstream tasks in the fine-tuning languages. The importance of surface similarity for transfer learning has also been highlighted in machine translation, where romanization of non-Latin scripts has been shown to improve the effectiveness of transfer learning when the pre-training and fine-tuning languages are related (Amrhein and Sennrich, 2020).

4 Method

4.1 Creating a Dataset

In order to conduct a controlled evaluation of our model after fine-tuning with multi-dialect Occitan data, we compile a parallel dataset comprising words from four Occitan dialects. This dataset allows us to explicitly compare the model's performance on the same content for each of the dialects. For the compilation of this dataset, we took inspiration from vocabulary themes in the book *Òc-ben! Première année d'Occitan—Livre d'éleve* (Arbousset et al., 2003). We included various functional and lexical words, alongside short multi-word expressions and conjugations of both regular and irregular verbs in several tenses. In cases where a dialect was found to have variant spellings of the same word, we created multiple entries for that word rather than arbitrarily choosing between spellings. That being said, given that there is not a standardized writing system for Occitan, the dataset by no means captures the full breadth of possibilities for spelling variations, nor does it include information about which variants are preferred or more frequently used by Occitan speakers. Our final parallel lexicon contains more than 2,200 entries in the Lengadocian, Lemosin, Provençau, and Gascon dialects. The dataset will be made available for use in academic research.

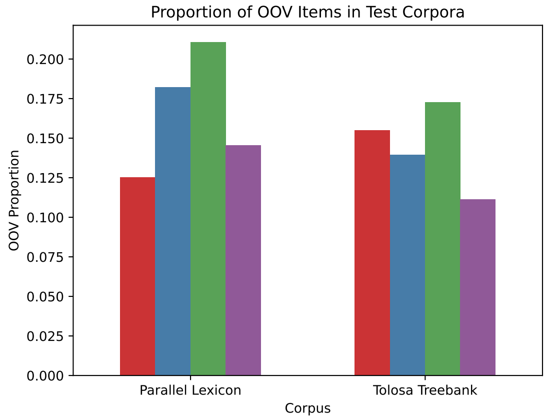


Figure 2: Proportion of vocabulary items in each evaluation corpus that did not appear in the fine-tuning dataset. Red: Lengadocian; Blue: Gascon; Green: Lemosin; Purple: Provençau.

4.2 Fine-Tuning mBERT

We incorporate Occitan dialects into a language model by fine-tuning the multilingual BERT (mBERT[1]; Devlin et al. (2019)). mBERT has been pre-trained with both a masked language modeling objective and a sentence prediction objective on Wikipedia data from 104 languages. Our fine-tuning data comprises data from two sources: the OcWikiDisc corpus (Miletic and Scherrer, 2022), which is compiled from Wikipedia discussion forums written in Occitan, and the WikiMatrix corpus (Schwenk et al., 2021), a corpus of parallel-sentence data mined from Wikipedia pages in 96 different languages. Annotation of a 100-sentence sample by the original authors of the OcWikiDisc corpus revealed that it contained several dialects of Occitan, though most of the sample came from the Lengadocian dialect. We specifically used the balanced OcWikiDisc corpus, which filtered the original dataset in a way designed to maximize the F1-score for language identification. From the WikiMatrix, we extracted the data from all language pairs containing Occitan and removed the parallel language data. Before combining with the OcWikiDisc data, we removed any duplicate lines from the dataset.

The final combined fine-tuning corpus contained 386,552 lines (10,941,124 tokens) of data in Occitan. Ten percent of the training data was used as validation dataset during the fine-tuning process of three epochs. The ocwikidisc_balanced corpus amounted to 756,922 (6.92%) of the total tokens in the corpus. For details on the proportion of each test corpora's out-of vocabulary items with respect to the fine-tuning corpus, see Figure 2.

5 Experiments

5.1 Analogy Representation

Background We conduct an intrinsic evaluation of the model's embedding space by assessing its representation of analogies. Using the parallel lexicon, we first created a dataset of analogies with approximately 35 data points per dialect. The data points were chosen to test linguistic relations similar to those presented by Mikolov et al. (2013b). Specifically, most of the relations were syntactic in nature, such as the relationship between infinitive verbs and conjugated forms, normal forms and diminutives or augmentatives, single and plural forms, and masculine and feminine forms. Ten of the data points for each dialect captured semantic relations, like antonym pairs and the relationship between capital city names and regions.

To test the model's representation of the analogies, we use two approaches described in Levy and Goldberg (2014). In the first approach, we use analogies in the form $a : b :: x : y$ and seek the word y. Specifically, we search for the word y in the given dialect's vocabulary whose embedding maximizes the following:

$$cos(y, b) - cos(y, a) + cos(y, x) \tag{1}$$

Similar to Levy and Goldberg (2014), we refer to Equation 1 as "3CosADD".

As a second means of evaluating the analogies, we implement the "3CosMUL" metric from the same work by Levy and Goldberg (2014). The authors found that this multiplicative approach to combining the query vectors' meanings outperformed the additive approaches above. In this approach, we search for the word y in a given dialect's vocabulary whose embedding maximizes the following:

$$\frac{cos(y, b)cos(y, x)}{cos(y, a) + \epsilon} \tag{2}$$

We set ϵ to 0.001—as in the original work—to avoid division by zero.

Results When solving the analogies as set forward in Equation 1, the accuracies of both the base mBERT model and our fine-tuned model are poor across all four dialects (see Table 1). There is an

[1] https://huggingface.co/bert-base-multilingual-cased

	3Cos-ADD	3Cos-MUL
Gascon	0.000 (-0.033)	**0.133** (+0.066)
Lengadocian	0.097 (+0.065)	**0.125** (+0.028)
Lemosin	0.069 (+0.035)	**0.103** (+0.034)
Provençau	0.103 (+0.034)	**0.138** (+0.035)

Table 1: Fine-tuned model's accuracy in analogy computation, measured with two criteria. Values in parentheses represent the change in score from the baseline.

Semantic	Fam : Minjar	*Hunger: Eat*
	Set : Beure	*Thirst : Drink*
Syntactic	Far : Fach	*Want (INF) : Want (PP)*
	Voler : Volgut	*Want (INF) : Want (PP)*

Figure 3: Examples of semantic and syntactic analogies from the Lemosin dataset with English translations in italics. INF: infinitive, PP: past participle.

overall increase in accuracy when using the 3Cos-MUL described in Equation 2. The largest such improvement occurs for Gascon.

Error Analysis As a means of better understanding the results on our analogy dataset, we calculate the accuracy separately for the syntactic and semantic relations in our analogy dataset. See Figure 3 for examples of each analogy type. The score on semantic analogies is 0.0 for all dialects, regardless of whether Equation 1 or Equation 2 was used to solve the analogies. In Table 2, we present the analogy scores for each dialect when only taking the syntactic relations into account.

This disparity in performance for semantic and syntactic relations may be attributable to our evaluation approach. Indeed, Levy and Goldberg (2014) note that there is an alternative formulation to the 3CosAdd approach called the "PairDirection" method. In their work on analyzing word embedding quality with analogy computations, Mikolov et al. (2013b) used this PairDirection method for evaluating their semantic analogies while using a method that was algebraically equivalent to 3CosADD for syntactic relations.

Beyond the method that we use to calculate the analogies, it would be interesting to experiment further with the specific pre-training tasks and architecture of our base model. For instance, the multi-network approach taken for pre-training sentence-BERT may offer a promising solution to the relatively poor representation of semantic relations in our fine-tuned model (Reimers and Gurevych, 2019).

	3CosADD	3CosMUL
Gascon	0.0000	**0.2000**
Lengadocian	0.1429	**0.1818**
Lemosin	0.1053	**0.1579**
Provençau	0.1579	**0.2105**

Table 2: Fine-tuned model's accuracy in analogy computation when only taking syntactic relations into account.

	Accuracy
Gascon	0.322 (+0.067)
Lemosin	0.291 (+0.051)
Provençau	**0.409** (+0.109)

Table 3: Fine-tuned model's accuracy in choosing a word's corresponding Lengadocian form ("Lengadocian Lexicon Induction"). Values in parentheses represent the change in score from the baseline.

5.2 Lengadocian Lexicon Induction

Background As another means of evaluating the fine-tuned model's representations of Occitan's dialects, we conduct a lexicon induction task that assesses the similarity of parallel words across the dialects. Bilingual lexicon induction is a common use case for multilingual embeddings (Woller et al., 2021; Mikolov et al., 2013a). Here, we aim to induce the Lengadocian lexicon using the other three dialects in our dataset.[2] For each word in Gascon, Provençau, and Lemosin, we find the Lengadocian word with the most similar embedding, again using cosine similarity. If the most similar word is an equivalent Lengadocian term for the other dialect's word, we score this as correct.

Results Accuracy scores for the fine-tuned model's performance on the Lengadocian lexicon induction task can be found in Table 3. Fine-tuning mBERT with multi-dialect Occitan data led to increases in performance on this task for all dialects. Interestingly, there is more disparity between dialects in this task compared to the 3Cos-MUL scoring of the analogy task. Whereas the fine-tuned model correctly selects the Lengadocian form of Provençau words in 40.9% of the cases, performance for selecting the Lengadocian forms of Lemosin words is correct in just 29.1% of cases.

Error Analysis To study the impact of dialectical variants' surface similarity on their representation, we stratify the results of the Lengadocian lexi-

[2] We choose to induce the Lengadocian lexicon because it is likely the best-represented dialect in our training data, but the procedure could be repeated to induce any of the other dialects' lexicons.

	Low	Med	High
Gascon	**0.4982**	0.2444	0.0917
Lemosin	**0.4172**	0.2744	0.0756
Provençau	**0.5671**	0.3766	0.1154

Table 4: Fine-tuned model's accuracy in choosing a word's corresponding Lengadocian form ("Lengadocian Lexicon Induction"), stratified by orthographic distance between target Lengadocian word and corresponding dialect word; low: LevDist=1; Med: LevDist in range [2,3], High: LevDist > 3.

con induction by the Levenshtein distance between the word in a given dialect and its counterpart in Lengadocian (see Table 4). These results indicate that the further apart two word forms are, the less similarly our model represents them, even though they are semantically similar.

This trend indicates that for cases where spelling differences are minimal, our fine-tuned model seems to model word pairs similarly. However, word pairs with low surface similarity are not represented well by the model. Thus, while our fine-tuned model may have learned to represent orthographical variation in Occitan well (i.e., variation of a few characters), it still struggles with dialectical variation at the level of whole lexical items. To some extent, Lemosin's relatively high proportion of OOV items with respect to the fine-tuning corpus (Figure 2) may explain the model's weaker performance in inducing the Lengadocian lexicon from Lemosin representations.

Table 5 contains examples of mistakes for the induction of the Lengadocian Lexicon from Provençau. Some mistakes are seemingly random, such as the Provençau word for "January" being closer to "Give" than the Lengadocian equivalent for January. However, an error such as the Provençau embedding for "Neighborhood" being most similar to the Lengadocian word "Social" shows some evidence of semantic consistency in our embeddings. Though this is just a single example, it shows that this method of evaluation for our embeddings is relatively strict. As emphasized by Glavas et al. (2019), if not done consistently and in the context of a more comprehensive analysis, bilingual lexicon induction is not necessarily an ideal evaluation of cross-lingual word embeddings.

5.3 Extrinsic Evaluation

Background Using the Tolosa Treebank, a multi-dialect dataset (Miletić et al., 2020), we train task heads for part-of-speech tagging and Universal Dependency[3] parsing. The Tolosa Treebank contains texts from the Occitan dialects Lengadocian, Lemosin, Provençau, and Gascon. We experiment with two training setups: In the first, we use data from all four dialects in the train, validation,[4] and test sets. In the second setup, we attempt to mirror a more realistic scenario for low-resource languages where less annotated data is typically available. To do this, we train the task heads with only Lengadocian data. Lengadocian was chosen because the authors of the OcWikiDisc corpus believed this dialect to be the best represented in the corpus, and it has the most data in the Tolosa Treebank. We then test the PoS-tagging and dependency parsing abilities of the model on all four dialects with test sets from the Tolosa Treebank. As with the intrinsic evaluations, we report results for both the baseline mBERT model and our fine-tuned mBERT model. We use the MaChAmp framework (van der Goot et al., 2021b) for the multitask fine-tuning.

Results Scores for the PoS taggers and dependency parsers are in Tables 6 and 7. In PoS tagging, accuracy is relatively high in both training scenarios. For Gascon, Lengadocian, and Provençau the fine-tuned model showed small improvements relative to the baseline mBERT, while the fine-tuned model performed slightly worse than the baseline for Lemosin. On the full test set with data from all four dialects, the model trained with PoS data from all four dialects reaches an accuracy of 94.1%, outperforming the model trained on only Lengadocian data. PoS-tagging performance is best for Lengadocian data in both training setups, although only by a small margin.

As for UD parsing, Gascon had the highest labeled attachment scores in both training conditions, while performance was again the worst for Provençau. Similar to PoS tagging, scores for Lemosin dependency parsing decreased with fine-tuning. Across the dialects, there was a wider range of UD parsing scores compared to scores for PoS tagging, but performance was best on average in the condition where all four dialects were used during training.

Error Analysis Despite observing similar results, we explore potential differences in the quality of the two part-of-speech taggers we trained. To do this, we visualize confusion matrices to illustrate how

[3] https://universaldependencies.org/

[4] Provençau and Lemosin are not included in the validation set due to the relatively small amount of data in these dialects.

Provençau (*EN*)	True Lengadocian	Lengadocian Term Selected in Evaluation (*EN*)	Provençau-Target Levenshtein Distance
Quartier (*Neighbourhood*)	Quartièr	Social (*Social*)	1
Fieu (*String*)	Fial	Lòc (*Place*)	2
Janvier (*January*)	Genièr	Dar (*Give (V.)*)	4

Table 5: Example errors from the Provençau–Lengadocian lexicon induction task. Column 2 contains the correct Lengadocian equivalent to the Provençau term; the incorrect Lengadocian term with the most similar embedding to the Provençau term is in column 3; English translations in parentheses.

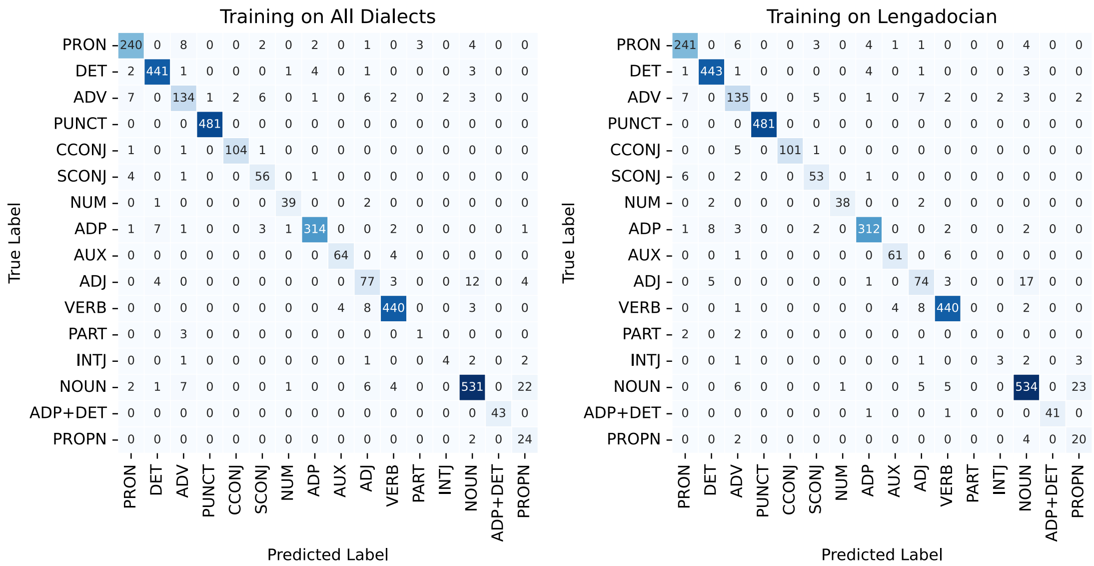


Figure 4: Confusion matrix for PoS taggers when trained on data from all dialects (left) and only Lengadocian (right).

	All Dialects	Lengadocian
Gascon	**0.941** (+0.005)	0.922 (+0.019)
Lengadocian	**0.946** (+0.008)	0.943 (+0.008)
Lemosin	**0.934** (-0.002)	0.927 (-0.005)
Provençau	0.927 (+ 0.000)	**0.932** (+ 0.012)
Full Test Set	**0.941** (+0.005)	0.936 (+0.007)

Table 6: PoS-tagging accuracy for the fine-tuned model; “All Dialect” PoS tagger used train and development data from all dialects. The “Lengadocian” PoS tagger was trained on only Lengadocian data. Values in parentheses represent the change in score from the baseline.

	All Dialects	Lengadocian
Gascon	**0.791** (+0.007)	0.755 (+0.035)
Lengadocian	**0.761** (+0.003)	0.742 (+0.008)
Lemosin	**0.691** (-0.007)	0.679 (-0.012)
Provençau	**0.603** (- 0.005)	0.579 (+0.003)
Full Test Set	**0.735** (+0.001)	0.714 (+0.008)

Table 7: Labeled attachment scores for Universal Dependency parsing with the fine-tuned model. “All Dialect” parser used train and development data from all dialects. The “Lengadocian” parser was trained on only Lengadocian data. Values in parentheses represent the change in score from the baseline.

each tagger performed on the full test set, which included data from all four studied dialects (see Figure 4). The results highlight one key failure of our “low-resource” method where we use a PoS tagger that was trained only on Lengadocian data for inference on all dialects. Specifically, it indicates that the Lengadocian-trained tagger never correctly classifies particles in the evaluation set. This is an expected limit of zero-shot attempts to do PoS tagging with Gascon, as Gascon is the only dialect of Occitan that carries enunciative particles that mark clause type. This problem was also encountered in previous work on Occitan PoS tagging, where the authors proposed explicit rules for tagging the particles as a solution (Vergez-Couret and Urieli, 2014). In the original dependency parsing experiments published with the Tolosa Treebank, the authors improve performance on Gascon evaluation data by including training data from Gascon in their model, an approach that also led to our highest-performing PoS tagger (Miletic et al., 2020a).

In the same paper on the Tolosa Treebank, Miletić et al. (2020) note that performance for dependency parsing was consistently worst for

Provençau. We obtained similar results for our dependency parsing experiments. As for our PoS taggers, when trained on all four dialects, performance was also worst on Provençau. Yet, the Provençau portion of the Tolosa Treebank has the lowest proportion of OOV items relative to the fine-tuning corpus, meaning we would generally expect its contents to be among the best represented of the dialects (Figure 2). To the contrary, performance on Provençau consistently outranks the other dialects for our intrinsic evaluations. We encourage further research on the specific morphosyntactic properties and orthographic tendencies of Provençau to clarify why—despite its relatively strong internal representation—it stands out as more difficult to tag in these tasks.

6 Discussion

Overall, our experiments on using non-standardized text data to fine-tune mBERT yielded mixed results. Fine-tuning mBERT with orthographically non-standard Occitan data led to little improvement in terms of performance on computing analogies, POS tagging, and dependency parsing. However, fine-tuning with the multi-dialect data consistently improved results in using one dialect's lexicon to induce the parallel lexicon of another dialect, Lengadocian.

Taken together, these results provide support for the idea that including data with dialectical lexical variation and non-standardized orthography in fine-tuning data is not necessarily harmful to model performance. I.e., the fine-tuning we carried out with multiple dialects of Occitan did not deprecate mBERT's baseline performance on downstream tasks like part-of-speech tagging. As normalizing data during preprocessing can pose a substantial burden for low-resource NLP, our results are encouraging in that they suggest that in some contexts, including orthographically inconsistent data from multiple dialects will not harm the model.

Previous work on fine-tuning language models has led to various conclusions about the effect of different types of "noise" in the fine-tuning data. For instance, it has been shown that fine-tuning on English data with synthetic spelling errors can reduce BERT's performance on downstream sentiment analysis (Kumar et al., 2020). While our results do not seem to indicate a negative effect of including non-standardized data in the fine-tuning data, simply including small amounts of data from multiple dialects of Occitan was not enough to increase mBERT's performance on downstream tasks (PoS tagging and UD parsing) with the dialects. Furthermore, as shown in the Lengadocian lexicon induction, the model failed to capture the similarity of parallel lexical items that have low surface similarities (i.e., high Levenshtein distance).

Some research has shown that high surface similarity between pre-training and fine-tuning data can result in better performance on downstream tasks such as PoS tagging and machine translation (Aepli and Sennrich, 2022; Amrhein and Sennrich, 2020). Bearing that in mind, our future efforts will look at whether increasing the surface similarity between the pre-training and Occitan fine-tuning data of our models will allow the model to better learn from the non-standardized data. Indeed, results from our Lengadocian induction task provide some further evidence that this may help, as the model represented parallel words from different dialects more similarly when their spelling was more similar. Increasing the surface similarity of the pre-training and fine-tuning data may involve using a model pre-trained only on languages that are more closely related to Occitan, such as Catalan, Spanish, and French. In the same vein as Aepli and Sennrich (2022), we may also explore injecting the pre-training data with noise in the form of characters or even replacing whole words with Occitan variants. Furthermore, Kumar et al. (2020) attribute BERT's detriment in performance when fine-tuned on noisy data to the model's tokenizer. Though out of the scope of the present work, we intend to focus future efforts on the overlap between the Occitan dialects' subtokens. Work on this matter may be particularly beneficial in understanding the results of our lexicon induction task.

7 Conclusion

In this work, we experimented with the capacity of a pre-trained encoder to represent dialect variation—both lexical and orthographic—in a low-resource language. In doing so, we aimed to test the extent to which cross-lingual transfer learning allows for effectively representing Occitan's dialects. While the experiments yielded no clear evidence that dialect representation was improved after fine-tuning, we can still interpret our findings as an indicator that orthographic normalization may not be necessary when fine-tuning large, multilingual models.

Limitations

A potential limitation of this study is that we only worked with one base model, mBERT. It is possible that experimenting with other language models, i.e., models only on languages closely related to Occitan, would have yielded different results, while also telling us more about the relative usefulness of massively multilingual models for low-resource languages.

Another limitation is our inability to better characterize the pre-training and fine-tuning data in the experiment. Indeed, while the authors of the OcWikiDisc corpus performed manual evaluations to determine the dialect make-up of a small sample of their corpus, the total number of data points in each dialect in the OcWikiDisc is not known (Miletic and Scherrer, 2022). Even less is known about the writing standard and dialect make-up of the WikiMatrix data which we also used for fine-tuning, meaning that overall, we cannot be sure that any variation between the dialects' results was not simply driven by a difference in the amount of data in each.

Further, our experimental setup is limited in that our UD parsers perform worse than the highest performing UD parsers trained on the Tolosa Treebank in Miletić et al. (2020). The authors note that their worst LAS scores come from a model that was also trained on UD data from languages closely related to Occitan. Along the same lines, it may be that our use of a large, multilingual language model to carry out the UD parsing is limiting the utility of the relatively small amounts of dialect-specific Occitan UD data.

Acknowledgements

We would like to extend our gratitude to Aleksandra Miletić for sharing the Tolosa Treebank, as well as to the anonymous reviewers for their insightful commentary. This work was supported by the Swiss National Science Foundation (project no. 191934).

References

Noëmi Aepli and Rico Sennrich. 2022. Improving zero-shot cross-lingual transfer between closely related languages by injecting character-level noise. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 4074–4083, Dublin, Ireland. Association for Computational Linguistics.

Khetam Al Sharou, Zhenhao Li, and Lucia Specia. 2021. Towards a better understanding of noise in natural language processing. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 53–62, Held Online. INCOMA Ltd.

Chantal Amrhein and Rico Sennrich. 2020. On Romanization for model transfer between scripts in neural machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2461–2469, Online. Association for Computational Linguistics.

Gilles Arbousset, Éric Astié, Dominique Decomps, Philippe Gardy, Patric Guilhemjoan, Jean-Marie Sarpoulet, Miquèla Stenta, Claire Torreilles, and Lise Boissinot. 2003. *Òc-Ben! Première Année d'Occitan: Livre de l'Élève*.

Rachel Bawden, Jonathan Poinhos, Eleni Kogkitsidou, Philippe Gambette, Benoît Sagot, and Simon Gabay. 2022. Automatic normalisation of early Modern French. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3354–3366, Marseille, France. European Language Resources Association.

P. Bec. 1995. *La Langue Occitane*, sixth edition. Presses universitaires de France.

Delphine Bernhard, Anne-Laure Ligozat, Myriam Bras, Fanny Martin, Marianne Vergez-Couret, Pascale Erhart, Jean Sibille, Amalia Todirascu, Philippe Boula de Mareüil, and Dominique Huck. 2021. Collecting and annotating corpora for three under-resourced languages of france: Methodological issues.

Philippe Blanchet. 2004. Provençal as a distinct language? sociolinguistic patterns revealed by a recent public and political debate.

Eugene Buckley. 2009. Phonetics and phonology in gallo-romance palatalisation. *Transactions of the Philological Society*, 107(1):31–65.

Ander Corral, Igor Leturia, Aure Séguier, Michäel Barret, Benaset Dazéas, Philippe Boula de Mareüil, and Nicolas Quint. 2020. Neural text-to-speech synthesis for an under-resourced language in a diglossic environment: the case of Gascon Occitan. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 53–60, Marseille, France. European Language Resources association.

Thales Felipe Costa Bertaglia and Maria das Graças Volpe Nunes. 2016. Exploring word embeddings for unsupervised textual user-generated content normalization. In *Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT)*, pages 112–120, Osaka, Japan. The COLING 2016 Organizing Committee.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Goran Glavas, Robert Litschko, Sebastian Ruder, and Ivan Vulic. 2019. How to (properly) evaluate cross-lingual word embeddings: On strong baselines, comparative analyses, and some misconceptions. *arXiv preprint arXiv:1902.00508*.

Pierre-Edouard Honnet, Andrei Popescu-Belis, Claudiu Musat, and Michael Baeriswyl. 2018. Machine translation of low-resource spoken dialects: Strategies for normalizing Swiss German. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Ankit Kumar, Piyush Makhija, and Anuj Gupta. 2020. Noisy text data: Achilles' heel of bert. In *WNUT*.

Olli Kuparinen, Aleksandra Miletić, and Yves Scherrer. 2023. Dialect-to-standard normalization: A large-scale multilingual evaluation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13814–13828, Singapore. Association for Computational Linguistics.

Omer Levy and Yoav Goldberg. 2014. Linguistic regularities in sparse and explicit word representations. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, pages 171–180, Ann Arbor, Michigan. Association for Computational Linguistics.

Massimo Lusetti, Tatyana Ruzsics, Anne Göhring, Tanja Samardžić, and Elisabeth Stark. 2018. Encoder-decoder methods for text normalization. In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018)*, pages 18–28, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Tomas Mikolov, Quoc V Le, and Ilya Sutskever. 2013a. Exploiting similarities among languages for machine translation. *arXiv preprint arXiv:1309.4168*.

Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. 2013b. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751, Atlanta, Georgia. Association for Computational Linguistics.

Aleksandra Miletic, Myriam Bras, Marianne Vergez-Couret, Louise Esher, Clamença Poujade, and Jean Sibille. 2020a. Building a Universal Dependencies treebank for Occitan. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2932–2939, Marseille, France. European Language Resources Association.

Aleksandra Miletic, Myriam Bras, Marianne Vergez-Couret, Louise Esher, Clamença Poujade, and Jean Sibille. 2020b. A four-dialect treebank for Occitan: Building process and parsing experiments. In *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 140–149, Barcelona, Spain (Online). International Committee on Computational Linguistics (ICCL).

Aleksandra Miletić, Myriam Bras, Marianne Vergez-Couret, Louise Esher, Clamença Poujade, and Jean Sibille. 2020. A four-dialect treebank for occitan: Building process and parsing experiments. In *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 140–149.

Aleksandra Miletic and Yves Scherrer. 2022. OcWikiDisc: a corpus of Wikipedia talk pages in Occitan. In *Proceedings of the Ninth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 70–79, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Aleksandra Miletić and Janine Siewert. 2023. Lemmatization experiments on two low-resourced languages: Low Saxon and Occitan. In *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*, pages 163–173, Dubrovnik, Croatia. Association for Computational Linguistics.

Annick Morin. 2008. The syntax of gascon clause-type particles. *Linguistica Atlantica*, 29:137–155.

Ivan Provilkov, Dmitrii Emelianenko, and Elena Voita. 2020. BPE-dropout: Simple and effective subword regularization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1882–1892, Online. Association for Computational Linguistics.

Ulrich Reffle. 2011. Efficiently generating correction suggestions for garbled tokens of historical language. *Natural Language Engineering*, 17(2):265–282.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Annette Rios and Richard Castro Mamani. 2014. Morphological disambiguation and text normalization for southern quechua varieties.

Holger Schwenk, Vishrav Chaudhary, Shuo Sun, Hongyu Gong, and Francisco Guzmán. 2021. WikiMatrix: Mining 135M parallel sentences in 1620 language pairs from Wikipedia. In *Proceedings of the*

16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, pages 1351–1361, Online. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Aarohi Srivastava and David Chiang. 2023. BERTwich: Extending BERT's capabilities to model dialectal and noisy text. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15510–15521, Singapore. Association for Computational Linguistics.

Gongbo Tang, Rico Sennrich, and Joakim Nivre. 2020. Understanding pure character-based neural machine translation: The case of translating Finnish into English. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4251–4262, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Rob van der Goot. 2021. Cl-monoise: Cross-lingual lexical normalization. In *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, pages 510–514.

Rob van der Goot, Alan Ramponi, Arkaitz Zubiaga, Barbara Plank, Benjamin Muller, Iñaki San Vicente Roncal, Nikola Ljubešić, Özlem Çetinoğlu, Rahmad Mahendra, Talha Çolakoğlu, Timothy Baldwin, Tommaso Caselli, and Wladimir Sidorenko. 2021a. MultiLexNorm: A shared task on multilingual lexical normalization. In *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, pages 493–509, Online. Association for Computational Linguistics.

Rob van der Goot, Ahmet Üstün, Alan Ramponi, Ibrahim Sharaf, and Barbara Plank. 2021b. Massive choice, ample tasks (MaChAmp): A toolkit for multi-task learning in NLP. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 176–197, Online. Association for Computational Linguistics.

Marianne Vergez-Couret and Assaf Urieli. 2014. Pos-tagging different varieties of Occitan with single-dialect resources. In *Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects*, pages 21–29, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.

Lisa Woller, Viktor Hangya, and Alexander Fraser. 2021. Do not neglect related languages: The case of low-resource Occitan cross-lingual word embeddings. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 41–50, Punta Cana, Dominican Republic. Association for Computational Linguistics.

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
June 20, 2024

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 macro-areas: 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.

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

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

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:

The Role of Adverbs in Language Variety Identification: The Case of Portuguese Multi-word Adverbs

Izabela Müller and Jorge Baptista
U. Algarve - FCHS
Campus de Gambelas, Faro (Portugal)
INESC-ID Lisboa – HLT
{belagrein,jorge.baptista}@inesc-id.pt

Nuno Mamede
U. Lisboa - IST, Lisbon (Portugal)
INESC-ID Lisboa – HLT
R. Alves Redol 9, Lisboa (Portugal)
Nuno.Mamede@inesc-id.pt

Abstract

This paper aims to assess the role of multi-word compound adverbs in distinguishing Brazilian Portuguese (PT-BR) from European Portuguese (PT-PT). For this study, a large lexicon of Portuguese multi-word adverbs (3,665) was annotated with diatopic information regarding language variety, which has not been available so far. The paper then investigates the distribution of this category in the DSL (Dialect and Similar Language) corpus of journalistic texts, representing Brazilian (PT-BR) and European Portuguese (PT-PT). Results indicate a substantial similarity between the two varieties, with a considerable overlap in the lexicon of multi-word adverbs. Additionally, specific adverbs unique to each language variety were identified. Lexical entries recognized in the corpus represent 18.2% (PT-BR) to 19.5% (PT-PT) of the lexicon, and approximately 5,700 matches in each partition. While many of the matches are spurious due to ambiguity with otherwise non-idiomatic, free strings, occurrences of adverbs marked as exclusive to one variety in texts from the other variety are rare.

1 Introduction

This study seeks to identify and contrast multi-word (compound) adverbs between the Brazilian (PT-BR) and European (PT-PT) varieties of Portuguese. Two key factors underpin this focus: Firstly, multi-word expressions often prove to be less ambiguous than single words, even when their meaning is idiomatic (non-compositional). Secondly, despite constituting a significant portion of lexicons in many languages, adverbs are frequently overlooked in Natural Language Processing, possibly due to their heterogeneous nature and lexical range. Furthermore, to the best of our knowledge, no assessment has been made until now, of the lexical distribution of language variety-specific multi-word adverbs in Portuguese. And even if such a distribution were skewed, no study seems to be available on the distribution in corpora of such language variety-specific multi-word adverbs. The goal of this paper is to provide a clear answer to both these issues.

While language variety identification can be a crucial task for dialect-sensitive NLP tasks, the main idea underlying this paper is *not* to evaluate the identification of the linguistic variety between PT-BR and PT-PT *per se* using multi-word adverbs, but to determine the impact of the adverbial expressions and the extent to which they are asymmetrical across both varieties. The aim is to achieve this through an analysis conducted on two comparable corpora of journalistic texts, one in PT-BR and the other in PT-PT, that have been used in previous DSL shared tasks (Tan et al., 2014).

This paper begins with an overview of the primary goals and the resources utilized. In Section 2, we look deeper into the notion of multi-word (compound) adverbs and discuss the ongoing research focused on developing a lexicon of multi-word adverbs in Brazilian Portuguese. Section 3 outlines the methodology used in this experiment, specifically addressing the asymmetry of adverbial constructions identified in both varieties of Portuguese. Section 4 presents the findings and their analysis. Lastly, Section 5 concludes the paper with final observations.

2 The Lexicon of Portuguese Multi-word Adverbs

Multi-word adverbs, also referred to as *compound adverbs*, are expressions composed of two or more words forming a single lexical unit with specific word combinations. While they generally adhere to syntactic rules for phrase formation, their structure is often “frozen”, meaning their components cannot be rearranged, inserted, or reduced (through ellipsis), and they typically exhibit idiosyncratic constraints on morphosyntactic variation (Gross,

Proceedings of the Eleventh Workshop on NLP for Similar Languages, Varieties, and Dialects (VarDial 2024), pages 99–106
June 20, 2024

1982, 1986b; Guimier, 1996). One key characteristic of multi-word adverbs is their lack of semantic compositionality, wherein the meaning of the expression cannot be deduced from the individual meanings of its elements, resulting in an idiomatic overall meaning, e.g.:

(1) *Aprendi isso* a duras penas.[1] 'I learned that *the hard way*.'

The example above demonstrates that the expression *a duras penas* 'the hard way' is a compound, multi-word adverb, as it adheres to the mentioned constraints regarding:[2] (a) word order: ?/°*a penas duras*; (b) reduction: **a penas*; (c) idiosyncratic restrictions on morphosyntactic variation: °*a dura pena*; and, finally, (d) limited insertions may be acceptable: *a* muito *duras penas*. In the *PtTenTenCorpus2020* (12.5 billion words) (Kilgarriff et al., 2014), (a) only 3 instances were found of the idiomatic string *a penas duras*, with the adjective in a post-nominal position; 7 other uses of this word order are all literal ('harsh penalties'), as indicated by '°'; these 10 instances are marginal, compared to 6,205 occurrences of *a duras penas*, with reversed order, all of which are idiomatic; (b) the reduced form (over 5.5 thousand occurrences) is never interpreted as an idiomatic expression, and it is usually a sub-string of a distributionally free phrase; (c) the singular form (21 instances) is always literal; and, finally, (d) only 2 instances were found with the adverb quantifier *muito* in the prenominal position, both idiomatic; and none after the noun.

Multi-word adverbial expressions have received a great deal of interest in the field of linguistic studies and have been the object of previous studies in various languages (Gross, 1996a; Di Gioia, 2001; Català, 2003; Laporte et al., 2008; Palma, 2009; Shudo et al., 2011; Català et al., 2020; Müller et al., 2022, 2023).

The lexicon of multi-word adverbs in Brazilian Portuguese is part of an ongoing study that focuses on identifying, classifying, and describing the lexical, syntactical, and semantic features of multi-word adverbs. The goal is to expand the list of European Portuguese multi-word adverbs (Palma, 2009) by incorporating Brazilian Portuguese adverbial compounds. This will result in a comprehensive lexicon, covering expressions specific to each Portuguese variety, as well as those common to both.

We adopt the theoretical-methodological framework of Lexicon-Grammar (Gross, 1975, 1981, 1996b), based on Harris (1991) Operator Transformational Grammar, along with the formal classification of compound adverbs as proposed by Gross (1986a). This classification system categorizes adverbs based on the internal sequence of their grammatical categories. Table 1 presents the current distribution of multi-word adverbs by formal classes and language variety. Currently, 68.3% of the lexicon incorporates multi-word adverbs shared between PT-PT and PT-BR. At the same time, 26.3% includes expressions exclusively found in PT-BR, and 5.5% is exclusive to PT-PT adverbial expressions.

Additionally, we apply the criteria proposed by Molinier and Levrier (2000) to classify single adverbs ending in *-ment* in French, according to their syntactic-semantic features. We believe these criteria are applicable to the description of *-mente* adverbs in Portuguese, as shown in Fernandes (2011), as well as multi-word adverbs in Portuguese (Palma, 2009; Català et al., 2020).

Molinier and Levrier (2000)'s framework outlines two primary categories of adverbs: those that modify the constituents of the sentence, and thus are considered *internal modifiers* (type M), and those that modify the entire sentence, known as *external modifiers* (type P). The authors further propose a nuanced sub-classification based on the adverbs' function and the relations they establish within a sentence, delineating various syntactic-semantic adverbial classes.

External modifiers are subdivided into three categories: (i) conjunctive adverbs (PC), (ii) disjunctive adverbs of style (PS), and (iii) disjunctive adverbs of attitude (PA). The latter category is further subdivided into four subclasses: (a) adverbs of *habit* (PAh), (b) *evaluative* adverbs (PAe), (c) *modal* adverbs (PAm), and (d) *subject-oriented* adverbs (PAs).

Adverbs that modify an internal constituent of the sentence are classified into six subclasses: (iv) *manner* adverbs (MV), (v) *subject-oriented manner* adverbs (MS), (vi) adverbs of *time* (MT), (vii) *point-of-view* adverbs (MP), (viii) *quantity* adverbs (MQ), and (ix) *focusing* adverbs (MF).

We adopted this general framework to classify and describe Portuguese adverbs. Furthermore, we introduce a new category, (x) *locative* adverbs

[1] *O Estadão*, from *PtTenTenCorpus*, id=11740543771

[2] In the examples, '*' indicates the string is unacceptable, '?' dubiously acceptable, or '°' acceptable but with literal/compositional meaning.

Class	Internal Structure	Examples	PT-PT	%	PT-BR	%	PT	%	Total	%
PC	*Prep C*	*em vão* 'in vain'	28	0.14	243	0.25	667	0.27	938	0.256
PDETC	*Prep Det C*	*pelo menos* 'at least'	57	0.29	218	0.23	522	0.21	797	0.218
PAC	*Prep Adj C*	*de má vontade* 'unwillingly'	11	0.06	46	0.05	231	0.09	288	0.079
PCA	*Prep C Adj*	*por maioria absoluta* 'by absolute majority'	22	0.11	70	0.07	268	0.11	360	0.098
PCDC	*Prep C1 de C2*	*por conta da casa* 'on the house'	21	0.11	83	0.09	207	0.08	311	0.085
PCPC	*Prep C1 Prep C2*	*da cabeça aos pés* 'head to toes'	46	0.23	105	0.11	265	0.11	416	0.114
PCONJ	*Prep C1 Conj C2*	*em verso e prosa* 'in verse and prose'	9	0.05	74	0.08	168	0.07	251	0.068
PF	frozen sub-clause	*dito isso* 'this said'	2	0.01	41	0.04	88	0.04	131	0.036
PV	*Prep V W*	*até dizer chega* lit.:'until say enough', 'a lot'	2	0.01	2	0.002	25	0.01	29	0.008
PJC	*Conj C*	*e por aí vai* 'and so on'	2	0.01	47	0.05	31	0.01	80	0.022
PACO	*<Adj>como C*	*<surdo>como uma porta* 'deaf as a door'			7	0.01	3	0.001	10	0.003
PVCO	*<V>como C*	*<trabalhar>como uma mula* 'word like a mule'			26	0.03	25	0.01	51	0.014
		Total	200	0.055	962	0.262	2,500	0.683	3,662	

Table 1: Formal classification of Portuguese multi-word adverbs. Codes for classes are conventional. Internal structure: adjective *Adj*, *C1* and *C2* lexical constants, conjunction *Conj*, Determiner *Det*, Preposition *Prep*, Verb *V*, undefined sequence of elements *W*. Distribution per variety: European Portuguese *PT-PT*, Brazilian Portuguese *PT-BR*, Common Portuguese *PT*. Zero values were removed.

(ML), which was not included in this framework before, even though it is not new to the study of adverbs. You can find more details about each category in (Müller et al., 2022, 2023).

Table 2 displays the distribution of the lexicon based on this syntactic-semantic classification across different language varieties. To the best of our knowledge, this lexicon represents the most extensive collection of multi-word adverbs available in Portuguese.

The predominant categories are *manner* adverbs (MV: 59.9%) and *time* adverbs (MT: 14.8%). Within the latter category, 76% are corresponded to *date* adverbs, indicating temporal locatives. Additionally, the recently introduced locative class (ML) from (Müller et al., 2023) accounts for 5.5%. Conjunctive adverbs (PC: 7%) and quantifying adverbs (MQ: 5.1%) are also noteworthy.

The assignment of language variety to the multi-word adverbs in the lexicon is based mostly on their distribution in the corpora, particularly (i) for PT-PT, the CETEMPúblico corpus (Rocha and Santos, 2000)[3] (ii) for PT-BR, the Corpus Brasileiro[4], with approximately 1 billion words; both (i) and (ii) are available through Linguateca[5]; and, (iii) for both variants, the Portuguese Web 2020 (ptTenTenCorpus20) (Wagner Filho et al., 2018; Kilgarriff et al., 2004), with 12,5 billion words (PT-PT: 893.2 million words, PT-BR: 8 billion words).

3 Methods

In order to assess the influence of multi-word adverbs on the two Portuguese varieties, we employed the European (PT-PT) and Brazilian (PT-BR) partitions of the Discrimination of Similar Languages (DSL) Corpus Collection (DSLCC, v.04) (Tan et al., 2014)[6]. These partitions were originally curated for the DSL task and served as the primary dataset for the shared tasks conducted as part of the NLP for Similar languages, Varieties and Dialects (VarDial) workshop (Zampieri et al., 2017). The PT-PT texts comprise 18,000 sentences with a total of 735,503 words, while the PT-BR texts also encompass 18,000 sentences and a slightly larger word count of 791,872. Table 3 shows the breakdown of the number of sentences, words and different words in each partition.

To process the corpora, we utilized the linguistic development platform Unitex (v.3.3)(Paumier et al., 2021).[7] The texts underwent pre-processing using the linguistic resources provided by the system, specifically the text segmentation tool and the simple-word dictionary. The lexicon of multi-word adverbs was also formatted into the DELA format compatible with Unitex and then applied to the corpora. For instance, consider the entry for the manner adverb *a duras penas* 'the hard way':

```
a duras penas.ADV+PAC+MV+PT+BR
```

In this format, each adverb entry consists of a string with a part-of-speech designation and a set of features, including its formal class, syntactic-semantic class, and the language varieties it pertains to (+PT and/or +BR). Adverbs not specific to a language variety are explicitly labeled with the features +NotPT (e.g., [*responder*] *de bate pronto* '(to reply) right away') or +NotBR (e.g., [*cair*] *de ratatulha* '(to fall) headlong').

[3] www.linguateca.pt/cetempublico
[4] http://corpusbrasileiro.pucsp.br/
[5] www.linguateca.pt/
[6] http://ttg.uni-saarland.de/resources/DSLCC/
[7] https://unitexgramlab.org/

Class	Examples	PT-PT	%	PT-BR	%	PT	%	Total	%
PC (conjunctive)	*afinal de contas 'after all'*	3	0.015	42	0.043	213	0.085	258	0.070
PS (disjunctive of style)	*com toda a franqueza 'in all honesty'*	1	0.005	10	0.010	54	0.022	65	0.018
PA (disjunctive of attitude)									
PAa (evaluative)	*por pura sorte 'by sheer luck'*					21	0.008	21	0.006
PAm (modal)	*com certeza 'certainly'*	1	0.005	10	0.010	25	0.010	36	0.010
PAs (subject-oriented)	*pelo meu lado 'for my part'*			4	0.004			4	0.001
PAh (habit)	*de costume 'usually'*			4	0.004	12	0.005	16	0.004
MV (manner)	*por amor à pátria 'for love of country'*	157	0.781	615	0.637	1,423	0.570	2,195	0.599
MS (subject-oriented mode)	*de boa fé 'in good faith'*	2	0.010	22	0.023	93	0.037	117	0.032
MT (time)									
MTd (date)	*a horas mortas 'at dead of night'*	24	0.119	80	0.083	307	0.123	411	0.112
MTf (frequency)	*dia sim dia não 'every other day'*	5	0.025	25	0.026	53	0.021	83	0.023
MTu (duration)	*anos a fio 'for years on end'*			12	0.012	35	0.014	47	0.013
MP (point of view)	*na prática 'in practice'*					4	0.002	4	0.001
MQ (quantifier)	*aos montes 'in abundance'*	5	0.025	60	0.063	119	0.048	185	0.051
MF (focalizer)	*em especial 'especially'*			3	0.003	17	0.007	20	0.005
ML (locative)	*nos confins do mundo 'at the ends of the earth'*	3	0.015	78	0.081	120	0.048	201	0.055
	Total	201	0.055	966	0.264	2,496	0.681	3,662	

Table 2: Syntactic-semantic classification of Portuguese multi-word adverbs. Codes for classes are conventional. Sub-classes of PA and MT are presented. Distribution per variety: European Portuguese *PT-PT*, Brazilian Portuguese *PT-BR*, Common Portuguese *PT*. Zero values were removed.

```
de bate pronto.ADV+PCA+MV+NotPT+BR
de ratatulha.ADV+PAC+MV+PT+NotBR
```

For the classification of adverbs according to the language variety, two linguists, native speakers of each variety, manually, separately and systematically annotated the lexicon entries, deciding whether they belonged to each other variety. Additionally, we also relied on corpus consultation *PtTenTen20* partitions of each language variety and controlled web search using domains **.pt** and **.br** to verify the occurrence of the adverbs in each variety. In a second moment, aspects of lexical variation (prepositions, determiners) were checked. Foremost, in the case of adverbs signaled to be common to both varieties, false-friends were detected by the authors, by elicitation of the meaning of those expressions. To this end, we also resource to these adverbs' use in real examples drawn from corpora, when the meaning was not clear or was apparently different from the expected meaning in one of the varieties – e.g. *toda vida* 'all life' as a locative (ML) adverb in PT-BR and not as a durative time adverb (MTd); or todo (o) dia 'all day' as a durative MTd in PT-PT instead of a frequency MTf adverb in PT-BR. As seen in these examples, it is often only at the syntactic-semantic classification that such differences arise.

(2) *É só chegar no hotel e seguir reto* toda a vida 'Just get to the hotel and go straight on ahead/'til the end(lit:. all [your] life) '

This approach allowed us to extract all instances of matched adverbs, particularly those with the `+NotBR` feature from the PT-BR partition of the corpus, and conversely, all adverbs marked as `+NotPT` from the PT-PT partition. In the following section, we present and discuss our findings.

4 Results

	DSLCC corpus	
	PT-PT	PT-BR
Sentences	18,000	18,000
Words	735,503	791,872
Different words	42,190	47,914
Adv lexical entries	715	668
PT-BR entries	629 (87.9%)	620 (92.8%)
NotPT entries	74 (10.3%)	46 (6.9%)
NotBR entries	12 (1.7%)	2 (0.3%)
Adv matches	5,695	5,700
NotPT/BR matches	517	2

Table 3: DLSCC Corpus: European (PT-PT) and Brazilian Portuguese (PT-BR) partitions. Results from lexical analysis.

From applying the lexicon of multi-word adverbs to each partition of the DSLCC corpus, the following results emerged, as depicted in Table 3. Although the word count in the PT-BR partition is marginally higher (+7.66%), the number of distinct lexical entries is slightly smaller (-7.04%).

Considering the Brazilian Portuguese (PT-BR) partition, the number of lexical entries found in the corpus (668) represents 18.23% of entries of the multi-word adverbs lexicon. These can be divided into exclusively Brazilian entries (46; 6.9%), exclusively European (2; 0.3%) and entries common to both varieties (620; 92.1%).

Moving now to the European Portuguese partition, the number of lexical entries found in the corpus comprises 715 adverbial entries, comparable to the size of the PT-BR corpus. Among these entries, 629 (87.9%) are common to both Brazilian and European Portuguese, while 74 (10.3%) are exclusive to European Portuguese, and 12 (1.7%) are not found in Brazilian Portuguese.

This breakdown illustrates the substantial lexical overlap between the multi-word adverbs of the two varieties. European Portuguese contains a slightly higher proportion of unique adverbs than Brazilian Portuguese. This overlap tends to make the use of adverbs a less-than-optimal linguistic device for the DSL task. In fact, as it will be seen from the observations made below, this overlap is even greater, as some entries, marked as exclusive from one variant (+NotPT), do occur in the PT-PT partition.

	ptTenTen2020 Corpus partition	
adverb	**PT-PT**	**PT-BR**
ao domicílio	254 284 377 41*10^{-6}	4 4 478 384*10^{-6}
a domicílio	7 873 842*10^{-6}	2,597 324 195 295*10^{-6}
	n = 893 179 245	*n* = 8 010 603 604

Table 4: Distribution of the multi-word adverbs *ao domicílio/a domicílio* 'to the domicile' in combination with the verbs *entregar* and *distribuir* 'deliver' in the ptTenTen2020 corpus; number of occurrences and ratio per million words; *n* is the number of tokens per each partition of the corpus.

The search in the BR corpus for entries with the +NotBR tag resulted in only 2 cases, which are illustrative of the phenomena found. Table 4 shows the distribution of the locative adverb *ao domicílio* / *a domicílio* 'to (the) domicile' in each partition of the ptTenTen2020 corpus in combination with the most frequently co-occurring verbs *entregar* and *distribuir* 'deliver', allowing for a 0 to 5-word window, in the ptTenTen2020 corpus.

From the data in this table, the expression *a_o domicílio* 'to_the domicile' (with the article *o* 'the') is deemed as predominantly used in PT-PT. In fact, in PT-BR, the corresponding expression is *a domicílio* 'to domicile', which lacks the article *o* 'the'. The single, spurious occurrence of this adverb constitutes a case of ambiguity:

(3) *Em relação à filiação partidária e* ao domicílio *eleitoral, a comissão manteve a legislação atual.* 'Regarding party affiliation and electoral domicile, the commission maintained the current legislation.´

The second case was *de facto*, 'in fact', which is the PT-PT orthographic form, while in PT-BR the correct spelling is *de fato*. The distribution of the two spellings in the same corpus, when the string is followed by a comma (usually a non-ambiguous context of the multi-word adverb), is shown in Table 5. This single occurrence suggests a spelling error. However, its analysis reveals another level of ambiguity:

(4) [. . .] *O governo de facto,* [. . .] *rechaça a volta do líder deposto ao poder.* 'The de facto government, [. . .], rejects the ousted leader's return to power.'

In this case, the *de facto/de fato* adverb is being used here as an adjectival modifier of *governo* 'government', and its meaning 'de facto', as shown in the translation, is that of a manner-like modifier. This is a clear contrast with the modal (PAm) value 'in fact', typically associated with the adverb.

	ptTenTen2020 Corpus partition	
adverb	**PT-PT**	**PT-BR**
de facto	50,270 56 282 095 99*10^{-6}	3,644 454 897 056*10^{-6}
de fato	1,878 2 102 601 477*10^{-6}	252,109 31 471 910 54*10^{-6}
	n = 893 179 245	n = 8 010 603 604

Table 5: Distribution of the multi-word adverbs *de facto* / *de fato* 'in fact'.

Besides, the distribution of the spellings shows that the distinction between the two varieties is often not a clear-cut divide. In this particular case, the adaptation to the orthographic reform[8] may have raised some level of uncertainty among language users.

The number of instances of +NotPT adverbs found in the PT-PT partition of the DSLCC corpus is significantly higher (517). For lack of space, only a few different cases will be mentioned here, to illustrate the general phenomena found.

Some cases correspond to real distinct expressions in each variety. For example, the adverb (*pagar*) *às prestações* [PT-PT]/ *à prestação* [PT-BR]/ 'in installments' is used with the plural form in PT-PT and in the singular PT-BR. All 4 matches

[8] http://www.portaldalinguaportuguesa.org/acordo.php

of the Brazilian adverb *à prestação* are spurious, and correspond to free prepositional phrases:

(5) [...] *aplicado* à prestação *de contas* [...]; *O acesso* à prestação *exige a assinatura prévia* [...]; [...] *quanto* à prestação *dos cuidados assistenciais* [...]; [...] *estar atento* à prestação *dos jogadores* [...]
'[...] applied to the rendering of accounts [...]; Access to the service requires a prior subscription [...]; [...] regarding the provision of assistance care [...]; [...] pay attention to the performance of the players [...]'

Another problem results from many compound adverbs being very short strings, and therefore highly ambiguous with other word combinations, including other multi-word expressions. Examples of such ambiguous, +NotPT strings are *à toda* 'full speed', *às avançadas* 'to the advanced', *de primeira* 'firstly', *na maior* 'confortably', *por cima* 'above', *por detrás* 'from behind'. Finally, several expressions have been marked as +NotPT but are, in fact, common to both varieties, e.g. *de há muito* [*tempo*] 'a long [time] ago'.

After a manual inspection, it was ascertained that out of 517 matches, (i) 112 (21.6%) were true-positives, that is, the multi-word adverbs were found in the PT-PT corpus though marked as +NotPT, hence the assignment of those expressions to a single variety needs careful revision; (ii) 405 matches (78.3%) were false-positives, that is, the matched string did not correspond to the multi-word adverb in the lexicon applied to the corpus. From these, however, 110 instances (21.5%) made part of longer multi-word expressions:

(6) *O IVA "super-reduzido", dos bens* de primeira *necessidade, irá permanecer em 4%.*
'The "super-reduced" VAT on basic necessities will remain at 4%.'

In these case, the compound noun *bens de primeira necessidade* 'basic necessities' overlaps the compound adverb *de primeira* [PT-BR] 'to start with'. recognizing the longer multi-word expression would have prevented these false-positive cases.

It should also be mentioned that many instances identified in both partitions of the corpus and signaled as belonging to the common Portuguese (+PT+BR) are, in fact, also spurious (false-positive) cases, for the same reasons as explained above. That is, the system identifies a sequence of words that resembles a dictionary-listed expression, but that does not align with the intended compound adverb. This discrepancy highlights the potential for ambiguity inherent in NLP processing, and requires deeper linguistic analysis of the ambiguous strings' context to improve precision.

Both partitions of the corpus are currently being annotated to delimit the targeted multi-word adverbial forms and tag them with their POS, formal and semantic class, as well as the language variety assignment. The goal is to build a reference corpus annotated for this category, aiming at improving parsing accuracy[9].

5 Conclusion

This paper introduces a lexicon of multi-word (MW, or compound) adverbs in Portuguese, examining their lexical distribution across Brazilian (PT-BR) and European (PT-PT) varieties. From a strictly lexical perspective, the majority of the lexicon pertains to Common Portuguese (68.1%), with exclusively Brazilian compound adverbs (26.4%) outnumbering those exclusive to the European variety (5.5%). However, these preliminary figures may require revision following the experiments conducted in this study.

This lexicon was utilized to annotate the European (PT-PT) and Brazilian (PT-BR) segments of a comparable corpus sourced from the Discrimination of Similar Languages (DSL) Corpus Collection (DSLCC, v.04) (Tan et al., 2014). The count of distinct adverb entries discovered in the corpus (PT-PT: 715 / PT-BR: 668), as well as the number of matches (PT-PT: 5,695 / PT-BR: 5,700), exhibits remarkable similarity.

The proportion of lexical entries attributed to Common Portuguese is notably high and comparable across both corpus partitions (PT-PT: 629 (87.9%) / PT-BR: 620 (92.8%)), although slightly larger in PT-BR. Conversely, the count of lexical entries exclusively associated with each variety in their respective partitions is relatively small (PT-PT: 74 (10.3%) / PT-BR: 46 (6.9%)), with a slightly higher proportion observed for European Portuguese entries.

On the contrary, the number of MW adverbs labelled as *not* belonging to either variety (+NotPT and +NotBR) and found within their respective partitions is arguably negligible (PT-PT: 12 (1.7%) / PT-BR: 2 (0.3%)), albeit marginally higher in PT-PT.

[9]The corpus of annotated sentences, and the list of matched MW adverbs' can be found in the link below, under a Creative Commons license: https://string.hlt.inesc-id.pt/wiki/Compound_Adverbs

Nevertheless, the frequency of such occurrences in each partition exhibits significant asymmetry.

In Brazilian texts, only two instances of +NotBR MW adverbs were identified. One of them (e.g., *ao domicílio* / *a domicílio* 'to (the) domicile') presents a case of ambiguity, as the phrase forming the MW adverb can also exist as a free sequence of words. The other instance is a misspelled word (*facto* / *fato* 'fact'), likely resulting from some uncertainty in applying the orthographic reform of the Portuguese language. By consulting the extensive corpus of *PtTenTen2020* (Wagner Filho et al., 2018; Kilgarriff et al., 2004), it was possible to determine: (i) the asymmetric distribution of each variant form and their true association with the PT-PT or PT-BR partition; and (ii) the clear-cut distribution of orthographic variants, alongside some ambiguity due to the imperfect application of the Portuguese orthographic reform.

Regarding the +NotPT adverbs found in the Portuguese partition of the corpus, surprisingly, a considerable portion (21.6%) were confirmed as true-positive instances of adverbs inaccurately marked as exclusive to the Brazilian variety, necessitating reassignment to European Portuguese. However, the majority of remaining instances (78.3%) were false positives, stemming from the ambiguity of the strings forming the multi-word adverb with other word combinations. Among these, 21.5% were even part of another multi-word expression (such as compound nouns or verbal idioms). Hence, there remains ample room for improvement in accurately identifying multi-word adverbs, particularly concerning their potential overlap with other, longer multi-word expressions, either previously identified or concurrently present.

In conclusion, multi-word adverbs in Common Portuguese constitute a significant portion of this lexical class (68%), representing the majority of all adverb entries discovered in comparable corpora (ranging from 87.9% to 92.8%). However, their sparse distribution in the corpus renders this segment of the language lexicon sub-optimal for the task of distinguishing dialects and similar languages.

In the near future, we aim to provide the two corpus partitions annotated with the newly identified multi-word adverbs. We believe that such a resource could then be utilized to enhance other dialect-sensitive natural language processing tasks.

References

Dolors Català. 2003. *Les adverbs composés: approches contrastives en linguistique appliquée.* Ph.D. thesis, Universitat Autònoma de Barcelona, Barcelona, Spain.

Dolors Català, Jorge Baptista, and Cristina Palma. 2020. Problèmes formels concernant la traduction des adverbes composés (espagnol/portugais). *Langue(s) & Parole*, 5:67–82.

M. Di Gioia. 2001. *Avverbi idiomatici dell'italiano. Analisi lessico-grammaticale.* l'Harmattan Italia, Torino.

Gaia Fernandes. 2011. Automatic Disambiguation of *-mente* ending Adverbs in Brazilian Portuguese. Master's thesis, Universidade do Algarve and Universitat Autònoma de Barcelona, Faculdade de Ciências Humanas e Sociais, Faro, Portugal.

Gaston Gross. 1996a. *Les expressions figées en français: noms composés et autres locutions.* Editions Ophrys.

Maurice Gross. 1975. *Méthodes en syntaxe.* Hermann, Paris.

Maurice Gross. 1981. Les bases empiriques de la notion de prédicat sémantique. *Languages*, 1(63):7–52.

Maurice Gross. 1982. Une classification des phrases figées du français. *Revue québécoise de linguistique*, 11(2):151–185.

Maurice Gross. 1986a. *Grammaire transformationnelle du français: 3 - Syntaxe de l'adverbe.* ASSTRIL, Paris.

Maurice Gross. 1986b. Lexicon-grammar. The representation of compound words. In *COLING 1986 Volume 1: The 11th International Conference on Computational Linguistics*.

Maurice Gross. 1996b. Lexicon-grammar. In Keith Brown and Jim Miller, editors, *Concise Encyclopedia of Syntactic Theories*, pages 244–259. Pergamon, Cambridge.

Claude Guimier. 1996. *French adverbs: the case of en* -ment *adverbs.* Editions Ophrys.

Zellig Sabbettai Harris. 1991. *Theory of Language and Information: a Mathematical Approach.* Clarendon Press, Oxford.

Adam Kilgarriff, Miloš Jakubíček, Jan Pomikálek, Tony Berber Sardinha, and Pen Whitelock. 2014. PtTenTen: A Corpus for Portuguese Lexicography. *Working with Portuguese Corpora*, pages 111–30.

Adam Kilgarriff, Pavel Rychlý, Pavel Smrž, and David Tugwell. 2004. The Sketch Engine. *Proceedings of the 11th EURALEX International Congress*, pages 105–116.

Éric Laporte, Takuya Nakamura, and Stavroula Voyatzi. 2008. A French corpus annotated for multiword expressions with adverbial function. In *Language Resources and Evaluation Conference (LREC). Linguistic Annotation Workshop*, pages 48–51.

Christian Molinier and Françoise Levrier. 2000. *Grammaire des adverbes: description des formes en* -ment. Droz, Genève.

Izabela Müller, Nuno Mamede, and Jorge Baptista. 2022. Bootstrapping a Lexicon of Multiword Adverbs for Brazilian Portuguese. In *International Conference on Computational and Corpus-Based Phraseology*, pages 160–174. Springer.

Izabela Müller, Nuno Mamede, and Jorge Baptista. 2023. Advérbios Compostos do Português do Brasil. *Revista da Associação Portuguesa de Linguística*, 1(10):230–250.

Cristina Palma. 2009. Estudo contrastivo português-espanhol de expressões fixas adverbiais. Master's thesis, Universidade do Algarve, Faculdade de Ciências Humanas e Sociais, Faro, Portugal.

Sébastien Paumier, Wolfgang Flury, Franz Guenthner, Eric Laporte, Friederike Malchok, Clemens Marschner, Claude Martineau, Cristian Martínez, Denis Maurel, Sebastian Nagel, et al. 2021. *UNITEX 3.3 User Manual*. Université de Paris Est /Institut Gaspard Monge.

Paulo A. Rocha and Diana Santos. 2000. CETEMPúblico: Um corpus de grandes dimensões de linguagem jornalística portuguesa. In *V Encontro para o processamento computacional da língua portuguesa escrita e falada (PROPOR 2000)(Atibaia SP 19-22 de Novembro de 2000) São Paulo: ICMC/USP*. ICMC/USP.

Kosho Shudo, Akira Kurahone, and Toshifumi Tanabe. 2011. A comprehensive dictionary of multiword expressions. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 161–170.

Liling Tan, Marcos Zampieri, Nikola Ljubešic, and Jörg Tiedemann. 2014. Merging comparable data sources for the discrimination of similar languages: The dsl corpus collection. In *Proceedings of the 7th Workshop on Building and Using Comparable Corpora (BUCC)*, pages 11–15, Reykjavik, Iceland.

Jorge A Wagner Filho, Rodrigo Wilkens, Marco Idiart, and Aline Villavicencio. 2018. The brWaC corpus: a new open resource for Brazilian Portuguese. In *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018).*, pages 4339–4344.

Marcos Zampieri, Shervin Malmasi, Nikola Ljubešić, Preslav Nakov, Ahmed Ali, Jörg Tiedemann, Yves Scherrer, and Noëmi Aepli. 2017. Findings of the vardial evaluation campaign 2017. In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, pages 1–15, Valencia, Spain. Association for Computational Linguistics.

NoMusic – The Norwegian Multi-Dialectal Slot and Intent Detection Corpus

Petter Mæhlum and **Yves Scherrer**
Language Technology Group
Department of Informatics
University of Oslo, Norway
pettemae@ifi.uio.no, yves.scherrer@ifi.uio.no

Abstract

This paper presents a new textual resource for Norwegian and its dialects. The NoMusic corpus contains Norwegian translations of the xSID dataset, an evaluation dataset for spoken language understanding (slot and intent detection). The translations cover Norwegian Bokmål, as well as eight dialects from three of the four major Norwegian dialect areas. To our knowledge, this is the first multi-parallel resource for written Norwegian dialects, and the first evaluation dataset for slot and intent detection focusing on non-standard Norwegian varieties. In this paper, we describe the annotation process and provide some analyses on the types of linguistic variation that can be found in the dataset.

1 Introduction

Over the last decades, various textual resources covering Norwegian dialects have been produced. This paper reports on the creation of yet another Norwegian dialect dataset which has some unique properties that set it apart from previous work.

As a starting point, we use the xSID corpus (van der Goot et al., 2021), which consists of natural prompts asked to digital assistants (e.g., *Is it going to rain today?*, *Change tomorrow morning's alarm to 6 am.*). A digital assistant will have to (a) recognize the *intent* of the prompt and (b) detect and classify the main *arguments*, also called *slots*, of the prompt. Solving these two tasks is commonly referred to as *spoken language understanding* (SLU) or *slot and intent detection* (SID).

The xSID corpus is already available in several low-resource and non-standard varieties (Aepli et al., 2023; Winkler et al., 2024) and consists of a text genre for which dialectal productions are natural. We have translated the English sentences of xSID into standard Norwegian Bokmål and into the dialects of eight native speakers of Norwegian who regularly write in these dialects. The slot and intent annotations were then semi-automatically transferred to the Norwegian translations.

The resulting dataset, which we call NoMusic (*NOrwegian MUlti-dialectal Slot and Intent detection Corpus*), has the following particularities compared to existing Norwegian dialect resources:

- It is a multi-parallel corpus, i.e., all translations have the same number of sentences with the same meanings.
- It is a natively written resource and does not consist of transcribed speech.
- It is openly available, as all the translations are created on purpose within the project.[1]

The corpus can be used for various purposes, both in dialectology and natural language processing, e.g.:

- to evaluate the robustness and cross-lingual and cross-lectal transfer capabilities of SLU systems, thanks to the slot and intent labels,
- to identify dialect-specific expressions,
- to investigate digital writing practices,
- to enable machine translation between different varieties of Norwegian.

In the following sections, we describe the data and the annotation process, and provide analyses of the observed linguistic variation.

2 Related Work

2.1 Dialect Corpora for Norwegian

The Norwegian language has two officially established written standards: Norwegian Bokmål and Norwegian Nynorsk. Bokmål is the more utilized of the two in terms of speakers, and is historically based on written Danish.

[1]The NoMusic dataset is integrated into the xSID repository https://github.com/mainlp/xsid, but it is also available on https://github.com/ltgoslo/NoMusic.

Proceedings of the Eleventh Workshop on NLP for Similar Languages, Varieties, and Dialects (VarDial 2024), pages 107–116

While there are cases of earlier dialectal writing, general acceptance of dialects in increasingly formal settings began in the 1970s (Bull et al., 2018, 235-238). Dialects are thus less stigmatized, even in writing, for example in social media.

Norwegian dialects have been researched both from dialectological and computational angles, and several textual resources have been created in recent years. Traditional dialectological corpora such as the Nordic Dialect Corpus (NDC, Johannessen et al., 2009)[2] or the LIA Norwegian Corpus (Hagen and Vangsnes, 2023)[3] typically consist of transcriptions of interviews conducted with a large number of informants. This setup does not lead to directly comparable texts because the different interviews will be of different lengths, cover different topics and contain different linguistic structures. Also, these transcriptions are typically made by trained annotators according to relatively strict guidelines; the resulting written representations are often quite different from "real-world" dialect writing, as they are meant to faithfully represent the spoken language, rather than the way users would write their own dialect in everyday communication. For example, NDC contains Bokmål glosses and phonemic spellings, but these do not necessarily match how the users of a particular dialect spell.

On the other hand, recent data collection efforts such as NorDial (Barnes et al., 2021, 2023) focused on identifying and annotating written dialect posts in social media. This does not address the problem of comparability, but even introduces other challenges: it is difficult to obtain a dense coverage of the different dialects used in Norway, and the resulting dataset may not be made publicly available due to copyright restrictions. It remains to be seen to what extent projects such as the Nordic Tweet Stream (NTS, Laitinen et al., 2018) provide a viable workaround to copyright and licensing questions.

2.2 Multi-Dialectal Corpora

A relatively common alternative strategy to create multi-dialectal corpora consists in asking dialect speakers to translate texts into their variety, either from the standard variety or from a third language like English.

The MADAR Corpus of Arabic Dialects (Bouamor et al., 2018) illustrates this approach: the authors use a fixed set of English sentences and have them translated by native Arabic dialect speakers into their variety. They use the Basic Travel Expressions Corpus (BTEC, Takezawa et al., 2007) as a starting point and obtain translations of 2000 sentences into 25 Arabic dialects.

The SwissDial corpus (Dogan-Schönberger et al., 2021) follows a similar strategy, resulting in 2500 sentences in 8 Swiss German dialects. The corpus contains both audio recordings and transcripts, making it suitable for speech processing applications. Moreover, the data is annotated on sentence level with topic and code-switching information.

In a related effort, the xSID corpus[4] (van der Goot et al., 2021; Aepli et al., 2023; Winkler et al., 2024) has been created to support the development of multilingual dialog systems. It consists of prompts to digital assistants and is annotated with intents and slots. The 800 prompts in xSID are originally in English and have been translated to 12 major languages and 4 low-resource varieties or dialects (as of version 0.5, with the latter being Bavarian German, South Tyrolean German, Swiss German, and Neapolitan). In contrast to the BTEC corpus used for MADAR, the xSID source data is freely available and provides additional sentence-level (intents) and chunk-level (slots) annotations for the SID task.

The DIALECT-COPA shared task held at VarDial 2024 (Chifu et al., 2024)[5] is based on a similar approach: it contains translations of the English causal commonsense reasoning corpus COPA (Roemmele et al., 2011; Ponti et al., 2020) into various South Slavic languages and dialects.

Most of the resources cited above are created by translation from (American) English. This can be problematic because the translators may not be sufficiently familiar with the North-American cultural references (music styles, holiday destinations, etc.) and/or linguistic expressions (e.g. date and time formats, imperial measurements) used in the original data. Furthermore, non-professional translators are prone to producing translationese, which can be perceived as unnatural and not representative of spontaneous dialect writing. We are aware of these limitations, but nevertheless find it the most practical and effective approach to create multi-dialectal annotated resources.

[2]https://tekstlab.uio.no/scandiasyn/

[3]https://tekstlab.uio.no/LIA/norsk/

[4]https://github.com/mainlp/xsid

[5]https://sites.google.com/view/vardial-2024/shared-tasks/dialect-copa

English	Set a reminder to go to the grocery store later
Danish	Sæt en påmindelse om at gå i supermarkedet senere
Bokmål	Sett på en påminnelse om å gå i butikken etterpå
A1	Minn mæ på at æ skal dra på butikken seinere.
A2	Sett enn påminnelse om å fære tel butikken seinar.
A3	Sett en alarm for å da te matbutikken seinere
A4	Sett en påminnelse om å gå te matbutikken seinar
A5	Sett en påminnelse for å gå t butikken seinar
A6	Sett en påminnelse om å stikke på butikken seinere.
A7	Sett på en påminnelse om å gå t butikken seinare
A8	Lag ein påminnelse om å gå på butikken seinere

Table 1: Examples of translations. The Danish translation is already part of xSID. The Norwegian dialect annotators are numbered *A1* to *A8* from North to South.

2.3 Spoken Language Understanding Datasets

The xSID corpus represents one of the few efforts to provide non-English datasets for the SLU/SID task. However, it only provides manually created validation and test sets. Training sets for non-English languages are available, but created automatically by machine translation. The only currently available SLU dataset that covers Norwegian is MASSIVE (FitzGerald et al., 2022). It provides training, validation and test sets for 51 languages, among which standard Norwegian Bokmål. The slot and intent label sets differ between xSID and MASSIVE, and we leave it to future work to investigate to what extent the two annotation standards can be harmonized meaningfully.

The NoMusic corpus is, to our knowledge, the first SID dataset that provides multiple alternative formulations of the same queries.[6] The alternatives show dialectal variation, but also different lexical and syntactic choices (see Section 4). This variety opens up new avenues for making both the training and the evaluation of SLU systems more robust.

3 Data and Annotation

The xSID corpus provides a development set of 300 sentences and a test set of 500 sentences. The NoMusic dataset consists of annotated translations of these sentences. It is produced in three phases:

1. Translate the English xSID sentences to standard Norwegian Bokmål and to the Norwegian dialects.
2. Annotate the Bokmål sentences with slots, using the English sentences as guides.
3. Annotate the dialectal sentences with slots, using the Bokmål sentences as guides.

The following sections describe these phases in detail.

3.1 Translation

We used the English xSID dataset as a starting point and produced translations to standard Norwegian Bokmål and to eight Norwegian dialects.[7] The dialect translations were made by university students who declared that they regularly write in their dialect.

The Bokmål translation was produced by one of the authors of the paper. While some dialects speakers normally use Nynorsk, the other written Norwegian norm, the choice of Bokmål is purely practical, and it is used as a means for more easily transferring the slot and intent labels, as well as functioning as a meta-language to which to compare the dialectal forms.

The translations were produced by editing .tsv files in a shared GitHub repository. The annotators had access to GitHub issues where they could discuss potential problems. An example sentence with all available translations is shown in Table 1.

3.2 Translation Guidelines

The translators were given simple instructions on how to translate, but were otherwise not controlled.

[6]The ITALIC dataset (Koudounas et al., 2023) provides audio files and transcripts of SLU prompts in various regional varieties of Italian, but it is only annotated with intents, not slots.

[7]Two additional dialect translations are in progress at the time of writing and will be added to the dataset when completed.

These guidelines mostly followed the ones from the xSID project, but deviated in some respects discussed here.

Time The xSID guidelines note that some languages that do not have pm/am equivalents might need to translate cases such as *7 pm* to *7 in the evening*. Our annotators were not given specific notes on these translations, but were generally asked to translate into natural written dialect. This has resulted in some variation. The 24-hour clock is widely used in Norway, but in the spoken language, the 12-hour clock is also used if the times are unambiguous. We see that three different strategies have been used by our annotators in these cases: 1) adding a temporal adverb (*om morran* 'in the morning', *ettermiddag* 'afternoon'), 2) leaving the time ambiguous, which often means directly translating the English time without pm or am. or 3) converting the time to the 24-hour clock (*4 pm* → *klokka 16* '16 o'clock'). At least 6 of the annotators convert to the 24 hour clock to some degree. There are also instances of confusion between am and pm in the translations, for example in one case 5pm was interpreted as 05:00 by one annotator.

Named Entities In the xSID guidelines it is noted that named entities are not to be translated, except for place names. While this has been the general tendency in our dataset, annotators were asked to translate the names of movies when an established Norwegian title exists, but otherwise not. There is also some confusion for certain named entities that contain translatable content, such as whether the *Theatres* part of *Cobb Theatres* should be translated or not. Some annotators have translated certain titles even in cases where there is no established Norwegian name.

Grammatical Mistakes Grammatical mistakes should be kept in the translations if possible, according to the xSID guidelines. We believe that this would have been difficult, as it is not obvious to decide how a certain mistake might map from one language to another. Our annotators were not specifically asked to keep mistakes from the English sentences. However, as discussed below, the informal nature of the writing has led to some spelling mistakes that are not reflections of the original English. It is difficult to distinguish between cases when deviations from normative writing are conscious representations of dialect, and when they are simply unintentional.

Capitalization and Punctuation Annotators were not asked to correct capitalization or punctuation, but were also not explicitly asked to ignore it; rather, they were asked to follow their usual dialect writing habits. As a result, we see different tendencies among annotators. Some diligently add it where needed, while some allow for variation in their translations. Table 1 is an example of this variation.

Abbreviations While there are generally few abbreviations, there are some spelling conventions that in the written language are similar to abbreviations, but that would not be detectable in the spoken language. The xSID guidelines discourage abbreviations that are not 'common in fluent discourse.' We see examples of abbreviations such as *min* 'minute', which might also be read in its abbreviatiated form, and we also commonly note the usage of shortened spelling conventions like writing *d* for Bokmål *det* 'it', or *t* for Bokmål *til* 'to', similar to the usage of *u* for *you* in English.

Avoiding Direct Translations The xSID guidelines point out that it is not necessary to directly translate certain things, exemplified by the ditransitive usage of *play*. We believe that this has been covered by asking the annotators to translate into natural-sounding dialectal Norwegian. Another example is the translation of the English polite marker *please*, which has been translated into a variety of ways in the data.

Possessive Determiners The xSID guidelines note that possessive determiners should be preserved and translated whenever possible, but the annotators were not explicitly asked to do this. Norwegian generally uses fewer possessive determiners than English. For example, four dialect and the Bokmål translations use a variation of *where I am now* or *here* to translate 'my current position': *her e e, her, der ej e no, her eg e nå, der jeg er nå*, perhaps due to a direct translation sounding a bit stilted.

3.3 Translator Demographics

Figure 1 shows the origin of the dialect translators (marked with *A1* to *A8*) in relation with the four major Norwegian dialect areas. It can be seen that three of the four main dialectal areas are represented in NoMusic, but that we lack translations from dialects representing Eastern Norwegian. This absence can be explained by there being less

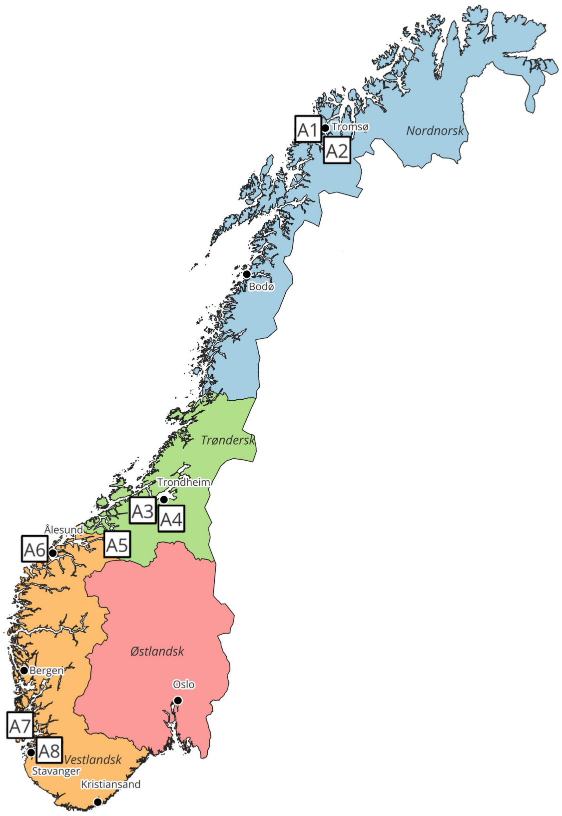


Figure 1: Map of Norway, with the four major dialect areas and the origins of the eight dialect annotators (*A1* to *A8*).

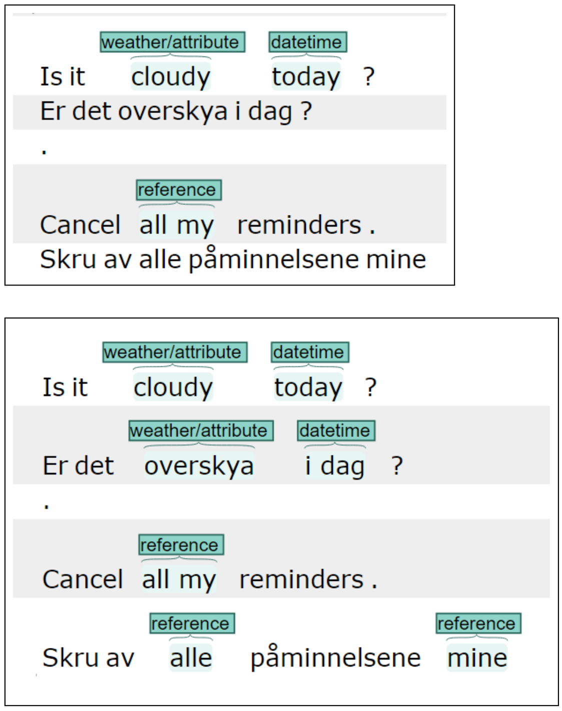


Figure 2: INCEpTION annotation interface showing the English-to-Norwegian annotation transfer. The upper part shows the initial state with pre-annotated English and unannotated Norwegian, the lower part shows the completed Norwegian annotations. Note the different number of labels.

perceived difference between the spoken language and the written language in Eastern Norway, as Bokmål is often associated with *Standard Østnorsk* 'Standard Eastern Norwegian', a commonly taught spoken variety.[8] Slåen (2022) describes written dialectal usage in the Northern reaches of the Eastern Norwegian dialectal area, but the tendency may be lower in and around Oslo.

As can be seen on the map, 2 translators speak Northern dialects, 3 central (Trøndersk), and 3 Western dialects. We had 6 female and 2 male translators. 6 translators were in the age range 20-24 and 2 in the range 25-29; all of them were university students on Bachelor's or Master's level.

3.4 Slot and Intent Annotations

Once the sentences are translated, they need to be labeled with slots and intents. Each sentence has a single intent, and the intent is not supposed to change across languages. Therefore, we automatically transfer the intent labels from English.

The slot labels are annotated manually in two steps, using the same procedure as for the original xSID corpus. In the first step we annotate the Bokmål version, using the annotated English sentence as a guide for each sentence. In the second step, the dialectal versions are annotated, using the already annotated Bokmål version as a guide.

We use the INCEpTION (Klie et al., 2018) platform for transferring the slot annotations. For the English-to-Bokmål step, we interleave annotated English sentences with their unannotated Bokmål translations. The annotation process is illustrated in Figure 2.[9] We note how the Norwegian syntax can lead to differences in the number of slot labels. In this case, the xSID guidelines state that consecutive reference labels specifically should be annotated as a single chunk, but as there are no discontinuous spans in the English data, we annotate them as two

[8] https://www.sprakradet.no/svardatabase/sporsmal-og-svar/oslodialekten/

[9] In order to upload the pre-annotated English sentences along with Bokmål, we merged the two and uploaded the resulting .txt file using the *plain text (one sentence per line)* setting. We then downloaded the UIMA CAS XMI file, which is INCEpTION's native format. Using the dkpro-cassis library (https://pypi.org/project/dkpro-cassis-tools/#description), we then added the English slot spans from the existing .conll files, and uploaded the resulting .XMI file. Annotations were added in a single token level layer.

Figure 3: Annotation of the dialect translations. Note how differences in spelling of *i dag* 'today' causes slight differences in labeling.

separate labels.

A similar process is used for the Bokmål-to-dialect annotation transfer step: the annotated Bokmål sentence is presented on top as a guide, with all dialectal translations following. See Figure 3 for an example.

4 Analysis

The dialectal translations differ in various respects from each other and from the standard version. In this section, we discuss different types of variation and their prevalence in the dataset, before briefly looking at how some of these features present themselves in the Nordic Dialect Corpus (NDC).

4.1 Variation in Translation

Unsurprisingly, the translations are largely similar in terms of word lengths and type-token ration, as reported in Table 2. We see that some annotators (A2, A6) have slightly longer sentences. The most striking difference is perhaps the lower number of types in English, but this could easily be attributed to the slightly higher morphological variation in Norwegian.

Annotator	Tokens	Types	Sent. length
A1	6200	1337	7.74
A2	6526	1360	8.15
A3	6282	1365	7.84
A4	6054	1346	7.56
A5	5955	1310	7.43
A6	6546	1350	8.17
A7	6004	1379	7.5
A8	6086	1366	7.6
Bokmål	6310	1392	7.88
English	6177	1245	7.71

Table 2: Tokens, types and average sentence lengths for the annotators, the Bokmål translations, and the original English.

4.2 Linguistic Variation

While there are many clear dialectal differences between the translators, that is not to say that all these differences are due to dialectal variation. For many sentences there are several possible translations, and there are also lexical or syntactic choices that do not necessarily have to be dialect-specific. For example, in Table 1, the verb 'to go' is expressed by *å gå, å dra, å fære* or *å stikke*, and 'grocery store' is translated by *butikken* or *matbutikken*. Before looking at dialectal features in the dataset, we discuss some more general features.

Spelling Annotators were asked to translate to their own dialect in a natural way. This has led to varying degrees of written expressions. In dialectal writing, the written forms naturally deviate from the established written norms, namely Bokmål or Nynorsk, but we would typically not expect deviations that cannot be explained by the dialectal features of the writer. We do see what we consider non-dialectal spelling deviations, or what would be spelling mistakes in a prescriptive setting. The frequency of these vary from annotator to annotator. In practice, this means that the corpus has some features of user-generated language that are not unique to dialectal writing.

Pronunciation Spelling One crucial difference between spoken dialect and written dialect is that not all words show indications of being associated with a dialect, and many words are left in their Bokmål or Nynorsk spelling, despite being pronounced differently from how most speakers would pronounce the normed spellings. In the NorDial

	A1	A2	A3	A4	A5	A6	A7	A8	NB
A1	568								
A2	239	612							
A3	319	235	573						
A4	317	224	313	609					
A5	310	233	294	334	589				
A6	312	228	273	276	265	570			
A7	280	194	258	291	281	277	582		
A8	255	192	223	264	246	270	315	632	
NB	313	217	271	292	316	301	304	287	675

Table 3: Lexical overlap between the dialect translations and the Bokmål (NB) translation. Words contained in the English dataset (mostly titles and names) are removed from the comparisons.

	Pearson's r	p-value
A1	−0.5329	0.1738
A2	−0.6885	0.0590
A3	−0.5516	0.1563
A4	−0.5792	0.1324
A5	−0.4962	0.2111
A6	−0.4454	0.2688
A7	−0.6159	0.1040
A8	−0.6069	0.1106

Table 4: Pearson correlation coefficients between lexical overlap and geographical distance.

corpus, the authors find that some sentences only contain a few words indicating dialect, although in spoken language all words would (Barnes et al., 2021). While most function words are written according to the pronunciation of a given dialect, many content words are not, despite obviously not following pronunciation rules. However, this varies from annotator to annotator in our dataset. An example is the word 'restaurant', whose spelling is kept in some cases by at least 6 annotators, while some use the spelling *resturant* or *resturang*. In the NDC, the spellings are *r[e/æ]s(s)t[u/o]ran(n)g(g)*.

Avoidance of Direct Translations As mentioned earlier, another source of variation is avoidance of direct translations, which can lead to syntactic and lexical variation. For example, when talking about weather predictions, it is quite common to use the auxiliary verb *skulle*, which indicates a planned action or a prediction, but it is in some cases also natural to use a more neutral feature with the composite auxiliary *komme til* 'will'. Both options are available in several dialects, and even the same user might alternative between these.

4.3 Lexical Overlap and Dialectal Features

We now examine the translations in terms of dialectal features and lexical overlap. Table 3 presents an overview of the lexical overlap between the translations. The diagonal shows the total number of types, reported in Table 2. We would expect annotators who come from dialectal areas in close proximity to exhibit higher overlap.

Table 4 shows the Pearson correlation coefficients between the lexical overlap (Table 3) and the geographical distances between the translators' origins. Correlations are computed for each annotator separately. The correlation coefficients indicate moderate to strong correlations,[10] but the p-values are too high to draw meaningful conclusions.

The clearest dialectal differences are observed in morphology. We will have a brief look at verbal, nominal and adjectival morphology, while acknowledging that this is only part of what constitutes dialectal variation in the dataset. Where attestations can be found, we look up corresponding forms in the NDC interface to inspect their distributions. Queries are done in Bokmål, and the reported phonological forms are compared to our dialectal writing.

Verbal Morphology One thing to observe in terms of verbal morphology is the infinitive. This is an oft-used dialectal feature, based on whether the dialect has infinitives (for consonant stem verbs) in -a, -e, -Ø (apocope) or a mix of these. For our annotators, we observe 5 patterns: infinitives ending in -e only (A1, A2 and A6), in -a only (A8), no ending (-Ø) (A5, A4), mixed -e and no ending (A3) and mixed -a and no ending (A7). Notably for A7 it seems like the apocope is only found in the verb *å vær*, but it is both consistent and frequent. According to the presentation of infinitives in Mæhlum and Røyneland (2023, p. 180), A1, A2 and A6 are all from typical e-infinitive areas, and A8 is from a typical a-infinitive area. A4 and A5 are theoretically both further south than the area typically associated with pure apocope. A3 is in the area for mixed infinitive, but A7's position in the South-West does not explain the form *å vær*. However, in the NDC, *vær* as an infinitive form is not infrequently observed in South and West Norway.

Nominal Morphology While there are not enough nouns to create a full overview of the writ-

[10]The correlations are negative because lexical similarity is compared with geographic distance.

ers' morphological systems, there is enough to give us indications. First of all, we get an impression of the gender system. Normally, both written Norwegian norms, Bokmål and Nynorsk, allow for a three-gender nominal inflection system, but to varying degrees. A three-gender system is obligatory in Nynorsk, while in Bokmål it is possible to conflate the masculine and feminine classes to a common gender (nor. *felleskjønn*). We see that all dialects mark feminine nouns to some extent, as all dialects use the feminine-specific singular definite marker -a (or -å) at least in some words (*boka 'the book', låta 'the tune', bogå 'the book'*), but not in all (*vermeldingen, vermeldinga* 'the weather forecast'). The use of the indefinite singular article *ei* 'a, an' is less frequent, as is also the case in Bokmål. While the masculine singular is invariably the same as in Bokmål, another difference between feminine and masculine nouns appears in the plural. Where some writers in our corpus have the same forms as in Bokmål for both genders (*filmer, stjerner* 'movies, stars'), we see that some writers have variant forms, which are still the same (*filma, stjerna*), while some make a distinction (*filma, stjerne; filmar, stjerner*). Some dialects have apocope in the plural definite (*filman*). Some annotators have the same forms for masculine and neuter nouns (*filma, minutta*), while others have the typical zero-ending that we also see in Nynorsk (*filma, minutt*).

Adjectival Morphology One notable feature for adjectives, is whether the neuter suffix -t is added to adjectives in -ig. This is done by the translator from Stavanger (A8), as in *tidligt* 'early'. This is confirmed to be a regional feature by the NDC, where corresponding forms are found in the area around Stavanger but not elsewhere in the country. We also see variation in the comparative forms, where three forms are found: *-ere (kaldere)* 'colder', *-are (kaldare)* and an apocopized version, *-ar (kaldar, kjørligar)* . While there are not many attestations of *kaldere*, we see that all attestations with the *-ere* ending are in Eastern Norway.

Lexicon While many words show clear dialectal influence, there are few cases where the annotators' lexical choices are markedly different from the standard language. One such example is the use of *bli å* lit. 'become to' as a future auxiliary.

Function Words Much of the variation seen between the translated material is in terms of function words: prepositions, pronouns, and determiners.

	I	ME	HOW	SOME	TO
A1	æ	mæ	kordan	nokka	til
A2	æ	mæ	kordan	nåkka	tel
A3	æ	mæ	koss	nokka	te
A4	æ	mæ	koss	nåkka	te
A5	e	me	koss	nokka	til
A6	ej	mej	kordan	nokke	til
A7	eg	meg	koss	noe	t(e)
A8	eg	meg	kordan	någe	te

Table 5: Selected pronouns and function words used by the different annotators.

In Table 5, we see five selected words that illustrate some of the variation between the translators. Looking at the pronominal variation, we get an idea of how distinctive some of these features are. The form *ej* (A6) 'I', is associated with an area between Ålesund and Bergen in the NDC, indicating that this is a quite distinguishing feature of A6's dialect. Otherwise it is only attested once close to Mo i Rana. Among the other words for *I*, both *æ* (A1-A4) and *e* (A5), are quite widespread in spoken Norwegian as reported in the NDC. The form *eg* (A7, A8) is more associated with the West, and is not found along the border to Sweden in the East. For the oblique forms, *mæ* (A1-A4) is quite widespread, except in the West and upper central areas, and *mej* (A6) is only registered in two locations: one on the Trøndersk/Vestnorsk border, and one in the Trøndersk area. The interrogative *Kordan* 'how' is mostly associated with Western and Northern Norwegian, while *koss*, also 'how', is associated with Southern, Central and upper Western Norway. The determiner *någe* 'some' is heavily associated with the Stavanger area, and is not found outside it except one attestation in Tromsø. *Noe* 'id.' is quite widespread, but not in the upper West. *Nokka* is associated with Trøndersk and Northern Norwegian, and a small cluster in the south in the NDC, while finally *nokke* is associated with the west.

4.4 Phonological Features

As the translators all report that they use dialectal writing in their daily lives, we see the translations as representative of at least some part of the written dialect of the area the translator represents, but this does not tell us to what degree the written language represents the spoken dialect of that area. However, some of these features can be inspected using NDC.

For example, a commonly used dialectal feature is the voicing of the ungeminated plosives /p/, /t/ and /k/ to /b/, /d/ and /g/. We see examples of this in our dataset in forms such as *søga* (Bokmål *søke*) 'search' and *bogå* (Bokmål *boka*) 'the book'. In NDC, forms of *søke* with voicing are only found in an area surrounding Kristiansand, while voiced forms of *boka* are found between Haugesund and Kristiansand.

5 Conclusion

We present a dataset of written dialectal Norwegian, which reflects various dialectal phenomena and is also annotated with slots and intents. The utterances are translations of the English validation and test sets of the xSID corpus (van der Goot et al., 2021).

Limitations

As discussed in Section 3.3, the geographical coverage of the translators is uneven, with Eastern dialects not represented at all in the corpus. This is due to linguistic factors, as discussed, and also to contingent factors related to the sample of qualified and interested students available during the project duration. We will consider extending the corpus if annotators from not yet covered areas become available.

Furthermore, as discussed in Section 2.2, the annotation workers are not professional translators and may find it difficult to produce natural and correct dialect writing in a translation setup. Moreover, certain cultural references and named entities may not be known well enough by our translators.

Finally, slot and intent detection models are typically applied to speech data in conjunction with an automatic speech recognition system. It could thus be useful to pair the dialectal transcripts with recorded speech. We currently do not offer speech recordings because our main goal was to create a resource for *written* dialectal Norwegian, but we may consider extending the dataset towards speech in the future.

Ethical Considerations

The translators were hired as student assistants and paid for the effective hours spent on the translation task (typically between 15 and 20 hours, not including slot annotation), according to the official salary schemes in use at the University of Oslo. The participation in the translation task was voluntary, and all translators agreed in writing that their productions may be publicly shared under the CC-BY-SA 4.0 licence.[11]

The English data used as source material is curated and does not contain any harmful content, to our knowledge.

Acknowledgements

The NoMusic project was granted funding from the TekstHub initiative at the University of Oslo.

We thank the many annotators who have contributed to this project: Andželika Andruškaite, Thomas Heim, Sanne Lima, Ada-Marie Sørensen Sneve, Elias Lynum Ringkjøb, Tonje Sandanger, Lilja Charlotte Storset, Ulrikke Strømsvold Tveit and Snorre Åldstedt.

We also thank Anders Næss Evensen for help with the system for uploading pre-annotated files to INCEpTION.

References

Noëmi Aepli, Çağrı Çöltekin, Rob Van Der Goot, Tommi Jauhiainen, Mourhaf Kazzaz, Nikola Ljubešić, Kai North, Barbara Plank, Yves Scherrer, and Marcos Zampieri. 2023. Findings of the VarDial evaluation campaign 2023. In *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*, pages 251–261, Dubrovnik, Croatia. Association for Computational Linguistics.

Jeremy Barnes, Petter Mæhlum, and Samia Touileb. 2021. NorDial: A preliminary corpus of written Norwegian dialect use. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 445–451, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden.

Jeremy Barnes, Samia Touileb, Petter Mæhlum, and Pierre Lison. 2023. Identifying token-level dialectal features in social media. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 146–158, Tórshavn, Faroe Islands. University of Tartu Library.

Houda Bouamor, Nizar Habash, Mohammad Salameh, Wajdi Zaghouani, Owen Rambow, Dana Abdulrahim, Ossama Obeid, Salam Khalifa, Fadhl Eryani, Alexander Erdmann, and Kemal Oflazer. 2018. The MADAR Arabic dialect corpus and lexicon. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Tove Bull, Espen Karlsen, Eli Raanes, and Rolf Theil. 2018. *Norsk språkhistorie*, volume 3. Novus, Oslo.

[11] https://creativecommons.org/licenses/by-sa/4.0/

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.

Pelin Dogan-Schönberger, Julian Mäder, and Thomas Hofmann. 2021. SwissDial: Parallel multidialectal corpus of spoken Swiss German. *CoRR*, abs/2103.11401.

Jack FitzGerald, Christopher Hench, Charith Peris, Scott Mackie, Kay Rottmann, Ana Sanchez, Aaron Nash, Liam Urbach, Vishesh Kakarala, Richa Singh, Swetha Ranganath, Laurie Crist, Misha Britan, Wouter Leeuwis, Gokhan Tur, and Prem Natarajan. 2022. Massive: A 1m-example multilingual natural language understanding dataset with 51 typologically-diverse languages.

Kristin Hagen and Øystein A. Vangsnes. 2023. LIA-korpusa – eldre talemålsopptak for norsk og samisk gjort tilgjengelege. *Nordlyd*, 2(47):119–130.

Janne Bondi Johannessen, Joel Priestley, Kristin Hagen, Tor Anders Åfarli, and Øystein Alexander Vangsnes. 2009. The Nordic Dialect Corpus - an advanced research tool. In *Proceedings of the 17th Nordic Conference of Computational Linguistics NODALIDA 2009*, volume 4 of *NEALT Proceedings Series*.

Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych. 2018. The inception platform: Machine-assisted and knowledge-oriented interactive annotation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 5–9. Association for Computational Linguistics. Event Title: The 27th International Conference on Computational Linguistics (COLING 2018).

Alkis Koudounas, Moreno La Quatra, Lorenzo Vaiani, Luca Colomba, Giuseppe Attanasio, Eliana Pastor, Luca Cagliero, and Elena Baralis. 2023. Italic: An italian intent classification dataset.

Mikko Laitinen, Jonas Lundberg, Magnus Levin, and Rafael Messias Martins. 2018. The Nordic Tweet Stream: A dynamic real-time monitor corpus of big and rich language data. In *Proceedings of the Digital Humanities in the Nordic Countries 3rd Conference*, pages 349–362.

Brit Mæhlum and Unn Røyneland. 2023. *Det norske dialektlandskapet: innføring i studiet av dialekter*. Cappelen Damm akademisk, Oslo.

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, Online. Association for Computational Linguistics.

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

Mari Slåen. 2022. "jeg synes det er mest vanlig å skrive slik jeg prater" en sosiolingvistisk studie av elevers dialektale skriving og holdninger i gudbrandsdalen. Master's thesis, Norwegian University of Science and Technology.

Toshiyuki Takezawa, Genichiro Kikui, Masahide Mizushima, and Eiichiro Sumita. 2007. Multilingual spoken language corpus development for communication research. In *International Journal of Computational Linguistics & Chinese Language Processing, Volume 12, Number 3, September 2007: Special Issue on Invited Papers from ISCSLP 2006*, pages 303–324.

Rob van der Goot, Ibrahim Sharaf, Aizhan Imankulova, Ahmet Üstün, Marija Stepanović, Alan Ramponi, Siti Oryza Khairunnisa, Mamoru Komachi, and Barbara Plank. 2021. From masked language modeling to translation: Non-English auxiliary tasks improve zero-shot spoken language understanding. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2479–2497, Online. Association for Computational Linguistics.

Miriam Winkler, Virginija Juozapaityte, Rob van der Goot, and Barbara Plank. 2024. Slot and intent detection resources for Bavarian and Lithuanian: Assessing translations vs natural queries to digital assistants. In *Proceedings of The 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*. Association for Computational Linguistics.

Understanding Position Bias Effects on Fairness in Social Multi-Document Summarization

Olubusayo Olabisi and **Ameeta Agrawal**
Portland State University
{oolabisi,ameeta}@pdx.edu

Abstract

Text summarization models have typically focused on optimizing aspects of quality such as fluency, relevance, and coherence, particularly in the context of news articles. However, summarization models are increasingly being used to summarize diverse sources of text, such as social media data, that encompass a wide demographic user base. It is thus crucial to assess not only the quality of the generated summaries, but also the extent to which they can fairly represent the opinions of diverse social groups. Position bias, a long-known issue in news summarization, has received limited attention in the context of social multi-document summarization. We deeply investigate this phenomenon by analyzing the effect of group ordering in input documents when summarizing tweets from three distinct linguistic communities: *African-American* English, *Hispanic-aligned* Language, and *White-aligned* Language. Our empirical analysis shows that although the textual quality of the summaries remains consistent regardless of the input document order, in terms of fairness, the results vary significantly depending on how the dialect groups are presented in the input data. Our results suggest that position bias manifests differently in social multi-document summarization, severely impacting the fairness of summarization models.

1 Introduction

As the use of natural language processing models gets more prevalent in various industries, academic and social settings, it is imperative that we assess not only the quality of these models but also their fairness when exposed to data originating from diverse social groups (Czarnowska et al., 2021). Text summarization models, in particular, facilitate the processing of large collections of a wide variety of text data by distilling documents into short, concise, and informative summaries while preserving the most relevant points from the source document (Nallapati et al., 2017; Zhang et al., 2018; Liu and Lapata, 2019). Multi-document summarization (MDS) is the task of generating a coherent summary from a set of input documents, usually centered around a topic, as opposed to single document summarization (SDS) which takes one document as input. The input in MDS consists of multiple documents, that may have been written by distinct users, varying in linguistic diversity, styles, or dialects.

MDS can be of type *extractive*, where the models extract the salient points directly from the source document to form the summary, or of type *abstractive* where the models generate summaries by rewriting salient information using novel words or phrases. In both cases, the resulting summary should be of good quality in terms of informativeness, coherence and relevance to the source document. At the same time, a good summary should be *unbiased* and should reflect the diversity of thoughts and perspectives present in the source documents.

The notion of fairness describes equal or fair treatment without favoritism or discrimination. However, plenty of evidence suggests intrinsic societal biases in language models (Bolukbasi et al., 2016; Bommasani et al., 2021; Deas et al., 2023). More specific to the task of summarization, fairness is measured by the ability of algorithms to capture the peculiarity in all represented groups (Shandilya et al., 2018; Dash et al., 2019; Keswani and Celis, 2021; Olabisi et al., 2022; Ladhak et al., 2023).

Conventionally, the documents in MDS are simply concatenated into one large collection of text as the input for the model. Prior research supports the existence of position bias, or lead bias, where the models rely excessively on the position of the sentences in the input rather than their semantic information (Lin and Hovy, 1997; Hong and Nenkova, 2014; Wang et al., 2019). This is a particularly common phenomenon in news summarization, where early parts of an article often contain the most salient information. While many

Proceedings of the Eleventh Workshop on NLP for Similar Languages, Varieties, and Dialects (VarDial 2024), pages 117–129

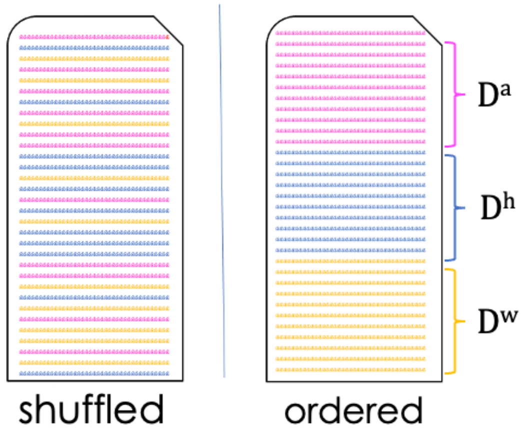


Figure 1: Illustration showing `shuffled` vs. `ordered` input for multi-document summarization consisting of documents from three diverse groups ($\mathcal{D}^a$, $\mathcal{D}^h$, $\mathcal{D}^w$) as indicated by the three colors. The `ordered` input is denoted as $\mathcal{O}^a$ when $\mathcal{D}^a$ documents appear first in the input.

algorithms exploit this fact in summary generation, it can have a detrimental effect when important information is spread throughout the input.

In non-news domains, weak or no position bias has been observed (Kedzie et al., 2018; Kim et al., 2019). Regardless of whether position bias is noted or not, previous investigations have quantified the *effects* of position bias mostly in terms of standard summarization metrics (e.g., ROUGE) which focus on the textual quality of the summary (Sotudeh et al., 2022; Scirè et al., 2023). In this work, we investigate the effects of position bias on the fairness of the generated summaries.

Specifically, we ask two questions: *(i)* Do the system summaries show any position bias when we vary the order of the input documents? *(ii)* What is the impact of position bias on the fairness of the system summaries?

For our experiments we use DivSumm, a summarization dataset of linguistically diverse communities representing three dialect groups (Olabisi et al., 2022). We explore the effects of position bias in the outputs of seven abstractive summarization models (and three extractive models) and under two investigation setups: `shuffled` (when the data is presented as randomly shuffled) and `ordered` (when the input documents are grouped according to their dialects). Figure 1 presents a schematic overview. The generated summaries are evaluated in terms of fairness, as well as metrics related to the textual quality.

The contributions of our work are as follows:

- We comprehensively investigate the phenomenon of position bias in the context of social multi-document summarization;
- We explore ten different summarization models, both abstractive and extractive;
- We contextualize and quantify the impact of position bias in terms of fairness and textual quality of generated summaries.

2 Related Work

In this section we present some notable prior research in two relevant areas. First, we discuss position bias in summarization, followed by works studying fairness in summarization.

Position Bias in Summarization Position bias can manifest in MDS scenarios just as it does in SDS scenarios because in MDS, the documents are typically concatenated into one long input and treated very much like a 'single' document. Several works have studied the substantial position bias (also known as lead bias), especially in the context of news summarization where the datasets and models prioritize selecting sentences from the beginning of an article (Lin and Hovy, 1997; Hong and Nenkova, 2014; Wang et al., 2019). Often the lead bias is so strong that the simple lead-k baseline or using the first k sentences of a news article to generate the summary can score higher than many other models (See et al., 2017). While some have suggested approaches for mitigating or countering lead bias (Grenander et al., 2019; Xing et al., 2021; Gong et al., 2022; Zhang et al., 2022), others have leveraged lead bias (Yang et al., 2020; Zhu et al., 2020; Padmakumar and He, 2021).

Interestingly, although position bias dominates the learning signal for news summarization or similar domains, it is less apparent in other domains where most non-news datasets show weak or no position bias (Kedzie et al., 2018; Jung et al., 2019; Kim et al., 2019; Sharma et al., 2019; Sotudeh et al., 2022; Scirè et al., 2023). Notably, none of these studies consider datasets where data originates from diverse social groups, which is the focus of our work.

Moreover, prior research studying the effect of position bias has quantified its impact exclusively in terms of textual quality, typically measured in terms of summarization metrics such as ROUGE, and others. To our knowledge, ours is the first work quantifying the impact of position bias in multidocument summarization in terms of fairness where data originates from diverse social groups.

Fairness in Summarization A significant amount of work has been done toward improving the textual quality of summaries but not so much in terms of enhancing the fairness of summaries, particularly in the context of diverse groups. Prior text summarization work has proposed fairness-preserving algorithms (Shandilya et al., 2018; Dash et al., 2019), bias mitigation models (Keswani and Celis, 2021) and fairness interventions for extractive and abstractive summarization (Olabisi et al., 2022). Furthermore, Ladhak et al. (2023) observed that name-nationality stereotypes propagate from pretraining data to downstream summarization systems and manifest as hallucinated facts.

3 Experimental Setup

Considering the extensive literature on fairness in natural language processing, which highlights significant disparities in the processing of data from different social groups, whether along the dimensions of gender or race or others, we are compelled to ask two questions:

1. What happens when the input data to be summarized is deliberately grouped according to the social groups, such as dialect groups in our case? (in Section 4) and,

2. How do the effects of position bias affect the fairness of generated summaries (Section 5).

Before exploring these questions, we first describe our experimental setup in this section.

3.1 Task Formulation

Considering a multi-document set of n topically-related documents $\mathcal{D} = \{d_1^{g_1}, ..., d_n^{g_r}\}$, where each document belongs to one of several diverse social groups $\mathcal{G} = \{g_1, ..., g_r\}$, the objective is to produce a summary $\mathcal{S}(\mathcal{D})$ that ideally exhibits both high textual quality and fairness. In this work, because of the original dataset design where the number of documents from each group is equal in the input, our investigation is concerned with the notion of equal representation. As such, a summary is considered to be fair when all groups $g_1, ..., g_r$ are equally represented in the output.

3.2 Dataset

For our experiments, we use the DivSumm dataset[1], an MDS dataset consisting of English tweets of three diverse dialects (*African-American* English, *Hispanic-aligned* Language, and *White-aligned* Language) (Olabisi et al., 2022), which was developed using a large corpus of tweets originally collected by Blodgett et al. (2016). The dataset includes 25 topically-related sets of documents (tweets) as input and corresponding human-written extractive and abstractive summaries. Each set $\mathcal{D}$ consists of 90 documents evenly distributed among the three dialects (i.e., 30 documents per dialect). A selection of dialect diverse tweets from DivSumm is presented in Table 3.

3.3 `Shuffled` and `Ordered`

To study the phenomenon of position bias in social multi-document summarization where documents originate from different social groups, we devise two distinct scenarios: `shuffled` and `ordered`, as depicted in Figure 1.

In the **shuffled** setting, documents appear randomly present in the input in no specific order. In fact, to ensure consistency, we retain the original order as presented in the DivSumm dataset which the annotators used to craft the summaries.

In the **ordered** setting, we perturb the input data by grouping documents from each social group together. When the subset of *White-aligned* Language tweets ($\mathcal{D}^w$) appears first, the input set is denoted as $\texttt{ordered}^{\texttt{white}}$ or, simply, $\mathcal{O}^w$. Similarly, when the subset of *African-American* English tweets ($\mathcal{D}^a$) come first, we denote that set as $\mathcal{O}^a$, and when the subset of *Hispanic-aligned* Language documents ($\mathcal{D}^h$) appears first, we denote that set as $\mathcal{O}^h$. Specifically, the input documents are ordered as follows:

$$\mathcal{O}^w = \{\mathcal{D}^w, \mathcal{D}^a, \mathcal{D}^h\}$$
$$\mathcal{O}^a = \{\mathcal{D}^a, \mathcal{D}^h, \mathcal{D}^w\}$$
$$\mathcal{O}^h = \{\mathcal{D}^h, \mathcal{D}^w, \mathcal{D}^a\}$$

These documents are summarized using several models described in the next section, allowing us to subsequently investigate the different summaries we generate – $\mathcal{S}(\mathcal{O}^w)$, $\mathcal{S}(\mathcal{O}^a)$, $\mathcal{S}(\mathcal{O}^h)$, and $\mathcal{S}$(`shuffled`) – which are obtained from four distinct sets of input documents – $\mathcal{O}^w$, $\mathcal{O}^a$, $\mathcal{O}^h$, and `shuffled`, respectively.

[1] https://github.com/PortNLP/DivSumm

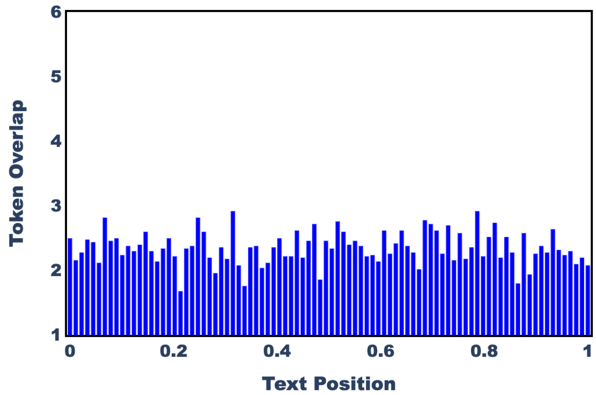


Figure 2: Average token overlap between human-written reference summaries and each document d_i using the DivSumm dataset. Text position on the x-axis has been normalized between 0 and 1.

3.4 Summarization Models

We study a total of seven abstractive models in our experiments. We also study three extractive models, the details and results of which are discussed in A. Following the setup of DivSumm, we generate summaries of 5 sentences per topic

The seven abstractive models included in our experiments are as follows:

- BART[2] (Lewis et al., 2019),
- T5 (Raffel et al., 2019),
- LED (Longformer Encoder-Decoder) (Beltagy et al., 2020),
- PEGASUS (Zhang et al., 2020),
- GPT-3.5,
- PRIMERA (Xiao et al., 2021), and
- CLAUDE (`Claude 3 Opus`).

GPT-3.5 and Claude were prompted with the following prompt – "Please summarize the following texts in only five sentences".

4 Position Bias in Social MDS

This section discusses position bias within three types of summaries: human-authored reference summaries of the DivSumm dataset, system summaries generated using the `shuffled` input, and system summaries generated using `ordered` inputs.

[2]Model checkpoints for BART, T5, LED, Pegasus, and Primera were accessed from `https://huggingface.co/models`.

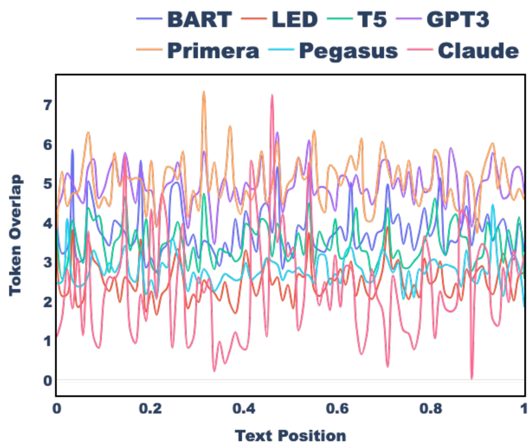


Figure 3: Average token overlap between `ordered` system-generated summaries by each abstractive summarization model and each document d_i in the input set $\mathcal{D}$ of DivSumm. Text position on the x-axis has been normalized between 0 and 1.

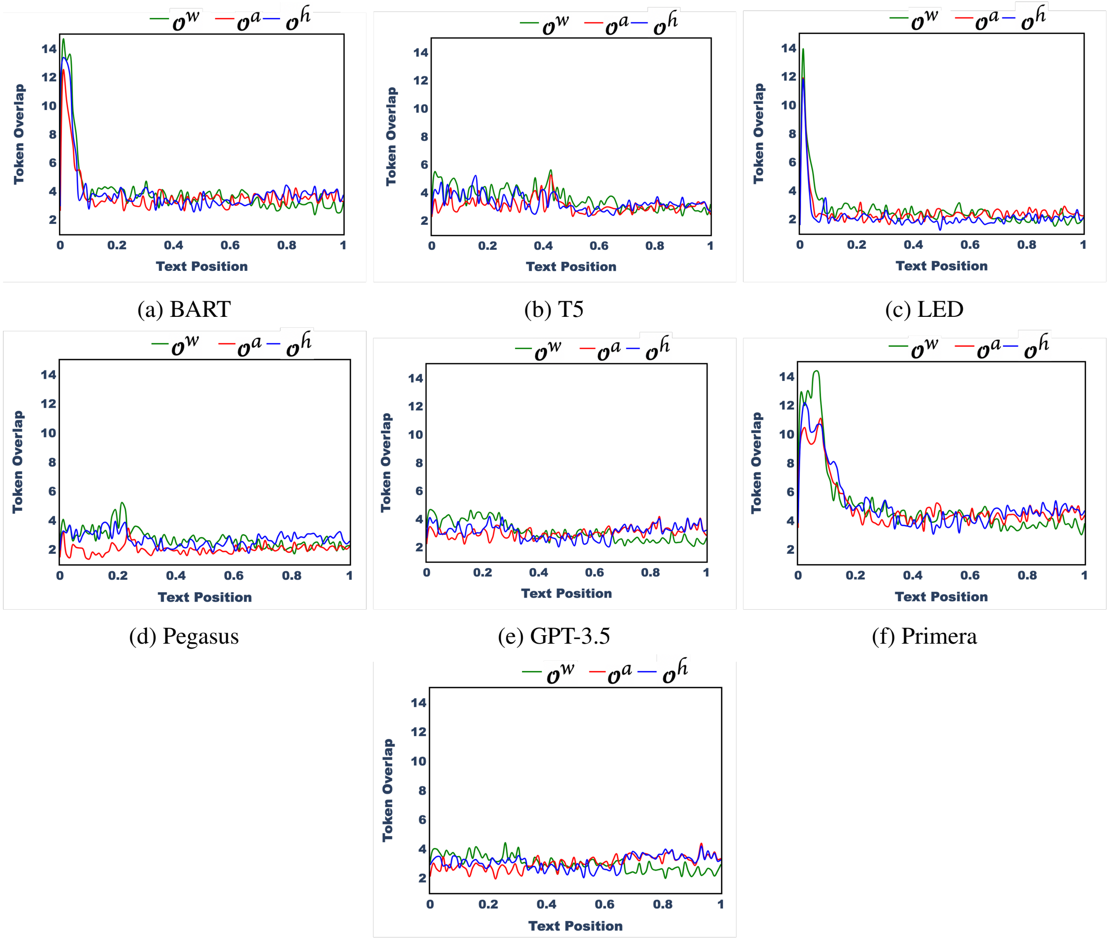
Following prior work on position bias, we calculate the overlap between the summaries and the input documents by computing the number of tokens shared between the summary and each document of the MDS topic set. That is, given the 90 documents in each topically-related input set, we get the overlap score for each document $(d_1, d_2, ..., d_{90})$ with respect to a summary, and report the average score over the entire dataset. A higher overlap score implies more semantic relationship between the summary and source document.

4.1 Position Bias in Human-Written Reference Summaries

To examine position bias in the summaries created by humans, we analyze both abstractive and extractive reference summaries of DivSumm dataset. Because the dataset contains two reference summaries per input, we report the average score. The results are presented in Figure 2 where no noticeable position bias is observed, and it is encouraging to note that the annotators were not influenced by the position of the documents in the input when producing their summaries.

4.2 Position Bias in System Summaries (`Shuffled`)

The results of position bias within model-generated summaries using `shuffled` inputs are presented in Figure 3. Similar to the human-written reference summaries, we observe no notable position bias suggesting that when summarizing randomly shuf-



(a) BART (b) T5 (c) LED

(d) Pegasus (e) GPT-3.5 (f) Primera

(g) Claude

Figure 4: Average token overlap between `ordered` system-generated summaries by each of the seven abstractive summarization models and each document d_i in the input set $\mathcal{D}$ of the DivSumm dataset. Text position on the x-axis has been normalized between 0 and 1.

fled data from various social groups, the models also do not exhibit any particular lead bias. This observation on DivSumm, a dataset of tweets, is consistent with trends observed in other social datasets (Reddit posts (Kim et al., 2019) and social user posts (Sotudeh et al., 2022)).

4.3 Position Bias in System Summaries (`Ordered`)

Now we discuss the results of position bias in system summaries that were generated using various `ordered` inputs: $\mathcal{O}^w$, $\mathcal{O}^a$, $\mathcal{O}^h$. Model-specific results are presented in Figure 4, where, interestingly, we now observe a **strong position bias in three out of seven abstractive models,** (BART, LED, and Primera), with up to 3 times higher token overlap in the beginning of the input document, as shown by the distribution. Three other models show weak position bias (T5, Pegasus, and GPT-3.5). This phenomenon diverges from traditional position bias, where models tend to favor earlier bits of text. *Instead, we notice that models favor earlier pieces of text only when the text exhibits some socially linguistic similarity*. These observations highlight the importance of more nuanced analysis when exploring position bias in summarization systems, especially when processing diverse social data.

5 Fairness and Textual Quality Amidst Position Bias

Having observed an instance of position bias, especially when input data is grouped according to dialect groups, the next natural question to ask is how does this position bias quantitatively impact the fairness and textual quality of the generated summaries. We briefly describe the evaluation metrics before discussing the main results.

Model	$\mathcal{O}^w$				$\mathcal{O}^a$				$\mathcal{O}^h$				shuffled			
	$\mathcal{D}^w$	$\mathcal{D}^a$	$\mathcal{D}^h$	ΔFair (↓)	$\mathcal{D}^w$	$\mathcal{D}^a$	$\mathcal{D}^h$	ΔFair (↓)	$\mathcal{D}^w$	$\mathcal{D}^a$	$\mathcal{D}^h$	ΔFair (↓)	$\mathcal{D}^w$	$\mathcal{D}^a$	$\mathcal{D}^h$	ΔFair (↓)
BART	**0.64**	0.41	0.45	0.23	0.41	**0.55**	0.40	0.15	0.44	0.42	**0.59**	0.17	**0.41**	**0.41**	0.40	0.01
LED	**0.47**	0.30	0.33	0.17	0.31	**0.43**	0.31	0.12	0.26	0.24	**0.36**	0.12	0.30	0.29	**0.35**	0.06
T5	**0.52**	0.39	0.48	0.13	0.39	**0.46**	0.43	0.07	0.40	0.41	**0.49**	0.09	0.37	**0.41**	0.40	0.04
PEGASUS	**0.34**	0.28	0.29	0.06	0.22	**0.25**	0.21	0.04	0.26	0.24	**0.32**	0.08	**0.32**	0.33	**0.32**	0.01
GPT-3.5	**0.47**	0.35	0.38	0.12	**0.38**	**0.38**	0.36	0.02	0.38	0.34	**0.41**	0.07	**0.40**	0.35	0.37	0.05
PRIMERA	**0.62**	0.41	0.45	0.21	0.42	**0.60**	0.44	0.18	0.45	0.44	**0.62**	0.18	**0.49**	0.48	0.50	0.02
CLAUDE	**0.39**	0.33	0.36	0.06	**0.37**	0.32	0.34	0.05	**0.36**	0.31	0.34	0.05	**0.37**	0.32	0.35	0.05
AVG	**0.49**	0.35	0.39	0.14	0.36	**0.43**	0.36	0.09	0.36	0.35	**0.45**	0.11	0.38	0.37	0.39	0.04

Table 1: ***Fairness.*** Similarity scores of summaries generated by ordered inputs ($\mathcal{O}^w$, $\mathcal{O}^a$, $\mathcal{O}^h$) and shuffled inputs compared to each group of documents ($\mathcal{D}^w$, $\mathcal{D}^a$, $\mathcal{D}^h$) across seven abstractive summarization models using the *DivSumm* dataset. The highest similarity scores are shown in bold.

5.1 Evaluation Metrics

Fairness (Gap): One way of measuring fairness is by estimating the amount of representation from each dialect group in the final summary by comparing the summary $\mathcal{S}$ to the set of documents from each group. Given that an unbiased summary should capture the perspectives across all groups, we evaluate summary fairness for both extractive and abstractive models using semantic similarity of the summary to each represented group. As an example, for input $\mathcal{O}^w$, we compare the final summary $\mathcal{S}(\mathcal{O}^w)$ to the document set of each dialect group: $\mathcal{D}^w$, $\mathcal{D}^w$, and $\mathcal{D}^h$. In other words, we compute $sim(i,j)$ where $i = \{\mathcal{S}(\mathcal{O}^w), \mathcal{S}(\mathcal{O}^a), \mathcal{S}(\mathcal{O}^h)\}$ and $j = \{\mathcal{D}^w, \mathcal{D}^a, \mathcal{D}^h\}$. Similarity can be estimated by many possible methods of obtaining semantic similarity. We use cosine similarity.

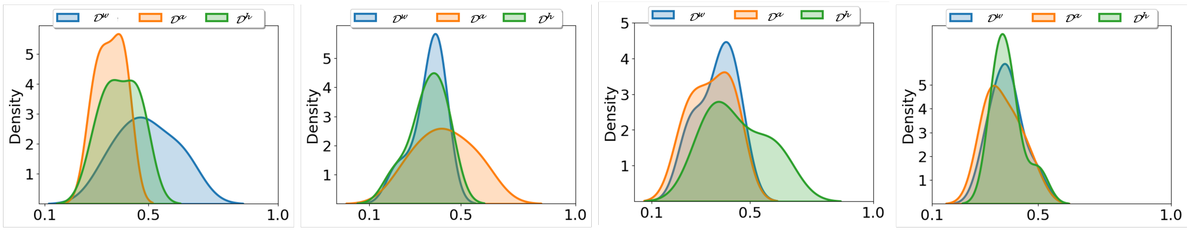
From these similarity scores, we can derive the **Fairness Gap (ΔFair)** by calculating the difference between the maximum and the minimum scores attributed to any of the groups (Olabisi et al., 2022). Intuitively, a summary that produces relatively similar representation scores across all groups can be considered as *fair* because it likely contains comparable representation from all groups such that no one group is significantly underrepresented.

Textual Quality: Four established metrics are used for assessing the quality of the summaries: ROUGE, BARTScore, BERTScore, and UniEval. **ROUGE** (Lin, 2004) calculates the lexical overlap between the model-generated summary and the reference summaries. For our experiments, we report the F1 scores of ROUGE-L which is the longest common subsequence between the two summaries. **BARTScore** (Yuan et al., 2021) leverages BART's average log-likelihood of generating the evaluated summary conditional on the source document. Since it uses the average log-likelihood for target tokens, the calculated scores are smaller than 0 (negative). We use the `facebook/bart-large-cnn` checkpoint. **BERTScore** (Zhang* et al., 2020) relies on BERT embeddings and matches words in system-generated summaries and reference summaries to compute token similarity. We use the `microsoft/deberta-xlarge-mnli` model and report the F1 scores. **UniEval** (Zhong et al., 2022) is a unified multi-dimensional evaluator that employs boolean question answering format to evaluate text generation tasks. We make use of `unieval-sum` which evaluates system-generated summaries in terms of four dimensions: coherence, consistency, relevance and fluency. Except for fluency, the rest are reference-free metrics. We report the overall score.

5.2 Results

Evaluating fairness. The results in Table 1 report the fairness scores for all seven models. **We clearly observe that ordering the input documents based on groups certainly favors the group that appears first.** This phenomenon is consistently observed in all three types of ordered sets, regardless of which particular dialect group's data is presented first. However, when the documents are presented as shuffled, no single group is over-represented and the summaries appear more balanced (ΔFair = 0.04).

The density plots in Figure 5 also show that the shuffled input set is the most balanced across all groups, unlike the ordered sets which are significantly skewed. Furthermore, amongst ordered



(a) $\mathcal{O}^w$ (b) $\mathcal{O}^a$ (c) $\mathcal{O}^h$ (d) shuffled

Figure 5: Density distribution of similarity scores between system-generated summaries and each group, across all summarization models for $\mathcal{O}^w$, $\mathcal{O}^a$, $\mathcal{O}^h$ and `shuffled` input sets. The outputs of `shuffled` inputs show very different and balanced distributions compared to the `ordered` inputs.

Model	**ROUGE-L**				**BARTScore**				**BERTScore**				**UniEval**			
	$\mathcal{O}^w$	$\mathcal{O}^a$	$\mathcal{O}^h$	Sh.	$\mathcal{O}^w$	$\mathcal{O}^a$	$\mathcal{O}^h$	Sh.	$\mathcal{O}^w$	$\mathcal{O}^a$	$\mathcal{O}^h$	Sh.	$\mathcal{O}^w$	$\mathcal{O}^a$	$\mathcal{O}^h$	Sh.
BART	**0.15**	0.14	0.14	**0.15**	-3.73	-3.74	-3.72	**-3.69**	**0.51**	0.50	**0.51**	0.50	0.46	0.46	**0.48**	0.44
T5	**0.15**	0.13	0.13	0.14	-3.76	-3.75	-3.74	**-3.72**	0.50	0.48	0.49	**0.51**	0.45	0.45	**0.47**	0.44
LED	**0.12**	0.11	0.10	**0.12**	-3.75	-3.79	-3.79	**-3.73**	0.44	0.40	0.39	**0.47**	0.44	0.44	**0.46**	0.43
PEGASUS	**0.14**	0.11	0.13	**0.14**	**-3.73**	-3.75	-3.76	**-3.73**	**0.47**	0.44	0.46	0.46	0.45	0.45	**0.47**	0.43
GPT-3.5	0.20	0.20	**<u>0.21</u>**	**<u>0.21</u>**	-3.64	-3.68	**<u>-3.62</u>**	-3.65	0.57	0.58	**<u>0.59</u>**	**<u>0.59</u>**	0.46	0.45	**<u>0.48</u>**	0.44
PRIMERA	**0.14**	0.12	0.13	0.13	-3.67	-3.68	**-3.63**	-3.64	0.51	0.49	**0.50**	0.49	0.45	0.46	**<u>0.48</u>**	0.44
CLAUDE	0.18	0.18	**0.19**	0.18	**-3.64**	**-3.64**	**-3.64**	-3.65	0.56	0.56	**0.57**	0.56	0.44	0.44	**0.46**	0.43
AVG	0.15	0.14	0.15	0.15	-3.70	-3.72	-3.70	-3.69	0.51	0.49	0.50	0.51	0.45	0.45	0.47	0.44

Table 2: ***Quality***. Results of `ordered` ($\mathcal{O}^w$, $\mathcal{O}^a$, $\mathcal{O}^h$) and `shuffled` (`Sh.`) approaches across seven abstractive summarization models showing ROUGE-L, BARTScore, BERTScore, and UniEval scores on the *DivSumm* dataset. The best scores are shown in **bold**, whereas the highest scores per metric are shown as <u>underlined</u>.

documents, the fairness gap is the largest when documents of White-aligned language are passed first (ΔFair = 0.14), and the smallest when documents of African-American English appear first (ΔFair = 0.09).

Evaluating textual quality. Table 2 presents the summary quality scores across all seven summarization models for the four sets of input. We clearly see that the scores of the `shuffled` approach are superior or comparable to the scores from the three input sets in the `ordered` approach, except in the case of UniEval. **This shows that with respect to quality, there is no significant difference whether documents are presented as ordered or shuffled.**

5.3 Discussion

Some samples of system summaries are presented in Table 3. The key findings of our study can be summarized as follows:

- We find no evidence of position bias in human-annotated reference summaries of DivSumm, a social MDS dataset of diverse groups. Same observation is made for the abstractive system-generated summaries obtained when the input documents are passed in randomly or shuffled.
- However, when the input is `ordered` based on dialect groups, we observe a significant position bias in the system summaries, with the summaries having higher overlap with the group that appears first in the input document.
- `Ordered` documents involving different dialects result in summaries that are significantly skewed in terms of fairness, with the group whose data appears first is clearly favored by the models. In contrast, `shuffled` documents show the least amount of fairness gap.
- In terms of quality, we observe that for all models and metrics, the scores for `ordered` and `shuffled` remain comparable, suggesting that ordering based on diverse groups has no noticeable effect on the quality of system-generated summaries.

Taken together, the findings of our study indicate that both the `ordered` and `shuffled` approaches yield comparable results in terms of textual quality, but highly disparate results in terms of fairness. This phenomenon is consistently observed in

Input Documents Set
d_1: **Hispanic** : The Grammys should have come out on Saturday so I won't stay up late today lol
d_2: **AA** : Wasn't it during the Grammys the last time Chris Brown slid Rhianna?
d_3: **White** : Feel free to join my lonesome self swiminngg at Grammys!!
d_4: **AA** : I've given up #DowntonAbbey for J.T.? This is serious #Grammys
d_5: **Hispanic** : oh lol thanks thought you were talking about the Grammys lol sorry lol
d_6: **Hispanic** : I don't even know if I am watching the right latin Grammys lol
d_7: **White** : "If I'm a hipster about anything, it's Kings of Leon. I listened to them before they won Grammys."
d_8: **White** : isn't performing at the Grammys? What's the point of even having the Grammys now?
d_9: **AA** : Imma get some ideas for Easter from these here Grammys . ____
d_{10}: **Hispanic** : Wow every celeb in the crowd is singing Bruno's lyrics. Awesome. #Grammys
...
...
...
d_{88}: **AA** : These Grammys need some life in it cause I'm so bored
d_{89}: **AA** : Got my easter outfit nd dress for the Grammys.. im set :)
d_{90}: **White** : "great, now she can have practice for whe she doesn't win at next year's Grammys"

Model	Summary
BART	I've given up DowntonAbbey for J.T.? This is serious Grammys. The Grammys lasted an extra 30 mins just for us to all get hazed by LL smh. Im not even watching the Grammys im too lazy to change the channel but isn't Justin doing a livestream or something at the same time. I'm gonna say yes.
GPT-3.5	Tina Campbell won't be attending the Grammys, and J Cole is being tipped for success. The Black Keys won five awards at the Grammys, and Eminem is expected to win Best Male Singer and Rapper. Rihanna and Chris Brown were together at the Grammys, and there is a lot of Twitter beef happening due to the event. An ad agency has the Grammys as their client, and a speech teacher asked someone to dress for the Grammys for a show. People are discussing the Grammys and celebrities attending.

Table 3: Example of a `shuffled` input set from the *DivSumm* dataset showing input documents with their respective dialects, and system summaries generated by the model with the lowest fairness gap (BART) and highest textual quality (GPT-3.5). Note that the dialect labels shown here are only for clearer presentation and are not part of the input to the model.

all abstractive models, suggesting that the models are not robust to fairly straightforward group-level data perturbations. These findings are important because they highlight a potential source of nuanced bias in the summarization models. The observation that ordering the input documents based on groups favors the group that appears first indicates a systematic bias in the models' behavior. The fact that the shuffled input set leads to more balanced summaries across all groups implies that the bias observed in the ordered sets can be mitigated by introducing randomness in the presentation of input data. This insight is crucial for understanding and addressing bias in summarization systems, especially in scenarios where fairness and equity are important considerations, such as in social data analysis or decision-making processes. Overall, this result sheds light on an important aspect of model behavior and informs strategies for improving the fairness and effectiveness of summarization models.

6 Conclusion

In this work, we investigate how position bias manifests in social multi-document summarization, specifically in scenarios where the input data is derived from three linguistically diverse communities. When presented with randomly shuffled input data, summaries generated by ten distinct summarization models exhibited no signs of position bias. However, a significant shift occurred when the input data was simply reordered based on social groups. In such instances, the models produced biased summaries, primarily favoring the social group that appeared earlier in the input sequence. In terms of the quality of generated summaries, however, there was no notable difference due to the order in which source documents were presented, whether shuffled or ordered. Our results suggest that position bias manifests differently in the context of social multi-document summarization. Furthermore, they highlight the need to incorporate randomized shuffling in multi-document summarization datasets particularly when summarizing documents from diverse groups to ensure that the resultant summaries are not only of high quality but also faithfully representative of the diversity present in the input data.

Ethical Considerations

Our findings and conclusions in this paper are based on an existing social media summarization dataset composed of tweets in English, primarily due to the lack of appropriate resources available to undertake such studies. Given the nature of naturally occurring data, it is possible that the data contains some offensive language. Hence, it is possible for the models to also generate summaries with offensive words. In addition to this, due to the constraint on tweet length, users are known to use acronyms and slangs that may have various meanings across different groups – this phenomenon is not accounted for in this study. Also, the existing dataset that we use in this work was originally collected from a corpus using geolocation and census data. This dialectal information used in categorizing users' languages should not be used as a representation of users' racial information. In this work, we evaluate summary fairness using proxy metrics such as semantic similarity to each represented group. The definition of fairness may vary for humans, and as such this should not be used as the gold standard.

Acknowledgments

We thank the anonymous reviewers as well as the members of PortNLP lab for their insightful comments. This research was supported by National Science Foundation grants (CRII:RI 2246174 and SAI-P 2228783).

References

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Su Lin Blodgett, Lisa Green, and Brendan O'Connor. 2016. Demographic dialectal variation in social media: A case study of african-american english. *arXiv preprint arXiv:1608.08868*.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Paula Czarnowska, Yogarshi Vyas, and Kashif Shah. 2021. Quantifying social biases in nlp: A generalization and empirical comparison of extrinsic fairness metrics. *arXiv preprint arXiv:2106.14574*.

Abhisek Dash, Anurag Shandilya, Arindam Biswas, Kripabandhu Ghosh, Saptarshi Ghosh, and Abhijnan Chakraborty. 2019. Summarizing user-generated textual content: Motivation and methods for fairness in algorithmic summaries. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–28.

Nicholas Deas, Jessi Grieser, Shana Kleiner, Desmond Patton, Elsbeth Turcan, and Kathleen McKeown. 2023. Evaluation of african american language bias in natural language generation.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Shuai Gong, Zhenfang Zhu, Jiangtao Qi, Chunling Tong, Qiang Lu, and Wenqing Wu. 2022. Improving extractive document summarization with sentence centrality. *PloS one*, 17(7):e0268278.

Matt Grenander, Yue Dong, Jackie Chi Kit Cheung, and Annie Louis. 2019. Countering the effects of lead bias in news summarization via multi-stage training and auxiliary losses. *arXiv preprint arXiv:1909.04028*.

Max Grusky, Mor Naaman, and Yoav Artzi. 2018. Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. *arXiv preprint arXiv:1804.11283*.

Kai Hong and Ani Nenkova. 2014. Improving the estimation of word importance for news multi-document summarization. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 712–721.

Taehee Jung, Dongyeop Kang, Lucas Mentch, and Eduard Hovy. 2019. Earlier isn't always better: Sub-aspect analysis on corpus and system biases in summarization. *arXiv preprint arXiv:1908.11723*.

Chris Kedzie, Kathleen McKeown, and Hal Daumé III. 2018. Content selection in deep learning models of summarization. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1818–1828, Brussels, Belgium. Association for Computational Linguistics.

Vijay Keswani and L Elisa Celis. 2021. Dialect diversity in text summarization on twitter. In *Proceedings of the Web Conference 2021*, pages 3802–3814.

Byeongchang Kim, Hyunwoo Kim, and Gunhee Kim. 2019. Abstractive summarization of Reddit posts with multi-level memory networks. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2519–2531, Minneapolis, Minnesota. Association for Computational Linguistics.

Faisal Ladhak, Esin Durmus, Mirac Suzgun, Tianyi Zhang, Dan Jurafsky, Kathleen Mckeown, and Tatsunori B Hashimoto. 2023. When do pre-training biases propagate to downstream tasks? a case study in text summarization. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3198–3211.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Chin-Yew Lin and Eduard Hovy. 1997. Identifying topics by position. In *Fifth conference on applied natural language processing*, pages 283–290.

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. *arXiv preprint arXiv:1908.08345*.

Rada Mihalcea and Paul Tarau. 2004. Textrank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing*, pages 404–411.

Derek Miller. 2019. Leveraging bert for extractive text summarization on lectures. *arXiv preprint arXiv:1906.04165*.

Ramesh Nallapati, Feifei Zhai, and Bowen Zhou. 2017. Summarunner: A recurrent neural network based sequence model for extractive summarization of documents. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31.

Olubusayo Olabisi, Aaron Hudson, Antonie Jetter, and Ameeta Agrawal. 2022. Analyzing the dialect diversity in multi-document summaries. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6208–6221.

Vishakh Padmakumar and He He. 2021. Unsupervised extractive summarization using pointwise mutual information. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2505–2512, Online. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.

Alessandro Scirè, Simone Conia, Simone Ciciliano, and Roberto Navigli. 2023. Echoes from alexandria: A large resource for multilingual book summarization. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 853–867, Toronto, Canada. Association for Computational Linguistics.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.

Anurag Shandilya, Kripabandhu Ghosh, and Saptarshi Ghosh. 2018. Fairness of extractive text summarization. In *Companion Proceedings of the The Web Conference 2018*, pages 97–98.

Eva Sharma, Chen Li, and Lu Wang. 2019. BIGPATENT: A large-scale dataset for abstractive and coherent summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2204–2213, Florence, Italy. Association for Computational Linguistics.

Sajad Sotudeh, Nazli Goharian, and Zachary Young. 2022. MentSum: A resource for exploring summarization of mental health online posts. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2682–2692, Marseille, France. European Language Resources Association.

Danqing Wang, Pengfei Liu, Ming Zhong, Jie Fu, Xipeng Qiu, and Xuanjing Huang. 2019. Exploring domain shift in extractive text summarization. *arXiv preprint arXiv:1908.11664*.

Wen Xiao, Iz Beltagy, Giuseppe Carenini, and Arman Cohan. 2021. Primera: Pyramid-based masked sentence pre-training for multi-document summarization. *arXiv preprint arXiv:2110.08499*.

Linzi Xing, Wen Xiao, and Giuseppe Carenini. 2021. Demoting the lead bias in news summarization via alternating adversarial learning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 948–954, Online. Association for Computational Linguistics.

Ziyi Yang, Chenguang Zhu, Robert Gmyr, Michael Zeng, Xuedong Huang, and Eric Darve. 2020. TED: A pretrained unsupervised summarization model with theme modeling and denoising. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1865–1874, Online. Association for Computational Linguistics.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. In *Advances in Neural Information Processing Systems*, volume 34, pages 27263–27277. Curran Associates, Inc.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*, pages 11328–11339. PMLR.

Shengqiang Zhang, Xingxing Zhang, Hangbo Bao, and Furu Wei. 2022. Attention temperature matters in abstractive summarization distillation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 127–141, Dublin, Ireland. Association for Computational Linguistics.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Xingxing Zhang, Mirella Lapata, Furu Wei, and Ming Zhou. 2018. Neural latent extractive document summarization. *arXiv preprint arXiv:1808.07187*.

Ming Zhong, Yang Liu, Da Yin, Yuning Mao, Yizhu Jiao, Pengfei Liu, Chenguang Zhu, Heng Ji, and Jiawei Han. 2022. Towards a unified multi-dimensional evaluator for text generation. *arXiv preprint arXiv:2210.07197*.

Chenguang Zhu, Ziyi Yang, Robert Gmyr, Michael Zeng, and Xuedong Huang. 2020. Make lead bias in your favor: Zero-shot abstractive news summarization. In *International Conference on Learning Representations*.

A Fairness in Extractive Models

We repeat the same experiments and analysis for extractive models to observe if they exhibit behavior similar to that observed in the abstractive models.

A.1 Summarization Systems

We study three summarization models in our experiments to generate summaries of 5 sentences per topic (multi-document set):

TextRank[3] (Mihalcea and Tarau, 2004), an unsupervised graph-based ranking method, determines the most important sentences in a document based on information extracted from the document itself.

Bert-Ext[4] (Miller, 2019), an extractive summarization model built on top of BERT (Devlin et al., 2018), uses k-means clustering to select sentences closest to the centroid as the summaries.

Longformer[5] (Beltagy et al., 2020) is a modification of the transformer architecture, using a self-attention operation that scales linearly with the sequence length.

A.2 Evaluation Metrics

In evaluating textual quality, We use the same four metrics used for the abstractive models. To estimate fairness (gap), in addition to semantic similarity used in evaluating the fairness of abstractive models, we consider **coverage** as well which measures the extent to which a summary is a derivative of the input text. Following previous literature (Dash et al., 2019; Keswani and Celis, 2021), we estimate group fairness via disparity in *extractive fragment coverage* (Grusky et al., 2018), which indicates the degree of surface-level text overlap by computing the percentage of words in the summary from each dialect group's collection of documents.

A.3 Results

While `shuffled` extractive models show no noticeable position bias in Figure 6, we observe a strong position bias using `ordered` inputs in two out of three extractive models (BERT and LongFormer), as shown in Figure 7 further highlighting the importance of exploring position bias in summarization of diverse social data.

[3] https://radimrehurek.com/gensim_3.8.3/summarization/summariser.html

[4] https://pypi.org/project/bert-extractive-summarizer/

[5] Model checkpoint for Longformer was accessed from https://huggingface.co/models

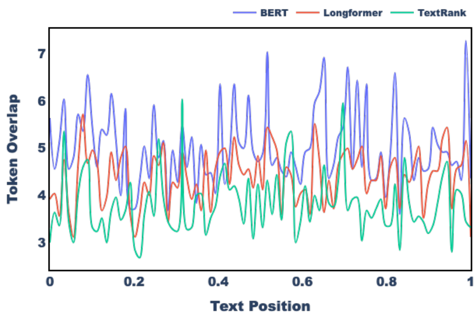
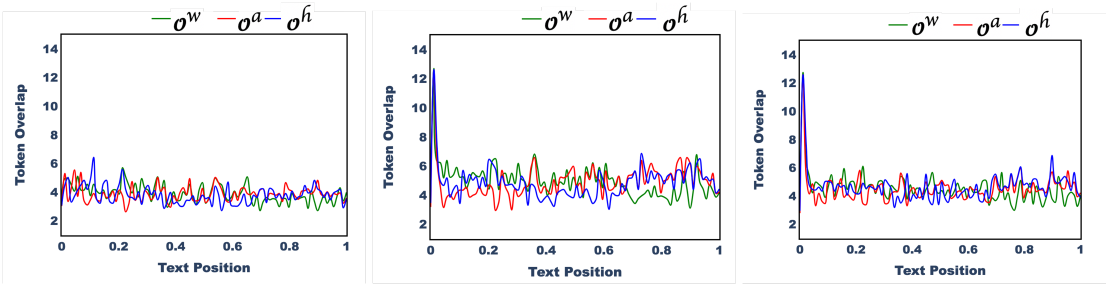


Figure 6: Average token overlap between `shuffled` system-generated summaries by each of the three extractive summarization models and each document d_i in the input set $\mathcal{D}$ of DivSumm. Text position on the x-axis has been normalized between 0 and 1.

Tables 4 and 5 show the fairness scores in terms of coverage and similarity, respectively, of extractive summaries. For all three models, we observe that the summaries generated using the `ordered` sets distinctly favor the group that appeared first in the input set of documents, while this phenomenon is absent from the `shuffled` set, where the results are much more evenly distributed across the three groups for all three models. Table 6 presents the quality scores along four metrics where, similar to abstractive models, little difference is noted between `ordered` and `shuffled` approaches.



(a) *TextRank* (b) *BERT* (c) *LongFormer*

Figure 7: Average token overlap between `ordered` system-generated summaries by each of the extractive summarization models and each document d_i in the input set $\mathcal{D}$ of DivSumm. Text position on the x-axis has been normalized between 0 and 1.

Model	$\mathcal{O}^w$				$\mathcal{O}^a$				$\mathcal{O}^h$				`shuffled`			
	$\mathcal{D}^w$	$\mathcal{D}^a$	$\mathcal{D}^h$	ΔFair	$\mathcal{D}^w$	$\mathcal{D}^a$	$\mathcal{D}^h$	ΔFair	$\mathcal{D}^w$	$\mathcal{D}^a$	$\mathcal{D}^h$	ΔFair	$\mathcal{D}^w$	$\mathcal{D}^a$	$\mathcal{D}^h$	ΔFair
TEXTRANK	**0.80**	0.72	0.76	0.08	0.70	**0.81**	0.74	0.11	0.72	0.73	**0.82**	0.10	0.74	0.76	**0.78**	0.04
BERT	**0.78**	0.69	0.77	0.09	**0.75**	0.74	0.73	0.02	0.78	0.69	**0.80**	0.11	**0.77**	0.74	0.76	0.03
LONGFORMER	**0.77**	0.72	0.73	0.05	0.70	**0.80**	0.71	0.10	0.73	0.72	**0.79**	0.07	0.72	**0.78**	0.77	0.06
AVG	0.78	0.71	0.75	0.07	0.72	0.78	0.73	0.07	0.74	0.71	0.80	0.09	0.74	0.76	0.77	0.05

Table 4: ***Fairness.*** Coverage scores of `ordered` and `shuffled` approaches compared to each group of documents ($\mathcal{D}^w$, $\mathcal{D}^a$, $\mathcal{D}^h$) for three extractive summarization models on *DivSumm* dataset. The highest scores are shown in bold.

Model	$\mathcal{O}^w$				$\mathcal{O}^a$				$\mathcal{O}^h$				`shuffled`			
	$\mathcal{D}^w$	$\mathcal{D}^a$	$\mathcal{D}^h$	ΔFair	$\mathcal{D}^w$	$\mathcal{D}^a$	$\mathcal{D}^h$	ΔFair	$\mathcal{D}^w$	$\mathcal{D}^a$	$\mathcal{D}^h$	ΔFair	$\mathcal{D}^w$	$\mathcal{D}^a$	$\mathcal{D}^h$	ΔFair
TEXTRANK	**0.57**	0.55	0.52	0.05	0.51	**0.54**	0.49	0.05	**0.55**	0.54	0.50	0.04	0.45	**0.46**	0.42	0.05
BERT	**0.61**	0.54	0.53	0.07	0.51	0.59	**0.61**	0.10	0.62	**0.63**	0.55	0.08	0.48	0.50	**0.52**	0.03
LONGFORMER	**0.58**	0.54	0.50	0.08	0.55	**0.56**	0.55	0.02	**0.54**	**0.54**	0.52	0.02	0.45	0.44	**0.47**	0.03
AVG	0.59	0.54	0.52	0.07	0.52	0.56	0.55	0.04	0.57	0.57	0.52	0.05	0.46	0.47	0.47	0.03

Table 5: ***Fairness.*** Semantic similarity scores of `ordered` and `shuffled` approaches compared to each group of documents ($\mathcal{D}^w$, $\mathcal{D}^a$, $\mathcal{D}^h$) across extractive summarization models on *DivSumm* dataset. The highest scores are shown in bold.

Model	ROUGE-L				BARTSCORE				BERTSCORE				UNIEVAL			
	$\mathcal{O}^w$	$\mathcal{O}^a$	$\mathcal{O}^h$	`Sh.`	$\mathcal{O}^w$	$\mathcal{O}^a$	$\mathcal{O}^h$	`Sh.`	$\mathcal{O}^w$	$\mathcal{O}^a$	$\mathcal{O}^h$	`Sh.`	$\mathcal{O}^w$	$\mathcal{O}^a$	$\mathcal{O}^h$	`Sh.`
TEXTRANK	**0.23**	0.21	0.22	**0.23**	-4.42	-4.42	-4.44	**-4.29**	0.55	0.54	0.55	**0.56**	0.46	0.46	**0.48**	0.44
BERT	**0.24**	**0.24**	0.23	0.21	**-4.28**	-4.33	-4.39	-4.71	**0.56**	**0.56**	**0.56**	0.55	0.47	0.46	**0.49**	0.45
LONGFORMER	**0.22**	0.21	**0.22**	0.20	-4.38	-4.44	-4.41	**-4.35**	**0.56**	0.55	**0.56**	**0.56**	0.46	0.46	**0.48**	0.45
AVG	0.23	0.22	0.22	0.22	-4.36	-4.40	-4.41	-4.45	0.56	0.55	0.55	0.56	0.46	0.46	0.48	0.45

Table 6: ***Quality***. Results of `ordered` and `shuffled` approaches across extractive summarization models showing ROUGE-L, BARTScore, BERTScore and UniEval scores on *DivSumm* dataset. The best scores are shown in **bold**, whereas the highest scores per metric are shown as underlined.

Can LLMs Handle Low-Resource Dialects? A Case Study on Translation and Common Sense Reasoning in Šariš

Viktória Ondrejová and **Marek Šuppa**
Department of Applied Informatics, Comenius University in Bratislava
NaiveNeuron
Cisco Systems
ondrejova18@uniba.sk, marek.suppa@fmph.uniba.sk

Abstract

While Large Language Models (LLMs) have demonstrated considerable potential in advancing natural language processing in dialect-specific contexts, their effectiveness in these settings has yet to be thoroughly assessed. This study introduces a case study on Šariš, a dialect of Slovak, which is itself a language with fewer resources, focusing on Machine Translation and Common Sense Reasoning tasks. We employ LLMs in a zero-shot configuration and for data augmentation to refine Slovak-Šariš and Šariš-Slovak translation models. The accuracy of these models is then manually verified by native speakers. Additionally, we introduce ŠarišCOPA, a new dataset for causal common sense reasoning, which, alongside SlovakCOPA, serves to evaluate LLM's performance in a zero-shot framework. Our findings highlight LLM's capabilities in processing low-resource dialects and suggest a viable approach for initiating dialect-specific translation models in such contexts.

1 Introduction

The recent explosion of development in the field of Large Language Models (LLMs) has offered an unprecedented set of capabilities in understanding, generating, translating and transforming text across a large number of contexts (Min et al., 2023). However, despite their wide-ranging applications, the effectiveness of LLMs in dialect-specific scenarios, particularly in languages with limited resources, remains a relatively unexplored domain. This gap in research presents a critical challenge, as dialects incorporate distinct linguistic traits and cultural subtleties, yet comprehensive large-scale datasets like newswire texts are not available for them.

This study aims to address this gap by focusing on Šariš, a Slovak dialect, with pronounced linguistic variety as shown in Table 1. As Slovak is a less-resourced language itself, it presents an interesting case for examining how large language models (LLMs) perform in specific dialect contexts where data is scarce. We focus on two key natural language processing (NLP) tasks: Machine Translation (MT) and Common Sense Reasoning (CSR), which we view as representative for assessing the model's ability to handle the complexities of real-world language.

English	I left the potatoes in the fridge.
Slovak	Nechal som zemiaky v chladničke.
Šariš	Ochabil som gruľe v chladňičke.
	Zochabil som bandurki v ľadňičke.

Table 1: An example of expressing a singular statement through various linguistic constructions in the Šariš dialect. Note that both of the listed examples were deemed valid and reasonable by a native speaker.

In terms of MT, we investigate how LLMs can aid in translating between Slovak and the Šaris dialect. Here the LLMs are first used in zero-shot setting, meaning that we assume that (to the best of our knowledge) the models are not directly trained with Šariš-specific data but are instead expected to apply their knowledge of Slovak to understand and translate Šariš. We use this approach both to evaluate the performance of LLMs on the Slovak → Šariš and Šariš → Slovak translation task as well as for data augmentation, which results in about 3,500 automatically translated Slovak-Šariš sentence pairs. These are then used to finetune a specific Slovak-Šariš translation model, whose performance is evaluated on a manually labelled test set.

Additionally, we further introduce a new dataset called ŠarišCOPA, designed to evaluate the model's performance in CSR tasks specifically in the Šariš dialect. This dataset is intended to complement an existing dataset for Slovak, SlovakCOPA, to compare how the models perform in understanding both the standard language and its dialect. In this case the LLM is first prompted to only output the CSR

Proceedings of the Eleventh Workshop on NLP for Similar Languages, Varieties, and Dialects (VarDial 2024), pages 130–139
June 20, 2024

classification directly while additional experiments with a prompt-specific "translate-test" approach are also evaluated.

Our contributions can thus be summarized as follows:

- We introduce the first Slovak-Šariš translation dataset and use it to finetune a Šariš specific Machine Translation model
- We manually evaluate the quality of the translations produced by the finetuned model, as well as leading LLMs
- We introduce the ŠarišCOPA dataset and use it to evaluate the common sense reasoning performance of LLMs in Šariš
- We experiment with various LLM prompting approaches for ŠarišCOPA, including translation to English and Slovak

We release the code and data associated with our experiments in the hopes of fostering possible future research in this area at `https://github.com/NaiveNeuron/saris`.

2 Slovak and its Dialects

Despite being a relatively small language in terms of the number of native speakers (roughly 5 million native speakers), Slovak has multiple dialects. In this work, we focus on the eastern part of Slovakia, where the majority of population speak in a multiple dialects from the Šariš, Spiš, Zemplín regions. Even though we categorize these dialects to distinct groups, their historical, phonetic and lexical features are intertwined. A substantial overlap exists in lexical terms between dialects, with minimal variance observed (Pavlíková, 2016). Additionally, instances occur where native speakers interchange words from different dialects within the same discourse. Given these linguistic dynamics, in pursuit of maximizing corpus size, we considered amalgamation of all 3 of the dialects eligible for extraction.

The **Šariš** dialect holds notable significance within the family circle of the Prešov region, where a substantial portion of the population consistently employs it in their daily interactions. Specifically, statistics published in (Vodičková, 2009) reveal that approximately 22.5% of the population, amounting to roughly 180 thousand speakers, within the Prešov region utilize the Šariš dialect as their primary mode of communication. From the broader perspective Šariš, as an Eastern Slovak dialect, is classified as "Vulnerable" by the UNESCO Atlas of the World's Languages in Danger (Moseley, 2010).

2.1 Šariš-Specific Challenges

The dialect lacks a formal codification, leading to an absence of definitive linguistic rules governing their usage in speech and writing. Consequently, dialectal variations manifest across different areas which can be as small as villages, resulting in multiple potential translations for a single word within the same dialect. An example of this phenomenon can be seen in Table 1.

Conversely, Eastern Slovakian dialects exhibit distinct features. Unlike standard Slovak, these dialects lack long vowels. The Slovak "ď" ([d^j] in IPA) and "ť" ([t^j] in IPA) are replaced by "c" ([ts] in IPA) and "dz" ([dz] in IPA), respectively. Most importantly, however, a majority of Eastern Slovakian dialects, including those of the Šariš region, do not include the vowel "y".

Another challenge arises from the fact that certain highly specific terms either cannot be adequately translated into Slovak or risk losing their intended meaning. Additionally, the Šariš dialect incorporates numerous archaic expressions that have fallen out of common usage, making them potentially incomprehensible to some speakers.

3 The Translation Task

Our aim with the translation task is to validate to what extent are the findings of (Gu et al., 2018) still relevant, which found that less than 13k sentence pairs are not enough to train a neural machine translation model to reasonable quality. To this end, we introduce the ŠarišSet corpora with the help of a LLM.

3.1 ŠarišSet

Creating a corpus for a new language presents substantial challenges. The ŠarišSet dataset, containing over 4,000 sentences in the Šariš dialect from Eastern Slovakia, was compiled from various online sources. To ensure a solid benchmark, a subset of 500 sentences received manual translation by three native speakers[1]. The bulk of the dataset was translated through a hybrid method combining prompt engineering with manual review of outputs

[1]Here, "native speaker" refers to someone fluent in the Šariš dialect with extensive exposure from childhood.

	Šariš	Slovak
vocabulary size	3560	3647
Q1	11	11
Median	16	17
Q3	23	23
Mean	18.98	19.23
SD	11.46	11.77

Table 2: The table shows the quantitative statistics of the ŠarišSet dataset as vocabulary size for the source and target languages, as well as the Q1, Q2, Median and Mean of the number of words per sentence. In addition, the standard deviation is displayed in the SD row.

from GPT-3.5-Turbo and GPT-4 (Achiam et al., 2023).

The Table 2 shows the aggregated statistics related to the dataset, such as its vocabulary size and quantitative statistics for the introduced dataset.

Extraction In order to gather data for a dialect of a low-resource language spoken by only a few tens of thousands of individuals, the conventional automated methodology proved unfeasible. With scarce online resources beyond traditional folk songs, the absence of suitable web pages for scraping presented a challenge. Šariš texts are predominantly confined to a handful of niche blogs and sporadic Facebook posts. To avoid the complexities of the Facebook interface, our focus was directed solely towards the identified blogs outlined in Appendix A, discovered through extensive online searches (mainly by searching a very specific word in the dialect), alongside the aforementioned folk songs which could be systematically scraped using the `scrapy` library in Python[2].

Throughout the scraping process, filtering criteria were implemented. The native speaker visually inspected the texts, reviewing the initial and final two sentences. If the sentences appeared plausible, with words in their proper positions and the structure intact, the text was kept and saved. The acquired data subsequently underwent a cleaning process via a script designed to remove duplicates, highly offensive language, extraneous characters, and segment the text into coherent sentences.

The final sentences originate from 133 various longer texts obtained from multiple blogs, together with more than 170 folk songs.

[2] https://scrapy.org/

Automatic Translation Given the laborious nature of manual translation, we opted to employ the GPT-3.5-Turbo and GPT-4 models for translating the remaining sentences, comparing their performance using various prompt engineering techniques.

Initially, we focused on the GPT-3.5-Turbo model, experimenting with three distinct prompts. The first prompt, applied to both models, was straightforward as we can see in Figure 1.

```
translate to Slovak
```

Figure 1: The first simplest prompt used for translation.

We further tested a more nuanced prompt, encouraging the model by stating that even an inaccurate translation would be beneficial (see Figure 2).

```
Please, try to translate this into Slovak, even an inaccurate version would help a ton
```

Figure 2: The second prompt used for translation.

Finally, we utilized a persona-based approach, directing the model to take on the role of a bilingual eastern Slovak youth proficient in translating dialects into Slovak. The prompt, visible in Figure 3, presented a scenario where the model was a native Šariš dialect speaker conversing with someone unfamiliar with it.

```
You're an eastern Slovak young man who has lived in one village his entire life. Though you are proficient in Slovak due to schooling, at home with your family, you speak in the eastern Slovak dialect known as Šariština. You've introduced a girl from central Slovakia, fluent in Slovak but unfamiliar with Šariština, to your family and need to provide the most accurate translation of this sentence into Slovak
```

Figure 3: The third prompt used for translation that employed the persona-based approach.

A selection of model responses is illustrated in Table 8.

We manually evaluated these results across 50 sentences, selecting the most suitable translation from the three generated ones. Surprisingly, the translations from Prompt 2 proved to be highly comparable to those from Prompt 3, despite the added narrative context. Ultimately, we chose the second prompt due to fewer instances of extraneous words in the final outputs.

When it finally came to the translating the reminder of ”ŠarišSet”, it was necessary to decide between using GPT-3.5-Turbo and GPT-4. Utilizing a similar approach as above, we evaluate the results of each on 50 sentences and concluded that GPT-4 is a better fit for this sort of a translation task and was used to translate the remaining 3,500 sentences using the Prompt 2 chosen before. The same prompt was then used for translation of the test set as well. Additional details on how these models were accessed can be found in Section B.

3.2 NLLB-Based Model

In the very first iteration we experimented with the mBART model (Tang et al., 2020), specifically the mBART-50 version that was created by multilingual fine-tuning. Perhaps owing to the fact that Slovak was not included in the languages it was pretrained on, the model tended to collapse to outputting a single word and not being useful at all.

As an alternative to the mBART model, we also experimented with the NLLB-200 model which was created as part of the No Language Left Behind project (Costa-jussà et al., 2022). The aim of this project is to provide open-source models ”capable of delivering evaluated, high-quality translations directly between 200+ languages – including low-resource languages”[3]. The list of 204 languages does not include Šariš but as opposed to mBART, it does include Slovak (which (Costa-jussà et al., 2022) even lists as being high resourced on page 15 in Table 1) and hence we opted to experiment with using it as the basis for the Šariš → Slovak and Slovak → Šariš translation models. We did so by adding a new ”pseudo language” tag `sar_Latn` to the model and finetuning it on the dataset introduced in Section 3.1. We finetuned the model, specifically its `nllb-200-distilled-600M` version[4], with the batch size of 16, 500 warm up steps and 20 000 training steps. Additionally, the maximum output length was set to 128.

[3]https://ai.meta.com/research/no-language-left-behind/

[4]https://huggingface.co/facebook/nllb-200-distilled-600M

	Š → S		S → Š	
	F	**A**	**F**	**A**
GPT-3.5-Turbo	2.96	3.15	1.02	1.23
GPT-4	**3.45**	**3.51**	1.17	1.57
NLLB	3.09	3.00	**3.16**	**3.80**

Table 3: The average fluency (F) and adequacy (A) obtained during evaluation of various models and translation directions. Š represents Šariš and S represents Slovak. The best result per each metric and language pair is boldfaced.

3.3 Evaluation

In our experimental framework, we utilize adequacy and fluency metrics (Chatzikoumi, 2020) to manually evaluate the outputs generated by the machine translation models. Each output, corresponding to a given source text, underwent assessment by an annotator on a graded scale ranging from 1 to 5, where the higher numbers represent better adequacy and fluency.

In terms of adequacy, we are primarily concerned with whether the output effectively conveys the same meaning as the input sentence. We evaluate whether any part of the original message is lost, added, or distorted during the translation process. Therefore, the rating of 5 signifies preservation of all semantic aspects from the source text, whereas a score of 1 indicates complete loss of meaning.

Regarding fluency, our focus lies in assessing whether the output exhibits fluent expression in the target language. This entails considerations of grammatical correctness and the use of idiomatic word choices to ensure that the translated text reads naturally and smoothly. Likewise, a fluency score of 5 indicates seamless language coherence in alignment with the intended target output, whereas a score of 1 suggests incomprehensibility.

During evaluation, we conducted comparisons between the translated sentences. If a text contained 1-2 errors (untranslated words, mismatched case ending and so on), it would receive a score of 4. Conversely, if the translated sentence exhibited only 1-2 accurately translated words and rest was implausible, it would be awarded a score of 1, and so forth.

The evaluation, conducted by a native speaker and detailed in Table 3, indicates that GPT-4 excelled in translating from Šariš to Slovak, while the NLLB model reported the best performance in the opposite direction. Notably, both GPT-3.5-Turbo

	PREMISE		CHOICE 1	CHOICE 2
sk en	*Vonku sa zotmelo.* It got dark outside.	R	*Z oblohy začali padať snehové vločky.* Snowflakes began to fall from the sky.	*Na oblohe sa objavil mesiac.* The moon became visible in the sky.
šr en	*Šľisknul som še na žemi.* I slipped on the floor.	C	*Kachľička bula prasknuta.* The tile was cracked.	*Kachľička bula morka.* The tile was wet.

Table 4: Examples of forward (Result [R]) and backward reasoning (Cause [C]) in the COPA, SlovakCOPA and ŠarišCOPA validation sets. Note that Šariš is denoted as šr in the list of languages.

and GPT-4 showed poor performance in translating from Slovak to Šariš, indicating a challenge in producing coherent Šariš output. Conversely, GPT-4's superior performance in translating to Slovak, surpassing even the fine-tuned NLLB model, underscores the importance of language-specific proficiency in LLM-based translation.

4 The Common Sense Reasoning Task

To gauge the effectiveness of natural language processing (NLP) systems in understanding different languages, it is crucial to employ various testing methods. Common sense reasoning evaluation is particularly significant, as it is a fundamental aspect of these systems, underscored by previous research (Davis and Marcus, 2015). The Choice Of Plausible Alternatives (COPA) serves as a notable benchmark, testing systems' ability to decipher cause-and-effect relationships in English sentence pairs (Roemmele et al., 2011). Due to its acclaim, COPA has been expanded into multiple languages through the XCOPA benchmark (Ponti et al., 2020) and adapted for Slavic languages such as Slovenian (Ljubešić et al., 2022a), Serbian (Ljubešić et al., 2022b), and Croatian (Ljubešić, 2021). Our study introduces the ŠarišCOPA dataset, focusing on the Šariš dialect.

4.1 ŠarišCOPA

The COPA framework is generally implemented as a binary classification challenge. Models must choose the more plausible scenario from two options, based on a given premise and question. This assessment distinguishes between cause and effect in scenarios: "cause" questions ask for the reason behind an event while "effect" questions seek the consequence of an event.

The ŠarišCOPA dataset, designed to test LLMs' common sense reasoning in Šariš, consists of 500 test and 100 validation triplets, each with a premise and two choices. It adapts the original English COPA, following XCOPA's translation methodology (Ponti et al., 2020), with the translation work carried out by native speakers from the ŠarišSet project. Additionally, we compare results with the SlovakCOPA dataset, created by a professional translator using a similar method. The format and examples of these datasets are displayed in Table 4.

4.2 Evaluation

Our evaluation of the SlovakCOPA and ŠarišCOPA datasets began with comparing native speaker labels to those from the original COPA dataset, revealing a 100% match in both cases.

Subsequently, we tested GPT-3.5-Turbo and GPT-4 on these datasets using specific prompts for the "cause" as well as the "efect" scenario. These prompts were inspired by the prompts used by "BENCHić - the benchmark for Bosnian, Croatian, Montenegrin, Serbian (and friends)"[5]. They were designed to minimize the amount of noise in the responses of LLMs and their full text can be seen below:

COPA Prompt: Cause

Given the premise "premise", and that we are looking for the cause of this premise, which hypothesis seems more plausible?
Hypothesis 1: "hypothesis1".
Hypothesis 2: "hypothesis2".
Please answer only with "1" or "2".

COPA Prompt: Effect

Given the premise "premise", and that we are wondering what happened as a result of this premise, which hypothesis seems more plausible?
Hypothesis 1: "hypothesis1".
Hypothesis 2: "hypothesis2".
Please answer only with "1" or "2".

[5] This benchmark can be found at `https://github.com/clarinsi/benchich/tree/main/copa`

As Table 5 shows, GPT-3.5-Turbo performed well on SlovakCOPA (76.6% accuracy) but struggled with ŠarišCOPA (55.4% accuracy, near random chance). GPT-4 showed remarkable performance on SlovakCOPA (96.6% accuracy) and significantly outperformed GPT-3.5-Turbo on ŠarišCOPA (79.8% accuracy), albeit with a 4.8% rate of unparseable responses, such as ”*As an AI language model, I'm unable to understand the premise and hypotheses because they are not in a recognizable language or a standard linguistic structure. Therefore, I can't determine which hypothesis is more plausible.*”.

We also tested a method where the model first translates the input into a more resource-rich language before making a prediction. This approach, inspired by the performance of GPT-4 in Šariš to Slovak translation and recent research on multilinguality in LLMs (Liu et al., 2024) and cross-lingual transfer (Ebing and Glavas, 2023), involved slightly modified prompts for translation into English and Slovak, which can be found below.

COPA Prompt: Cause with translation

Given the premise ”premise”, and that we are looking for the cause of this premise, which hypothesis seems more plausible?
Hypothesis 1: ”hypothesis1”.
Hypothesis 2: ”hypothesis2”.

First translate the premise and the hypotheses to English, then answer only with ”Prediction: 1” or ”Prediction: 2”.

COPA Prompt: Effect with translation

Given the premise ”premise”, and that we are wondering what happened as a result of this premise, which hypothesis seems more plausible?
Hypothesis 1: ”hypothesis1”.
Hypothesis 2: ”hypothesis2”.

First translate the premise and the hypotheses to English, then answer only with ”Prediction: 1” or ”Prediction: 2”.

The results, labeled ”+ translate en” and ”+ translate sk” in Table 5, showed that translating to English improved GPT-3.5-Turbo's performance on SlovakCOPA (from 76.6 to 88.0) and ŠarišCOPA (from 55.4 to 71.0), with a slight increase for GPT-4 on ŠarišCOPA (from 79.8 to 82.0). Translating to Slovak yielded less pronounced improvements. Interestingly, the number of unparseable responses increased, including ”*The text provided is not in a recognizable language, therefore it cannot be translated or used to make a prediction.*” in English and ”*The premise and hypotheses are already in Slovak, but they are written in a dialect or with many spelling mistakes, making them difficult to understand. Therefore, it is impossible to make a prediction.*” in Slovak, hinting at GPT-4's ability to recognize Šariš as a Slovak dialect.

Model	SlovakCOPA	ŠarišCOPA
GPT-3.5-Turbo	76.6 (0.0)	55.4 (0.0)
+ translate en	**88.0** (0.2)	**71.0** (0.4)
+ translate sk		70.0 (0.4)
GPT-4	**96.6** (0.0)	79.8 (4.8)
+ translate en	**96.6** (0.0)	**82.0** (8.6)
+ translate sk		81.6 (8.8)

Table 5: The accuracy of GPT 3.5 Turbo and GPT 4 on the SlovakCOPA and ŠarišCOPA datasets. The number in parentheses denotes the number of responses that we were unable to parse. The best performing model in a specific model family on a particular dataset is boldfaced.

5 Discussion

This study aimed to assess the proficiency of large language models (LLMs) in processing the Šariš dialect, a low-resource variant of Slovak. Our investigation, detailed in Section 3, showcased GPT-4's ability to translate between Slovak and Šariš, albeit with varying success, particularly in Šariš-targeted translations. Enhancing the NLLB model with GPT-4's Šariš translations significantly improved its performance, outstripping GPT-3.5-Turbo in Slovak to Šariš translation accuracy and surpassing both GPT iterations in the opposite direction. This indicates that leveraging LLMs for initial translations can create a solid foundation for building effective translation tools for underrepresented dialects, as demonstrated by our results with just 3,500 sentences.

Furthermore, as detailed in Section 4, translation plays a crucial role in the Common Sense Reasoning Task. Having models translate inputs to English or Slovak before making inferences improved the outcomes for both GPT-3.5-Turbo and GPT-4, with English translations being marginally more effective. Intriguingly, GPT-4 occasionally declined to make predictions, identifying inputs as specific Slovak dialects or variants, indicating its potential in dialect recognition, despite limitations in dialect generation.

In summary, our experiments illustrate that LLMs have the potential to be instrumental in handling dialects with scarce resources. By integrating strategic prompting with LLMs, we cannot only enhance model performance but also empower subsequent models trained on the data produced, setting a promising direction for future research in NLP for low-resource dialects.

6 Related Work

Machine translation for low-resource languages and dialects has been an active area, often leveraging transfer learning from high-resource languages (Tars et al., 2021; Maimaiti et al., 2019). Dialect translation has been studied for Arabic (Harrat et al., 2019), German (Honnet et al., 2018), Portuguese (Costa-jussà et al., 2018) and French, Croatian, Serbian and Malay (Lakew et al., 2018) dialects, finding substantial data in the dialect language is beneficial.

The Choice of Plausible Alternatives (COPA) dataset (Roemmele et al., 2011) has been widely used to evaluate commonsense causal reasoning in English, and has further been translated into 11 languages, including resource-poor languages like Haitian Creole as part of XCOPA (Ponti et al., 2020) and separately into Slavic languages as well (Ljubešić, 2021; Ljubešić et al., 2022a,b). Analysis has found translate-test approaches can boost performance over zero-shot cross-lingual transfer (Artetxe et al., 2023), aligning with our findings. Our ŠarišCOPA dataset provides a new test for reasoning in a low-resource dialect context.

While Slovak is considered a lower-resource language compared to major world languages, there has been some prior work on developing NLP tools and resources. This includes machine translation systems focused on European languages (Popel, 2018), pre-trained language models like SlovakBERT (Pikuliak et al., 2022) and annotated datasets for tasks like named entity recognition (Suba et al., 2023) and question answering (Hládek et al., 2023). However, work specifically targeting Slovak dialects like Šariš has been very limited. Perhaps the closest work to ours would be (Darjaa et al., 2018) in which the authors conduct a preliminary analysis on the distinguishability of Slovak dialects in spoken language and introduce the Sound Archive of Slovak Dialects – roughly 150 hours of recordings which include all basic Slovak dialects. To the best of our knowledge, our work is the first to investigate the use of Natural Language Processing specifically on texts in Slovak dialects.

7 Conclusion

This study assesses LLMs' abilities in translating and understanding the Šariš dialect through machine translation and common sense reasoning tasks, introducing the ŠarišCOPA dataset. While LLMs show proficiency in translating from Šariš to Slovak, reverse translations pose challenges. The inclusion of translation as a preprocessing step improved common sense reasoning performance, particularly notable when comparing results on ŠarišCOPA with SlovakCOPA. These findings highlight the potential and limitations of LLMs in processing and reasoning in low-resource dialects. The code and data associated with our experiments can be found at `https://github.com/NaiveNeuron/saris`.

Limitations

Data Scarcity Despite our efforts, the amount of Šariš data we could obtain remains very limited compared to standard benchmarks for high-resource languages. The ŠarišSet corpus contains only around 4,000 sentences, and ŠarišCOPA has just 600 examples. This scarcity makes it difficult to fully assess LLM capabilities and prevents training extremely high-performing dialect-specific models from scratch. Obtaining more in-domain data would strengthen future analyses.

Human Evaluation Our human evaluations of translation quality and the ŠarišCOPA dataset drew upon a limited number of native Šariš speakers. While we took care to involve highly proficient speakers, from multiple parts of the Šariš region, inherent subjectivity in such evaluations means the ratings may not fully generalize. A larger evaluation involving more speakers would increase confidence. Additionally, no standard evaluation datasets exist for Šariš, preventing benchmarking against prior work.

Model Limitations The prominent LLMs like GPT-3, GPT-4, and NLLB that we evaluated are large models trained primarily on text from high-resource languages. While their pretraining data likely contained little-to-no examples of low-resource dialects like Šariš, it is difficult to claim that with certainty – particularly for models that are not publicly released, which further hinders the reproducibility of our experiments.

Reasoning About Dialect While our ŠarišCOPA probe provides a window into LLM's commonsense reasoning abilities for the dialect, the examples come from a single constructed dataset. Drawing broader conclusions about general language understanding of Šariš from this limited test would be an overreach. More comprehensive benchmarks probing other core language skills are needed.

Acknowledgements

This research was partially supported by grant APVV-21-0114. Last but not least, we would like to thank our annotators, in particular Pavol Hudák, Daniel Haraksim, Matej Novotný and Matúš Kobyl'an from Technical University of Košice (TUKE) for their help with the Šariš to Slovak and English to Šariš translation and Terézia Šuppová for her help with English to Slovak translation.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Mikel Artetxe, Vedanuj Goswami, Shruti Bhosale, Angela Fan, and Luke Zettlemoyer. 2023. Revisiting machine translation for cross-lingual classification. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6489–6499, Singapore. Association for Computational Linguistics.

Eirini Chatzikoumi. 2020. How to evaluate machine translation: A review of automated and human metrics. *Natural Language Engineering*, 26(2):137–161.

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Marta R. Costa-jussà, Marcos Zampieri, and Santanu Pal. 2018. A neural approach to language variety translation. In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018)*, pages 275–282, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Sakhia Darjaa, Róbert Sabo, Marián Trnka, Milan Rusko, and Gabriela Múcsková. 2018. Automatic recognition of slovak regional dialects. In *2018 World Symposium on Digital Intelligence for Systems and Machines (DISA)*, pages 305–308. IEEE.

Ernest Davis and Gary Marcus. 2015. Commonsense reasoning and commonsense knowledge in artificial intelligence. *Commun. ACM*, 58(9):92–103.

Benedikt Ebing and Goran Glavas. 2023. To translate or not to translate: A systematic investigation of translation-based cross-lingual transfer to low-resource languages. *ArXiv*, abs/2311.09404.

Jiatao Gu, Hany Hassan, Jacob Devlin, and Victor O.K. Li. 2018. Universal neural machine translation for extremely low resource languages. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 344–354, New Orleans, Louisiana. Association for Computational Linguistics.

Salima Harrat, Karima Meftouh, and Kamel Smaili. 2019. Machine translation for arabic dialects (survey). *Information Processing & Management*, 56(2):262–273.

Daniel Hládek, Ján Staš, Jozef Juhár, and Tomáš Koctúr. 2023. Slovak dataset for multilingual question answering. *IEEE Access*, 11:32869–32881.

Pierre-Edouard Honnet, Andrei Popescu-Belis, Claudiu Musat, and Michael Baeriswyl. 2018. Machine translation of low-resource spoken dialects: Strategies for normalizing Swiss German. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Surafel Melaku Lakew, Aliia Erofeeva, and Marcello Federico. 2018. Neural machine translation into language varieties. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 156–164, Brussels, Belgium. Association for Computational Linguistics.

Chaoqun Liu, Wenxuan Zhang, Yiran Zhao, Anh Tuan Luu, and Lidong Bing. 2024. Is translation all you need? a study on solving multilingual tasks with large language models.

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

Mieradilijiang Maimaiti, Yang Liu, Huanbo Luan, and Maosong Sun. 2019. Multi-round transfer learning for low-resource nmt using multiple high-resource languages. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 18(4).

Bonan Min, Hayley Ross, Elior Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heintz, and Dan Roth. 2023. Recent advances in natural language processing via large pre-trained language models: A survey. *ACM Comput. Surv.*, 56(2).

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

Michaela Pavlíková. 2016. Východoslovenské nářečí v psaném textu. SUPERVISOR: prof. PhDr. Marie Krčmová, CSc.

Matúš Pikuliak, Štefan Grivalský, Martin Konôpka, Miroslav Blšták, Martin Tamajka, Viktor Bachratý, Marian Simko, Pavol Balážik, Michal Trnka, and Filip Uhlárik. 2022. SlovakBERT: Slovak masked language model. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 7156–7168, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. Xcopa: A multilingual dataset for causal commonsense reasoning. *arXiv preprint arXiv:2005.00333*.

Martin Popel. 2018. CUNI transformer neural MT system for WMT18. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 482–487, Belgium, Brussels. Association for Computational Linguistics.

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*.

David Suba, Marek Suppa, Jozef Kubik, Endre Hamerlik, and Martin Takac. 2023. WikiGoldSK: Annotated dataset, baselines and few-shot learning experiments for Slovak named entity recognition. In *Proceedings of the 9th Workshop on Slavic Natural Language Processing 2023 (SlavicNLP 2023)*, pages 138–145, Dubrovnik, Croatia. Association for Computational Linguistics.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning.

Maali Tars, Andre Tättar, and Mark Fishel. 2021. Extremely low-resource machine translation for closely related languages. *CoRR*, abs/2105.13065.

Zuzana Vodičková. 2009. Šarišská nárečová lexika.

A Data Sources

The list below outlines the domains which were ultimately used for extraction of the ŠarisSet. The majority of the sentences were obtained from various local newspapers, blogs and folk tales found on the following internet pages:

- https://www.obeckrivany.sk/
- https://bandzone.cz/
- https://www.ilonas.net/valal/
- https://prerag.sk/
- https://blog.pravda.sk/

Similarly, for obtaining the texts of various Šariš folk songs, the following domains were scraped:

- https://narecie.sk/
- https://www.videorohal.com/

B Details on Accessing GPT-3.5-Turbo and GPT-4

All models were accessed via the AzureOpenAI endpoints[6], with the API version being set to `2023-07-01-preview` and the temperature=0 to aid reproducibility.

C GPT-4 Translations

The examples of good and bad translations from Šariš to Slovak obtained from the GPT-4 model oaired with the Prompt 2 can be seen in Table 6 and Table 7, respectively.

D Translation Prompts

In Table 8 we can see the original sentence in Šariš and its Slovak manual translation; together with the translations obtained from GPT-3.5-Turbo model using Prompts 1, 2 and 3.

[6] `https://learn.microsoft.com/en-us/azure/ai-services/openai/reference`

Šariš dialect	**Slovak**
”Ta ňeznam,” - hvarim ja jej, ”ja ňemam kedi, hibaľ na večar?”	”Ja neviem,” hovorím jej, ”ja nemám kedy, možno večer?”
Dok sme buľi malki, naša mama nam veľo času pošvecovala.	Keď sme boli malí, naša mama nám venovala veľa času.
Počali sme medži sobu bešedovac.	Začali sme medzi sebou rozhovor.
Dovidzeňa i ščešľivo - skričal som jej.	”Dovidenia a šťastlivo,” zakričal som na ňu.
Šofer po nas vriskal, naj śe sciśime, bo ho to ruši.	Vodič na nás kričal, aby sme prestali vrieskať, pretože ho to rušilo.

Table 6: Examples of a bad translation from Šariš dialect to Slovak obtained from the GPT-4 model.

Šariš dialect	**Slovak**
Ona vichovala parobkom frajirku.	Ona obdaruvala chlapca drahým darčekom.
Taže me ňemaľi aňi dze śedzec, aňi dze spac.	Teda ma nemajú ani na seďať, ani na spať.
Jak me prišľi na chižu, ta me muśeľi stolok podkladac.	Keď ma priviezli na pohotovosť, museli mi podkladať stôl.

Table 7: Examples of a good translation from Šariš dialect to Slovak obtained from the GPT-4 model.

	Translations
Šaris	Heslo jej babi bulo: ”ked na dakoho ňemožeš patric zos očmi, popatri na ňho zos šercom”
Slovak	Heslo jej babky bolo: ”Keď na niekoho nemôžeš pozrieť s očami, pozri naňho so srdcom”.
Prompt 1	Heslo jej babičky bolo: ”Keď na niekoho nemôžeš hľadieť očami, pozri naňho srdcom.”
Prompt 2	Heslo jej babičky bolo: ”Keď na niekoho nemôžeš pozerať očami, pozri naňho srdcom.”
Prompt 3	Heslo jej babičky bolo: ”Keď na niekoho nemôžeš pozerat z očí, popozri na neho z srdcom”.
Šariš	Poneže še M. ľubel chvaľic, i ponosovac, zos veľima stvarami, vše ked prichodzel.
Slovak	Pretože sa M. rád chválil aj povyšoval s veľkými svárami všade kde prišiel.
Prompt 1	Áno, napríklad, oni sa hnevali na M., keď im niečo prekladal, lebo, áno, všetko rozumeli.
Prompt 2	Pretože sa M. ľúbil chváliť a pýšiť sa mnohými vecami, keďže prichádzal.
Prompt 3	Tak, napríklad, oni sa hnevali na M., keď im niečo prekladal, pretože, áno, všetko rozumeli.
Šariš	”Ta ňeznam,”- hvarim ja jej, ”ja ňemam kedi, hibaľ na večar?”
Slovak	”Tak neviem” - hovorím jej ja, ”ja nemám kedy, možno večer?”
Prompt 1	”Ta neviem,”- pýtam sa jej, ”ja nemám plán, ideš večerať?”
Prompt 2	”Ta neviem,” - hvarím jej, ”ja nemám mačku, alebo ísť večer?”
Prompt 3	”Neznamená to,” hovorím jej, ”ja nemám práve teraz čas, ideme večerať?”

Table 8: Comparison of Šariš sentences with Slovak translations and translations from Prompts 1, 2, and 3 obtained from GPT-3.5-Turbo.

Experiments in Multi-Variant Natural Language Processing for Nahuatl

Robert Pugh and **Francis M. Tyers**
Indiana University, Bloomington
Department of Linguistics
pughrob@iu.edu, ftyers@iu.edu

Abstract

Linguistic variation is a complicating factor for digital language technologies. This is particularly true for languages that lack an official "standard" variety, including many regional and minoritized languages. In this paper, we describe a set of experiments focused on multi-variant natural language processing for Nahuatl, an indigenous Mexican language with a high degree of linguistic variation and no single recognized standard variant. Using small (10k tokens), recently-published annotated datasets for two Nahuatl variants, we compare the performance of single-variant, cross-variant, and joint training, and explore how different models perform on a third Nahuatl variant, unseen in training. These results and the subsequent discussion contribute to efforts of developing low-resource NLP that is robust to diatopic variation. We share all code used to process the data and run the experiments.[1]

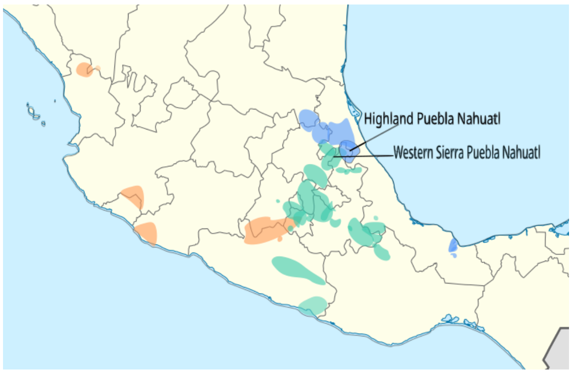


Figure 1: A map approximating the location of many of the Nahuatl variants spoken in Mexico. The colors correspond to the division defined in Kaufman (2001), blue for the Eastern branch, Turquoise for the Central branch, and Orange for the Western branch. We label the two variants for which we have training data in the form of UD treebanks. Importantly, this map is an approximation, and does not claim to represent every Nahuatl variant.

1 Introduction

Linguistic variation, though a ubiquitous feature of human language, is a complicating factor for digital language technologies. While natural language processing (NLP) has made significant advances in recent years, the "dialect gap," which refers to the drop in performance of NLP systems on non-standard linguistic varieties, remains (Kantharuban et al., 2023). In many cases, non-standard, low-resource variants are similar or related to a more uniform, standard variety with a greater number of linguistic resources. One popular approach to remedy this problem is to leverage a high-resource standard variant in concert with data augmentation methods to train models on a similar non-standard variant (Zampieri et al., 2020).

However, the case of a related, high-resource standard variant is not the only linguistic situation that speakers and writers of non-standard variants find themselves in. On the contrary, there are numerous distinct dialect situations across the world. In a treatment of such scenarios in Europe, Auer (2011) identifies a useful typology for thinking about the diversity of language situations with respect to standard languages and dialectal variation. Relevant to the present paper, this typology includes *exoglossic diglossia* or "Type 0", which describes a situation of multiple non-standard variants without any endoglossic standard. In these cases, if a standard variety does exist it is viewed as imported or significantly different from the vernacular dialects.

In the absence of a spoken or written standard variety ("Type 0"), in particular when there is little available annotated linguistic data for the non-standard varieties, developing digital language technologies robust to diatopic language variation is a particularly important and valuable objective.

[1] https://github.com/Lguyogiro/multidialectal-nlp-nahuatl

Proceedings of the Eleventh Workshop on NLP for Similar Languages, Varieties, and Dialects (VarDial 2024), pages 140–151
June 20, 2024

Nahuatl, a group of approximately 30 language varieties spoken in Mexico and Central America (Described in further detail in Section 2), fits the "Type 0" characterization quite well, given that there are a large number of recognized varieties and no single standard[2]. There also exists a vast body of literature in the language written primarily in historic Nahuatl varieties from the early colonial era, known as "Classical Nahuatl" (Gingerich; León-Portilla, 1985), to which speakers of contemporary Nahuatl varieties have little exposure.[3]

While these aspects of the Nahuatl language situation make it an interesting candidate for NLP research, they are not unique to Nahuatl. In fact, numerous indigenous language in Latin America fit the characterization of having many diatopically-diverse variants, no single contemporary standard, and a colonial-era written canon. Other examples include the Zapotec (Foreman and Lillehaugen, 2017; Flores-Marcial et al.; Hilts, 2003) and Quechuan (Luykx et al., 2016; Durston, 2008; Escobar, 2011) languages.

The present work evaluates a number of approaches to multi-variant NLP for Nahuatl. We leverage recently-published, relatively small Universal Dependencies corpora in two Nahuatl variants and compare monolingual model performance with that of cross-lingual and jointly-trained models, as well as the impact of leveraging multi-variant, unlabeled data by adding an auxiliary task during training.

Our goal in this effort is two-fold: (1) to set the stage for high-quality NLP models that support speakers of any variety of Nahuatl, leveraging their similarities, and (2) to inform similar efforts involving other languages in a similar dialect situation.

2 The Nahuatl Language Complex

Nahuatl is a polysynthetic, agglutinating Uto-Aztecan language spoken throughout Mexico and Mesoamerica. The Mexican Government's *Instituto Nacional de Lenguas Indígenas* (INALI) recognises 30 distinct Nahuatl varieties (INALI, 2009), with highly-variable levels of linguistic similarity and mutual intelligibility. Furthermore, linguistic similarity and mutual intelligibility is not always correlated with geographic distance, a fact that is due in part to multiple waves of migration of Nahuatl speakers leading speakers of different varieties to end up in close proximity to one another (Canger, 1988; Kaufman, 2001; Beekman and Christensen, 2003).

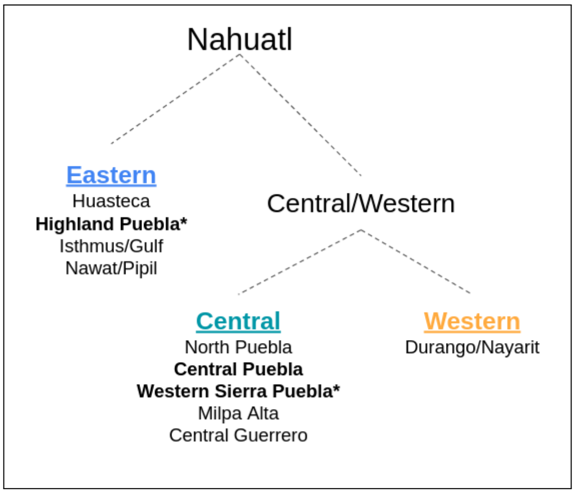


Figure 2: An abbreviated diagram of the sub-classification of Nahuatl variants, offering a glimpse at the taxonomic relationship between the variants we investigate here. The classification largely follows Kaufman (2001) using the same color-coding scheme in Figure 1. The variants used in this paper are bolded, and the two for which we have annotated training data are marked with an asterisk. The classification of the Central Guerrero variant follows (Lastra, 1986).

Dialectological research on Nahuatl dates back to at least (Lehmann, 1920). More recently, researchers largely converge around the dialect sub-classifications presented in (Lastra, 1986), (Canger, 1988), and (Kaufman, 2001) which, while not identical, agree on a number of important points, namely on the existence of Eastern Nahuatl varieties, which are thought to correspond to one wave of early migration, Central Nahuatl varieties, corresponding to the Nahuatl spoken in the valley of Mexico and in what is now Mexico City, and Western varieties, including Nayarit/Durango Nahuatl.

There is no unanimous consensus about the classification of Nahuatl variants, but for a number of cases there is widespread agreement (e.g. Pipil/Nawat of El Salvador and Sierra Puebla, or Highland Puebla, Nahuatl belonging to the Eastern group). Pharao Hansen (2014) provides some additional recommendations for the sub-classification between Eastern and Central/Western groups based on a survey of linguistic evidence. Nahuatl variants

[2]Alternatively, the label of "Pluricentric" (Clyne, 2012) may also be considered appropriate, though this typically refers to multiple standard, national languages, which is not the case of Nahuatl

[3]Interestingly, Sullivan (2011) describes a course with Nahuatl-speaking students focused on reading classical Nahuatl manuscripts, and notes that the students could read and understand it with little difficulty.

can differ at essentially every level of linguistic structure: Lexicon (e.g. *totoltetl* vs. *teksistli* "egg"), phonology (e.g. *e* vs. *i* (Canger and Dakin, 1985), *t–tl–l*, word-initial *e-* vs. *ye-*), morphology (e.g. the presence or absence of the "antecessive" *o-* for verbs in the past, the presence or absence of the perfective *-ki* suffix), and syntax (e.g. relative clauses (Pharao Hansen, 2015), and the order of certain adverbs with respect to verbs).

Additionally, since the invasion of Mexico in the 16th Century by the Spanish, Nahuatl has had close contact with Spanish, resulting in both in extensive "material borrowing" (Matras and Sakel, 2007) such as loanwords and new phonemes, but also a non-trivial amount of morphosyntactic "pattern borrowing" like syntactic calque, such as a development of the periphrastic future, and the development of adpositions from relational nouns (Farfán, 2008; Olko et al., 2018).

3 Related Work

Research on linguistic variation in NLP has recently become an important topic in the field, with now ten iterations of the Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial) (Scherrer et al., 2023), which has over the years included a number of important and relevant shared tasks, such as similar language detection (Aepli et al., 2022) and cross-lingual parsing (Zampieri et al., 2017). Scherrer and Rambow (2010) explores approaches to NLP for the Swiss German dialect area that leverage geographic information, weighting rules using knowledge about the distribution of variant features in different regions. Also working on the Swiss German dialect continuum, Aepli (2018) evaluates syntactic parsing approaches including annotation projection, for which a parallel corpus with standard German was compiled, and delexicalized parsing.

These two approaches, annotation projection (Hwa et al., 2005; Agić et al., 2016) and delexicalized parsing (Zeman and Resnik, 2008), are common methods for cross-lingual parsing of related languages. More recently, the use of multilingual embedding representations used with neural network architectures has been shown to be quite effective for multilingual parsing (Ammar et al., 2016), particularly with pretrained transformer language models such as multilingual BERT (Devlin et al., 2019), as demonstrated in, e.g. in Kondratyuk and Straka (2019). Abdul-Mageed et al. (2021) build a language-specific transformer (with a large volume of data), reporting improved performance on multiple NLP tasks for a number of Arabic dialects.

One straightforward approach to multi-variant parsing is cross-lingual model transfer, wherein a model is trained on one variety (typically the higher-resource, standardized variety), and used on a different, related variety (Zampieri et al., 2020). Alternatively, work on two Norweigian standard languages, Bokmål and Nynorsk, found that simply combining the training data for closely-similar languages produces better results than straightforward model-transfer (Velldal et al., 2017).

While Nahuatl dialectology has a rich tradition in the field of linguistics (see Section 2), computational work addressing linguistic variation in Nahuatl is harder to come by. Efforts in this area include Farfan (2019)'s detailed analysis of the similarities of contemporary Nahuatl writing (from multiple variants) with Classical Nahuatl using a finite-state morphological analyzer built for the latter language, and Pugh and Tyers (2021), which found that simple, character-based language models, when evaluated across variants, track well with variant groupings and mutual intelligibility.

4 Data

We use recently published, linguistically-annotated datasets for two Nahuatl varieties: Highland Puebla Nahuatl (alternatively Sierra Puebla Nahuat, ISO-639: `azz`) (Pugh and Tyers, 2024) and Western Sierra Puebla Nahuatl (alternatively Zacatlán-Ahuacatlán-Tepetzintla Nahuatl, ISO-639: `nhi`) (Pugh et al., 2022), both spoken in the *Sierra Norte* region of the state of Puebla. Each of these datasets contains approximately 10,000 tokens, annotated using the Universal Dependencies (UD) (Nivre et al., 2020) framework for multiple levels of analysis: lemmatization, part-of-speech tagging, morphological analysis, and syntactic parsing.[4]

With respect to dialectal classification, Highland Puebla Nahuatl is clearly identifiable as an Eastern Nahuatl variety, and its place within the Nahuatl sub-classification is generally agreed-upon in the literature. Western Sierra Nahuatl's place is a bit trickier, in that it has a number of Central isoglosses, but also shares some features with the Eastern varieties (e.g. having /i/ where central vari-

[4]A quantitative comparison of the two treebanks can be found in Pugh and Tyers (2024).

Use	Source	Variant	Annotation	Tokens	Sents
train/eval	azz treebank	azz	UD	10,088	1,260
train/eval	nhi treebank	nhi	UD	10,132	909
train only	Axolotl	azz, nci, nhm, nhn, nhw	unlabeled	182,174	13,519
eval only	Casanova stories	ncx	UD	2,355	200

Table 1: A breakdown of the datasets used in the paper and their total sizes. For the treebanks, which make up the data for the bulk of the experiments, we divided up the dataset 10 times into 90/10 splits in order to perform 10-fold cross validation. The variant labels listed with the Axolotl corpus are approximations based on an analysis of the 30 text sources that the sentences come from. The "Casanova stories" is a sample of texts from a larger collection, generously provided Joe Campbell.

eties have /e/). (Sasaki, 2015) provides a detailed comparison of Nahuatl variants spoken in Puebla's Northern Sierra, including Highland Puebla and Western Sierra Nahuatl. Table 2 provides an example of the differences between the variants.

It is worth noting that, though they are relatively distinct genealogically, these two variants are spoken in some adjacent communities and are in contact in areas such as Tetela de Ocampo, an azz-speaking municipality where some nhi-speakers go for commerce and school. It is therefore possible that these two varieties have more common features than any random selection of two variants. That being said, the two variants are distinguished by multiple isoglosses, e.g. the /t/-/tl/ distinction and use of the antecessive /o-/ in the past tense.

The nhi corpus consists of a combination of short stories, personal narratives, and grammar examples, and contains samples representing some linguistic diversity within the variant group (see Pugh et al. (2022) for specifics). The azz corpus, on the other hand, is more homogeneous, with the majority of the data coming from a single town and being largely of a single genre, namely descriptions of plants and their medicinal/culinary use.

For one experiment, we supplement the UD tree data with unlabeled Nahuatl text from the Axolotl corpus (Gutierrez-Vasques et al., 2016), a Nahuatl-Spanish parallel corpus with over 10k sentences.

Finally, we collect and annotate a small sample (about 2k tokens) of texts from the Central Puebla Nahuatl (ISO-639: ncx), a Central Nahuatl variety. The sample ("Casanova stories") is taken from a collection of short stories from Gonzalez-Casanova and prepared by Joe Campbell. We annotate the sample with the UD schema, but ignore morphological analyses due to the time-intensive nature of such annotation. This small dataset is used to evaluate our models' performances on a Nahuatl variety not seen during training.

4.1 Orthography

Numerous orthographic standards have been proposed over the years for written Nahuatl (using the Latin alphabet), but there is no real consensus. Often, written Nahuatl may be in a one-off orthography, and not necessarily consistent within a given text. Our data represents a variety of orthographies, and we normalize it using a finite-state transducer from the Py-Elotl Python package[5]. As the target orthography, we use one of the norms proposed by the Summer Institute of Linguistics (SIL) for Nahuatl, which uses 's' for /s/, 'c/qu' for /k/, 'tz' for /ts/, and 'u' for /w/. This decision is largely arbitrary. our motivation for choosing this instead of, for example, the INALI standard orthography (INALI, 2018), is the former's greater similarity to Spanish spellings (e.g. the graphemes "w" and "k" in Spanish are seen primarily only in loanwords). Since Nahuatl texts typically contain many Spanish words, and given the fact that the multilingual BERT model we use in our experiments was trained on a large amount of Spanish data, we chose to use an orthography that reflects Spanish spellings in order to better leverage the representations in the BERT model[6]. We use the normalized forms in all of the experiments in order to remove orthographic variation as a variable.

[5] https://github.com/ElotlMX/py-elotl

[6] Another option that would achieve the same goal would have been the ACK orthography, the only difference being the latter's lack of "s", which is relatively common in contemporary Spanish orthography. The quantification of orthographic similarity, and the extent to which orthography plays a role in Nahuatl parser performance using multilingual pretrained language models is a topic that we leave for future work.

azz	nhi	en
Tepos **teyin** tepaleuia mah **ica** se quita **teyin** amo u**e**li se quita ica se ixtololo.	Tipos**tl** **tlen** tepaleuia **ica** mo sequita **tlen** amo u**i**li sequita ica se ixtololo.	"Instrument that helps people see what cannot be seen with an eye."
Ocs**e**pa ti**qui**yolitijkej.	Ocs**i**pa oti**c**yolitihkeh.	"We started it up again."

Table 2: Example of two parallel sentences in azz and nhi. The azz text was taken from the corresponding treebank, and was translated by a speaker of nhi. Some specific differences are bolded, and include the raising of short /e/ in azz to /i/ in nhi, the *tl-t* isogloss, the absence of the antecessive *o-* on past tense verbs in azz, and a word-order difference with respect to the relational noun *ica* "with (instrumental)". The differences described here are by no means exhaustive.

Train	Eval	OOV%
nhi	nhi	$38\% \pm 3$
	azz	$81\% \pm 1$
	ncx	80%
azz	nhi	$83\% \pm 1$
	azz	$31\% \pm 3$
	ncx	87%
nhi + azz	nhi	$37\% \pm 3$
	azz	$30\% \pm 2$
	ncx	76%

Table 3: The percentage of out-of-vocabulary (OOV) tokens for the experiment configurations. When the Eval variant is nhi or azz, the experiments involve 10-fold cross-validation, so we average the OOV percentages over the folds and include the standard deviation. When calculating OOV percentage for the ncx data, we use the first fold of the training data. These numbers help give an initial impression of the difficulty of the different parsing tasks. Specifically, we see that, unsurprisingly, other-variant Eval datasets have substantially higher OOV percentages than same-variant Eval data.

5 Experiments

For all of the experiments described in this section, we use the MaChAmp toolkit (van der Goot et al., 2021) to fine-tune contextual subword embeddings from the pretrained multilingual BERT (mBERT) model[7] on each UD task. The model leverages multi-task learning, such that all of the tasks share encoder parameters, but each has its own unique decoder: a transformation-rule classifier (Straka, 2018) for lemmatization, a softmax layer on the contextual embeddings for part-of-speech tagging and morphological analysis, and a deep biaffine parser for dependency parsing (Gardner et al., 2018). During training, the best model is selected by summing the accuracy metrics of these tasks.

Due to the relatively low total volume of labeled data, we report results of 10-fold cross-validation.

5.1 Monolingual

We first evaluate the monolingual ("Mono" in Table 4), i.e. single variant, performance of the two Nahuatl variants, which serves as a benchmark for comparison with subsequent models. Intuitively, we expect these models to perform best on their respective variants, but be less robust when faced with multi-variant data.

5.2 Cross-Variant

Secondly, in order to get a sense of how challenging multi-variant NLP actually is for Nahuatl, we test zero-shot, cross-variant model transfer ("Cross" in Table 4, i.e. training on one variant and evaluating on the other. The motivation behind this experiment is the recognition that, it could be the case that many Nahuatl variants are similar enough to one another that there is no real need for special efforts targeted at multi-variant NLP for the language. If this were the case, we would expect zero-shot, cross-variant performance to be comparable with that of a monolingual model.

Recognizing that a major limitation of our dataset is the fact that it only represents two out of 30 Nahuatl variants, we annotated a small sample of short stories in a third variant, Central Puebla variety (ncx). We evaluate zero-shot, cross-lingual experiments on this dataset, as well as the performance of models jointly trained on both nhi and azz training sets. The objective of this experiment is to provide better a sense of the multi-variant capabilities of a model trained on limited data representing only a small set of Nahuatl varieties.

[7]We use the bert-base-multilingual-cased model.

Var.	Experiment	N	Lemma	UPOS	Morph.	UAS	LAS
azz	Mono	1,134	0.92 ± 0.02	0.94 ± 0.01	0.85 ± 0.02	0.84 ± 0.02	0.77 ± 0.03
	Cross	818	0.68 ± 0.02	0.68 ± 0.02	0.39 ± 0.01	0.67 ± 0.02	0.47 ± 0.03
	Joint Adj.	976	0.89 ± 0.01	0.93 ± 0.01	0.75 ± 0.02	0.81 ± 0.02	0.73 ± 0.02
	Joint	1,952	0.92 ± 0.01	0.95 ± 0.01	0.82 ± 0.03	0.85 ± 0.02	0.77 ± 0.02
	Joint+MLM	1,952	0.92 ± 0.01	0.95 ± 0.01	0.82 ± 0.02	0.85 ± 0.02	0.78 ± 0.02
nhi	Mono	818	0.82 ± 0.02	0.93 ± 0.01	0.67 ± 0.02	0.83 ± 0.02	0.74 ± 0.02
	Cross	1,143	0.65 ± 0.02	0.65 ± 0.02	0.44 ± 0.01	0.64 ± 0.02	0.42 ± 0.01
	Joint Adj.	976	0.79 ± 0.02	0.91 ± 0.02	0.60 ± 0.02	0.81 ± 0.02	0.71 ± 0.02
	Joint	1,952	0.82 ± 0.02	0.93 ± 0.01	0.67 ± 0.02	0.84 ± 0.02	0.76 ± 0.02
	Joint+MLM	1,952	0.82 ± 0.02	0.93 ± 0.01	0.68 ± 0.01	0.85 ± 0.02	0.76 ± 0.03

Table 4: Accuracy of a neural, multi-task UD parsing model in various training configurations. Each result is the average performance over 10 folds, followed by the standard deviation of the performance distribution. Note that, given the distribution overlap, not much can be said about the difference in performance of most of these experiments with the exception of the the cross-variant experiments, which consistently under-perform both monolingual (single-variant) and jointly trained models. Mono=Monolingual; Cross=Cross-variant (e.g. train on azz and predict on nhi); Joint=trained on the concatenation of both variants' corpora; Joint Adj.=like the Joint model, but only use half of the data from each variant during training; Joint w/ MLM=same as Joint, but with an additional masked language modeling task. "N" is the number of sentences in the training data for each experiment.

5.3 Joint Training

We train a model on the concatenation of the training data from the two Nahuatl variants, and evaluate its performance on each individual variant's evaluation data ("Joint" in Table 4). Ideally, given sufficient training data, the model can learn to implicitly detect the variant of an input text and, since a single set of model parameters is used for both variants, benefit from the similarities and increased coverage of Nahuatl linguistic features. Alternatively, it is plausible that the diatopic variation could add unhelpful noise during training.

By combining the training sets from two variants, we are also in effect doubling the training data size. To get a sense of how variant diversity in training effects model performance, while controlling for training data volume, we also experiment with combining just half of each of the nhi and azz training sets (*Joint Adj.* in Table 4).

5.4 Adding an Auxiliary Task During Training

We have emphasized that there is little available annotated Nahuatl text. However, there is a sizable amount of unlabeled text available that we can leverage to potentially improve system performance. We experiment with the Axolotl corpus (Gutierrez-Vasques et al., 2016), a parallel (Nahuatl-Spanish) collection of over 10,000 sentences of Nahuatl from multiple regions and time periods, including a large volume of colonial-era Classical Nahuatl.

Specifically, we perform the same multi-task approach described above, with an additional masked language modeling task using the Axolotl data ("Joint+MLM" in Table 4).

Since part of the azz treebank comes from the Axolotl corpus, we remove all text from source before creating this datasets in order to avoid data leakage.

6 Results and Discussion

The results of our experiments can be seen in Table 4. All results report the average and standard deviation of the performance on 10 folds.

6.1 Monolingual and Cross-variant Performance

Comparing the monolingual model performances, we note that the azz model performs either the same or better than the nhi model on nearly every task. This is likely due to the aforementioned greater homogeneity of the linguistic samples and genre in the azz treebank.

Secondly, the performance drops significantly from the monolingual models to the cross-variant models. Given the linguistic differences between the two variants, not to mention other differing characteristics of the corpora, this is largely expected. These results suggest the importance of focusing on developing multi-variant capabilities in Nahuatl NLP, since these data appear to be different enough to impede straightforward cross-variant transfer, at least with small data volume. Upon collecting more annotated data, it would be valuable to also evaluate the monolingual models on same-variant, different-genre data in order to tease apart the influence of linguistic variation and other sources of variation in the corpora.

6.2 Analyzing Multi-Variant Performance

In analyzing the results of these experiments, we are most interested in the *multi-variant* performance. For "Joint" experiments, where the training data of both training variants is concatenated, the multi-variant performance is the combined performance on both variants. These models can be compared with a monolingual model evaluated on both variants (e.g. the monolingual result on `azz` and the cross-variant result on `nhi`).

For both variants, the jointly-trained model (See the "Joint" rows in Table 4) performs on par with two respective monolingual models, despite not having explicit language labels. For some tasks, the jointly-trained model has a higher average performance (taking error into consideration, however, the difference is not robust).

While the high performance of the jointly-trained model may be due to exposing the model to linguistic diversity during training, an important caveat is that the jointly-trained experiment has twice the volume of training data as the monolingual or cross-variant experiments. In order to investigate the extent to which data volume alone (versus, e.g. more robust learning during training) can explain the good multi-variant performance of the jointly-trained model, we also performed a volume-adjusted joint training experiment by combining half of the training set from each variant.

The results of this experiment and a comparison with the full jointly-trained model, are listed in Table 5. Unsurprisingly, here we see a dip in performance compared to the full jointly-trained model. Nonetheless, the volume-adjusted jointly-trained model still shows better multi-variant performance than the monolingual equivalent (monolingual cross-variant), supporting the utility of diverse training data.

6.3 The Effect of an Additional Training Task

Even for the model trained on the concatenation of datasets, the total available training data volume of is low (barely over 2k sentences) compared to so-called "high-resource" scenarios. Since no Nahuatl variant nor any genetically- or aerially-related language (with the exception of perhaps Spanish) was included in the multilingual BERT training data, we are interested in how we might be able to use additional unlabeled Nahuatl data, even if from different varieties or time periods, to improve the mBERT representations for Nahuatl.

We investigated whether training on an additional task, MLM using the Axolotl data, improves the model's Nahuatl representations, impacting parser performance. However, we do not see a significant impact: results for all tasks were still within the estimated margin of error (1 standard deviation of the 10-fold results) when compared to the jointly-trained model with no auxiliary tasks.

7 Performance of Monolingual and Multi-Variant Models on a Third, Unseen Nahuatl Variant

We evaluate the different trained models on parsing text from the unseen `ncx` variant. Performance on this unseen variant text are reported in Table 6.

As with the two-variant experiments listed in 4, the jointly-trained model, which is trained on the concatenation of the full `nhi` training data and the full `azz` training data, achieves the top performance on all tasks. Unlike the two-variant experiments, however, here the volume-adjusted jointly-trained model (trained on half of the `nhi` training data concatenated with half of the `azz` training data) does not out-perform both monolingual models. Instead, we see that the monolingual model trained on `nhi` data performs comparably to the volume-adjusted joint model on all tasks.

One plausible explanation is that the differences in performance between the two mono-variant models is due to a combination of variant similarity and genre overlap in the corpora. Namely, since *nhi* and *ncx* are both Central Nahuatl variants, they share a number of linguistic features, such as the presence of the /tl/ phoneme and the use of the antecessive *o-* on verbs in the past tense. For example, both *nhi* and *ncx* tokenize the antecessive suffix *o-* and tag it as AUX (e.g. *o niquitac*, "I saw it", which in `azz` is just *niquitac*). The *azz* corpus does not have any instances of this, since this variety does not mark past tense verbs with the antecessive, and instead the only instances of the word *o* are the Spanish conjunction meaning "or". As a result, the anteces-

Exp.	N	Lemma	UPOS	Morph.	UAS	LAS
Joint	1,952	0.86 ± 0.01	0.94 ± 0.01	0.75 ± 0.02	0.85 ± 0.01	0.77 ± 0.01
Joint Adj.	976	0.83 ± 0.01	0.92 ± 0.01	0.67 ± 0.01	0.81 ± 0.01	0.72 ± 0.01
azz alone	1,134	0.76 ± 0.01	0.79 ± 0.01	0.64 ± 0.01	0.74 ± 0.01	0.59 ± 0.01
nhi alone	818	0.74 ± 0.03	0.80 ± 0.01	0.53 ± 0.02	0.75 ± 0.02	0.60 ± 0.02

Table 5: Comparing the multi-variant performance of different training configurations. The "azz and nhi alone" experiments use a monolingual model to parse multi-variant evaluation data. The "Joint" experiment trains a model on the concatenation of nhi and azz training data, leading to twice the training data volume as the other experiments. The "Joint Adj." experiment similarly trains on multi-variant data, but subsamples data from each variant to control for the possibility of data volume in and of itself being responsible for improved performance. "N" is the number of sentences in the training data for each experiment.

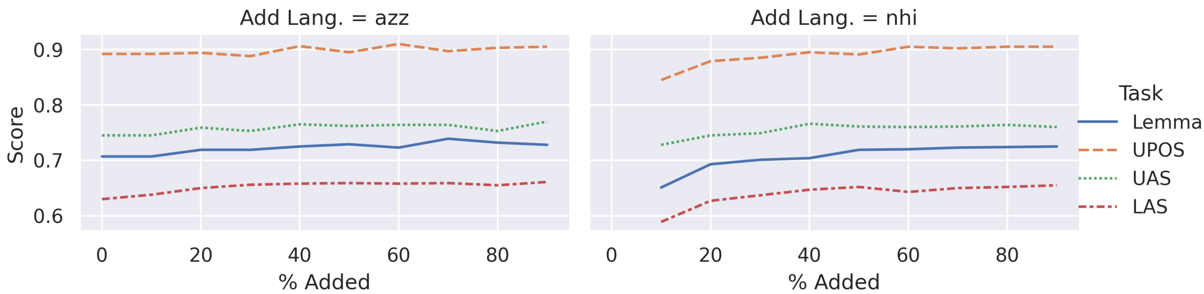


Figure 3: A plot of how the performance on the ncx data changes for the different tasks changes as we move from a monolingual model to a jointly-trained multi-variant model. As noted, ncx is linguistically-similar to nhi, and as such, adding azz data (the left plot) provides very minimal improvements, most of which seem to happen only once we've added 50% of the azz data. In the right plot, we see a larger improvement by just adding a small amount (the biggest marginal improvement happens when going from 0% to 20%) of nhi data to the azz-trained model.

Exp.	Lemma	UPOS	UAS	LAS
Joint	0.73	0.92	0.77	0.68
Joint Adj.	0.7	0.89	0.73	0.62
nhi alone	0.71	0.89	0.75	0.63
azz-ified nhi	0.58	0.83	0.73	0.57
azz alone	0.62	0.64	0.63	0.36
nhi-ified azz	0.62	0.79	0.68	0.52

Table 6: Performance of different models on an unseen Nahuatl variant, Central Puebla Nahuatl (ncx). Due to time constraints, we did not annotate the morphological analyses in this data, and thus do not report the performance.

sive in the ncx data is never correctly analyzed by the azz-only model.

To approach a better understanding of this proposed explanation of the results, we make copies of the monolingual datasets, altering the word forms with respect to both the antecessive *o-* and the /t/-/tl/ isogloss. That is, we make the nhi data more azz-like by removing the antecessive tokens and converting all instances of "tl" to "t". We expect a model trained on this version of the data to underperform on the ncx data since there is now a discrepancy in two prominent isogloss values. Likewise, we alter the azz add the antecessive token to all verbs with the morphological feature Tense=Past, and replace "t" with "tl" in positions that correspond to the latter segment in Nahuatl in general.[8] We expect a model trained on this dataset to perform better on the ncx data than the real azz data, since it has more common dialectal isoglosses.

The results show that the monolingual model trained on the nhi-ified azz data does indeed perform quite a bit better than that trained on the original azz data. Likewise, the model trained on the azz-ified nhi data peforms worse than that trained on the original nhi data. This shows the importance of dialectal similarity, even in the form of a pair of simple isoglosses. This result, while intuitive, is instructive for future work, since it indicates that, in the absence of more training data, variant-based

[8]The process of converting "t" to "tl" in the wordforms and lemmas was done via manual annotation.

data augmentation may be effective in increasing system performance.

It is also worth noting that, even after changing the isogloss values in the two datasets, the model trained on nhi data still outperforms that trained on the azz data when evaluating on the held-out Central variety, ncx. This fact indicates that morphological and syntactic factors are also at play. Furthermore, we also recognize the possible influence of genre on the performance differences.

With respect to genre and style, the unseen ncx text, a pair of short stories, more closely reflects the nhi corpus, which itself is largely made up of short stories, whereas the azz corpus consists almost entirely of transcriptions of recorded monologues describing the medicinal and culinary uses of plants. Findings such as those by Wang and Liu (2017), that a small but significant effect of genre on syntactic patterns such as adjacent dependency rate and dependency direction, may partially explain the much lower UAS and LAS performance by the azz model.

7.1 Learning Curve Experiment

To get a better sense of how adding different-variant data changes model performance on the ncx evaluation set, we perform a learning curve experiment for each variant, progressively adding 10% of the other variant's training data. The results of this experiment can be seen in Figure 3, plotting how the performance changes as we transition from a monolingual to a jointly-trained model by randomly adding data from the other variant. The azz model improves substantially with the addition of just a small amount (20%) of nhi data, and continues to improve as more data is added. The nhi-only model, on the other hand, improves only gradually with the addition of azz.

8 Future work

Revisiting the map in Figure 1, where we see that only two of Nahuatl variants have annotated treebanks, we recommend that the top priority for developing multi-variant NLP for Nahuatl be the continued collection of annotated corpora in additional variants and from diverse domains. Once more data is made available, we plan to empirically investigate the role of linguistic and genre similarity in multi-variant parsing using a variety of similarity metrics. For example, with an annotated test set for an additional Eastern Nahuatl variant, such as Huasteca Nahuatl, or azz sentences from a more diverse set of genres, further experiments could help shed light on the relative impact of genre and variant.

We also hope to explore other approaches to pretraining/auxiliary tasks in order to improve multi-variant parsing, such as building a language-specific pretrained model as described in Gessler and Zeldes (2022).

Finally, Nahuatl's long-standing contact with Spanish, a language with a significant number of annotated resources, offers a promising avenue of investigation of the extent to which Spanish data can be leveraged to improve NLP performance for Nahuatl.

9 Concluding Remarks

We reported the results of a series of experiments on UD parsing for Nahuatl, with a specific emphasis on multi-variant capabilities. We found, perhaps unsurprisingly, that the more examples of a given variant there are in the training data, the better the resulting model can perform on that variant. The multi-variant model performed as well as or better than two separate monolingual models, suggesting that having more data from diverse variants leads to a more robust model. Interestingly, we also found that a model's performance can be improved by superficially altering other-variant training data based on Nahuatl isoglosses. Though a number of points are still left to be investigated more thoroughly, this report serves as a first in-depth exploration of shallow Nahuatl NLP with the currently available datasets.

Acknowledgments

We are very grateful to Joe Campbell for his permission to use the Casanova stories, and to the anonymous reviewers for their helpful feedback.

References

Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. 2021. ARBERT & MARBERT: Deep bidirectional transformers for Arabic. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7088–7105, Online. Association for Computational Linguistics.

Noëmi Aepli. 2018. *Parsing approaches for swiss german*. Ph.D. thesis, University of Zurich.

Noëmi Aepli, Antonios Anastasopoulos, Adrian-Gabriel Chifu, William Domingues, Fahim Faisal, Mihaela Gaman, Radu Tudor Ionescu, and Yves Scherrer. 2022. Findings of the VarDial evaluation campaign 2022. In *Proceedings of the Ninth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 1–13, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Željko Agić, Anders Johannsen, Barbara Plank, Héctor Martínez Alonso, Natalie Schluter, and Anders Søgaard. 2016. Multilingual projection for parsing truly low-resource languages. *Transactions of the Association for Computational Linguistics*, 4:301–312.

Waleed Ammar, George Mulcaire, Miguel Ballesteros, Chris Dyer, and Noah A. Smith. 2016. Many languages, one parser. *Transactions of the Association for Computational Linguistics*, 4:431–444.

Peter Auer. 2011. *26 Dialect vs. standard: a typology of scenarios in Europe*, pages 485–500. De Gruyter Mouton, Berlin, Boston.

Christopher S Beekman and Alexander F Christensen. 2003. Controlling for doubt and uncertainty through multiple lines of evidence: A new look at the mesoamerican nahua migrations. *Journal of Archaeological Method and Theory*, 10:111–164.

Una Canger. 1988. Subgrupos de los dialectos nahuas (1988). In J. Kathryn Josserand and Karen Dakin, editors, *Smoke and Mist: Mesoamerican Studies in Memory of Thelma D. Sullivan.Part. Oxford: BAR International Series 402 (Ii)*, volume 402 of *BAR International*, pages 473–98. BAR, Oxford.

Una Canger and Karen Dakin. 1985. An inconspicuous basic split in nahuatl. *International Journal of American Linguistics*, 51(4):358–361.

Michael Clyne. 2012. *Pluricentric languages: Differing norms in different nations*, volume 62. Walter de Gruyter.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Alan Durston. 2008. Native-language literacy in Colonial Peru: The question of mundane Quechua writing revisited. *Hispanic American Historical Review*, 88(1):41–70.

Anna María Escobar. 2011. Spanish in contact with Quechua. *The handbook of Hispanic sociolinguistics*, pages 321–352.

J.I.E. Farfan. 2019. *Nahuatl Contemporary Writing: Studying Convergence in the Absence of a Written Norm*. University of Sheffield.

José Antonio Flores Farfán. 2008. *The Hispanisation of modern Nahuatl varieties*, pages 27–48. De Gruyter Mouton, Berlin, New York.

Xóchitl Flores-Marcial, Moisés García Guzmán, Felipe H. Lopez, George Aaron Broadwell, Alejandra Dubcovsky, May Helena Plumb, Mike Zarafonetis, and Brook Danielle Lillehaugen. *Caseidyneën Saën – Learning Together: Colonial Valley Zapotec Teaching Materials*.

John Foreman and Brook Danielle Lillehaugen. 2017. Positional verbs in colonial valley zapotec. *International Journal of American Linguistics*, 83(2):263–305.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer. 2018. AllenNLP: A deep semantic natural language processing platform. In *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, pages 1–6, Melbourne, Australia. Association for Computational Linguistics.

Luke Gessler and Amir Zeldes. 2022. MicroBERT: Effective training of low-resource monolingual BERTs through parameter reduction and multitask learning. In *Proceedings of the The 2nd Workshop on Multilingual Representation Learning (MRL)*, pages 86–99, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Willard P Gingerich. A bibliographic introduction to twenty manuscripts of classical nahuatl literature. *Latin American Research Review*, 10(1):105–125.

Ximena Gutierrez-Vasques, Gerardo Sierra, and Isaac Hernandez Pompa. 2016. Axolotl: a web accessible parallel corpus for spanish-nahuatl. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4210–4214.

Craig Hilts. 2003. From taxonomy to typology: The features of lexical contact phenomena in atepec zapotec-spanish linguistic contact.

Rebecca Hwa, Philip Resnik, Amy Weinberg, Clara I. Cabezas, and Okan Kolak. 2005. Bootstrapping parsers via syntactic projection across parallel texts. *Natural Language Engineering*, 11:311 – 325.

INALI. 2009. *Catalogo De Las Lenguas Indigenas Nacionales: Variantes Linguisticas De Mexico Con Sus Autodenominaciones Y Referencias Geoestadisticas*. Instituto Nacional de Lenguas Indigenas, México, D.F.

INALI. 2018. Breviario: Norma ortográfica del idioma náhuatl, méxico. (conforme al avance preliminar de la norma de escritura de la lengua náhuatl a nivel nacional).

Anjali Kantharuban, Ivan Vulić, and Anna Korhonen. 2023. Quantifying the dialect gap and its correlates across languages. In *Findings of the Association for Computational Linguistics: EMNLP 2023*. [to appear].

Terrence Kaufman. 2001. The history of the nawa language group from the earliest times to the sixteenth century: some initial results.

Dan Kondratyuk and Milan Straka. 2019. 75 languages, 1 model: Parsing Universal Dependencies universally. *arXiv preprint arXiv:1904.02099*.

Yolanda Lastra. 1986. *Las areas dialectales del nahuatl moderno*. Universidad Nacional Autonoma de Mexico, Instituto de Investigaciones Antropologicas.

Walter Lehmann. 1920. Die sprachen zentral-amerikas in ihren beziehungen zueinander sowie zu sudamerika und mexiko, 1/2. *Zentral-Amerika, Teil I.*

Miguel León-Portilla. 1985. Nahuatl literature. In *Supplement to the Handbook of Middle American Indians, Volume 3: Literatures*, pages 7–43. University of Texas Press.

Aurolyn Luykx, Fernando García Rivera, and Félix Julca Guerrero. 2016. Communicative strategies across quechua languages. *International Journal of the Sociology of Language*, 2016(240):159–191.

Yaron Matras and Jeanette Sakel. 2007. Investigating the mechanisms of pattern replication in language convergence. *Studies in Language. International Journal sponsored by the Foundation "Foundations of Language"*, 31(4):829–865.

J. Nivre, M.-C. de Marneffe, F. Ginter, J. Hajič, C. D. Manning, S. Pyysalo, S. Schuster, F. Tyers, and D. Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 4027–4036.

Justyna Olko, Robert Borges, and John Sullivan. 2018. Convergence as the driving force of typological change in Nahuatl. *STUF-Language Typology and Universals*, 71(3):467–507.

Magnus Pharao Hansen. 2014. The East-West split in Nahuan dialectology: Reviewing the evidence and consolidating the grouping. In *Friends of Uto-Aztecan Workshop*.

Magnus Pharao Hansen. 2015. Dialectal variation in contemporary Nahuatl relative clause formation. AIILS Seminar.

Robert Pugh, Marivel Huerta Mendez, Mitsuya Sasaki, and Francis Tyers. 2022. Universal Dependencies for western sierra Puebla Nahuatl. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5011–5020, Marseille, France. European Language Resources Association.

Robert Pugh and Francis Tyers. 2021. Investigating variation in written forms of Nahuatl using character-based language models. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 21–27, Online. Association for Computational Linguistics.

Robert Pugh and Francis M. Tyers. 2024. A Universal Dependencies Treebank for Highland Puebla Nahuatl. In *2024 Annual Conference of the North American Chapter of the Association for Computational Linguistics*.

Mitsuya Sasaki. 2015. A view from the Sierra : the Highland Puebla area in Nahua dialectology. , 36(TULIP):153–165.

Yves Scherrer, Tommi Jauhiainen, Nikola Ljubešić, Preslav Nakov, Jörg Tiedemann, and Marcos Zampieri. 2023. Tenth workshop on nlp for similar languages, varieties and dialects (vardial 2023). In *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*.

Yves Scherrer and Owen Rambow. 2010. Natural language processing for the swiss german dialect area. In *KONVENS*, pages 93–102.

Milan Straka. 2018. Udpipe 2.0 prototype at CoNLL 2018 ud shared task. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 197–207.

John Sullivan. 2011. The IDIEZ project: A model for indigenous language revitalization in higher education. *Collaborative Anthropologies*, 4(1):139–154.

Rob van der Goot, Ahmet Üstün, Alan Ramponi, Ibrahim Sharaf, and Barbara Plank. 2021. Massive choice, ample tasks (MaChAmp): A toolkit for multi-task learning in NLP. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 176–197, Online. Association for Computational Linguistics.

Erik Velldal, Lilja Øvrelid, and Petter Hohle. 2017. Joint UD Parsing of Norwegian Bokmål and Nynorsk. In *Proceedings of the 21st Nordic Conference on Computational Linguistics*, pages 1–10.

Yaqin Wang and Haitao Liu. 2017. The effects of genre on dependency distance and dependency direction. *Language Sciences*, 59:135–147.

Marcos Zampieri, Shervin Malmasi, Nikola Ljubešić, Preslav Nakov, Ahmed Ali, Jörg Tiedemann, Yves Scherrer, and Noëmi Aepli. 2017. Findings of the VarDial evaluation campaign 2017. In *Proceedings of the fourth workshop on NLP for similar languages, varieties and dialects (VarDial)*, pages 1–15.

Marcos Zampieri, Preslav Nakov, and Yves Scherrer. 2020. Natural language processing for similar languages, varieties, and dialects: A survey. *Natural Language Engineering*, 26(6):595–612.

Daniel Zeman and Philip Resnik. 2008. Cross-language parser adaptation between related languages. In *Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages*.

Highly Granular Dialect Normalization and Phonological Dialect Translation for Limburgish

Andreas Simons, Stefano De Pascale, Karlien Franco
QLVL, KU Leuven
simonsandreasjc at gmail.com, {stefano.depascale, karlien.franco}@kuleuven.be

Abstract

We study highly granular dialect normalization and phonological dialect translation on Limburgish, a non-standardized low-resource language with a wide variation in spelling conventions and phonology. We find improvements to the traditional transformer by embedding the geographic coordinates of dialects in dialect normalization tasks and use these geographically-embedded transformers to translate words between the phonologies of different dialects. These results are found to be consistent with notions in traditional Limburgish dialectology.

1 Introduction

We argue in this paper that encoding geographic coordinates as continuous parameters into transformer-based architectures allows for the improvement of normalization tasks between closely related varieties and reveals new methods in handling spatially-determined language variation.

In most tasks on multilingual data or closely related varieties, the varieties are treated on a coarse level (Dabre et al., 2020; Wu et al., 2021), without meaningfully encoding their relation to one another. The main idea behind encoding the relation between the different varieties is that knowledge transfer will take place between closely related varieties, therefore allowing for a solution to the issue of imbalanced data and a more generalized and continuous treatment of the studied varieties.

In this work we encode the geographic coordinates of approximately 1000 locations within the Limburgish language area - whose language varieties we will refer to as dialects from now on - by appending them as additional dimensions after the positional encoding in the original transformer architecture (Vaswani et al., 2023). This geographically-embedded transformer is then trained to normalize single dialect words following various different spelling conventions to a single phonetic-like spelling convention. The geographic embedding also enables the transformer to translate between any pair of Limburgish dialects on a highly granular level. We therefore separately consider the task of phonological dialect translation.

1.1 A Short Introduction to Limburgish

Limburgish is a West-Germanic language spoken by at least a million[1] native speakers in Belgium, the Netherlands and Germany. Limburgish partially underwent the High German consonant shift and has some unique features such as 3 grammatical genders, tonality, a gerund and a subjunctive in some dialects. Due to its geography and history it remained relatively isolated from both the Dutch and German standardization processes (Belemans and Keulen, 2004). Nowadays, Limburgish does not have a standard language and is superseded in official domains by the Dutch, German and French standard languages in different parts of the language area. As a result of this, Limburgish retains a complex phonology that varies continuously throughout its spoken area.

At the same time, Limburgish has been going through an atypical codification process where various standardized spelling conventions have existed since the 19th century, but often codified for individual towns. Its speakers consider all Limburgish dialects to be equally important, yet distinct varieties in what has been called a *multidialectal space* (Assendelft, 2019). This results in Limburgish texts featuring variation not only in terms of their native speakers' phonologies, but also in terms of the chosen spelling conventions. Additionally, Limburgish is one of the more extreme low-resource languages among the Germanic language family (Blaschke et al., 2023), making it very difficult to work with

[1]As per Ethnologue (2024), no elaborate estimates are known as the language only enjoys some official recognition in the Netherlands and the French-speaking community of Belgium (Limburgish Academy, 2024).

Proceedings of the Eleventh Workshop on NLP for Similar Languages, Varieties, and Dialects (VarDial 2024), pages 152–162
June 20, 2024 ©2024 Association for Computational Linguistics

in most Natural Language Processing tasks.

Due to the structure of the used dataset (see Section 3), we will only consider the Limburgish dialects spoken in Belgium and the Netherlands, although there is a priori no linguistic reason to separate the dialects in Germany from the ones in Belgium and the Netherlands.[2]

2 Related Work

2.1 Limburgish NLP

NLP research on Limburgish is scarce: Nguyen and Cornips (2016) developed dialect identification for Limburgish using the Limburgish Wikipedia as a corpus. This is the only available corpus for Limburgish apart from very limited web crawl and Ubuntu localization files corpora (Blaschke et al., 2023). Michielsen-Tallman et al. (2017) is working on a Limburgish corpus which is not publicly accessible yet, and Meta's No Language Left Behind included the Maastricht dialect through its *FLORES-200* dataset (NLLB Team et al., 2022), which is now included in some applications on Hugging Face. Franco et al. (2019a,b) previously applied a statistical approach to study lexical diversity and the influence of geography on loanwords in Limburgish using the WLD (Section 3).

2.2 Dialect Normalization

Methods related to normalizing dialects using machine learning and neural approaches have been studied by Pettersson et al. (2014); Scherrer and Ljubeić (2016); Bollmann and Søgaard (2016); Honnet et al. (2018); Lusetti et al. (2018); Partanen et al. (2019); van der Goot (2021). To the best of our knowledge, no dialect normalization task has been considered where the geographic coordinates are explicitly embedded in the transformer architecture with the goal of improving knowledge transfer. Neither has such a smooth, highly granular geographic normalization task been studied. Scherrer (2011) previously studied continuous variation of Swiss-German through a statistical word generation approach. Ramponi and Casula (2023) introduced a coordinate-tagged variety corpus for Italy using Tweets and studied highly granular language identification, which was previously also considered on other languages by Han et al. (2016); Gaman et al. (2020); Chakravarthi et al. (2021).

[2]Limburgish is typically demarcated between the major Uerdinger and Benrather isoglosses within West-Germanic (Goossens, 1965). This region extends into Germany, where it is also known as Südniederfränkisch.

2.3 Dialect Translation

Character or syllable-based dialect machine translation have been considered for Swiss-German (Honnet et al., 2018), and for Japanese (Abe et al., 2018). To the best of our knowledge, no approach has considered a smooth, highly granular dialect translation task of our magnitude or considered the direct embedding of geographic coordinates for the purpose of knowledge transfer between dialects in dialect machine translation.

3 Data

The dataset used for this work is the digitized (van Hout et al., 2024) version of the *Woordenboek van de Limburgse Dialecten* (Dictionary of the Limburgish Dialects) or *WLD* (Weijnen et al., 1983-2008), an onomasiological dictionary of Limburgish, covering the dialects spoken in the Belgian and Dutch provinces of Limburg and the north of Liège. The dictionary is onomasiological in the sense that it is indexed along semantic concepts such as agrarian (e.g. ploughing, cattle), professional (e.g. bakery, mining) and general concepts (e.g. health, religion). Per semantic concept, it groups all variants per cognate, geotagged with their exact town of origin. The structure of this dictionary therefore allows us to study the phonological and orthographic variation of the Limburgish lexicon and how they interact with geography.

The WLD contains approximately 17k concepts, featuring 139k cognates and a total of 1.7M Limburgish words spread over approximately 1000 locations. About half of the entries follow a high-quality *morpho-phonological* spelling, a combination of standard Dutch orthography, the International Phonetic Alphabet (IPA) and some custom diacritics. This part of the WLD was carefully reviewed by its original curators (Weijnen et al., 1983-2008). The remainder follows various spelling conventions from local dictionaries or standardized conventions such as the Veldeke spelling (Bakkes et al., 2003).

The locations are tagged with *kloeke* codes, geographic tags that refer to all locations in Belgium, the Netherlands, northern France and western Germany. We converted these kloeke codes to geographic coordinates, which were then normalized to unit intervals. Entries corresponding to locations that were clearly outside the Limburgish area were omitted. We carried out some preprocessing and cleaning steps on the data such as deleting entries

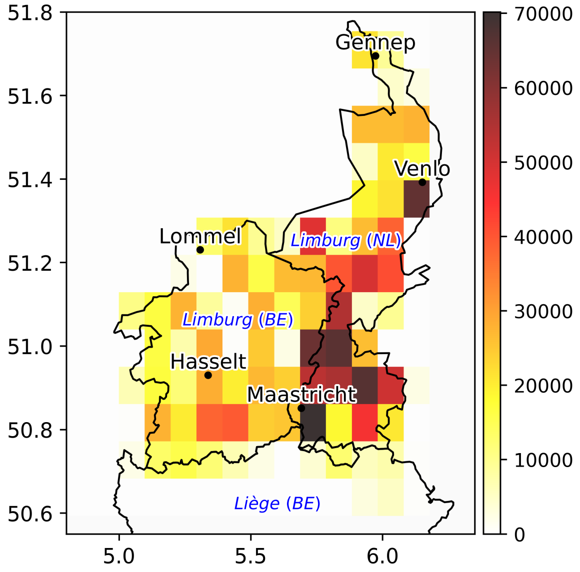


Figure 1: 2D frequency histogram of the geographic spread of entries in the WLD.

with overly noisy characters (as a result of poor digitization), splitting sentence entries into individual words and omitting superfluous characters such as punctuation marks. The geographic spread of the resulting data are shown in Fig 1. Finally, we applied some manually curated rules to resolve predictable digitization errors, such as converting incorrect ASCII characters.

3.1 Task-Specific Datasets

Since we did not have access to curated datasets or parallel corpus data for Limburgish or any of its dialects, two new datasets needed to be generated from the WLD for the normalization and phonological dialect translations tasks.

For the **normalization task**, we aimed to train a model that correctly converts the characters of any of the Limburgish (conventional) spelling systems to the high quality morpho-phonological standards that constitute approximately half the WLD. These standards closely resemble IPA, and enable further study of Limburgish text data, which is often blurred by localized spelling conventions.

We therefore split the dataset into words following an accurate phonetic spelling convention and words with local or other conventional spelling conventions. This is achieved using manually selected filters containing typical conventions in the Veldeke spelling and other local conventions that are not known within the WLD's phonetic system, such as the use of *ieë*, *aa* or *äö*. Any words that do not contain any n-grams which are exclusively used in conventional spelling are then assumed to be in phonetic notation, which a manual inspection confirms. This results in an approximately equal split in normalized-unnormalized data. For each unnormalized entry, cognates of nearby dialects are then selected as the normalized equivalents. By defining a nearby dialect as being within a 0.5 km radius, we ensure that the variation is more likely due to spelling conventions rather than a change in phonology between the two dialects. Typically, the phonologies of Limburgish dialects stay consistent within such a radius, unless major isoglosses are crossed. This results in a dataset of 118k unnormalized-normalized pairs, with an average pair distance of 0.39 km. An example of such a pair is given below (with both words originating from the dialect of Echt).

> *kroedwès- krutweš* (without translation; a folkloristic herb)

For the task of **phonological dialect machine translation**, we again chose a character-based approach and trained a model to translate the phonology between any dialects, as this is the largest source of variation in Limburgish apart from spelling variation. We only used the part of the WLD dataset that is already normalized (approximately 805k entries), therefore avoiding arbitrary spelling conventions, and generated a new dataset. We paired each word in this normalized subset with all cognates from other dialects in the same dataset. Very frequent words such as *in* (English: in) or *van* (English: of) were omitted, as were words that are rare in other dialects (a frequency of <10). For each family of cognates, we undersampled the available cognates due to the imbalance in geographic representativeness of the data (see Fig 1). The undersampling was done by weighting the geographic frequency distribution of the cognates according to a 2D Gaussian kernel smoothing with a bandwidth factor of 0.2 and then undersampling by 70%.

As the entries in the new dataset grow quadratically with the number of cognates, a 10% subset is sampled of all pairs in the new dataset, resulting in a phonological dialect translation dataset of 20.2M entries. For example, the word *šo.l* (originating from Bree) is paired with 85 different cognates from other dialects:

> *sxo:l* (Grote-Brogel), *šǭl* (Kanne), *šuəl* (Kerkrade), *šo.əl* (Valkenburg), *sxo.l* (Nederweert) . . .

4 Methods

4.1 Encoding

We tried two encoding methods: a simple one-hot encoding and an experimental method using phonological vectors from the PanPhon library (Mortensen et al., 2016) in Python. The main idea was that encoding 24 articulatory features is more meaningful and compact for data that varies greatly phonologically. The phonological encoding was surprisingly outperformed by a simple one-hot encoding in all experiments. This was likely due to the high complexity of the phonology of some dialects. For example, the dialect of Weert has 28 vowels over 5 heights, for which PanPhon's binary vowel height system is insufficient.

We opted for a 92-dimensional one-hot encoding, corresponding to all unique characters that remained after the preprocessing and cleaning steps. Due to the complexity of Limburgish phonology, many special diacritics are featured to realize the correct vowels or tonality. These diacritics are represented as separate characters. All words above 10 characters were omitted, and shorter words were padded to 10-dimensional vectors.

4.2 Coordinate Embedding in Transformer

The modification to the traditional transformer architecture was done in Tensorflow's functional API v2.12 (Abadi et al., 2015), Keras v2.15 (Chollet and et al., 2015), and KerasNLP v0.4.1 (Watson et al., 2022). Typically when discrete language tokens are used for multilingual models (and therefore different from our approach), this is done at the tokenizer level, thereby increasing the input vocabulary dimension or dimension of the input embedding. We instead pass the geographic coordinates as two extra dimensions directly after the positional encoding, resulting in a similar number of weights and training complexity when compared to the traditional transformer architecture. The (rescaled) coordinates are only appended to the first dimension of the embedded vector (after the input embedding and positional encoding) to keep the data sparse and can be visualized as

$$\begin{bmatrix} e_{1,1} & e_{1,2} & \cdots & e_{1,9} & e_{1,10} \\ \vdots & \vdots & & \vdots & \vdots \\ e_{N,1} & e_{N,2} & \cdots & e_{N,9} & e_{N,10} \\ y & 0 & \cdots & 0 & 0 \\ x & 0 & \cdots & 0 & 0 \end{bmatrix},$$

where $e_{i,j}$ represent the floats of the embedded vector after the input embedding and positional encoding, N the dimension of the embedding vector space and x, y the rescaled geographic coordinates.

For the encoder block, the coordinates corresponding to the input word are embedded. For the decoder block, the first input (using autoregression) is the embedded start token with the coordinates of the target word appended as two extra dimensions.

4.3 Evaluation

We consider two evaluation metrics::

Levenshtein ratio: the Levenshtein ratio between two words s_1 (reference) and s_2 (hypothesis) is defined as (Bachmann, 2021)

$$1 - \frac{\text{Levenshtein distance}(s_1, s_2)}{\text{len}(s_1) + \text{len}(s_2)} \tag{1}$$

where the Levenshtein distance between s_1 and s_2 is defined as the number of single-character insertions, deletions, and substitutions required to transform s_1 into s_2. The Levenshtein ratio is a character-based measure of similarity between two words, normalized for the lengths of the words (unlike the typical Levenshtein distance). Two identical words have a Levenshtein ratio of 1, the minimum ratio is 0.

CharacterF: *character n-gram F-score* or *chrF* (Popović, 2015) is the character-based machine translation equivalent of the traditional F-score. As it relies on character n-grams, it is more sensitive towards morpho-syntactic phenomena. The chrF score between two words s_1 (reference) and s_2 (hypothesis) is defined as (Popović, 2015)

$$\text{chrF} = 2\frac{\text{chrP} \cdot \text{chrR}}{\text{chrP} + \text{chrR}} \tag{2}$$

where chrP is the percentage of n-grams from s_2 that can be found in s_1 and chrR the percentage of n-grams from s_1 in s_2. We use 3-grams as these correspond closely to human judgment (Popović, 2015).

We initially expect both these metrics to undervalue the performance in this task; the WLD is very rich in diacritics and it is undesirable that the predicted normalized word is penalized for using a diacritic that is phonologically very close, but not identical to the expected diacritic. To mitigate this, we will also compute these metrics after stripping the diacritics, and thus only considering the ASCII characters. An example of how these metrics behave can be found in Table 1.

Unnormalized - Normalized (translation)	ChrF	Levenshtein ratio	ChrF (no diacritics)	Levenshtein ratio (no diacritics)
vief - *viêf* (five)	0	0.75	1	1
kroedwès - *krutweš* (herb)	0	0.4	0.18	0.53
waere - *wĕre* (to become)	0	0.8	0.4	0.8
aafdoeë - *āfdūə* (to mow grass)	0	0.33	0.25	0.33
schoppen - *sxopə* (to kick)	0	0.46	0	0.46

Table 1: A sample of the normalization dataset and their evaluation according to the selected metrics.

We observe that chrF is generally much too strict, while chrF with diacritics removed is significantly more tolerant. The Levenshtein ratio seems more tolerant than the non-diacritic chrF, while the non-diacritic Levenshtein ratio seems the most tolerant metric. A more in-depth manual analysis showed that the non-diacritic chrF metric corresponded closest to human judgement.

An inherent difficulty of working with Limburgish data is discerning the variation caused by differences in phonology from the variation caused by different spelling conventions, which is also a barrier for any other language that varies phonologically and orthographically (usually the case for non-standardized languages or families of dialects). To establish some baselines, we determined a lower boundary for all four metrics by measuring them on the unnormalized-normalized word pairs in the dataset, reflecting the accuracy when the same input were to be predicted. We also determined an upper boundary by estimating the inherent variation in spelling conventions: we computed the four metrics for all cognates within a radius of 6 km of each unnormalized word in the dataset. The assumption is that most remaining variation will then be due to differences in orthography, rather than phonology. This baseline therefore indicates the maximally attainable values for these metrics. We found the following lower and upper boundaries:

	ChrF	ChrF no diac.	Lev.	Lev. no diac.
Lower	0.112	0.242	0.599	0.710
Upper	0.440	0.589	0.751	0.84

Table 2: Expected lower and upper boundaries for the evaluation metrics.

Due to a lack of any curated data for Limburgish and the inherent variation in the data, these boundaries and a manual analysis in Section 6 are our best available approaches for evaluation, for a more elaborate discussion we refer to Section 8.

4.4 Normalization Task

To test whether embedding geographic coordinates improves the traditional transformer architecture, we first ran a hyperparameter search on the task using a traditional transformer without coordinate embedding. Using the traditional transformer for this task is possible since the target words follow the (relatively) uniform morpho-phonological spelling of the WLD and no dialect or spelling variation is required from the decoding part. We split the data in a $80 - 10 - 10$ train, validation, test dataset and varied stacking of encoder and decoder blocks from $1 - 5$, the embedding dimension from $1 - 1024$, the latent dimension from $1 - 1024$ and the number of attentions heads in each block from $1 - 16$ using the Optuna library (Akiba et al., 2019). We used the Adam training method and a Sparse Categorical Crossentropy metric and ran 100 iterations using Optuna's *Tree Parzen Estimator*. The optimized set of parameters was then used to train the traditional transformer and the geographically-embedded transformer and compare their performance on the test set.

The optimized traditional transformer has a total of 5.1M parameters, the geographically-embedded transformer has 5.2M parameters due to the extra 2 dimensions after the positional encoding step. These additional parameters only allow for a heterogeneous interaction between the coordinates and the embedded characters in the attention mechanism, and do not allow for any further inference of information in the attention mechanism that could otherwise be associated with having slightly more parameters.

4.5 Phonological Dialect Translation Task

Unlike the previous task, performance in the phonological dialect translation cannot be readily compared to the traditional transformer architecture as it does not natively allow for variation of the target dialect. We therefore only considered the geographically-embedded transformer.

We again used an $80-10-10$ train, validation, test dataset split but did not run a hyperparameter search due to resource constraints. We instead used standard parameter values such as an embedding dimension of 256, a latent dimension of 1024, 8 attention heads, and no stacked encoder or decoder blocks, resulting in a total of 7.6M parameters. For the optimizer and loss we again opted for Adam with a Sparse Categorical Crossentropy loss.

5 Results

5.1 Normalization Task

The hyperparameter search for the traditional transformer architecture yielded the following parameters: 2 layers of stacked encoder/decoder blocks, an embedding dimension of 150, a latent dimension of 1000, and 7 attention heads, resulting in a total of 5.1M parameters. The evaluation metrics on the test sets for the traditional transformer and the architecture with geographic coordinates embedded can be found in Table 3. We also present some representative examples of the geographically-embedded transformer's performance when normalizing the test set, along with the non-diacritic chrF metric in Table 5.

coords	ChrF	ChrF no diac.	Lev.	Lev. no diac.
no	0.353	0.506	0.713	0.817
yes	**0.363**	**0.516**	**0.718**	**0.821**

Table 3: The evaluation metrics on the test set for the traditional transformer and the transformer with geographic coordinates embedded.

5.2 Phonological Dialect Translation Task

The evaluation of the geographically-embedded transformer on the phonological dialect translation task can be found in Table 4. We again present some representative examples of the translation task with their corresponding non-diacritic chrF metrics and the locations of the input and target dialects in Table 6.

ChrF	ChrF no diac.	Lev.	Lev. no diac.
0.407	0.485	0.687	0.736

Table 4: The evaluation metrics on the test set for the phonological translation task.

6 Discussion

6.1 Normalization Task

As can be seen in Table 3, the geographically-embedded transformer outperforms the normal transformer according to all metrics that we measured. The results are statistically significant ($p < 0.001$) according to two-sided Wilcoxon hypothesis tests. The improvements to the traditional transformer's performance are most prominent in ascending order of 'tolerance' of the metrics, as we could have expected. When comparing these results with our established lower and upper boundaries (Table 2), we again find that the upper boundaries are approached more closely by the more tolerant metrics. A geographic analysis of the evaluation metrics showed that there is no geographic bias, as the performance is relatively homogeneously spread.

Manually analyzing a sample of the geographically-embedded transformer's predictions (Table 5), we find that the model generally succeeds in correctly normalizing various Limburgish spelling conventions to a phonetic spelling. For example, in entry 3 (*daavekot* → *dāvəkọ̜t*), the long *aa* is normalized to *ā*, the *e* to the schwa and the *o* to the correct Limburgish phoneme.

The model also abides by well-known notions in Limburgish dialectology: in entry 1 (*sjnaps* → *snaps*), the *sj* is normalized to an *s*, even though this is a neologism derived from High German, showing that the model correctly applies the Panninger isogloss within Limburgish that is associated with the *s*→*ʃ* rule (Bakkes et al., 2007). In entry 7 (*kool* → *kiəl*), the unnormalized word uses the Dutch phoneme *o* which does not occur for that word in Limburgish, but the model correctly predicts *iə*.

In other instances such as as 5 and 13, the model predicts normalized words that are more accurate than the original target normalizations. This is due to the fact that we generated this dataset ourselves without a very elaborate manual curation, as we did not have access to a curated or golden standard dataset. Despite inaccuracies in the generated dataset, the model has generalized well to avoid conventional spelling: in entry 3 an *ò* is included in the target, which is not part of the phonetic notation used in the WLD. The model instead correctly normalized this phoneme to ọ̜. The evaluation met-

	Unnormalized word	Prediction	Target	ChrF (no diac.)	Translation
1	*sjnaps*	*snaps*	*snàps*	1.0	schnaps (drink)
2	*zeik*	*zẹ̢i̯.k*	*zẹ̢i̯.k*	1.0	fecal sludge
3	*daavekot*	*dāvəkọ̢t*	*dāvəkòt*	1.0	dovecote
4	*sjollek*	*šoląk*	*šọ̢lək*	1.0	type of apron
5	*volle*	*vọ̢lə*	*vọ̢l*	0.667	full
6	*strooie*	*stroi̯ə*	*strōən*	0.5	of straw (material)
7	*kool*	*kiəl*	*kīəl*	1.0	cabbage
8	*hèndichə*	*hendix*	*hendixe*	0.889	handy
9	*kwartsche*	*kwartse*	*kwē˂rtšə*	0.2	quarter
10	*hiemël*	*hi:məl*	*hi:məl*	1.0	heaven
11	*lintteeke*	*lintēkə*	*lentēkə*	0.6	scar
12	*sjei*	*šẹ̢i̯*	*šẹ̢i̯*	1.0	vagina (horse)
13	*tweede*	*twēdə*	*de*	0	second
14	*preuvə*	*prèùve*	*prēūvə*	0.75	to taste
15	*áfzéttə*	*ifzetə*	*afzɛtə*	0.75	to rip off/defraud

Table 5: A random sample of the geographically-embedded transformer's performance on the test set.

	Input	Prediction	Target	ChrF (no diac.)	Translation	Locations
1	*špat*	*spat*	*spat*	1.0	osteoarthritis (horse)	Moresnet→Meeuwen
2	*rempəl*	*rimpels*	*rumpels*	0.6	wrinkles	Meterik→Blerick
3	*moder*	*mojər*	*mojər*	1.0	mother	Bocholt→Millen
4	*xeld*	*xẹ̢̄lt*	*xēld*	0.5	money	Gulpen→Moelingen
5	*bọ̢tərham*	*botəram*	*botəram*	1.0	sandwich	Achel→Blitterswijck
6	*werk*	*werk*	*werk*	1.0	work	Sittard→Beverst
7	*bẹ̢səl*	*bẹ̢səl*	*bẹ̢.səl*	0.286	bushel (hay)	Beverst→Munsterbilzen
8	*wẹ̢x*	*wex*	*wex*	1.0	road	Landen→Venray
9	*họ̢̄tə*	*họ̢wtə*	*họ̢wtən*	0.857	wooden	Genk→Neeroeteren
10	*briə.kə*	*brẹ̢̄.kə*	*brẹ̢̄.kə*	1.0	to spread manure	Jesseren→Nerem
11	*kát*	*kat*	*kat*	1.0	cat	Wijchmaal→Blitterswijck
12	*slẹ̢i̯p*	*slẹ̢i̯.p*	*slẹ̢i̯.p*	1.0	field drag	Neeritter→Bocholt
13	*wilde*	*wel*	*wøl*	0	wild	Ophoven→Lummen
14	*nak*	*nek*	*nek*	1.0	neck	Munsterbilzen→Horst
15	*hūs*	*hōēs*	*hôêês*	0.4	house	Gruitrode→Lottum

Table 6: A random sample of the geographically-embedded transformer's performance on the phonological dialect translation test set.

rics penalize these instances, even though they are desirable for our purposes.

The model fails in a few instances in predicting the correct normalization: in 8 the wrong gender/plural is predicted, in 9 (pred.: *kwartse*, target: *kwē˂rtšə*) and in 15 (pred.: *ifzetə* target: *afzɛtə*) some characters are incorrectly normalized. In entry 14, conventional spelling is used in the prediction (*prèùvə*).

From this manual analysis and the close correspondence to the estimated upper boundaries we can conclude that the geographically-embedded architecture is appropriate for normalization of Limburgish spelling to phonetic notation and an improvement over the traditional transformer architecture for this purpose.

6.2 Phonological Dialect Translation Task

The evaluation metrics (Table 4) approach the estimated upper boundaries, but not as closely as in the normalization task. There is again no geographic bias as the evaluation metrics are homogeneously spread over the studied language area. A manual analysis of a sample of some predictions (Table 6) shows that the model succeeds in correctly translating the phonology of various words from one Limburgish dialect to another: in for example entry 1 (*špat* → *spat*), the *s*→ʃ rule associated with the Panninger isogloss is correctly applied. In entry 10 (*briə.kə* → *brē̜.kə*), the correct sound ē̜ is translated for Nerem, even though the place of origin Jesseren is only 10 km removed and uses the sound *iə*.

In most instances, the translation matches the expected target. In some entries such as 2 and 13 the model corrects the input: entry 2 is an incorrect singular as the other cognates are plurals and the missing pluralization is therefore an error in the dataset, entry 13 is likewise a transcription error in the WLD and should have been *wel*.

In entries such as 4, 9, and 15 the model also performs better than the dataset and correctly predicts phonetic notation: *t* instead of *d* for entry 4 (pred.: *xē̜lt* target: *xēld*), no end-*n* for entry 9 (pred.: *ho̜wtə* target: *ho̜wtən*) and no conventional Limburgish spelling in the case of entry 15 (pred.: *hōēs* target: *hôêês*).

In some instances the model fails to predict the correct diacritics, such as in entry 7 (pred.: *be̜səl* target: *be̜.səl*), but most of the low evaluation metrics correspond to instances where the model predicted more desirable results than the dataset provided as target words.

From this manual analysis and the evaluation metrics we can conclude that phonological dialect translation was successfully achieved using the geographically-embedded transformer.

6.3 Language Variation Maps

The phonological translation model allows us to specify any input and target coordinates for the translation task; by fixing an input word and coordinates and varying the target coordinates we can generate highly granular language variation maps that show the phonological variation of the input word/phoneme. These maps can be used to extensively compare the model's predictions to traditional dialectological maps, or to study generalizations the model has learned due to the inherent language variation. In Fig 2 we provide such a map with all variation of *school* for fixed location Bree with input word *šo.l*. Overlaid on this map we indicate the natural variation of Limburgish phonology using dots, sourced from the WLD. We notice the close correspondence to the natural phonological

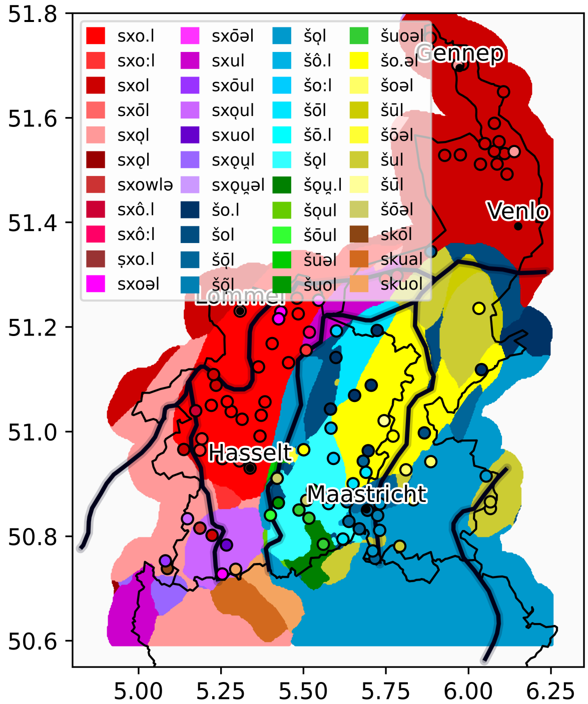


Figure 2: Generated variation map for *school* (*šo.l*) with Bree as fixed input. The overlaid dots correspond to the natural variation as found in the WLD. The color scheme purposely groups closely related phonemes. The black lines indicate the major isoglosses in the Limburgish area (Bakkes et al., 2007).

variation, in particular the clear separation between variants starting with *sx* and *š*. This isophone is part of the Panninger side line, an important isogloss in the Limburgish language area (in black, right next to Hasselt and converging in the north with the other isoglosses).

6.4 Availability

The trained models, datasets, figures and GIS data will be made available on `github.com/AndreasJCSimons/LimburgishNLP`.

7 Conclusion

We found that embedding geographic coordinates after the positional encoding allows us to normalize highly phonologically and orthographically varying data more accurately than the traditional transformer architecture. Additionally, we found that this geographic embedding allows us to translate the phonology of words between any Limburgish dialects and to generate language variation maps that can be compared to traditional dialectology or to study generalized phonological patterns that the model has implicitly learned.

8 Limitations

This work is limited by the lack of properly curated datasets and methodologies to evaluate the performance of dialect normalization and translation tasks, which hinders a more accurate evaluation of the used methods. We therefore had to evaluate the trained models using a manual analysis and estimates for an expected upper boundary on some evaluation metrics, given the inherent phonological and orthographic variation in the data. In Subsections 6.1 and 6.2 it is clear that in some instances the dataset is of low quality. However, due to the size of the dataset it is likely that the model has generalized beyond the low-quality entries: this can be seen both in the manual analysis where the model corrects wrong targets (even though it is penalized by the loss function) as well as in the language variation maps of Subsection 6.3 where the model has correctly learned Limburgish sound changes.

Another limitation is that the normalization and phonological dialect translation tasks only took the spelling and phonology of the words into account and not their semantics. While this rarely resulted in inaccurate predictions, a more elaborate normalization or translation scheme should take semantic information into account, as this can sometimes be tied to phonological patterns. For example, High German loanwords such as *sjnaps* (Table 5) are typically not subject to internal Limburgish sound changes and remain invariant.

Finally, the data of the WLD is not fully synchronous: it contains older dialect surveys such as the data from the Willems survey (19th century), SGV (1914), and ZND (from 1922 onwards). Additionally, data was collected in Belgian Limburg from the 1960s onwards to match missing data with respect to Dutch Limburg (Weijnen et al., 1983-2008). This means that the data collection occurred during a period of a major linguistic shift: between 1950 and 1980 a period of hyperstandardization occurred in Belgium that sought to promote *Algemeen Beschaafd Nederlands* (General Civilized Dutch) and stigmatize any other languages or language variation (Hoof and Jaspers, 2012). We also did not have access to any data from beyond the Dutch-German border, even though there is linguistically no reason to separate the dialects spoken between the Uerdinger and Benrather lines in Germany from the dialects in Belgium and the Netherlands.

9 Ethics Statement

This work complies with the ACM Code of Ethics and Professional Conduct (`https://www.acm.org/code-of-ethics`) with particular attention to articles 1.1 and 1.4: many underserved languages and language communities exist, and language variation and diversity is itself an exercise in low-resource NLP. By contributing to the research of non-standardized languages, low-resource languages or methods in NLP that can handle language variation, we hope to provide instruments that may be beneficial to disadvantaged languages communities.

The data used in this work, the *Woordenboek van de Limburgse Dialecten*, was manually processed over many years using dialect surveys and native speakers, who have been anonymized in the final dataset. Regardless, we are aware that by the very nature of this research, i.e. highly granular geographic analysis of language variation using methods in Deep Learning, we are studying phenomena that are tied to a person's native dialect, upbringing and socioeconomic situation. It is worrying that in recent years this has been abused for purposes of surveillance. For example, language variation has been used in dialect identification software by countries to evade privacy regulations during asylum procedures (European Digital Rights et al., 2021).

References

Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2015. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. Software available from tensorflow.org.

Kaori Abe, Yuichiroh Matsubayashi, Naoaki Okazaki, and Kentaro Inui. 2018. Multi-dialect neural machine translation and dialectometry. In *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation*, Hong Kong. Association for Computational Linguistics.

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A Next-generation Hyperparameter Optimization Framework.

Brenda Assendelft. 2019. De codificatie van het Limburgs. *Taal en Tongval*, 71(1):1–30.

Max Bachmann. 2021. Levenshtein module. Accessed: 10/03/2024.

Pierre Bakkes, Rob Belemans, Georg Cornelissen, Ronny Keulen, Ton Van de Wijngaard, Herman Crompvoets, and Frans Walraven. 2007. Riek van klank: inleiding in de Limburgse dialecten.

Pierre Bakkes, Herman Crompvoets, Jan Notten, and Frans Walraven. 2003. Spelling 2003 voor de Limburgse dialecten. Accessed: 10/03/2024.

Rob Belemans and Ronny Keulen. 2004. *Belgisch-Limburgs*. Taal in stad en land. Lannoo.

Verena Blaschke, Hinrich Schuetze, and Barbara Plank. 2023. A survey of corpora for Germanic low-resource languages and dialects. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 392–414, Tórshavn, Faroe Islands. University of Tartu Library.

Marcel Bollmann and Anders Søgaard. 2016. Improving historical spelling normalization with bi-directional LSTMs and multi-task learning. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 131–139, Osaka, Japan. The COLING 2016 Organizing Committee.

Bharathi Raja Chakravarthi, Gaman Mihaela, Radu Tudor Ionescu, Heidi Jauhiainen, Tommi Jauhiainen, Krister Lindén, Nikola Ljubešić, Niko Partanen, Ruba Priyadharshini, Christoph Purschke, Eswari Rajagopal, Yves Scherrer, and Marcos Zampieri. 2021. Findings of the VarDial evaluation campaign 2021. In *Proceedings of the Eighth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 1–11, Kiyv, Ukraine. Association for Computational Linguistics.

François Chollet and et al. 2015. Keras. `https://keras.io`.

Raj Dabre, Chenhui Chu, and Anoop Kunchukuttan. 2020. A Comprehensive Survey of Multilingual Neural Machine Translation.

Ethnologue. 2024. Limburgish. Accessed: 10/03/2024.

European Digital Rights, Access Now, Migration, Technology Monitor, Platform for International Cooperation on Undocumented Migrants, and Statewatch. 2021. Uses of AI in migration and border control: A fundamental rights approach to the Artificial Intelligence Act. Accessed: 10/03/2024.

Karlien Franco, Dirk Geeraerts, Dirk Speelman, and Roeland van Hout. 2019a. Concept characteristics and variation in lexical diversity in two Dutch dialect areas. *Cognitive Linguistics*, 30(1):205–242.

Karlien Franco, Dirk Geeraerts, Dirk Speelman, and Roeland van Hout. 2019b. Maps, meanings and loanwords: The interaction of geography and semantics in lexical borrowing. *Journal of Linguistic Geography*, 7(1):14–32.

Mihaela Gaman, Dirk Hovy, Radu Tudor Ionescu, Heidi Jauhiainen, Tommi Jauhiainen, Krister Lindén, Nikola Ljubešić, Niko Partanen, Christoph Purschke, Yves Scherrer, and Marcos Zampieri. 2020. A report on the VarDial evaluation campaign 2020. In *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 1–14, Barcelona, Spain (Online). International Committee on Computational Linguistics (ICCL).

Jan Goossens. 1965. Die Gliederung des Südniederfränkischen. In *Rheinische Vierteljahrsblätte*, pages 79–94.

Bo Han, Afshin Rahimi, Leon Derczynski, and Timothy Baldwin. 2016. Twitter geolocation prediction shared task of the 2016 workshop on noisy user-generated text. In *Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT)*, pages 213–217, Osaka, Japan. The COLING 2016 Organizing Committee.

Pierre-Edouard Honnet, Andrei Popescu-Belis, Claudiu Musat, and Michael Baeriswyl. 2018. Machine translation of low-resource spoken dialects: Strategies for normalizing Swiss German. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Sarah Van Hoof and Jürgen Jaspers. 2012. Hyperstandaardisering. *Tijdschrift voor Nederlandse Taal- en Letterkunde*, 128(1):97–125.

Limburgish Academy. 2024. Limburgish Language. Accessed: 10/03/2023.

Massimo Lusetti, Tatyana Ruzsics, Anne Göhring, Tanja Samardžić, and Elisabeth Stark. 2018. Encoder-decoder methods for text normalization. In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018)*, pages 18–28, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Yuri Michielsen-Tallman, Ligeia Lugli, and Michael Schuler. 2017. A Limburgish Corpus Dictionary: Digital Solutions for the Lexicography of a Non-standardized Regional Language.

David R. Mortensen, Patrick Littell, Akash Bharadwaj, Kartik Goyal, Chris Dyer, and Lori Levin. 2016. PanPhon: A resource for mapping IPA segments to articulatory feature vectors. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3475–3484, Osaka, Japan. The COLING 2016 Organizing Committee.

Dong Nguyen and Leonie Cornips. 2016. Automatic detection of intra-word code-switching. In *Proceedings of the 14th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 82–86, Berlin, Germany. Association for Computational Linguistics.

NLLB Team et al. 2022. No Language Left Behind: Scaling Human-Centered Machine Translation.

Niko Partanen, Mika Hämäläinen, and Khalid Alnajjar. 2019. Dialect text normalization to normative standard Finnish. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 141–146, Hong Kong, China. Association for Computational Linguistics.

Eva Pettersson, Beáta Megyesi, and Joakim Nivre. 2014. A multilingual evaluation of three spelling normalisation methods for historical text. In *Proceedings of the 8th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH)*, pages 32–41, Gothenburg, Sweden. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Alan Ramponi and Camilla Casula. 2023. DiatopIt: A corpus of social media posts for the study of diatopic language variation in Italy. In *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*, pages 187–199, Dubrovnik, Croatia. Association for Computational Linguistics.

Yves Scherrer. 2011. Morphology Generation for Swiss German Dialects. In *Systems and Frameworks for Computational Morphology*, pages 130–140, Berlin, Heidelberg. Springer Berlin Heidelberg.

Yves Scherrer and Nikola Ljubešić. 2016. Automatic normalisation of the Swiss German ArchiMob corpus using character-level machine translation. In *Conference on Natural Language Processing*.

Rob van der Goot. 2021. CL-MoNoise: Cross-lingual lexical normalization. In *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, pages 510–514, Online. Association for Computational Linguistics.

Roeland van Hout, Nicoline van der Sijs, Erwin Komen, Henk van den Heuvel, and et al. 2024. Elektronisch Woordenboek van de Limburgse Dialecten (e-WLD). Accessed: 10/03/2024.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention is all you need.

Matthew Watson, Chen Qian, Jonathan Bischof, François Chollet, and et al. 2022. Kerasnlp. `https://github.com/keras-team/keras-nlp`.

Antonius Weijnen, Jan Goossens, Pieter Goossens, Joep Kruijsen, Har Brok, Jo Kokkelmans, Herman Crompvoets, Jan van Schijndel, Jos Molemans, Joke Verbeek, Miet Ooms, Ton van de Wijngaard, J. Busch, Ronny Keulen, and Mariëtte Lubbers. 1983-2008. *Woordenboek van de Limburgse dialecten*. Van Gorcum/Gopher Publishers, Assen/Amsterdam/Maastricht/Utrecht.

Liwei Wu, Shanbo Cheng, Mingxuan Wang, and Lei Li. 2021. Language tags matter for zero-shot neural machine translation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3001–3007, Online. Association for Computational Linguistics.

Multilingual Identification of English Code-Switching

Igor Sterner
Department of Computer Science and Technology
University of Cambridge, UK
is473@cam.ac.uk

Abstract

Code-switching research depends on fine-grained language identification. In this work, we study existing corpora used to train token-level language identification systems. We aggregate these corpora with a consistent labelling scheme and train a system to identify English code-switching in multilingual text. We show that the system identifies code-switching in unseen language pairs with absolute F_1 measure 2.3-4.6% better than language-pair-specific SoTA. We also analyse the correlation between typological similarity of the languages and difficulty in recognizing code-switching.

1 Introduction

Code-switching is when bilinguals alternate between languages at the sentence or word level. Increasing attention is being placed on computational approaches to code-switching, driven by six code-switching workshops to date (Solorio et al., 2014; Molina et al., 2016; Aguilar et al., 2018b; Solorio et al., 2020, 2021; Winata et al., 2023). In part, this line of research is due to the rise in the use of code-switching on social media (Jose et al., 2020), potentially as a result of language contact (Gardner-Chloros, 2020).

Language technology users now expect automatic speech recognition systems, text-to-speech engines, generative models etc. to handle code-switching as a natural form of language. But even SoTA large language models (LLMs) perform poorly on zero-shot NLP tasks with code-switching data (Zhang et al., 2023). They are outperformed by smaller fine-tuned models. Further, Yong et al. (2023) report acceptability judgements of LLM-generated code-switching, showing few generations are acceptable. Despite the prevalence of code-switching in spoken and online discourse, code-switching is likely a linguistic phenomenon severely underrepresented in the training data of models like those in the GPT family (Brown et al., 2020). The availability of code-switching data has therefore become a common barrier to address the limitations of existing NLP tools on code-switching input. A tool required to address this barrier is fine-grained and multilingual language identification systems.

In this paper, we develop a fine-grained tool that distinguishes words between English and any other language.[1] We make our models and code available.[2]

2 Background

There are many works aimed at identifying languages in documents at more fine-grained levels, e.g. the word-level (Lyu and Lyu, 2008; Solorio et al., 2014; Mave et al., 2018; Zhang et al., 2018; Nguyen et al., 2021; Hidayatullah et al., 2022; Hegde et al., 2024) or even sub-word level (Mager et al., 2019; Sabty et al., 2021). Figure 1b shows the annotation scheme of one German–English work which aims for very fine-grained classification.

Approaches to compile code-switching corpora traditionally involved collecting spoken recordings of bilinguals (Myers-Scotton, 1992; Deuchar, 2009; Nguyen and Bryant, 2020) or more recently generating synthetic code-switching (Chang et al., 2019; Gupta et al., 2020; Rizvi et al., 2021). Manually collecting recordings is an expensive, arduous and lengthy task; meanwhile, synthetic code-switching is inherently limited in the code-switching phenomena it exhibits. But with automatic code-switching identification systems, much larger corpora of naturally occurring code-switching have begun to be collected (Nayak and Joshi, 2022; Sterner and

[1] In practice, the choice to focus only on languages switched with English was as a result of data availability.

[2] Code-switching identification: https://huggingface.co/igorsterner/AnE-LID, binary named entity recognition: https://huggingface.co/igorsterner/AnE-NER, code: https://github.com/igorsterner/AnE

Proceedings of the Eleventh Workshop on NLP for Similar Languages, Varieties, and Dialects (VarDial 2024), pages 163–173
June 20, 2024

Teufel, 2023; Wintner et al., 2023). The source of such data is large troves of social media posts.

Such corpora offer the potential to test various theories of code-switching, theories of when humans code-switch and why. An example is the triggering hypothesis (Clyne, 1980), which suggests that shared lexical items (e.g. named entities) are triggers of code-switching. Broersma and De Bot (2006) and Broersma (2009) found statistical significance between such lexical triggers and code-switching points for a small handful (c. 100) of switch points in recorded corpora. Soto et al. (2018) test on the larger Spanish–English spoken corpus of Deuchar (2009), but limit their study to a small list of cognates. Wintner et al. (2023) test on data of three different language pairs (Arabic–English, Spanish–English and German –English) with a total of 648,498 switch points from almost 10M tokens of mostly automatically language-identified social media content. They found statistical correlation suggesting switch points tend to be close to the shared lexical items. For this to be possible, substantial effort was invested to build word-level language identifiers specific to each of the three language pairs they explore (Aguilar et al., 2020; Shehadi and Wintner, 2022a; Osmelak and Wintner, 2023). Corpus-linguistic approaches to code-switching will continue to depend on the quality of such fine-grained language identification tools.

Existing code-switching identification systems are language-specific; they distinguish between a fixed number of (typically two) specified languages in a training corpus. This approach fails to support code-switching research in lower-resource languages, where annotated training data is either not available or available at much smaller scales. To collect more data for low-resource language pairs requires an identification system, a circular problem. This circular problem applies more generally to language identification. But it is especially challenging in code-switching because code-switching sentences are often only found in seas of spoken/written data of primarily monolingual sentences.

3 Existing Corpora

Large language-identified corpora of code-switching with English only exist in a small set of language pairs, namely Hindi–English (Singh et al., 2018), Spanish–English (Molina et al., 2016; Aguilar et al., 2018a), Nepali-English (Solorio et al., 2014), German–English (Osmelak and Wintner, 2023) and Arabic–English (Shehadi and Wintner, 2022b). Smaller corpora of code-switching of low-resource language pairs also exist, e.g. Indonesian–English (Barik et al., 2019), Turkish–English (Yirmibeşoğlu and Eryiğit, 2018) and Vietnamese–English (Nguyen and Bryant, 2020). These corpora are derived from posts on social media platforms such as Twitter and Reddit, except for the Vietnamese–English corpus which is of spoken code-switching.

Of the corpora, there is a variation in the labelset used to classify the words. The variation is centred around the annotation of shared words and words of mixed morphology. Example labelsets, alongside the frequency of words of each label, are given in Tables 1 and 2.

In many public corpora of low-resource language pairs, code-switching is identified at a coarser-grained level. These corpora only include labels for each of the two languages, and sometimes a third label for all tokens not of the two languages. Meanwhile, higher-resource language pairs include the identification of named entities, or more generally shared words, mixed words and foreign words not of either the two languages in question. The labelset proposed by (Molina et al., 2016, the second shared task on language identification in code-switching) includes 'lang1', 'lang2', 'other', 'ne' (named entity), 'fw' (foreign word), 'mixed', 'unk' and 'ambiguous' labels. This labelset was adapted from Solorio et al. (2014, the first shared task) which has the same labels except without 'fw' or 'unk'.

Hindi–English, Spanish–English and Nepali–English code-switching datasets have been brought together in the LinCE benchmark (Aguilar et al., 2020), under a language identification (LID) task for code-switching data. They use the labels of Molina et al. (2016) or Solorio et al. (2014).

In addition to code-switching identification, LinCE includes a benchmark for named entity recognition (NER) in code-switching data. The code-switching examples in the LID and NER benchmarks are different.

In the Denglisch corpus of Osmelak and Wintner (2023), German–English code-switching is identified at a more fine-grained level. Figure 1b displays the fine-grained labels they annotate words for, demonstrating the number of linguistic phenomena in code-switching inter-play. In their work, they use 100% of their human-annotated data in the

	1	2	punct	EOS	EOP	4b	3a	3a-D	3a-E	3a-AD	4a	3a-AE	url	4c	3-O	3c-C
Train (3364)	24134	23598	9016	3351	2976	899	460	405	311	210	206	191	184	99	96	80
Dev (420)	2621	3093	992	420	212	58	51	37	46	26	18	18	15	10	2	11
Test (421)	3125	2914	1082	417	279	80	62	43	35	26	19	21	10	9	12	9

	3c-M	4d	4b-D	3-D	3b	4	3-E	4d-D	3c-EC	4e-E	4b-E	4d-E	3c	3	3c-EM
Train (3364)	76	71	65	60	58	48	46	41	22	15	14	14	12	7	5
Dev (420)	13	10	4	0	13	1	9	7	1	0	3	5	2	0	1
Test (421)	9	7	5	0	16	2	6	3	5	1	0	2	0	0	0

(a) Label frequencies

- 1 English
- 2 German
- 3 Overlaps
  - 3a Named Entities
    - 3a-E English Origin
    - 3a-D German Origin
    - 3a-AE Adapted to English
    - 3a-AD Adapted to German
  - 3c Merge-Words
    - 3c-C Compounds
    - 3c-M Morphology
    - 3c-EC Entity Compounds
    - 3c-EM Entity Morphology
  - 3b Ambiguous Words
    - 3-E Untranslatable English
    - 3-D Untranslatable German
    - 3-O Untranslatable Other
- 4 Neutral
  - 4a Foreign
  - 4b Numbers
    - 4b-E English only
    - 4b-D German only
  - 4c Smiley
  - 4d Interjections
    - 4d-E English only
    - 4d-D German only
    - 4e-E English abbr.
  - <url> URL
  - <punct> Punctuation
  - <EOS> End of Sentence
  - <EOP> End of Paragraph

(b) Annotation scheme. Source: Osmelak and Wintner (2023)

Table 1: Details of the Denglisch corpus of German–English code-switching (Osmelak and Wintner, 2023)

	lang1	lang2	other	ne	fw	mix	unk	amb
Train (4823)	54720	19134	14017	6069	398	33	10	8
Dev (744)	8942	3303	2210	837	29	5	2	1
Test (1854)	20635	7487	5369	2432	106	14	5	32

(a) Hindi–English (Singh et al., 2018)

	lang2	lang1	other	ne	amb	unk	mix	fw
Train (21030)	111422	77843	53851	4725	263	210	27	22
Dev (3332)	14787	16618	7810	769	37	32	3	2
Test (8289)	42850	31916	20311	2059	100	80	17	8

(b) Spanish–English (Molina et al., 2016)

	lang2	lang1	other	ne	mix	amb
Train (8451)	49936	38827	29847	3146	90	72
Dev (1332)	8385	5557	4653	452	13	11
Test (3228)	19881	14009	11321	1268	48	32

(c) Nepali–English (Solorio et al., 2014)

	id	un	en
Test (825)	11200	5917	5608

(d) Indonesian–English (Barik et al., 2019)

	t	e
Test (377)	3941	1489

(e) Turkish–English (Yirmibeşoğlu and Eryiğit, 2018)

	@vie	@eng	@non
Test (3313)	16974	7219	614

(f) Vietnamese–English (Nguyen and Bryant, 2020)

Table 2: Code-switching identification corpora, with frequencies of labels. lang1 is always English.

cross-validation setup. Their data can be collapsed to have a labelset similar to the data in LinCE.

The Arabic–English code-switching dataset contains labels for 'Shared Other' words, which are less simple to adapt to the LinCE labelset, likely requiring some further annotation.

For low-resource language pairs, Turkish–English (Yirmibeşoğlu and Eryiğit, 2018) includes only binary labels (Turkish and English), Indonesian–English (Barik et al., 2019) adds an 'other' (or 'unknown' as they called it) category for named entities, punctuation and other non-language units. The Vietnamese–English CanVEC corpus includes the same three categories, but their data is semi-automatically annotated; a human only corrects words not contained in wordlists of either language, and words in both wordlists.

SoTA language identification performance for the high-resource language pairs is displayed on the LinCE benchmark leaderboard.[3] As of 23 April 2024, the best system is the XLM-RoBERTa language model (Conneau et al., 2020) fine-tuned separately for classification on each of the language pairs. This is an anonymous submission and no reference is given to the exact training setup. There is no existing language identification baseline on the Vietnamese–English corpus, likely because it is semi-automatically annotated data. For Indonesian–English and Turkish–English, SoTA language identification performance remains from the original works; both using conditional random field (CRF) classifiers. Like Osmelak and Wintner (2023) do for German–English, these systems also use 100% of their corpora in the cross-validation setup. They release no separate test set.

The disparity in size and labelset of these code-switching corpora has presented a challenge to research in this field. The best code-switching iden-

[3] https://ritual.uh.edu/lince/leaderboard

tification systems are language pair-specific. This has left low-resource language pairs behind in code-switching research. In addition, there is no baseline for research on new language pairs to evaluate against.[4]

4 AnE

Our goal is to develop an Any-English (AnE) code-switching identification system, which we reformulate as the task of identifying English code-switching in a sea of text of other languages. English here encompasses the many local varieties of English present in the aforementioned corpora of code-switching. We aim to achieve our goal by matching up the labelsets of existing corpora with this task in mind. A key challenge we face is that some corpora distinguish named entities, whilst others do not.

To alleviate this challenge, we will train two classifiers:

1. **Code-switching identification -** one will distinguish between *English*, other languages (hereinafter *notEnglish*), words that mix English and another language within the word (*Mixed*) and other words such as punctuation, emojis and mentions (*Other*).
2. **Binary named entity recognition -** the other will make a binary distinction as to whether a word is part of a named entity or not.

The data we searched for to train these classifiers broadly fits into three categories. The first is corpora that classify the language of the words but also have a named entity class (LID+NER). These corpora are labelled with the previously discussed labelsets of Molina et al. (2016) or Solorio et al. (2014). The second is corpora that only classify the language of the words as L1 or L2 (LID). Some of these corpora also have an 'other' category which includes named entities/punctuation/emojis etc., and some simply remove 'other' words from the data by manual means. The third is derived from the task of named entity recognition on code-switching text; such corpora include the named entity labels and classes in BIO (Ramshaw and Marcus, 1995) format (NER).

We preprocess the corpora as follows.

[4]Except by prompting LLMs, of which only the largest models perform well (Zhang et al., 2023). This is currently a subpar and prohibitively expensive solution.

	English	notEnglish	Mixed	Other
Train (4823)	54720	19550	33	14017

(a) Hindi–English (LinCE-LID, Singh et al., 2018)

Train (21030)	77843	111917	27	53851

(b) Spanish–English (LinCE-LID, Molina et al., 2016)

Train (33611)	78588	199723	45	110015

(c) Spanish–English (LinCE-NER, Aguilar et al., 2018a)

Train (8451)	38827	50008	90	29847

(d) Nepali–English (LinCE-LID, Solorio et al., 2014)

Train (3364)	24725	24865	195	16525

(e) German–English (Osmelak and Wintner, 2023)

Table 3: Collapsed LID training data statistics

	I	O
Train (4823)	6069	88320

(a) Hindi–English (Singh et al., 2018)

Train (21030)	4725	243638

(b) Spanish–English (Molina et al., 2016)

Train (8451)	3146	118772

(c) Nepali–English (Solorio et al., 2014)

Train (3364)	1577	65193

(d) German–English (Osmelak and Wintner, 2023)

Train (1243)	2222	17806

(e) Hindi–English (Singh et al., 2018)

Train (33611)	11722	385055

(f) Spanish–English (Aguilar et al., 2018a)

Table 4: Binary NER training data statistics

- **LID+NER** Each corpora becomes two subcorpora. In the first, the language other than English, foreign words, ambiguous words and unknown words all become *nonEnglish*. The English, Mixed and Other tags stay as *English*, *Mixed* and *Other*. Named entities receive a special ID to be ignored in all training updates. In the second subcorpora, named entities become a generic inside (*I*) and all other labels become outside (*O*).
- **LID** All labels are taken directly, which always includes *English* and *notEnglish*. *Other* is also taken if included in the data. There were no *Mixed* labels in any of these corpora.

- **NER** All B or I labels of any type become an inside (*I*) label. All outside (*O*) labels stay.

Table 3 gives statistics for the output from the collapse of the corpora into our LID scheme, and Table 4 for the collapse into binary NER. Statistics for the LID-only category of corpora follow directly from Table 2 (d)-(f).

5 Experiment

5.1 Experimental Setup

Systems and Baselines Our AnE system is an ensemble of the language identification (AnE_{LID}) and binary named entity recognition (AnE_{NER}) classifiers. The ensemble is achieved by classifying words based on each classifier separately, and then overwriting labels of words AnE_{NER} predicts to be named entities. For the high-resource language pairs, these labels make up the *NamedEntity* class. For the low-resource language pairs with an *Other* category, they are moved into there. For languages without any *Other* category, AnE_{NER} is not used and AnE_{LID}'s *Other* labels become *NotEnglish*. The low-resource language pair copora also do not include a *Mixed* category, so such predicted words become *NotEnglish*.

For the LinCE benchmark language pairs, and German–English, we train separate baseline classifiers using only data from each single language pair. This baseline corresponds to reproducing the SoTA (anonymous) system on the LinCE leaderboard. For the low-resource language pairs, we test on 100% of the data. Therefore, we are unable to train a baseline system. Instead, we use the best-performing system from the original works as baseline, even though these were trained in the cross-validation setup.[5]

All classifiers are single-layer perceptron classification heads on XLM-RoBERTa (large).

Data We use the data described in Section 4. In particular, we use the provided splits from the LinCE benchmark (Aguilar et al., 2020), which includes three language pairs. We also mix in Denglisch (German–English) data from Osmelak and Wintner (2023). In their work, they train with the cross-validation setup. We instead split their data into train/dev/test with splits 80:10:10%.

We balance the training data between these four language pairs by up-sampling until all language pairs contribute the same number of training sentences.

We also evaluate on 100% of the three low-resource language pair corpora, namely Indonesian–English, Turkish–English and Vietnamese–English. We remind the reader that the Vietnamese–English corpus is different to all other corpora in that (a) it is a corpus of spoken code-switching and (b) it is only silver-standard data.

Training We train all systems for 3 epochs with a learning rate of 1e-5 and a batch size of 32. All parameters are updated using a cross-entropy loss criterion and the Adam optimizer (Kingma and Ba, 2014). We use weight decay = 0.01 for the optimizer with $\beta = (0.9, 0.999)$ and ϵ =1e-8. For the named entity tokens without language subcategorization, as described in Section 4, losses are zeroed. These hyperparameters were chosen based on recommendations from prior work (e.g., Devlin et al., 2019). No hyperparameter tuning was performed.

When training the baseline systems, we continue training for additional epochs until the same number of sentences are seen as in the up-sampled AnE data for the language pair in question. We found validation accuracy monotonically increases and plateaus by the end of training; there was no evidence of overfitting despite this extended training setup.

Metrics We will compare the performance of the AnE system against baseline by computing precision (P), recall (R) and weighted-average F_1 measure. All the measures are word-based. The LinCE submission portal generates P, R and F_1 metrics for each label, and an overall weighted F_1 measure.[6] We also use weighted-average F_1 measure for other evaluations.

XLM-RoBERTa uses byte-pair encoding for subword tokenization. If there is more than one unique subword label for a given word, we select the most frequent label. In the event of a tie, we select the label which appeared first. This detail is likely to particularly affect the classification of mixed-morphology words, which will often be split into subwords. Further investigation of this effect is beyond the scope of this work.

[5]Therefore, numbers are not directly comparable. But either way, our system is not favoured as it does not have any training data for these language pairs.

[6]No overall P or R is provided.

	English			notEnglish			Mixed		
	P	R	F_t	P	R	F_t	P	R	F_t
Hindi–English		*(20635)*			*(7487)*			*(14)*	
hi-en only	98.39	98.47	98.43	95.77	96.66	96.21	47.37	64.29	54.55
AnE	98.32	98.49	98.40	94.24	96.61	95.41	61.54	57.14	59.26
Spanish–English		*(42850)*			*(31916)*			*(17)*	
es-en only	98.22	98.94	98.58	98.98	99.22	99.10	0.00	0.00	0.00
AnE	98.51	98.62	98.57	99.01	99.13	99.07	54.55	35.29	42.86
Nepali–English		*(19881)*			*(14009)*			*(48)*	
ne-en only	96.34	96.90	96.62	98.27	98.07	98.17	54.29	39.58	45.78
AnE	96.71	96.25	96.48	97.78	98.40	98.09	62.50	41.67	50.00
German–English		*(3134)*			*(2978)*			*(23)*	
de-en only	98.76	99.23	99.00	99.39	98.93	99.16	78.26	78.26	78.26
AnE	97.13	99.27	98.19	99.06	98.62	98.84	63.64	60.87	62.22
	Named Entity			**Other**			**Ambiguous**		
Hindi–English		*(2432)*			*(5369)*			*(32)*	
hi-en only	90.18	89.14	89.66	99.16	98.96	99.06	0.00	0.00	0.00
AnE	91.77	87.54	89.60	98.38	98.44	98.41	0.00	0.00	0.00
Spanish–English		*(2059)*			*(20311)*			*(100)*	
es-en only	87.76	82.18	84.88	99.82	99.78	99.80	0.00	0.00	0.00
AnE	77.45	81.40	79.37	99.81	99.82	99.82	0.00	0.00	0.00
Nepali–English		*(1268)*			*(11321)*			*(32)*	
ne-en only	73.07	74.68	73.87	97.63	97.32	97.48	8.33	3.12	4.55
AnE	72.52	75.95	74.19	97.68	97.06	97.37	0.00	0.00	0.00
German–English		*(187)*			*(1877)*				
de-en only	90.67	93.58	92.11	100.00	99.63	99.81	-	-	-
AnE	88.54	90.91	89.71	99.89	96.70	98.27	-	-	-
	Unknown			**Foreign Word**			**Overall**		
Hindi–English		*(5)*			*(106)*			*(36080)*	
hi-en only	0.00	0.00	0.00	87.10	50.94	64.29	-	-	97.33
AnE	0.00	0.00	0.00	0.00	0.00	0.00	-	-	96.86
Spanish–English		*(80)*			*(8)*			*(97341)*	
es-en only	50.00	5.00	9.09	0.00	0.00	0.00	-	-	98.58
AnE	0.00	0.00	0.00	0.00	0.00	0.00	-	-	98.44
Nepali–English								*(46559)*	
ne-en only	-	-	-	-	-	-	-	-	96.76
AnE	-	-	-	-	-	-	-	-	96.66
German–English								*(8199)*	
de-en only	-	-	-	-	-	-	99.03	99.02	99.03
AnE	-	-	-	-	-	-	98.17	98.15	98.15

Table 5: Results for the LID task for language pairs in the training data

5.2 Results

Table 5 gives test results on the four language pairs included in the training data. Mixing the data to train one AnE model does not result in a large change in performance compared to the separate baseline models. Overall F_1 measures for the baseline and AnE are 97.33/96.86% for Hindi–English, 98.58/98.44% for Spanish–English, 96.76/96.66% for Nepali–English, and 99.03/98.15% for German–English. AnE is numerically worse for all language pairs, but only by a small margin of less than 1% absolute F_1.

For the first three, which are all from LinCE, the differences are all less than 0.5%. For German–English, it is slightly larger (-0.88%). We collapsed the labels for German–English to match the LinCE evaluation labels where possible. But there may be some differences between the LinCE data and the German–English data scheme. This may be a cause of the slightly greater drop in the overall performance of AnE for this language pair.

AnE also does not have predictive classes 'Ambiguous', 'Unknown' or 'Foreign Word'. There are few (all < 106) words in these categories in the test data. Nevertheless, the AnE system scores zero for all these categories, which may be another reason for the small numerical drop in overall performance compared to the baselines.

The separate baseline classifiers perform near-identical to the anonymous SoTA reported on the LinCE leaderboard.

In terms of evaluating our approach of separating out the binary NER task, the results show that AnE_{NER} in the AnE ensemble is near-identical to the baseline where 'Named Entity' is simply a label amongst the other labels. In particular, named entity F_1 measure for the baseline and AnE is recorded at 89.66/89.60% for Hindi–English, 84.88/79.37% for Spanish–English, 73.87/74.19% for Nepali–English and 92.11/89.71% for German–English. The reduced performance in Spanish–English compared to the baseline can be attributed to a substantially worse precision (77.45 vs. the baseline 87.76). AnE_{NER} was trained on both the

	English			notEnglish			Other			Overall		
	P	R	F_t	P	R	F_t	P	R	F_t	P	R	F_t
Indonesian–English		*(5608)*			*(11200)*			*(5917)*			*(22725)*	
id-en SoTA	89.90	84.42	87.07	88.13	96.22	91.99	94.99	83.96	89.14	90.70	87.38	88.86
AnE	86.86	97.63	91.93	95.44	94.86	95.15	97.13	86.83	91.69	93.76	93.45	93.45
Turkish–English		*(1489)*			*(3941)*						*(5430)*	
tr-en SoTA	91.7	92.2	91.9	97.2	96.8	97.0	-	-	-	95.7	95.5	95.6
AnE_{LID}	94.16	98.46	96.26	99.41	97.69	98.54	-	-	-	97.97	97.90	97.91
Vietnamese–English		*(7219)*			*(16974)*			*(614)*			*(24807)*	
AnE	90.64	95.07	92.80	98.17	95.82	96.98	60.72	65.96	63.23	95.05	94.86	94.93

Table 6: Zero-shot LID results. SoTA results from Barik et al. (2019); Yirmibeşoğlu and Eryiğit (2018). No existing Vietnamese–English SoTA.

training sets of the LID task we are evaluating here, and collapsed NER training splits. It is possible that introducing these NER datasets brings a conflict of annotation guidelines. Alternatively it is possible that introducing additional data here was simply not necessary. Either way, we have shown here that AnE_{NER} is an optional NER module in our system that performs on-par with baseline.

We now proceed to evaluate performance on low-resource language pairs, for which AnE is not explicitly fine-tuned on any code-switching data. Table 6 gives results in P, R and F_1 measure on the Indonesian–English, Turkish–English and Vietnamese–English language pairs.

Zero-shot AnE outperforms SoTA classifiers fine-tuned directly (in the cross-validation setup) for the language pairs. F_1 measures in all categories are improvements over SoTA. For Indonesian–English code-switching, AnE is evaluated at overall $F_1 = 93.45\%$, outperforming the previous SoTA of 88.86%. The same holds for Turkish–English, where AnE_{LID} is evaluated at $F_1 = 97.9\%$ compared to the previous SoTA of 95.6% (significant figures/digits reduced to match reported SoTA).

Overall F_1 for Vietnamese–English is 94.93, but this is severely affected by the low score in the 'Other' category of $F_1 = 63.23\%$. This is because their 'X' category (which we collapse to 'Other') represents language-neutral words, rather than named entities/punctuation/emojis as our 'Other' here is targeted at. There is a mismatch here. Such labelled words only arise as a result of human intervention in their semi-automatic language identification process, which may be a factor. Another factor is that the Vietnamese–English data is the only corpus originating from recordings. Many of these ambiguous words arise from the conversational discourse setting not present in social media, like interjections and fillers. AnE is not able to effectively handle such words. We have still set baseline performance for Vietnamese–English code-switching identification.

We hypothesize the good zero-shot performance of AnE may be attributed to two factors. The first is the multilingual pre-training of XLM-RoBERTa. Monolingual training data in Indonesian, Turkish and Vietnamese as well as English is included in the multilingual pre-training data of XLM-RoBERTa. This may contribute to the performance of AnE in distinguishing these languages from English. The second factor is the way we formulated this task: distinguishing English from not English. This task formulation aimed to be independent of the other language.

This zero-shot evaluation shows that AnE performs well at identifying code-switched English amongst words of other languages not seen in the fine-tuning data. We consider this a good result, and it means AnE can be a baseline system for future research on code-switching between any language and English. It can also be a tool to quickly gather more data for low resource language pairs in code-switching research.

5.3 Correlation With Typological Similarity

We finish by investigating the connection between language typology and difficulty in recognizing code-switching. To this end, we used lexical similarity as a measure of linguistic similarity. We note that our language identification task is mostly lexical, in identifying the language of individual words. But for lexically similar languages it cannot be solved perfectly even with an ideal lexicon; interlingual homographs, words with the same surface form in two languages but different meanings, are one example for why.

A challenge in this investigation is what measure of difficulty in recognizing code-switching to use. We found earlier that overall F_1 score is heavily affected by how named entities and other words are annotated. Meanwhile, annotation schemes also

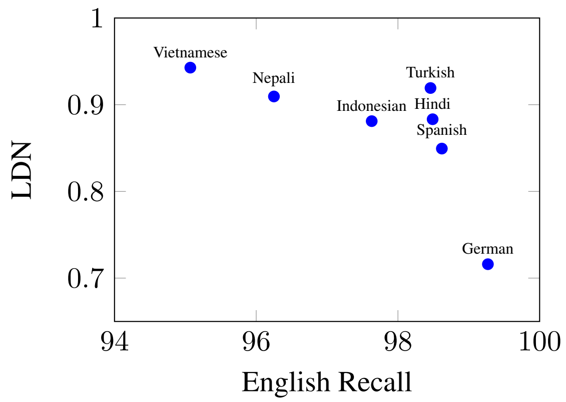


Figure 1: Normalized Levenshtein distance (LDN) between the ASJP wordlists in each of the languages, against English recall of AnE.

lead to some large variations between precision and recall. In the end, we decided to use the recall of English words as our measure. Ultimately, we anticipate this is the measure of most interest to those searching for code-switching in large corpora.

Figure 1 presents a plot of this recall measure against a measure of lexical similarity for each of the language pairs we explored in this work. Lexical similarity is computed as a distance using the ASJP corpus (Wichmann et al., 2022) and associated methods of Müller et al. (2010), but we acknowledge that no measure will be perfect.[7] We find a strong (ρ=-0.82) and significant (p=0.02) Spearman's correlation between lexical distance and English recall. But the correlation is negative, indicating that lexically similar languages are easier to distinguish. This is a surprising result, as one would expect that lexically dissimilar languages could be distinguished near-perfectly with a hashtable. We posit that an explanation may reside in the monolingual pre-training of the language model. It is plausible that the model learns representations that better distinguish lexically similar languages. An alternate hypothesis is that this correlation arises from the volume of pre-training data in each of the languages. It is also possible that the correlation is just a facet of the difficulty of each code-switching corpus we investigate.

6 Conclusion

In this work, we have presented a system (AnE) that distinguishes English words and words of other languages in multilingual text. On high-resource language pairs, the system underperforms language-pair-specific SoTA by a numerically small margin (always less than 1% absolute F_1). Meanwhile, it outperforms SoTA on low-resource language pairs, even though it was not trained on any code-switching of these language pairs. Analysis of our results revealed a negative correlation between lexical similarity and difficulty in recognizing code-switching, a surprising result which we leave to future work for further exploration. We believe our work bridges some of the resource-gap in code-switching research. We make it possible to compile new large-scale code-switching corpora of currently underrepresented language pairs. AnE is also a new and competitive baseline in code-switching identification research between any language and English.

Limitations

The main limitation of this work is in the language pairs AnE is able to support. The main motivation for this work was to make the most of existing high-resource code-switching data to support research on lower-resource language pairs in code-switching. We achieved this, but only for language pairs where one language is English.

There are of course many code-switching language pairs that do not involve English. But we found the data is not available today to train an AnE-type system to support those lines of research. For example, we would have wished to train a system that distinguished between code-switching of different language families, e.g. *Romance* vs. *notRomance*.

Ethics Statement

The Turkish–English and Vietnamese–English corpora we used were made available to us on our request to the authors of those works. The latter is a corpus of spoken code-switching, and hence comes with additional privacy constraints. We do not train any systems on that data, only using it for evaluation. All other corpora are publically available.

Acknowledgements

We are grateful to Andreas Vlachos for fruitful discussions and thank our four reviewers for their comments.

[7]We use the numbers released here.

References

Gustavo Aguilar, Fahad AlGhamdi, Victor Soto, Mona Diab, Julia Hirschberg, and Thamar Solorio. 2018a. Named Entity Recognition on Code-Switched Data: Overview of the CALCS 2018 Shared Task. In *Proceedings of the Third Workshop on Computational Approaches to Linguistic Code-Switching*, pages 138–147, Melbourne, Australia. Association for Computational Linguistics.

Gustavo Aguilar, Fahad AlGhamdi, Victor Soto, Thamar Solorio, Mona Diab, and Julia Hirschberg, editors. 2018b. *Proceedings of the Third Workshop on Computational Approaches to Linguistic Code-Switching*. Association for Computational Linguistics, Melbourne, Australia.

Gustavo Aguilar, Sudipta Kar, and Thamar Solorio. 2020. LinCE: A Centralized Benchmark for Linguistic Code-switching Evaluation. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 1803–1813, Marseille, France. European Language Resources Association.

Anab Maulana Barik, Rahmad Mahendra, and Mirna Adriani. 2019. Normalization of Indonesian-English code-mixed Twitter data. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, pages 417–424, Hong Kong, China. Association for Computational Linguistics.

Mirjam Broersma. 2009. Triggered codeswitching between cognate languages. *Bilingualism: Language and Cognition*, 12(4):447–462.

Mirjam Broersma and Kees De Bot. 2006. Triggered codeswitching: A corpus-based evaluation of the original triggering hypothesis and a new alternative. *Bilingualism: Language and cognition*, 9(1):1–13.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Ching-Ting Chang, Shun-Po Chuang, and Hung-Yi Lee. 2019. Code-Switching Sentence Generation by Generative Adversarial Networks and its Application to Data Augmentation. In *Proc. Interspeech 2019*, pages 554–558.

Michael G Clyne. 1980. Triggering and language processing. *Canadian Journal of Psychology/Revue canadienne de psychologie*, 34(4):400.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Margaret Deuchar. 2009. The miami corpus: Documentation file.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Penelope Gardner-Chloros. 2020. Contact and code-switching. *The handbook of language contact*, pages 181–199.

Deepak Gupta, Asif Ekbal, and Pushpak Bhattacharyya. 2020. A semi-supervised approach to generate the code-mixed text using pre-trained encoder and transfer learning. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2267–2280.

Asha Hegde, F Balouchzahi, Sharal Coelho, Shashirekha H L, Hamada A Nayel, and Sabur Butt. 2024. Coli@fire2023: Findings of word-level language identification in code-mixed tulu text. In *Proceedings of the 15th Annual Meeting of the Forum for Information Retrieval Evaluation*, FIRE '23, page 25–26, New York, NY, USA. Association for Computing Machinery.

Ahmad Fathan Hidayatullah, Atika Qazi, Daphne Teck Ching Lai, and Rosyzie Anna Apong. 2022. A systematic review on language identification of code-mixed text: Techniques, data availability, challenges, and framework development. *IEEE Access*, 10:122812–122831.

Navya Jose, Bharathi Raja Chakravarthi, Shardul Suryawanshi, Elizabeth Sherly, and John P. McCrae. 2020. A survey of current datasets for code-switching research. In *2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS)*, pages 136–141.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Dau-Cheng Lyu and Ren-Yuan Lyu. 2008. Language identification on code-switching utterances using multiple cues. In *Ninth Annual Conference of the International Speech Communication Association*.

Manuel Mager, Özlem Çetinoğlu, and Katharina Kann. 2019. Subword-level language identification for intra-word code-switching. *arXiv preprint arXiv:1904.01989*.

Deepthi Mave, Suraj Maharjan, and Thamar Solorio. 2018. Language Identification and Analysis of Code-Switched Social Media Text. In *Proceedings of the Third Workshop on Computational Approaches to Linguistic Code-Switching*, pages 51–61, Melbourne, Australia. Association for Computational Linguistics.

Giovanni Molina, Fahad AlGhamdi, Mahmoud Ghoneim, Abdelati Hawwari, Nicolas Rey-Villamizar, Mona Diab, and Thamar Solorio. 2016. Overview for the second shared task on language identification in code-switched data. In *Proceedings of the Second Workshop on Computational Approaches to Code Switching*, pages 40–49, Austin, Texas. Association for Computational Linguistics.

André Müller, Søren Wichmann, Viveka Velupillai, Cecil H Brown, Pamela Brown, Sebastian Sauppe, Eric W Holman, Dik Bakker, Johann-Mattis List, Dmitri Egorov, et al. 2010. Asjp world language tree of lexical similarity: Version 3 (july 2010). *Retrieved*, 10(19):2015.

Carol Myers-Scotton. 1992. Comparing codeswitching and borrowing. *Journal of Multilingual and Multicultural Development*, 13(1-2):19–39.

Ravindra Nayak and Raviraj Joshi. 2022. L3Cube-HingCorpus and HingBERT: A code mixed Hindi-English dataset and BERT language models. In *Proceedings of the WILDRE-6 Workshop within the 13th Language Resources and Evaluation Conference*, pages 7–12, Marseille, France. European Language Resources Association.

Li Nguyen and Christopher Bryant. 2020. CanVEC - the canberra Vietnamese-English code-switching natural speech corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4121–4129, Marseille, France. European Language Resources Association.

Li Nguyen, Christopher Bryant, Sana Kidwai, and Theresa Biberauer. 2021. Automatic language identification in code-switched hindi-english social media text. *Journal of Open Humanities Data*, 7.

Doreen Osmelak and Shuly Wintner. 2023. The denglisch corpus of German-English code-switching. In *Proceedings of the 5th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*, pages 42–51, Dubrovnik, Croatia. Association for Computational Linguistics.

Lance Ramshaw and Mitch Marcus. 1995. Text chunking using transformation-based learning. In *Third Workshop on Very Large Corpora*.

Mohd Sanad Zaki Rizvi, Anirudh Srinivasan, Tanuja Ganu, Monojit Choudhury, and Sunayana Sitaram. 2021. GCM: A toolkit for generating synthetic code-mixed text. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 205–211, Online. Association for Computational Linguistics.

Caroline Sabty, Islam Mesabah, Özlem Çetinoğlu, and Slim Abdennadher. 2021. Language identification of intra-word code-switching for arabic–english. *Array*, 12:100104.

Safaa Shehadi and Shuly Wintner. 2022a. Identifying code-switching in arabizi. In *Proceedings of the The Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 194–204.

Safaa Shehadi and Shuly Wintner. 2022b. Identifying code-switching in Arabizi. In *Proceedings of the The Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 194–204, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Kushagra Singh, Indira Sen, and Ponnurangam Kumaraguru. 2018. Language identification and named entity recognition in Hinglish code mixed tweets. In *Proceedings of ACL 2018, Student Research Workshop*, pages 52–58, Melbourne, Australia. Association for Computational Linguistics.

Thamar Solorio, Elizabeth Blair, Suraj Maharjan, Steven Bethard, Mona Diab, Mahmoud Ghoneim, Abdelati Hawwari, Fahad AlGhamdi, Julia Hirschberg, Alison Chang, and Pascale Fung. 2014. Overview for the first shared task on language identification in code-switched data. In *Proceedings of the First Workshop on Computational Approaches to Code Switching*, pages 62–72, Doha, Qatar. Association for Computational Linguistics.

Thamar Solorio, Shuguang Chen, Alan W. Black, Mona Diab, Sunayana Sitaram, Victor Soto, Emre Yilmaz, and Anirudh Srinivasan, editors. 2021. *Proceedings of the Fifth Workshop on Computational Approaches to Linguistic Code-Switching*. Association for Computational Linguistics, Online.

Thamar Solorio, Monojit Choudhury, Kalika Bali, Sunayana Sitaram, Amitava Das, and Mona Diab, editors. 2020. *Proceedings of the The 4th Workshop on Computational Approaches to Code Switching*. European Language Resources Association, Marseille, France.

Victor Soto, Nishmar Cestero, and Julia Hirschberg. 2018. The role of cognate words, pos tags and entrainment in code-switching. In *Interspeech*, pages 1938–1942.

Igor Sterner and Simone Teufel. 2023. TongueSwitcher: Fine-grained identification of German-English code-switching. In *Proceedings of the 6th Workshop on Computational Approaches to Linguistic Code-Switching*, pages 1–13, Singapore. Association for Computational Linguistics.

Søren Wichmann, Eric W. Holman, and Cecil H. Brown. 2022. The asjp database (version 20). http://asjp.clld.org/.

Genta Winata, Sudipta Kar, Marina Zhukova, Thamar Solorio, Mona Diab, Sunayana Sitaram, Monojit Choudhury, and Kalika Bali, editors. 2023. *Proceedings of the 6th Workshop on Computational Approaches to Linguistic Code-Switching*. Association for Computational Linguistics, Singapore.

Shuly Wintner, Safaa Shehadi, Yuli Zeira, Doreen Osmelak, and Yuval Nov. 2023. Shared Lexical Items as Triggers of Code Switching. *Transactions of the Association for Computational Linguistics*, 11:1471–1484.

Zeynep Yirmibeşoğlu and Gülşen Eryiğit. 2018. Detecting code-switching between turkish-english language pair. In *Proceedings of the 2018 EMNLP Workshop W-NUT: The 4th Workshop on Noisy User-generated Text*, pages 110–115.

Zheng Xin Yong, Ruochen Zhang, Jessica Forde, Skyler Wang, Arjun Subramonian, Holy Lovenia, Samuel Cahyawijaya, Genta Winata, Lintang Sutawika, Jan Christian Blaise Cruz, Yin Lin Tan, Long Phan, Long Phan, Rowena Garcia, Thamar Solorio, and Alham Aji. 2023. Prompting multilingual large language models to generate code-mixed texts: The case of south East Asian languages. In *Proceedings of the 6th Workshop on Computational Approaches to Linguistic Code-Switching*, pages 43–63, Singapore. Association for Computational Linguistics.

Ruochen Zhang, Samuel Cahyawijaya, Jan Christian Blaise Cruz, Genta Winata, and Alham Aji. 2023. Multilingual large language models are not (yet) code-switchers. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12567–12582, Singapore. Association for Computational Linguistics.

Yuan Zhang, Jason Riesa, Daniel Gillick, Anton Bakalov, Jason Baldridge, and David Weiss. 2018. A fast, compact, accurate model for language identification of codemixed text. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 328–337, Brussels, Belgium. Association for Computational Linguistics.

Studying Language Variation Considering the Re-Usability of Modern Theories, Tools and Resources for Annotating Explicit and Implicit Events in Centuries Old Text

Stella Verkijk
Vrije Universiteit Amsterdam
Huygens Institute
s.verkijk@vu.nl

Pia Sommerauer and **Piek T. J. M. Vossen**
Vrije Universiteit Amsterdam
p.sommerauer@vu.nl
p.t.j.m.vossen@vu.nl

Abstract

This paper discusses the re-usability of existing approaches, tools and automatic techniques for the manual annotation and automatic extraction of events in a challenging variant of centuries old Dutch in documents of the Dutch East India Company. We describe our annotation process and provide a thorough analysis of different versions of manually annotated data and the first automatic results from two fine-tuned Language Models. The paper studies to what extent we can use NLP theories and tasks formulated for modern English to design an annotation task for early modern Dutch and to what extent we can use NLP models and tools built for modern Dutch (and other languages) on early modern Dutch. We believe these analyses give us insight into how to deal with the large variation that language shows in describing events, and how this variation may differ across domains. We release the annotation guidelines, annotated data, and code (https://github.com/StellaVerkijk/VarDial2024).

1 Introduction

Event extraction is a well-researched but very challenging task in Natural Language Processing (NLP). Though there are many datasets, systems and ontologies created for event extraction, there is little consensus on how to create a robust system for heterogeneous material. This problem is amplified when the texts are centuries old and the context is to a large extent unknown.

In this paper, we study a use case of annotating early modern Dutch texts for event trigger detection and classification. These texts originate from the Dutch East India Company (VOC) archives. This corpus of handwritten communications within the VOC holds a vast amount of information on trade, culture, business, slavery and early globalisation, which took place across much of the Indian Ocean World in the 17th and 18th centuries. The complete corpus consists of twenty-five million pages. It has been hard to conduct historical research with this corpus, because of its size, and because not many people can read the handwritten text (see Figure 1[1]).

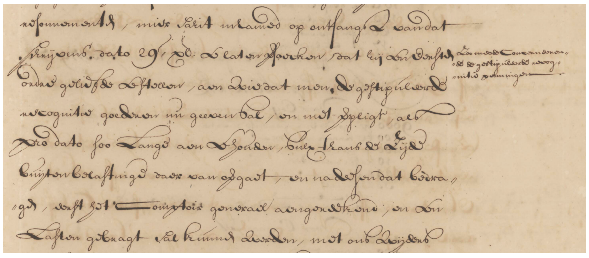


Figure 1: Snippet of the VOC archives

This paper describes the creation of a small annotated dataset that serves as a starting point for an automatic system that labels the archival material and enables a human-computer interaction solution: we are developing an event reconstruction pipeline to support a (re)search interface for historians.

The challenging nature of event extraction is showcased in recent results as reported in Hong et al. (2018) where deep learning systems achieve f-scores in the seventies for English, but drastically drop in performance when tested on data from a slightly different domain. Hong et al. (2018) also show how reaching high recall is a persistent problem in event extraction. We hypothesize that the main reason for this is because there is so much variation in how language is used to refer to events. We note that high recall is essential when building software that should support a search engine. In our case, we start our task unknown to the type and degree of variation in the language used, since much of the corpus' content and form remains unstudied. Although this poses considerable challenges, we can utilize event extraction as a looking glass

[1]National Archive, The Hague, The Netherlands, 1.04.02 (Archive of the VOC), inventory no. 1812, p. 33. https://www.nationaalarchief.nl/onderzoeken/archief/1.04.02/invnr/1812/file/NL-HaNA_1.04.02_1812_0803

Proceedings of the Eleventh Workshop on NLP for Similar Languages, Varieties, and Dialects (VarDial 2024), pages 174–187
June 20, 2024 ©2024 Association for Computational Linguistics

through which we study the variation in the language.

In order to create any automatic system for event extraction, we have to ask ourselves the question: What kind of variation of Dutch are we dealing with? The subjects discussed and the way they are discussed might be vastly different from other early modern sources. We are looking at two centuries of history of a huge organization that in the early 17th century had more operations in Asia than all other European nations combined (Lucassen, 2004). There were no spelling conventions, countless different clerks writing, summarizing or translating texts, and intricate political and cultural conventions to adhere to in the language.

We find that i) the challenging nature of event extraction as a task, ii) the fact that the performance of automatic solutions highly depends on how similar the domain they were trained on was and iii) the complexity of the language we work with itself specifically call for a tailored solution. In our case this begins with defining a new annotation task and subsequently fine-tuning language models pre-trained on different varieties of Dutch.

Our contributions are the following. Firstly, we discuss and illustrate the complexity of interpreting the language in this specific corpus, providing deep analyses of examples of our data. Secondly, we evaluate the re-usability of existing tools and resources by employing them on these examples, showing how models trained on modern language struggle with the variation present in our data. Thirdly, we present a new annotation approach where annotators work in teams and annotations are guided by an ontology specifically built for our data. We provide agreement analyses at different stages of the annotation process and show how our approach leads to an inter-annotator agreement (IAA) of 84% for trigger detection, 86% for classification (of 80+ event types) and 72% for the combined task of detection and classification. We also provide first insights of automatic solutions fine-tuned on the annotated data. Lastly, we publish our annotated datasets, containing a thoroughly analysed test set with annotations adjudicated by four historians and a linguist.

2 Related Work

Various English datasets have been annotated with events. While these approaches yielded valuable insights, none of the existing annotation schemes satisfies the needs of our use case. The main limitation lies in the selection of events annotated. Some of the proposed schemes only cover event types that refer to an event's aspectuality (distinguishing between state, process, action etc.) such as in Saurí et al. (2006) (as used in for example TempEval-3 (UzZaman et al., 2013)), ISO-TimeML (Pustejovsky et al., 2010) and THYME-TimeML (Styler IV et al., 2014) (as used in SemEval-2016: Clinical TempEval (Bethard et al., 2016)). Other datasets contain semantically more informative event types like TRANSPORT, but still only represent one corner of a modern Western world, such as ACE (Walker et al., 2006) and a light-weight version of ACE, ERE (Chen et al., 2023), both created to represent a limited number of event types of interest to the military, the latter created to make annotation easier and more consistent (Aguilar et al., 2014). FrameNet (Baker et al., 1998) is too specific for our purposes, requiring specialised linguistic knowledge about frame semantics not relevant for historical analysis. PropBank (Kingsbury and Palmer, 2003) and VerbNet (Schuler, 2005) are overly driven by syntax and lexica. Existing lexical and syntax-driven approaches do not fit our purposes because we are dealing with text that has no clear sentence boundaries (see Section 4) and for which we have very limited lexical semantic resources.

There has also been extensive research in the field of event-centric ontologies. However, the event classes they contain are mostly not representative for an early modern Dutch world (e.g., SUMO (Pease et al., 2002), DOLCE (Borgo et al., 2022)). For example, SUMO has a class for *PoliticalRevolution*, but none for *Mutiny* or a revolt that does *not* result in overthrowing of government. Also, while it has an entity class for *HumanSlave*, it does not feature an event like *Enslaving*. Still, we can draw on the way general ontologies include certain axioms, such as the Brandeis Semantic Ontology (BSO) (Pustejovsky et al., 2006) and the Rich Event Ontology (REO) (Bonial et al., 2021) that explicitly incorporate qualia relations. Even more relevant in this respect is the Circumstantial Event Ontology (CEO) (Segers et al., 2017), which includes pre-, during- and post- states of events, to incorporate *weak* causality. One event possibly causes a second when the post-state of the first equals a pre-state of the second. Pustejovsky (2021) urges to embed the state-change model from AI within the compositional model of semantics adopted in linguis-

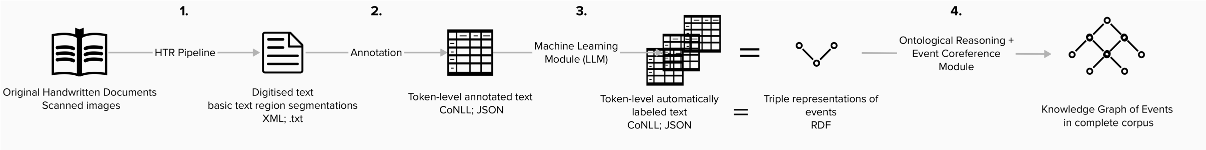


Figure 2: Event Reconstruction pipeline

tics. Verkijk and Vossen (2023) take Pustejovsky's and Segers' frameworks as a starting point and model Static Events (states) as logically inferred post-conditions of Dynamic Events (changes), e.g. an event like killing leads to a post-state of someone being dead. For this ontology, a team of historians identified and formulated event classes that are relevant for historical research in the VOC archive.

Recent state-of-the-art event extraction models using distributional embedding representations do not acquire an F-score above 0.77 for trigger detection and 0.74 for trigger detection + event classification on the ACE dataset (Hong et al., 2018). Hong et al. (2018) also show that systems trained on the broadcast news and newswire parts of the dataset and evaluated on the webblogs drop in F1 performance with 19.5-22 percentage points. This shows that even when adhering to the same annotation scheme, a difference in domain heavily influences performance. Finally, Hong et al. (2018) show that many systems demonstrate large gaps between precision and recall, where precision is almost always higher. We speculate that the variation in describing events is much larger than expected and requires other approaches than those offered by traditional NLP. We therefore expect that an end-to-end neural system for our use case can only partially reconstruct events and needs to be augmented with richer and more explicit semantics to connect the dots.

3 Approach

In order to support event-centric search in the archives, we aim to build an event-centric Knowledge Graph (KG). There are several steps that have to be undertaken to reach this end product: Figure 2 shows the most important steps in this pipeline. The handwritten documents first have to go through Handwritten Text Recognition (HTR) in order to become digitised (see Section 4). We then perform manual annotation on the digitised text. We plan to perform and experiment with some data augmentation at a later stage of the project, i.e. create synthetic training data. Finally, we fine-tune a Language Model (LM) to automatically annotate the rest of the corpus. This will provide us triple representations of events, which we gather in a KG. Through ontological reasoning we filter and complement our KG.

For the last step, we utilise the event ontology described by Verkijk and Vossen (2023). This event ontology is made for VOC archival material and models Static Events (SEs) as logical implications of Dynamic Events (DEs). For example, the election of a new person as king or raja in a certain region implies their status of *being a leader* from the moment of the election onward. Similarly, the Agent of a *Leaving* event *is no longer at the place* it left from the time of the event onward. These post-states of events are automatically inferable through the ontology. The ontology also features a taxonomic structure of DEs, allowing for generalizations like a *Leaving* event being a type of *Translocation* event. Such generalizations capture the variation in the data and language.

The ontology features 65 dynamic and 18 static events. Of the DEs, 50 of them imply a SE as a post-condition. There are only two SEs that cannot be inferred from the occurrence of a DE. The class *Dynamic Event* branches out in two classes that do not have any subclasses and five broad classes that branch out into more fine-grained subclasses. Those five classes are *SocialStatusChange*, *ChangeOfPossession*, *SocialInteraction* (with subclasses like *Mutiny, StartingAConflict*), *Translocation* and *InternalChange* (with subclasses like *Dying, Increasing, FallingIll*). The taxonomic structure of DEs is four steps at the deepest level.

The choice to create new manually annotated data following the event classes of Verkijk and Vossen's (2023) ontology was motivated by results of preliminary experiments we performed where we tried to use existing resources for automatic event detection (see Section 5). Furthermore, the ontology forms a closed world that guides annotations, where the richer semantics steer annotators to look for specific information in the text. It also enables us to alleviate some of the annotation labour:

The possibility of automatic inference allows us to infer unexpressed information. We expect that the automatic extraction of SEs will help solve a recall gap in future automatic labelling systems.

4 Data

4.1 The Corpus and its Contents

The corpus is the collection of *Overgekomen Brieven en Papieren* (Received Letters and Papers, OBP) within the VOC archive. The OBP contains the *Generale Missiven* (General Missives) and a large and varied collection of documents on which these missives are based. The General Missives are reports from the VOC's central administration (Council of India) in Batavia to the board (Gentlemen Seventeen) in the Dutch Republic. They contain accounts of all things current for the VOC world over almost two centuries, including, for example, detailed overviews of historical events, as well as social, political, economic and ecological developments. These narrative accounts begin with a brief introduction and then report in long sentences events that are more broadly described in the other documents that make up the OBP. In the margins are small summaries, which we call marginalia (see Figure 1).

The OBP spans the period 1610-1796 and contains around seven million handwritten pages. While the OBP has an average document length of 28 pages, a General Missive is on average 207 pages. For annotation, we selected pages from different types of documents from a range of different years. The annotated data we are releasing upon publication comprise 62 pages of 6 different documents. They include parts of General Missives, original missives, letters, journals and notes and they span a period of 151 years (1626-1777).

Throughout this paper we will refer to Example (1), the transcribed text of Figure 1, to illustrate the complexity of interpreting our data through an event annotation task. Example (1) is a snippet of a paragraph that spans four pages[1]. Within the paragraph, there is no indication of the end or beginning of a sentence. The text is written as one long description of happenings in a specific place at a specific time, which is a very common way of writing in our corpus. In order to illustrate the type of language used, we offer a word-by-word translation to English in (1b). For a paraphrased and more readable version and its translation to English, see Appendix A. Event triggers are printed in boldface. Corresponding event classes from our annotation scheme are *Getting; Request; SocialInteraction; Giving; HavingInPossession; ForcingToAct; FinancialTransaction; FinancialTransaction*.

(1a) **Original source**

'*(...) op* ***ontfangst*** *van dat „schrijvens, dato 29„e xb: te laten* ***versoecken****, dat hij ten Eersten* ***ordre*** *geliefde te stellen, aen wie dat men, de gestipuleerde recognitie goederen nu* ***geeven*** *bal, en niet* ***verpligt****, als pro dato soo Lange* ***aen tehouden****, sulx thans de zijde buijten* ***belastinge*** *daer van overgaet, en na deesen dat bedragen, eerst het comptoir generael* ***aengereekend; en ten Lasten gebragt*** *sal kunnen werden, (...)*[2]

(1b) **Literal translation**

(...) on ***reception*** *of that writing, date 29 xb to be* ***requested*** *that he firstly an* ***order*** *would like to establish, to who that one, the before identified taxable goods now shall* ***give****, and not* ***obliged****, if per the date so long* ***to hold****, so that the silk free of* ***tax*** *there from go off, and after these the amounts, first the local office general* ***charged; and debited*** *will be, (...)*

4.2 Data Processing: HTR

For HTR we use Loghi[3]. As mentioned, the archive contains handwritings of a vast amount of different people living in time periods that can differ more than a hundred years and there were no spelling conventions in early modern Dutch. On top of that, defining reading order and separating main text from marginalia is very challenging. Because of this, we are often dealing with very noisy output. For example, (1) showcases a character misclassification that transforms a verb into a noun. The transcribed 'bal' (*ball*) in 'aan wie men de gestipuleerde recognitie goederen nu geeven **bal**" (*to whom one the identified taxable goods now give* ***ball***) represents 'sal' (*shall*) in the original text. The untouched transcription of (1) is given in Appendix A, also showcasing how the HTR pipeline mis-identifies text regions, complicating the annotation process. Different transcription conventions in the different ground truth sets that Loghi was trained on, especially for punctuation, affect the transcriptions and make the Character Error Rate (CER) currently quite high (>10 percent). However,

[2]HTR errors related to region detection have been taken out of this example for clarity reasons

[3]https://github.com/knaw-huc/loghi

Loghi's HTR quality on our data using a classifying tool that was created for our data[4] shows that the HTR quality of a large majority of our corpus is what domain experts on the project classify as 'good'. Loghi is still under development and we expect to have digitized data of sufficient quality for our Event Extraction pipeline in the future.

5 Early Modern Dutch and Existing Tools

Since we have seen how unique our corpus is, we can expect existing NLP tools and models to perform poorly on our data. Also, we expect a high degree of variation, so even in cases where modern Dutch is similar, it is extremely hard to generalise. For example, early modern Dutch contains lexical items that no combination of subtokens in a Dutch Language Model (LM) trained on modern data can approximate to represent. An example of this is the word 'natgierig' (having alcoholic tendencies being described as an illness), which would be split in 'nat' (wet) and 'gierig' (greedy). We conducted several preliminary experiments to assess to what degree existing resources could be used to process our data, which we will discuss in this section.

5.1 Predicate Mapping

With a large amount of data to be annotated with a large amount of event classes, it is good practice to adopt heuristic methods to narrow down trigger and type candidates and automatically pre-label to help annotators (Wang et al., 2020). In the NewsReader pipeline (Vossen et al., 2016), events were extracted by linking them to FrameNet frames with the Predicate Matrix (PM) (Lopez de Lacalle et al., 2016). This matrix links entries in WordNet, VerbNet, PropBank, and FrameNet in different languages. As an experiment we tried to apply this approach on our data. We first extracted possible predicates with dependency parsing with spaCy, after which we automatically annotated the possible predicates with the corresponding lemmas and POS-tags by mapping them to a historic Dutch lexicon created by the Institute for Dutch Language (INT) made for OCR and OCR-postcorrection (for the period from 1550 to around 1970)[5]. We proceeded to select the set of lemmas with a verb POS-tag annotation of a mid-frequency range (occurring between 5 and 15 times in the corpus we had available at that time). We then provided those with translations to modern Dutch lemmas manually, using a dictionary that covers Dutch word meanings over several ages (Woordenboek der Nederlandsche Taal, WNT)[6]. Those translations were mapped to the PM and the corresponding FrameNet frames were extracted. We performed a small error analysis of this experiment which showed that the PM produced more false positives (126) than true positives (95)[7]. These results indicate that using existing resources for pre-annotation poses too many issues; we expect that developing our own lexicon for pre-annotation will be more fruitful.

5.2 Zero-shot POS-tagging

In order to see to what extent several LMs are familiar with lexical and syntactic aspects of early modern Dutch sentences, we tested their zero-shot POS-tagging accuracy on sample (1). Measuring zero-shot performance can give us insight into which models are best suited to fine-tune on our event extraction task. We do this by masking each token in the sample one by one and asking the models to fill the masked token each time. We then manually label the predicted tokens with POS-tags and compare these to the gold labels. Gold labels as well as the labelling of the predictions was done by an expert linguist. We also test two Dutch spaCy models.

The LMs we compare are RobBERT (Delobelle et al., 2020), trained on modern Dutch, XLM-R (Conneau et al., 2019), a multilingual RoBERTa model, which outperformed Dutch LMs in an entity labeling task on early modern Dutch in a study by Arnoult et al. (2021), and two versions of GysBERT (Manjavacas and Fonteyn, 2022), a LM pre-trained on historical Dutch. The first version of GysBERT was trained on 7.1B tokens spanning almost 500 years of Dutch data (up to 20th-century Dutch). Early modern Dutch was underrepresented in the training data. The second version of GysBERT was pre-trained in exactly the same way but with the inclusion of 1.3B extra tokens from early modern Dutch datasets, of which 940M tokens from our HTR'ed VOC archival material.

As we can see in Table 1, all scores are low. The second version of GysBERT outperforms all other models but not the best performing spaCy model. It is noteworthy that GysBERT outperforms XLM-R

[4] https://github.com/LAHTeR/htr-quality-classifier

[5] https://taalmaterialen.ivdnt.org/download/tstc-int-historische-woordenlijst/

[6] https://ivdnt.org/woordenboeken/woordenboek-der-nederlandsche-taal/

[7] The full report of this experiment can be found here

and RobBERT but has severely lower performance than GysBERT-v2.

model	accuracy
spacy_sm	.61
spacy_lg	**.66**
RobBERT	.38
XLM-R	.38
GysBERT	.46
GysBERT-v2	.65

Table 1: Zero-shot performance on POS-tagging of sample text in early modern Dutch

Looking at the individual predictions[8], we see a trend where XLM-R predicts very general tokens (adverbs, adpositions, determiners, pronouns, auxiliary verbs, conjunctions). This makes sense since XLM-R is trained to carry general information about several languages and is expected to be stronger when fine-tuned. This is worth further investigation. Parsing example (1) with the best performing model at this task, the largest spaCy model, still shows many issues, for example with 'versoucken' (requesting) being labeled as a noun, 'buijten' (free/outside) and 'bedragen' (amounts) as verbs and 'belastinge' (tax) as an adjective.

The results indicate that existing models seem to have encountered a diverging lexicon and syntactic structure in their training data. Though a base understanding of syntactic structure is necessary for any meaningful NLP task, we want to investigate whether existing models can be useful for other aspects of linguistic modelling.

5.3 Fill-mask for Events

In order to see whether LMs perform better at a semantically relevant task, we check how they fill masked event triggers (to control for cases where a model for example predicts the verb 'receive' in the place of the noun 'reception').

We used all LMs to fill masked event triggers in example (1) and provide results in Tables 11 and 12 in Appendix B. RobBERT, GysBERT and XLM-R all show very poor results. XLM-R and RobBERT do not predict the right token in any of their top 5 predictions for any masked event trigger, nor a token that has a similar meaning, and GysBERT only once. Noteworthy is that XLM-R predicts Dutch words in almost all cases both in this task and zero-shot POS-tagging. It therefore recognises this version of the language as Dutch.

Also telling is the fact that RobBERT never predicts any token with a confidence score of above 0.39; for GysBERT this is even lower, namely 0.27. GysBERT-v2 outperforms all models by far.

Existing resources and tools show unpromising results when confronted with our data. Even a model trained on historical Dutch but not on the VOC letters (GysBERT) is enormously outperformed by the exact same model but in which the VOC letters were included in the pretraining (GysBERT-v2). Additionally, pre-annotation methods using existing resources and heuristics also fail. We therefore argue for a new annotation scheme that captures the information we want to extract by clearly establishing i) our model of the world and ii) the way we deal with the variation in sense and reference, since the language in our corpus is often vague and woolly.

6 Annotation

6.1 Task

Annotators are presented with a document and are instructed to label any token or span of tokens that refers to an event that corresponds to one of the 83 event classes described in the ontology of Dynamic and Static events (Verkijk and Vossen, 2023). Apart from event trigger detection and classification, our annotators also labeled participants of each event. Which participants could be annotated for each specific event was specified in our event wiki.[9]

One of the most challenging parts of this task is deciding what it means for a string of tokens to refer to an event class (trigger detection). In order to facilitate the labelling of explicitly described events (directly referring to an event class) as well as implicitly described events (indirectly referring to an event class), we adopt two types of reference. A (span of) tokens either isOfType <eventclass> or evokes <eventclass>. We adopt this distinction from Postma et al. (2020) and Remijnse and Minnema (2020), who propose a very similar distinction for FrameNet annotation. The distinction is important for our annotation task because of the vague language in our data. For example, in (1), 'requested' directly refers to our event class *Request*, while 'order' is a noun that directly refers to an intangible entity, while it evokes a type of *SocialInteraction*. Also, 'to hold' directly refers to keeping something, but evokes *HavingInPosses-*

[8] https://github.com/StellaVerkijk/VarDial2024

[9] https://github.com/globalise-huygens/nlp-event-detection/wiki

sion and also *BeingAtAPlace*. Indirect referrals are essential to extract and model as much important information as possible (i.e., that fits in our model of the world, i.e. the predefined event classes).

The combination of the difficulties of the HTR'ed handwritten language we work with, as illustrated in Section 4.2, the linguistic and historical knowledge needed to annotate, and the inclusion of implicit reference annotation makes our task very challenging. We tried out different annotation settings in order to see what best practices are, which we will describe in the following section.

6.2 Annotation Settings

All annotations were performed by expert historians. They annotated individually in the first setting. Agreement was analysed on annotations of a General Missive of 1628[10], where we noticed that agreement in trigger detection was very low. We asked individual annotators to check each other's annotations, which we will refer to as the *check-task*. This check-task consisted of the following: each annotator was presented individually with all spans annotated by the other annotators as triggering an event, but not by them. They were asked to indicate whether they would now, reviewing it for a second time, also label it with an event class, and if so, with which one. We saw that they often agreed with each other's mention detection – hence, annotators were initially missing event triggers, most probably due to the demanding nature of the task. We further adjudicated the document we performed this experiment on into a test set, which meant we discussed each possible annotation among all annotators and an expert linguist after the check-task. We calculated precision and recall scores for trigger detection (no classification) before and after the check-task compared to the final test set. We see a steep increase in recall scores after the check-task (see Tables 7 and 8 in Appendix B). We therefore performed all further annotations in teams of two, so that annotators can discuss annotations and correct each other. We performed two more annotation rounds in this team setting. After each round, we sharpened the annotation guidelines, taking into account continuous feedback and questions.

[10]National Archive, The Hague, The Netherlands, 1.04.02 (Archive of the VOC), inventory no. 1092, folio 1, r. https://www.nationaalarchief.nl/onderzoeken/archief/1.04.02/invnr/1092/file/NL-HaNA_1.04.02_1092_0017

6.3 Ontological Resolutions

In order to compensate for the difficulty of the annotation task and provide a valuable IAA analysis, we also analyse results after performing two types of automatic resolutions.

Taxonomic resolutions We resolve disagreements on direct subclasses of the same class. E.g., when one annotator labels a token as a trigger for *Leaving* and another labels it for *Voyage*, it is resolved to a *Translocation* annotation (the superclass of *Leaving, Voyage, Arriving and Transportation*). If one annotator uses the superclass (*Translocation*) and another a direct subclass (*Transportation*), it is also resolved to the superclass.

Implicative resolutions The second type of resolution has to do with the implications built in the ontology, modeling how some dynamic events automatically imply a change in state, hence the occurrence of a static event (Section 3). Any event trigger label disagreements where one annotator chose a dynamic event and the other a related static event (e.g., one annotator chose *Attacking* and the other *BeingAtConflict*), the annotation was counted as an agreement and resolved to the static event (*BeingAtConflict*). This was done because the static event is the most conservative meaning (there are often multiple dynamic events that share the same static event as implication).

6.4 IAA Evaluation

Agreement on event mention detection + classification among annotators or annotator teams, presented in Table 2, was calculated with

$$\bar{A} = \frac{1}{2}(\frac{A^{xy}}{S^x} + \frac{A^{xy}}{S^y})$$

where A^{xy} is the number of spans both teams labeled with the same event class (using span overlap, not exact span matching), S^x is the total number of spans annotated with an event class by one annotator team and S^y is the total number of spans annotated with an event class by the other annotator team. We then calculate the ratio of agreed upon annotations out of all annotations made by one of the teams. We calculate this ratio for both teams and then take the average of the ratios as our agreement score. We decide to use a simpler calculation than Cohen's Kappa (Cohen, 1960), which includes a chance of accidental agreement in the calculation. Since we have many class types, chance of accidental agreement is quite low. Given that the Kappa score is not transparent and sensitive to skewed distributions, it is more informative to consider a simple ratio. We also provide results on only class agreement in Table 2. These results were calculated by comparing how often two annotators agreed on the class label, only considering those spans that received a class label from both annotators. In the first annotation round, there were four individual annotators. In the second annotation round, there were three teams of two, of which one new to the task. In the third round there were four, of which also one new to the task. The scores in Table 2 are average scores: for individual comparisons, see Appendix B, Tables 4 to 6.

	Det. + Class.			Class.		
	R1*	R2	R3	R1*	R2	R3
Before Resolution	.32	.49	**.57**	.59	.70	.68
After Resolution	.48	.55	**.72**	.91	.78	.86

Table 2: Average agreement scores on event detection + classification (Det. + Class.) and class agreement scores (Class.) between individuals* or teams in different annotation rounds (R = Round) (partial span overlap)

The results show that agreement increased with each round, indicating the task became more clearly defined through several rounds of discussion and reflection. The high score in classification in Round 1 can be explained through the low score in detection: the most obvious event triggers are also easiest to classify. In Round 2, trained teams annotated a total of 79 and 62 triggers respectively, whereas the untrained team annotated a total of only 21 triggers. In Round 3, trained teams annotated a total of 139, 147 and 151 triggers and the untrained team 141. Agreement score on only event trigger detection for the last round was 84% (see Table 9 in Appendix B), while in Round 2 this score was 63% comparing only trained teams and 45% including the untrained team. Note that class agreement is high in spite of a large selection of event classes (more than 80). It is hard to compare our results to IAA scores of other annotated datasets (like ACE) because they either do not evaluate trigger detection, cover much fewer event types, or report on different metrics. Wang et al. (2020) report a Cohen's Kappa score for trigger and type annotation of 38.2% and 42.7% respectively for crowd-source annotation with 168 event types in their contemporary English dataset MAVEN using pre-annotation with heuristics. See Table 10 in Appendix B for an overview of the annotated data we are releasing.

7 Automatic Baselines

The annotations should serve as training data for software that supports event-centric search in the VOC archives. In order to establish a baseline for this, we fine-tuned XLM-R and GysBERT-v2 on our event trigger detection task. Although XLM-R showed disappointing results in our preliminary experiments, it has shown to outperform general LMs at NLP tasks on historical Dutch (Arnoult et al., 2021) and there might be ways to leverage its general knowledge of language in the fine-tuning phase. For this experiment we split the development data we currently have available ('Dev' in Table 10) into a train set of 171KB and a test set of 22KB (json format). We fine-tuned both LMs on a token classification task for event mention detection (binary BIO classification). Since results with fine-tuned versions with early stopping showed low scores, we decided to try the grokking principle (Power et al., 2022; Murty et al., 2023) and evaluate several versions of fine-tuned models trained for increasing amounts of epochs, thereby training far beyond overfitting.

	XLM-R	GysBERT-v2
epochs	**P/R**	**P/R**
6	0 / 0	.31 / .06
9	.35 / .24	.20 / .08
12	.40 / .36	.26 / .14
20	.40 / **.43**	.35 / .16
50	.54 / .32	.42 / .20
150	.47 / .32	**.55** / .22

Table 3: Precision and recall scores of fine-tuned models on event trigger detection

Table 3 shows precision and recall scores on token level, which were obtained by mapping the model's prediction of the first sub-token to the complete token. The results show that GysBERT-v2 learns earlier from our data than XLM-R, which is in line with the results of our zero-shot experiments (Section 5). Surprisingly, XLM-R surpasses GysBERT-v2 in recall, and, for several epoch settings, also in precision. GysBERT-v2 eventually reaches slightly higher precision. The results indicate potential to leverage different LMs for different aspects of our task. Users could leverage different LMs at different levels of the system, allowing them to chose a model that suits their needs.

8 Discussion & Conclusion

This paper motivated a newly defined event annotation task by on the one hand discussing existing literature and theories on event extraction and on the other hand experimenting with existing tools. We show that the early modern Dutch used in the archives of the VOC is different from modern Dutch to such an extent that it calls for a tailored solution. We presented our bespoke annotation scheme and showed that a reasonable IAA could be reached by taking into account annotator needs and following an ontology that allows for the grouping of event classes through inference where necessary. Experiments with baseline automatic solutions for a VOC event-centric search engine show that we need to do more research into what kind of training strategies are needed for this task, and whether grokking can be a solution. Results seem to indicate that both more general LMs and more domain-specific LMs can be useful for different purposes. Future research should include a thorough comparison of different LMs, such as GysBERT and GysBERT-v2. We also aim to create more manually annotated data, develop a domain-specific lexicon for pre-annotation and experiment with automatic data augmentation techniques.

Acknowledgements

This research falls under the GLOBALISE project, funded by the Dutch Research Council (NWO) under project number 175.2019.003. We want to thank our annotators: Kay Pepping, Brecht Nijman, dr. Lodewijk Petram, dr. Manjusha Kuruppath, Femke Brink, Renate Smit, Pascal Konings and Philipp Huber. The code used to fine-tune XLM-R and GysBERT-v2 was an adapted version of code written by dr. Sophie Arnoult. Finally, we thank dr. Lodewijk Petram for proofreading.

References

Jacqueline Aguilar, Charley Beller, Paul McNamee, Benjamin Van Durme, Stephanie Strassel, Zhiyi Song, and Joe Ellis. 2014. A comparison of the events and relations across ACE, ERE, TAC-KBP, and FrameNet annotation standards. In *Proceedings of the second workshop on EVENTS: Definition, detection, coreference, and representation*, pages 45–53.

Sophie I. Arnoult, Lodewijk Petram, and Piek Vossen. 2021. Batavia asked for advice. pretrained language models for named entity recognition in historical texts. In *Proceedings of the 5th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 21–30.

Collin F. Baker, Charles J Fillmore, and John B Lowe. 1998. The Berkeley FrameNet project. In *COLING 1998 Volume 1: The 17th International Conference on Computational Linguistics*.

Steven Bethard, Guergana Savova, Wei-Te Chen, Leon Derczynski, James Pustejovsky, and Marc Verhagen. 2016. Semeval-2016 task 12: Clinical tempeval. In *Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016)*, pages 1052–1062.

Claire Bonial, Susan W Brown, Martha Palmer, and Ghazaleh Kazeminejad. 2021. The rich event ontology. *Computational Analysis of Storylines: Making Sense of Events*, page 47.

Stefano Borgo, Roberta Ferrario, Aldo Gangemi, Nicola Guarino, Claudio Masolo, Daniele Porello, Emilio M Sanfilippo, and Laure Vieu. 2022. Dolce: A descriptive ontology for linguistic and cognitive engineering. *Applied ontology*, 17(1):45–69.

Song Chen et al. 2023. DEFT English Light and Rich ERE annotation LDC2023T04. *Philadelphia: Linguistic Data Consortium.*

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Pieter Delobelle, Thomas Winters, and Bettina Berendt. 2020. Robbert: a dutch roberta-based language model. *arXiv preprint arXiv:2001.06286*.

Yu Hong, Wenxuan Zhou, Jingli Zhang, Guodong Zhou, and Qiaoming Zhu. 2018. Self-regulation: Employing a generative adversarial network to improve event detection. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 515–526.

Paul Kingsbury and Martha Palmer. 2003. PropBank: the next level of treebank. In *Proceedings of Treebanks and lexical Theories*, volume 3. Citeseer.

Maddalen Lopez de Lacalle, Egoitz Laparra, Itziar Aldabe, and German Rigau. 2016. Predicate matrix: automatically extending the semantic interoperability between predicate resources. *Language Resources and Evaluation*, 50:263–289.

Jan Lucassen. 2004. A multinational and its labor force: The dutch east india company, 1595–1795. *International Labor and Working-Class History*, 66:12–39.

Enrique Manjavacas and Lauren Fonteyn. 2022. Nonparametric word sense disambiguation for historical languages. In *Proceedings of the 2nd International*

Workshop on Natural Language Processing for Digital Humanities, pages 123–134.

Shikhar Murty, Pratyusha Sharma, Jacob Andreas, and Christopher D Manning. 2023. Grokking of hierarchical structure in vanilla transformers. *arXiv preprint arXiv:2305.18741*.

Adam Pease, Ian Niles, and John Li. 2002. The suggested upper merged ontology: A large ontology for the semantic web and its applications. In *Working notes of the AAAI-2002 workshop on ontologies and the semantic web*, volume 28, pages 7–10.

Marten Postma, Levi Remijnse, Filip Ilievski, Antske Fokkens, Sam Titarsolej, and Piek Vossen. 2020. Combining conceptual and referential annotation to study variation in framing. In *Proceedings of the International FrameNet Workshop 2020: Towards a Global, Multilingual FrameNet*, pages 31–40, Marseille, France. European Language Resources Association.

Alethea Power, Yuri Burda, Harri Edwards, Igor Babuschkin, and Vedant Misra. 2022. Grokking: Generalization beyond overfitting on small algorithmic datasets. *arXiv preprint arXiv:2201.02177*.

James Pustejovsky. 2021. The role of event-based representations and reasoning in language. *Computational Analysis of Storylines: Making Sense of Events*, page 23.

James Pustejovsky, Catherine Havasi, Jessica Littman, Anna Rumshisky, and Marc Verhagen. 2006. Towards a generative lexical resource: The brandeis semantic ontology. In *LREC*, pages 1702–1705.

James Pustejovsky, Kiyong Lee, Harry Bunt, and Laurent Romary. 2010. Iso-timeml: An international standard for semantic annotation. In *LREC*, volume 10, pages 394–397.

Levi Remijnse and Gosse Minnema. 2020. Towards reference-aware FrameNet annotation. In *Proceedings of the International FrameNet Workshop 2020: Towards a Global, Multilingual FrameNet*, pages 13–22.

Roser Saurí, Jessica Littman, Bob Knippen, Robert Gaizauskas, Andrea Setzer, and James Pustejovsky. 2006. TimeML annotation guidelines version 1.2. 1.

Karin Kipper Schuler. 2005. *VerbNet: A broad-coverage, comprehensive verb lexicon*. University of Pennsylvania.

Roxane Segers, Tommaso Caselli, and Piek Vossen. 2017. The circumstantial event ontology (ceo). In *Proceedings of the Events and Stories in the News Workshop*, pages 37–41.

William F. Styler IV, Steven Bethard, Sean Finan, Martha Palmer, Sameer Pradhan, Piet C. De Groen, Brad Erickson, Timothy Miller, Chen Lin, Guergana Savova, et al. 2014. Temporal annotation in the clinical domain. *Transactions of the association for computational linguistics*, 2:143–154.

Naushad UzZaman, Hector Llorens, Leon Derczynski, James Allen, Marc Verhagen, and James Pustejovsky. 2013. Semeval-2013 task 1: Tempeval-3: Evaluating time expressions, events, and temporal relations. In *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 1–9.

Stella Verkijk and Piek Vossen. 2023. Sunken ships shan't sail: Ontology design for event reconstruction in the dutch east india company archives. In *Proceedings of the Fourth Conference on Computational Humanities Research.*

Piek Vossen, Rodrigo Agerri, Itziar Aldabe, Agata Cybulska, Marieke van Erp, Antske Fokkens, Egoitz Laparra, Anne-Lyse Minard, Alessio Palmero Aprosio, and German Riga. 2016. NewsReader: using knowledge resources in a cross-lingual reading machine to generate more knowledge from massive streams of news. *Knowledge-Based Systems*.

Christopher Walker et al. 2006. ACE 2005 multilingual training corpus LDC2006T06. *Philadelphia: Linguistic Data Consortium*.

Xiaozhi Wang, Ziqi Wang, Xu Han, Wangyi Jiang, Rong Han, Zhiyuan Liu, Juanzi Li, Peng Li, Yankai Lin, and Jie Zhou. 2020. Maven: A massive general domain event detection dataset. *arXiv preprint arXiv:2004.13590*.

A Example texts

- **Original source, parts of marginalia in boldface and in brackets**[11]
  '(...) op ontfangst van dat „schrijvens, dato 29„e xb: te laten versoecken, dat hij ten Eersten **[zomee conconeren„]** ordre geliefde te stellen, aen wie dat men, de gestipuleerde recognitie goederen nu geeven bal, en niet verpligt, als pro dato soo Lange aen tehouden, sulx thans de zijde buijten belastinge daer van overgaet, en na deesen dat bedra„ „gen, eerst het comptoir generael aengereekend; en ten Lasten gebragt sal kunnen werden(...)'[1]

- **Literal translation**
  (...) on reception of that writing, date 29 xb to be requested that he firstly would like to be put an order, to who that one, the beforely identified taxable goods shall give, and not be obliged, if per the date so long to hold, so that the silk free of tax there be shipped off, and after these the amounts, first the local office general charged; and debited will be, (...)

[11]'zomee conconeren„' is one line in a marginalium that originally reads 'zo meede concerneerende(...)', meaning *also concerning...*.

- **Paraphrased Dutch**
  (...) op ontvangst van die brief, heeft hij op 29 december verzocht om instructies te krijgen aan wie dat men de besproken belastbare goederen zal geven, zodat hij niet gedwongen is de goederen zo lang ter plaatse te laten blijven dat hij daardoor belasting zal moeten betalen over de zijde, wat het lokale comptoir zal worden aangerekend, (...)

- **Paraphrased English**
  (...) on receiving the letter, he requested instructions on the 29th of December as to whom the goods should be given to, so that he will not be forced to keep the goods for such a long time that he would be forced to pay taxes for the silk, for which the regional office would be charged, (...)

B IAA results, data characteristics, fill-mask results

	T1/T2	T2/T1	T1/T3	T3/T1	T2/T3	T3/T2	avg-tr	avg
Before Resolution	.46	.58	.19	.71	.26	.76	.52	.49
After Resolution	.52	.66	.22	.81	.27	.81	.59	.55

Table 4: Agreement between seperate teams in Round 2 on event trigger detection + classification. avg-tr: only trained teams, avg: including the untrained team

	T4/T5	T5/T4	T4/T6	T6/T4	T4/T7	T7/T4	T5/T6	T6/T5	T5/T7	T7/T5	T6/T7	T7/T6	avg
BR	.52	.51	.60	.56	.57	.60	.53	.57	.57	.60	.60	.58	.57
AR	.68	.67	.72	.68	.71	.76	.74	.77	.71	.74	.74	.74	.72

Table 5: Agreement between separate teams in Round 3 on event trigger detection + classification. avg: including the untrained team

	Ann1	Ann2	Ann3	Ann4
Before Resolution				
Ann1	x	.36	.40	.31
Ann2	.26	x	.29	.22
Ann3	.43	.43	x	.34
Ann4	.25	.25	.25	x
After Resolution				
Ann1	x	.59	.49	.54
Ann2	.43	x	.38	.43
Ann3	.54	.59	x	.50
Ann4	.44	.47	.37	x

Table 6: Agreement between individual annotators on event trigger detection + classification (Round 1).

	P	R	n
Ann1	.85	.55	131
Ann2	.95	.35	75
Ann3	.81	.43	108
Ann4	.85	.44	105

Table 7: Precision and recall scores per annotator on event trigger detection before check-task. Gold = test set. n = true + false positives

	P	R	n
Ann1	.83	.81	199
Ann2	.85	.80	192
Ann3	.86	.70	166
Ann4	.82	.71	176

Table 8: Precision and recall scores per annotator on event trigger detection after check-task. Gold = test set. n = true + false positives

T4/T5	T4/T6	T4/T7	T5/T6	T5/T7	T6/T7
.82	.84	.83	.85	.83	.86

Table 9: Event trigger detection agreement in Round 3

	Pages	Docs	Agreement	years
Dev	57	5	59%	1626-1777
Test	5	1	100%	1628

Table 10: Characteristics of annotated data currently processed and of acceptable quality, which includes data annotated by the two trained teams in Round 2 and the adjudicated test set. The third annotation round is currently still in process

	RobBERT				XLM-R			
Masked token	Prediction	Probability	L2C?	FW?	Prediction	Probability	L2C?	FW?
ontfangst	grond	0.26	no	no	grond	0.55	no	no
(reception)	basis	0.13	no	no	basis	0.01	no	no
Getting	een	0.09	no	yes	Grund	0.01	no	no
	straffe	0.05	no	no	aanleiding	0.01	no	no
	elk	0.02	no	yes	grond	0.01	no	no
versoucken	weten	0.15	no	no	weten	0.77	no	no
(requesting)	zien	0.05	no	no	zien	0.04	no	no
Request	staan	0.04	no	no	wissen	0.02	no	no
	toe	0.04	no	yes	merken	0.02	no	no
	zeggen	0.03	no	no	horen	0.02	no	no
ordre	is	0.05	no	no	de	0.21	no	yes
(order/instruction)	om	0.03	no	yes	,	0.07	no	yes
SocialInteraction	,	0.02	no	yes	een	0.04	no	yes
	heeft	0.02	no	no	is	0.03	no	no
	bekent	0.01	no	no	in	0.02	no	yes
geeven	te	0.10	no	yes	te	0.04	no	yes
(giving)	,	0.05	no	yes	reeds	0.02	no	no
Giving	niet	0.04	no	yes	al	0.02	no	no
	kan	0.02	no	no	betalen	0.02	no	no
	sal	0.02	no	no	,	0.02	no	yes
verpligt	meer	0.07	no	no	meer	0.12	no	no
(obliged/ to oblige)	anders	0.02	no	no	langer	0.07	no	no
ForceToAct	is	0.02	no	no	,	0.05	no	yes
	ook	0.01	no	yes	zoo	0.02	no	yes
	zijnde	0.01	no	no	zo	0.02	no	yes
belastinge	,	0.39	no	yes	,	0.30	no	yes
(tax)	en	0.09	no	yes	d	0.04	no	yes
FinancialTransaction	de	0.08	no	yes	s	0.04	no	yes
	ende	0.06	no	yes	en	0.04	no	yes
	daer	0.02	no	no	der	0.03	no	no
aengereekend	is	0.12	no	no	,	0.08	no	yes
(charged)	wordt	0.03	no	no	aan	0.03	no	yes
FinancialTransaction	eert	0.02	no	no	zal	0.03	no	no
	e	0.02	no	yes	naar	0.02	no	no
	int	0.01	no	yes	dient	0.01	no	no

Table 11: Top 5 predicted tokens per model with probability scores. L2C = whether the predicted token is linkable to the corresponding event class. FW = whether the predicted token is a function word.

	GysBERT				GysBERT-v2			
Masked token	Prediction	Probability	L2C?	FW?	Prediction	Probability	L2C?	FW?
ontfangst	ordre	0.17	no	no	ontfang	0.14	yes	no
(reception)	copie	0.10	no	no	antwoorde	0.10	no	no
Getting	grond	0.04	no	no	antwoord	0.08	no	no
	last	0.04	no	no	grond	0.06	no	no
	ende	0.03	no	yes	dato	0.04	no	no
versoucken	weten	0.27	no	no	versoeken	0.70	yes	no
(requesting)	weeten	0.24	no	no	dienen	0.06	no	no
Request	volgen	0.15	no	no	weten	0.04	no	no
	blijken	0.04	no	no	weeten	0.03	no	no
	verstaan	0.03	no	no	versoecken	0.03	yes	no
ordre	,	0.07	no	yes	ordre	0.82	yes	no
(order/instruction)	soo	0.05	no	yes	ordres	0.07	yes	no
SocialInteraction	dat	0.03	no	yes	vast	0.06	no	no
	vast	0.03	no	no	order	0.01	yes	no
	daer	0.02	no	no	uijt	0.00	no	yes
geeven	te	0.13	no	yes	soude	0.04	no	no
(giving)	doen	0.02	no	no	toe	0.03	no	yes
Giving	sal	0.02	no	no	kan	0.02	no	no
	geeven	0.02	yes	no	moet	0.02	no	no
	,	0.01	no	yes	sal	0.02	no	no
verpligt	anders	0.17	no	no	anders	0.41	no	no
(obliged/ to oblige)	meer	0.14	no	no	meer	0.09	no	no
ForceToAct	deselve	0.05	no	yes	langer	0.09	no	no
	die	0.03	no	yes	verder	0.05	no	no
	om	0.03	no	yes	deselve	0.05	no	no
belastinge	,	0.09	no	yes	,	0.05	no	yes
(tax)	cours	0.05	no	no	verwagting	0.05	no	no
FinancialTransaction	die	0.04	no	yes	verantwoording	0.04	no	no
	ende	0.02	no	yes	factuur	0.04	yes	no
	##waerts	0.02	no	yes	gebruijk	0.03	no	no
aengereekend	is	0.07	no	no	belast	0.13	yes	no
(charged)	,	0.02	no	yes	gebragt	0.08	no	no
FinancialTransaction	gebracht	0.02	no	no	overgebragt	0.08	no	no
	overgegeven	0.02	no	no	verantwoord	0.07	yes	no
	gehouden	0.02	no	no	toegesonden	0.04	no	no

Table 12: Top 5 predicted tokens per model with probability scores. L2C = whether the predicted token is linkable to the corresponding event class. FW = whether the predicted token is a function word.

Language Identification of Philippine Creole Spanish: Discriminating Chavacano From Related Languages

Aileen Joan Vicente
De La Salle University
Philippines
aileen_vicente@dlsu.edu.ph

Charibeth Cheng
De La Salle University
Philippines
charibeth.cheng@dlsu.edu.ph

Abstract

Chavacano is a Spanish Creole widely spoken in the southern regions of the Philippines. It is one of the many Philippine languages yet to be studied computationally. This paper presents the development of a language identification model of Chavacano to distinguish it from languages that influenced its creolization using character-level Convolutional Neural Networks (CNN). Unlike studies that discriminated similar languages based on geographical proximity, this paper reports a similarity based on a language's creolization. We established the similarity of Chavacano and its related languages, Spanish, Portuguese, Cebuano, and Hiligaynon, from historical accounts and lexical similarity based on the number of common words in the corpus for all languages. We report an accuracy of 93% for the model generated from a CNN using ten filters with a filter width of 5. The training experiments reveal that increasing the filter width, number of filters, or training epochs is unnecessary even if the accuracy increases because the generated models present irregular learning behavior or may have already been overfitted. This study also demonstrates that the character features extracted from CNN, similar to n-grams, are sufficient in identifying Chavacano. Future work on the language identification of Chavacano includes improving classification accuracy, especially for short or code-switched texts for practical applications such as social media sensors for disaster response and management.

1 Introduction

Language Identification (LI) is the task of deciding which natural language a particular text is written in. The research in this field aims to mimic the ability of humans to recognize these languages. LI enables many natural language applications and language processing (NLP) tasks. For example, automatic machine translation applications must identify the text's language before translating it into English. It can be used for document collections where the languages of the documents are unknown beforehand (Jauhiainen et al., 2019), such as in the case of crawling the web as part of corpus-building.

Many LI systems and studies target English and other major languages spoken worldwide. It is especially understandable since large repositories of language texts exist for these languages. There are also initiatives to identify low-resource languages such as Uralic languages (Jauhiainen et al., 2020) and Austronesian languages (Dunn and Nijhof, 2022). However, many other low-resource languages do not have enough digital resources for extensive research. While LI is generally considered a solved task, the work on LI for low-resource languages persists due to the widespread use of the Internet and the development of applications based on natural language understanding, such as chatbots. Selamat and Akosu (2016) argued that the inability to identify a language makes the language invisible in any multilingual environment, such as in the case of Chavacano, the Philippines' Creole Spanish.

Chavacano is one of those under-researched, low-resource languages. Websites with automatic translations identify Chavacano as Spanish, given the former's similarity with the latter. Chavacano's lexicon is predominantly Spanish (Lipski and Santoro, 2007) but with orthographic shifts.

Languages can differ in many ways. They may use different sounds, other writing systems, different vocabulary, or put words together to form a sentence differently. For similar languages, however, such as language variants and dialects, discriminating between them remains challenging (Zampieri et al., 2014) and is one of the bottlenecks of state-of-the-art language identification systems.

Proceedings of the Eleventh Workshop on NLP for Similar Languages, Varieties, and Dialects (VarDial 2024), pages 188–196
June 20, 2024

This paper reports the language identification of Philippine Creole Spanish. Unlike the similar languages investigated in the Discriminating between Similar Languages (DSL) shared tasks, whose language similarities are mostly due to geographic proximity, this study investigates the identification of a Creole, i.e., Chavacano, among its related languages.

This study brings forward the unique characteristics of Philippine Creole Spanish (PCS) as an amalgamation of foreign and native languages. In the case of Chavacano, it is the complex intermixing of Spanish, Portuguese, Cebuano, and Hiligaynon during centuries of colonization, migration, and trade. The linguistic features of Chavacano that combine elements of multiple language sources make it a linguistically rich and unique variety.

The language identification of Chavacano is expressed as a character-level sentence classification that discriminates among similar, related languages and where the languages are considered as the target classes.

The remainder of this paper is organized as follows. Section 2 presents the linguistic properties of Chavacano and its similarities with related languages. Section 3 introduces related works implementing CNN for language identification. Section 4 gives a detailed overview of the steps to build the language identification model. In particular, Section 4.2 provides an overview of char-CNN, the character-level Convolutional Neural Network used to train the model. In Section 5, we report and analyze our experimental results, while Section 6 concludes this paper and gives some directions for future research.

2 Chavacano: Philippine Creole Spanish

The Philippine Creole Spanish, collectively known as Chavacano, comprises three major dialects spoken in Ternate, Cavite, and Zamboanga (Lipski, 2001). Both the Ternate and Cavite dialects are classified as the Manila Bay PCS. Ternateño was the oldest Spanish-based Creole in the Philippines, and Caviteño was an off-shoot. Zamboangueño, on the other hand, comprises the largest group of Chavacano speakers in Zamboan a City and neighboring towns and cities in Mindanao. In this study, we refer to the variant Zamboangueño, as it is the only thriving variant. Aside from the population of speakers, Zamboangueño is actively used in blogs, news, and social media that can be used as digital resources.

The formation of Chavacano in Zamboanga resulted from historical and cultural interactions in the Philippines during the Spanish colonial period from 1565 to 1898. Chavacano belongs to the Creole family of languages of Spanish descent (Eberhard et al., 2023).

The language started to develop during the Spanish garrison in Zamboanga, beginning with the absorption of grammatical and lexical structures from Manila Bay PCS in the 18th century. Manila Bay PCS is said to have been influenced by the Portuguese language (University of Hawai'i Press, 1975; Lipski, 2001). Ilonggo or Hiligaynon later influenced Chavacano as Iloilo became a stopover for ships from Manila to Zamboanga. Later in the 20th century, immigration from the Central Visayan region to southwest Mindanao added some Visayan or Cebuano items to the language. Given this history, Chavacano is described as a "contact vernacular that has undergone numerous remakings by an ever-changing population that has never given up their native languages" (Lipski, 1992). It is easy to see that Chavacano's words are predominantly Spanish, but an inspection of usage tells us that they are not entirely Spanish.

Over three centuries of Philippine history influenced the morphology, grammar, and syntax of Chavacano (Lipski and Santoro, 2007). It has retained its Austronesian foundation, evidenced by the Verb-Subject-Object word order, with many alternative possibilities (Lipski, 1992). The Philippine languages belong to the Austronesian language family. This contrasts Spanish's Subject-Verb-Object word order (Lee, 2017).

The lexicon of Chavacano is largely Spanish (Lipski and Santoro, 2007) but with orthographic shifts. It has experienced several stages of relexification to include lexical items of Philippine origin from regional Visayan (Cebuano), Ilonggo (Hiligaynon), and occasionally Tagalog (Lipski, 2001). It has also adopted a heavy English lexical transfer (Lipski, 1992) over time.

Chavacano words are spelled using the alphabet of the word's traced etymology (DepEd-IX, 2016). For example, the Spanish-derived words *zacate* (grass) and *mañana* (tomorrow) are spelled using the Spanish alphabet, the *Abecedario*. In contrast, the Chavacano words of local origin, like *kanila* (them) and *kanamon* (us), are spelled using the Philippine alphabet system. The letter *r* in the Spanish verbs like *comer* (to eat), *bailar* (to dance) are dropped in Chavacano, i.e., *come*, *baila*. In general, Chavacano words are spelled the way they are pronounced. It is also interesting to note that the Spanish writing utilizes diacritics that are not necessarily applied in Chavacano.

In summary, Chavacano began as a hybrid pan-Philippine contact language whose Spanish items had already been filtered through Philippine languages and which was, therefore, a Philippine language in the structural sense at every point of its existence (Lipski, 2001).

3 Related Works

Jauhiainen et al. (2019) assert that from a computational perspective, the algorithms and features used to discriminate between languages, language varieties, and dialects are identical. Hence, the choice of features and algorithms depends on the researcher and the data used for the study.

Both discriminative and generative algorithms have been explored in more recent LI studies. Hidden Markov Models and Latent Dirichlet Allocation are the common generative methods used. Decision trees, support vector machines, neural networks, and ensembles are widely used discriminative models.

Characters are the building blocks of a language's writing system. Although most languages follow an alphabetic system, the languages still differ in character combinations and orthography. Hence, characters and their combinations have been widely used in LI.

An example of character combinations is n-grams. Character n-grams are widely used character sequences that may capture a language's orthography (Simões et al., 2014). Character n-grams are sequences (consecutive or overlapping) of characters of length *n*. The frequency of these n-grams has been used as feature vectors for most LI research involving discriminative methods.

Using CNN for LI is seen as a means of automatically extracting character features from text for classification. Zhang et al. (2015) was among the first to introduce character-level CNN for text classification. In this case, text is seen as a kind of raw signal at the character level where CNN extracts features (Zhang et al., 2015; Kim et al., 2016). The successful application of Zhang et al. (2015) and Kim et al. (2016) also sparked interest in CNN for LI. Guggilla (2016), Belinkov and Glass (2016), Jaech et al. (2016b), Jaech et al. (2016a), Ali (2018a), Ali (2018b), Chung et al. (2019) are among those who have successfully implemented CNN for LI. It has grown in acceptance in LI because it eliminates the need to extract or handcraft features separately, such as feature engineering.

4 Methodology

4.1 Data Preparation

The corpus used in the study is mixed-domain. The monolingual Hiligaynon and Cebuano sentences were taken from the PH-MNMT corpus (Coronia, 2022), which consists of web-scraped articles and bible translations. The Spanish and Portuguese sentences were mainly taken from the DSL Corpus Collection (Tan et al., 2014), which consists of news articles. Additional sentences for Spanish and Portuguese were taken from Bible translations as well.

On the other hand, the Chavacano sentences were collected from print sources (de Saint Exupéry (Author) and De Los Reyes (Translator), 2018) and online sources (Herrera; Zamboanga News Online; Wycliffe Bible Translators, Inc.).

The corpus contains 107,500 sentences with 21,500 sentences for each language (Table 1).

The raw sentences used in the corpus are made available at `https://github.com/ajvicente/cbk-li`.

The Spanish and Portuguese sentences were tokenized using tokenizers specific to the language. Cebuano and Hiligaynon, on the other hand, were tokenized using English-based tokenizers. Punctuation and numerical literals were later removed

Language	Source Domains	No. of Sentences	Traning Data	Testing Data	Validation Data
Chavacano	bible translations, blogs, book, feature articles	21,500	18,000	2,000	1,500
Cebuano	bible translations, web-scraped documents	21,500	18,000	2,000	1,500
Hiligaynon	bible translations, web-scraped documents	21,500	18,000	2,000	1,500
Spanish	bible translations, news articles	21,500	18,000	2,000	1,500
Portuguese	bible translations, news articles	21,500	18,000	2,000	1,500
		107,500	90,000	10,000	7,500

Table 1: Chavacano and Related Languages Corpus

from the data set. The texts were converted to lowercase after all unnecessary characters had been removed. The alphabet of the corpus contained 46 characters. Digraphs such as *ch*, *ng*, *rr*, and *lh* are counted as single characters. Characters with diacritics are also counted separately.

Language	Unique Words	Overlap Words with Chavacano	Unique Characters	Sentence Length (Max Characters)
Chavacano	5,740		36	424
Cebuano	26,486	981	27	639
Hiligaynon	20,832	941	27	574
Spanish	50,836	2,580	42	3,654
Portuguese	38,790	1,357	45	4,846

Table 2: Corpus Statistics

There are 5,740 unique Chavacano words in the corpus. Of these, 44.95% overlap with Spanish, 23.64% with Portuguese, 17.09% with Cebuano, and 16.39% with Hiligaynon (Table 2). Most of the shared words or overlaps are content words.

The small number of unique words in the Chavacano corpus is due to shorter sentence fragments in Chavacano and because most of the sentence fragments in the dataset were sourced from bible translations. Unlike the Cebuano, Hiligaynon, Spanish, and Portuguese datasets were primarily sourced from news articles and web texts covering more topics than bible translations. Hence, there is a greater variety of words in the related languages.

4.1.1 Character Encoding

The characters for each word in the corpus are sequentially encoded as in the work of Zhang et al. (2015). Encoding is based on an alphabet dictionary of size $m = 47$ that consists of the 46 common alphabet characters in the corpus and the space as the word delimiter. Each character is then quantized using 1-of-m encoding (or one-hot encoding). A fixed sentence length of $l = 1000$ characters is set. This value was empirically identified to cover all the words in the Chavacano sentence fragments. Shorter sentences are padded, while longer sentences (especially for Spanish and Portuguese) are truncated.

The labels are similarly one-hot encoded over five language classes: Chavacano, Cebuano, Hiligaynon, Spanish, and Portuguese.

4.1.2 Data Split

Training, Validation, and Test sets were extracted from the corpus using stratified sampling to ensure that all language classes are represented proportionally in each data set. 18,000 sentences per language are used for training, 2,000 for validation, and 1,500 for testing.

4.2 charCNN: Character-based Convolutional Neural Network

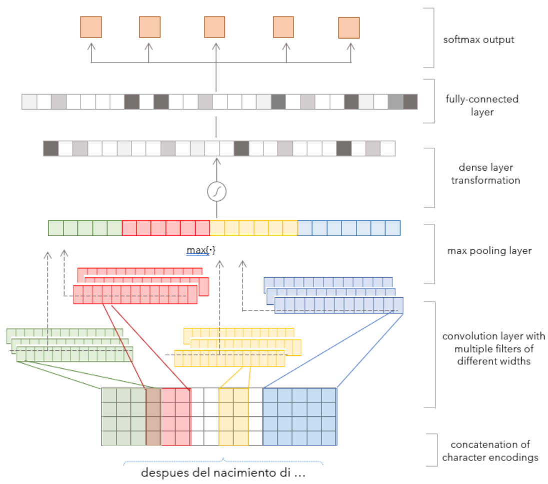


Figure 1: charCNN Network Architecture adapted from Kim et al. (2016)

Following the work of Kim et al. (2016) and Zhang et al. (2015), a simple convolutional neural network was used to extract features from the training data and then fed to a dense layer for classification. Figure 1 illustrates the neural network architecture.

4.2.1 Convolution Layer

Based on Kim et al. (2016), a 2D convolution is applied between the input sentence $\mathbf{C}^s$ and a filter $\mathbf{H}$ $\in \mathbf{R}^{m \times w}$ where the filter width $w \in \{2, 3, 4, 5, 6\}$. With each filter, a feature vector $f^s \in \mathbf{R}^{(l-w)+1}$ is generated where the $i - th$ element of f is given by:

$$f^s(i) = \langle \mathbf{C}^s[*, i : i + w - 1]\mathbf{H} \rangle \tag{1}$$

where $\langle \mathbf{A}, \mathbf{B} \rangle = Tr(\mathbf{AB}^T)$ is the Frobenius inner product.

Characters, as used in the study, correspond to signals in images, videos, and sounds (Zhang et al., 2015) that are typical inputs in CNN-based tasks.

4.2.2 Pooling Layer

The maximum value in f^s is extracted at the pooling layer as the feature corresponding to the filter $\mathbf{H}$ when applied to the sentence $\mathbf{C}^s$. According to Kim et al. (2016), in this process, the filter essentially picks out a character n-gram whose size of the n-gram corresponds to the filter width.

Given that multiple filters h are used in the study, then the representation of the input sentence is a concatenation of max pooling layers in the form $y^s = [y_1^s, ..., y_h^s]$.

A bias is added, and a non-linear transformation (tanh) is applied.

4.2.3 Dense Layer

A dense layer of 512 units followed by a dropout at 0.5 is added to the convolutional network before concluding with a softmax layer of 5 units to represent each of the five language classes. The categorical cross-entropy loss is used to fit the model. The model is optimized with Adam optimizer using a learning rate of 0.001.

4.3 Model Evaluation

Loss and accuracy metrics are collected during training (validation) and testing to evaluate the model's performance. The validation step during training uses the validation dataset to assess the model's performance during training. Model testing is performed after training using unseen data to simulate real-world scenarios. Ideally, the accuracy and loss values during validation and testing should be close enough to ascertain that the model does not overfit or underfit the data. The use of overall accuracy in this study is sufficient, given that the data is balanced for all language classes.

Several experiments that involved changes in the number of filters and combining filter widths are also conducted to arrive at optimized network parameters.

5 Results

Various training configurations using the number of filters (5, 10, and 15), range and combination of filter widths (2, 3, 4, 5, and 6), and number of epochs (10, 20, and 30) were experimented on in this study. The following sections report the result of such experiments and insights from the language identification modeling of Chavacano.

5.1 Experiments

The results of the experiments on various training configurations based on the number of filters, filter widths, and epochs show that the accuracy of the model naturally increases with increasing number of filters, filter widths, and epochs, as shown in Figure 2 and Figure 3.

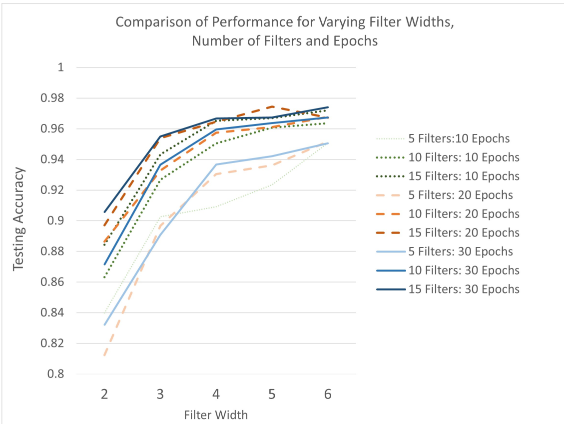


Figure 2: Comparison of Model Performances for Varying Filter Widths

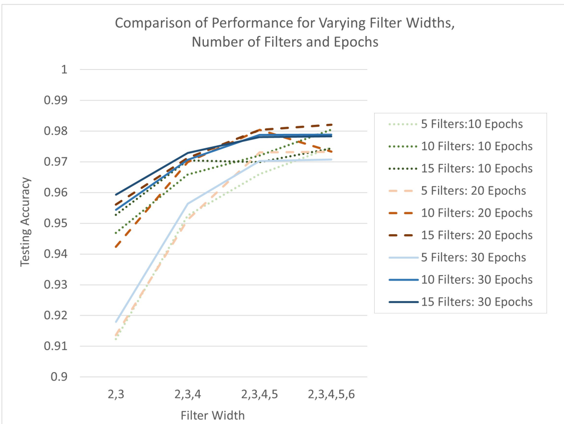


Figure 3: Comparison of Model Performances for Combined Filter Widths

The comparison in Figure 2 also shows that there is generally a sharp increase in performance using filter width 2 to 4 with the increasing number of filters and epochs, after which a slight and steady increase in the performance is observed except for the degradation of performance of the model learned at 15 filters and 20 epochs.

On the other hand, the combined filter widths in Figure 3 show similar behavior in the increase in accuracy until the combined filter widths of 2, 3, 4, and 5.

Figures 2 and 3 show that increasing the number of filters and the number of times these are seen during training does not necessarily contribute to a better model.

In the same way, a comparison of training and validation losses also reveals that although increasing the number of filters and the number of epochs increases validation accuracy, the model's training performance seemed irregular, as shown in an example in Figure 4.

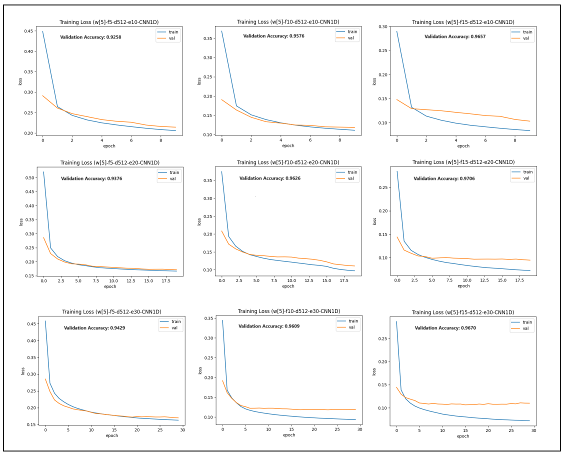


Figure 4: Comparison of Training/Validation Losses for Filter Width = 5.

The comparison shows that the divergence in the training and validation losses increases as the number of filters and epochs increases. This behavior indicates that the models may have already picked up noise in the data and overfit.

Finally, based on the model accuracy and variance of training and validation losses, the model generated using 10 filters with a filter width of 5 and trained in 20 epochs, earning a validation accuracy of 0.9376, is chosen as the best model among all training configurations.

5.2 Error Analysis

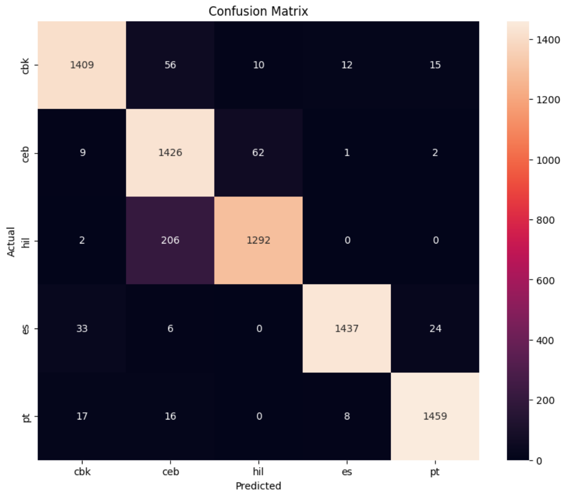


Figure 5: Confusion Matrix based on Model Testing

The confusion matrix in Figure 5 reveals that Chavacano can be confused with Cebuano, Hiligaynon, Spanish, and Portuguese. The related languages are also often mistaken for Chavacano.

It is observed that Hiligaynon and Cebuano, both local languages, are mostly confused with each other and that Hiligaynon is only confused with Cebuano and rarely Chavacano.

On the other hand, Chavacano is mostly confused with Cebuano, followed by Portuguese, Spanish, and Hiligaynon. Interestingly, Chavacano exhibits a greater overlap with Spanish and Portuguese when compared to Cebuano and Hiligaynon. Yet, Chavacano is mostly confused with the local language, Cebuano. This behavior may be attributed to Chavacano's orthography. Despite following the Spanish's Abecedario (DepEd-IX, 2016), Chavacano does not use many of the diacritics used by Spanish and Portuguese in writing.

The model confuses Spanish and Portuguese with Chavacano more than the local languages. In the case of Spanish, 20 of 33 (61%) misclassifications do not contain diacritics, and the rest of the 13 sentences only contained at most three characters with diacritics. For Portuguese, all 17 sentences that are misclassified did not contain diacritics.

The error analysis also revealed that 63% (35 of 56) of the Chavacano sentences misclassified as Cebuano were single-word sentence fragments. The longest misclassified sentence consists of 11 words. This result indicates that the model may be unable to correctly classify short sentences, significantly since most words overlap with other languages. Language identification involving short texts continues to be a challenging task for many languages (Jaech et al., 2016b; Jauhiainen et al., 2019).

The misclassification of Chavacano to Hiligaynon, Spanish, and Portuguese also share the same observation, albeit not as short as the Cebuano misclassifications. All misclassified sentences fall within less than 30% of the maximum number of words in the language's corpus.

6 Conclusions and Recommendations for Future Work

6.1 Conclusion

The experiments show that the language identification of Chavacano does not require a complex and deep CNN network. The model can already learn to discriminate the language from among its related languages using 10 filters with a filter width of 5. The hyperparameter search reveals that because the related languages share common characters to a large extent, it is vulnerable to overfitting. With the performance at 93%, the model can be used in the future to develop web applications to collect Chavacano documents.

This study demonstrates the viability of character features, specifically those generated by a convolutional neural network, to identify related languages. Instead of manually extracting n-gram features, this study demonstrates an end-to-end system of training a language identification model using neural networks.

The study also gleaned the orthographical similarities between Chavacano and Cebuano despite the latter being predominantly Spanish in cognates, although further studies need to be undertaken to establish this relatedness. Diacritics was also considered a contributing factor in discriminating Chavacano from Spanish and Portuguese.

6.2 Recommendations for Future Work

This paper presents a benchmark study for Chavacano LI that can be used as a baseline for future works. Further experimentation is recommended, including using other learning algorithms, such as SVM, or deep learning models, such as Transformers. In addition, the study uses mixed domains in training. The effect of the dataset domain in training needs to be experimented as this has been one of the issues in discriminating similar languages.

This preliminary work on Chavacano opens many other opportunities to understand and document Chavacano computationally and study Creole languages. The next step of this project is to implement the network design to discriminate Chavacano in natural settings, i.e., no preprocessing and within the context of multilingual documents. Based on the results, the language identification study can be extended to improve the classification of Chavacano in shorter, maybe code-switched, sentences such as those coming from Tweets to be used for practical applications such as social media sensors for disaster monitoring and management or more natural translation from code-switched sentences.

Limitations

While most language identification of related languages worked on dialects or variants, this study is limited to the related languages of Creole. The similarity is based on the languages' lexical, syntactical, and morphological influence on Chavacano. Another limitation is using CNN as the only model experimented with in the study. Experiments with other models to improve LI for Chavacano are encouraged as future works.

References

Mohamed Ali. 2018a. Character level convolutional neural network for Arabic dialect identification. In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018)*, pages 122–127, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Mohamed Ali. 2018b. Character level convolutional neural network for German dialect identification. In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018)*, pages 172–177, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Yonatan Belinkov and James Glass. 2016. A character-level convolutional neural network for distinguishing similar languages and dialects. In *Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial3)*, pages 145–152, Osaka, Japan. The COLING 2016 Organizing Committee.

Tonglee Chung, Bin Xu, Yongbin Liu, Chunping Ouyang, Siliang Li, and Lingyun Luo. 2019. Empirical study on character level neural network classifier for chinese text. *Engineering Applications of Artificial Intelligence*, 80:1–7.

Jeremy Dale Coronia. 2022. ph-mnmt-dataset. `https://huggingface.co/datasets/ecridale/ph-mnmt-dataset`. Accessed February 2023.

Antoine de Saint Exupéry (Author) and Robin De Los Reyes (Translator). 2018. *El Principe Niño: Der kleine Prinz - Zamboangueño Chabacano*, Creole edition. Edition Tintenfaß, Neckarsteinach, Germany.

DepEd-IX. 2016. *Zamboanga Chavacano Orthography*. Local Government of Zamboanga City: Philippines.

Jonathan Dunn and Wikke Nijhof. 2022. Language identification for Austronesian Languages. In *Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022)*, pages 6530–6539, Marseille, France. European Language Resources Association (ELRA).

David Eberhard, Gary Simons, and Charles Fenning, editors. 2023. *Ethnologue: Languages of the World*, 26th edition. SIL International, Dallas, Texas.

Chinnappa Guggilla. 2016. Discrimination between similar languages, varieties and dialects using CNN- and LSTM-based deep neural networks. In *Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial3)*, pages 185–194, Osaka, Japan. The COLING 2016 Organizing Committee.

Jerome Herrera. Bien chabacano. `https://bienchabacano.blogspot.com/`. Accessed February 2023.

Aaron Jaech, George Mulcaire, Shobhit Hathi, Mari Ostendorf, and Noah A. Smith. 2016a. Hierarchical Character-Word Models for Language Identification. In *Proceedings of The Fourth International Workshop on Natural Language Processing for Social Media*, pages 84–93, Austin, TX, USA. Association for Computational Linguistics.

Aaron Jaech, George Mulcaire, Mari Ostendorf, and Noah A. Smith. 2016b. A Neural Model for Language Identification in Code-Switched Tweets. In *Proceedings of the Second Workshop on Computational Approaches to Code Switching*, pages 60–64, Austin, Texas. Association for Computational Linguistics.

Tommi Jauhiainen, Heidi Jauhiainen, Niko Partanen, and Krister Lindén. 2020. Uralic language identification (ULI) 2020 shared task dataset and the wanca 2017 corpora. In *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 173–185, Barcelona, Spain (Online). International Committee on Computational Linguistics (ICCL).

Tommi Jauhiainen, Marco Lui, Marcos Zampieri, Timothy Baldwin, and Krister Lindén. 2019. Automatic language identification in texts: A survey. *Journal of Artificial Intelligence Research*, 65(1):675–682.

Yoon Kim, Yacine Jernite, David Sontag, and Alexander M. Rush. 2016. Character-aware neural language models. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, page 2741–2749. AAAI Press.

James F. Lee. 2017. Word order and linguistic factors in the second language processing of spanish passive sentences. *Hispania*, 100(4):580–595.

John Lipski. 2001. The place of Chabacano in the Philippine linguistic profile. *Sociolinguistic Studies*, 2(2):119–163.

John Lipski and Maurizio Santoro. 2007. Zamboangueño creole spanish. In John Holm and Peter Patrick, editors, *Comparative creole syntax. Parallel outlines of 18 creole grammars*, Westminster Creolistics Series 7, pages 373–398. Battlebridge, London. Much information is based on Forman (1972).

John M. Lipski. 1992. New thoughts on the origins of zamboangueño (philippine creole spanish). *Language Sciences*, 14(3):197–231.

Ali Selamat and Nicholas Akosu. 2016. Word-length algorithm for language identification of under-resourced languages. *Journal of King Saud University - Computer and Information Sciences*, 28(4):457–469.

Alberto Simões, José João Almeida, and Simon D. Byers. 2014. Language Identification: a Neural Network Approach. In *3rd Symposium on Languages, Applications and Technologies*, volume 38 of *OpenAccess Series in Informatics (OASIcs)*, pages 251–265, Dagstuhl, Germany. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

Liling Tan, Marcos Zampieri, Nikola Ljubešic, and Jörg Tiedemann. 2014. Merging comparable data sources for the discrimination of similar languages: The dsl corpus collection. In *Proceedings of the 7th Workshop on Building and Using Comparable Corpora (BUCC)*, pages 11–15, Reykjavik, Iceland.

University of Hawai'i Press. 1975. Chabacano (Philippine Creole Spanish). *Oceanic Linguistics Special Publications*, (14):210–216.

Wycliffe Bible Translators, Inc. El nuevo testamento. https://worldbibles.org/language_detail/eng/cbk/Chavacano. Accessed February 2023.

Zamboanga News Online. Comentarios desde zamboanga. http://comentariosdesdezamboanga.blogspot.com/. Accessed February 2023.

Marcos Zampieri, Liling Tan, Nikola Ljubešić, and Jörg Tiedemann. 2014. A report on the DSL shared task 2014. In *Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects*, pages 58–67, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems*, volume 28, pages 649–657. Curran Associates, Inc.

Data-Augmentation-Based Dialectal Adaptation for LLM

Fahim Faisal, Antonios Anastasopoulos
Department of Computer Science, George Mason University
{ffaisal,antonis}@gmu.edu

Abstract

This report presents GMUNLP's participation to the Dialect-Copa shared task at VarDial 2024 (Chifu et al., 2024), which focuses on evaluating the commonsense reasoning capabilities of large language models (LLMs) on South Slavic micro-dialects. The task aims to assess how well LLMs can handle non-standard dialectal varieties, as their performance on standard languages is already well-established. We propose an approach that combines the strengths of different types of language models and leverages data augmentation techniques to improve task performance on three South Slavic dialects: Chakavian, Cerkno, and Torlak. We conduct experiments using a language-family-focused encoder-based model (BERTić) and a domain-agnostic multilingual model (AYA-101). Our results demonstrate that the proposed data augmentation techniques lead to substantial performance gains across all three test datasets in the open-source model category. This work highlights the practical utility of data augmentation and the potential of LLMs in handling non-standard dialectal varieties, contributing to the broader goal of advancing natural language understanding in low-resource and dialectal settings.[1]

1 Introduction

Recent advancements in large language models (LLMs) have led to remarkable performance on a wide range of natural language understanding tasks, particularly in standard languages. However, the effectiveness of these models on non-standard dialectal varieties remains an open question (Faisal et al., 2024). The Dialect-Copa shared task, introduced by Ljubešić et al. (2024), aims to bridge this gap by evaluating the commonsense reasoning capabilities of LLMs on South Slavic dialects.

Commonsense reasoning, as originally proposed by Gordon et al. (2011), requires models to make plausible inferences based on everyday knowledge and understanding of the world. Extending this task to dialects poses unique challenges, as models must capture the nuances and variations specific to these language varieties. The Dialect-Copa shared task provides a platform to explore the adaptability and generalization capabilities of LLMs in this context.

In this GMUNLP submission, we explore the potential of data augmentation techniques in enhancing the performance of language models on dialectal commonsense reasoning tasks. Our approach harnesses the power of state-of-the-art LLMs to generate synthetic training data, which we combine with the provided training dataset. By employing a diverse set of language models, we aim to quantify the performance gains achievable through data augmentation. Specifically, we utilize three categories of language models to maximize dialectal task performance: (1) smaller language models that are well-suited for low-resource settings and can be easily customized, (2) mid-size language models that strike a balance between task-specific performance and language understanding capabilities, and (3) closed-source language models that generate high-quality synthetic task data to further enhance the performance of the other two categories of language models.

We achieved the highest scores across all three test datasets in the open-source model category. In addition, our solution performed on par with the GPT-4 zero-shot iterative prompting approach employed by one of the teams, demonstrating the competitiveness of the proposed approach against state-of-the-art closed-source models. Furthermore, we achieved substantial performance improvements for the small-scale, language-family-focused model BERTić by combining it with our data augmentation strategy, showcasing the effectiveness of our approach in boosting the performance of language models tailored for low-resource settings.

The remainder of this paper is organized as

[1]Code and data are publicly available: https://github.com/ffaisal93/dialect_copa

Proceedings of the Eleventh Workshop on NLP for Similar Languages, Varieties, and Dialects (VarDial 2024), pages 197–208
June 20, 2024

follows: Section 2 provides an overview of the Dialect-Copa shared task and dataset, Section 3 describes our methodology and experimental setup, Section 4 presents our results and analysis, and Section 5 concludes the paper and discusses future directions.

2 The Dialect-Copa shared task

Task Information In the Dialect-Copa shared task, a premise sentence is provided along with a question that can be either a cause or an effect. The objective is to build a classifier that selects the most plausible response from two candidate answer choices based on the given premise and question. To illustrate, consider the following training example in English, where the task is to identify the most plausible cause:


```
{"premise": "My body cast a shadow over the grass.",
"choice1": "The sun was rising.",
"choice2": "The grass was cut.",
"question": "cause", "label": 0, "idx": 0}
```


The Dialect-Copa dataset consists of such cause-effect examples across 8 languages and dialects, challenging models to perform commonsense reasoning in non-standard language varieties.

code	language	train	val.	test
en	English	400	100	
sl	Slovenian	400	100	
sl-cer	Cerkno	400	100	500
hr	Croatian	400	100	
hr-ckm	Chakavian	-	-	500
sr	Serbian	400	100	
sr-trans	Serbian (transliterated)	400	100	
sr-tor	Torlak	400	100	
sr-tor-trans	Torlak (transliterated)	400	100	500
mk	Macedonian	400	100	
mk-trans	Macedonian (transliterated)	400	100	

Table 1: Dialect-Copa dataset statistics for different languages and their dialectal varieties.

Languages The Dialect-Copa dataset encompasses training and validation data in 7 languages, including English, 6 moderately resourced South Slavic languages, and two related micro-dialects. The test dataset features these two micro-dialects along with an additional previously unseen dialect. The three dialects in the test set are as follows:

1. The Cerkno dialect of Slovenian, spoken in the Slovenian Littoral region, specifically in the town of Idrija.

Base model	Fine-tuning (FT)/Prompting	Epoch	Acc. (%)	
			en	hr
Aya-101	4 shot (2 cause, 2 effect)	-	**80**	**75**
MaLA-500	4 shot (2 cause, 2 effect)	-	50	–
Llama2-CHAT (7B)	4 shot (2 cause, 2 effect)	-	75	50
BERT	FT (eval. lang)	3	**66**	55
mBERT	FT (eval. lang)	3	55	57
XLM-R	FT (eval. lang)	3	54	54
BERTić	FT (eval. lang)	3	48	**64**

Table 2: Preliminary evaluation results on the English and Croatian validation set for different base models.

2. The Chakavian dialect of Croatian from the northern Adriatic, particularly from the town of Žminj.

3. The Torlak dialect, spoken in southeastern Serbia, northeastern North Macedonia, and northwestern Bulgaria, with the specific test instances coming from the town of Lebane.

Cerkno and Torlak dialects are present in all three dataset splits (training, validation, and test) whereas, the Chakavian dialect is intentionally held out from the training and validation splits and is exclusively encountered during the test phase. Each dialect in the test dataset comprises 500 instances. Table 1 presents the detailed statistics of the Dialect-Copa dataset, providing an overview of the distribution of instances across languages and dialects.

3 Experimental Phases

In this section, we report different phases of our experiments. We step by step perform experiments to choose appropriate base models followed by data augmentation, combination and task-specific model tuning.

Phase 1: Model Selection In the preliminary phase of our experiments, we conduct a series of trials to identify base language models that demonstrate strong performance on language understanding tasks in a multilingual context. To achieve this, we fine-tune widely-used encoder-based models, such as BERT (Devlin et al., 2019), mBERT, and XLM-R (Conneau et al., 2020), on the English and Croatian subsets of the Dialect-Copa training dataset. Additionally, we explore the potential of more recently open-sourced large language models (LLMs) of varying sizes, such as LLaMA-2 (Touvron et al., 2023), Aya-101 (Üstün et al., 2024) and MaLA-500 (Lin et al., 2024), to gauge their effectiveness on the task.

Data Identifier	Description	Covered Language
[lang]-train	Original Dialect-Copa training data	en, hr, mk, sl, sl-cer, sr, sr-tor
[lang]-trans	Transliterated (Cyrillic → Latin) training data	mk, sr, sr-tor
[lang]-claude	Providing grammar rules and few-shot Croatian-Chakavian examples to generate synthetic parallel hr-ckm-train examples given the hr-train examples	hr-ckm
[lang]-gpt4	Additional synthetic English training data generated by GPT-4 (Whitehouse et al., 2023)	en
[lang]-reverse	Reverse-augmentation on [lang]-train, [lang]-trans, and [lang]-claude data	en, hr, mk, sl, sl-cer, sr, sr-tor
[lang]-nllb	Machine translation of en-gpt4 source data to other languages using the NLLB-6B model	hr, mk, sl, sr

Table 3: Training data augmentation approaches

Our key observations from this preliminary phase are as follows:

→ BERT, mBERT, and XLM-R exhibit comparable performance on the Croatian subset, achieving an accuracy of around 55%(+/-) after 3 epochs of in-language fine-tuning. However, the monolingual English BERT model surpasses the multilingual models on the English subset when fine-tuned for the same number of epochs.

→ BERTić (Ljubešić and Lauc, 2021), a transformer-based model pre-trained on Bosnian, Croatian, Montenegrin, and Serbian languages, aligns well with the target languages of the Dialect-Copa test set. Fine-tuning BERTić on the Croatian subset yields a notable performance improvement of approximately 12 percent (i.e. 7 percentage points) compared to the aforementioned multilingual models.

→ Employing 4-shot prompting with the LLaMA-2 7B parameter model results in better performance on the English subset. However, for the Croatian subset, LLaMA-2 generates random inferences. This finding aligns with expectations, as LLaMA-2 is primarily an English-centric model and not inherently multilingual. In an effort to address the multilingual limitations of LLaMA-2, Lin et al. (2024) proposed MaLA-500, a multilingual adaptation of the model that underwent fine-tuning using a causal language modeling objective. However, after this adaptation, MaLA-500 produces random-level inferences on the English subset.

→ Aya-101, a 13B parameter mt5-xxl-based model (Xue et al., 2021) instruction-tuned in 101 languages. It shows superior performance both in English and Croatian.

Based on these preliminary findings, we select the two best-performing models, Aya-101 and BERTić , for further experimentation in the subsequent phases of our study. We report our preliminary experimental findings in Table 2. The results of our preliminary experiments are summarized in Table 2.

Phase 2: Data Augmentation To address the limited size of the Dialect-Copa training dataset, which consists of only 400 instances per language, we employ various data augmentation techniques to expand the available training data. This step is crucial in mitigating the data scarcity bottleneck and improving the models' ability to generalize across diverse dialectal variations. By augmenting the training data, we aim to provide a more representative dataset for task-specific fine-tuning and instruction tuning of our selected language models. The data augmentation approaches we explore include:

→ The test dataset primarily contains instances written using the Latin script. Hence, we transliterate the Macedonian (mk) dataset from Cyrillic to Latin script to maintain consistency with the already available Serbian, and Torlak transliterated datasets.

→ For each instance in the training data, we swap the premise and the correct answer choice, effectively transforming cause examples into effect examples and vice versa, thereby doubling the number of training instances. For example consider the following premise and two 'effect' choices:

premise: I poured water on my sleeping friend.
choice1: My friend awoke. ✓
choice2: My friend snored. ×

Now our proposed reverse-augmentation

Setting	Description	Data Combination
o	All **o**riginal dialect-copa training data mixed together	[en, hr, mk, sl, sl-cer, sr, sr-tor]-train [sr, sr-tor]-trans
otrsl	Combining all **o**riginal, **t**ransliterated as well as **r**everse-augmented and **s**ynthetic training data (only **l**atin script ones)	[en, hr, sl, sl-cer]-train [sr, sr-tor, mk]-trans [en, hr, mk-trans, sl, sl-cer, sr-trans, sr-tor-trans]-reverse en-gpt4, hr-ckm-claude [hr, sl, mk-trans, sr-trans]-nllb
otrslc	Combining all original, transliterated as well as reverse-augmented and synthetic training data (Both latin and cyrillic script)	all available training data
$\texttt{otrsl}_{\texttt{mk-hr-ckm}}$	Selective otrsl setting with upsampled data count by repetition for mk, hr and hr-ckm	hr-train, mk-trans, hr-ckm-claude [hr-train, mk-trans, hr-ckm-claude]-reverse [hr, mk-trans]-nllb
$\texttt{otrsl}_{\texttt{hr-ckm}}$	Selective otrsl setting with upsampled data count by repetition for hr and hr-ckm	hr-train, hr-ckm-claude [hr-train, hr-ckm-claude]-reverse hr-nllb
$\texttt{otrsl}_{\texttt{sl-cer}}$	Same as previous but for sl and sl-cer	[sl, sl-cer]-train [sl, sl-cer]-reverse sl-nllb
$\texttt{otrsl}_{\texttt{sr-tor}}$	Same as previous but for sr and sr-tor	[sr, sr-tor]-trans [sr-trans, sr-tor-trans]-reverse sr-nllb-trans
$\texttt{otrslc}_{\texttt{sr-tor}}$	Same as previous but we include both transliterated as well as Cyrillic script data	[sr, sr-tor]-train, [sr, sr-tor]-trans [sr, sr-trans, sr-tor, sr-tor-trans]-reverse sr-nllb, sr-nllb-trans
$\texttt{otrsl}_{\texttt{mix}}$	Cross-lingual mix and match using all data from otrsl setting	[en, hr, sl, sl-cer]-train [sr, sr-tor, mk]-trans [en, hr, mk-trans, sl, sl-cer, sr-trans, sr-tor-trans]-reverse en-gpt4, hr-ckm-claude [hr, sl, mk-trans, sr-trans]-nllb
$\texttt{otrsl}_{\texttt{mix-mk-hr-ckm}}$	Cross-lingual mix and match using all data from $\texttt{otrsl}_{\texttt{mk-hr-ckm}}$ setting	hr-train, mk-trans, hr-ckm-claude [hr-train, mk-trans, hr-ckm-claude]-reverse [hr, mk-trans]-nllb
$\texttt{otrsl}_{\texttt{mix-hr-ckm}}$	Cross-lingual mix and match using all data from $\texttt{otrsl}_{\texttt{hr-ckm}}$ setting	hr-train, hr-ckm-claude [hr-train, hr-ckm-claude]-reverse hr-nllb
$\texttt{otrslc}_{\texttt{mix-testset}}$	Cross-lingual mix and match using all data from otrslc setting except English	all available training data except English

Table 4: After performing data augmentation, we create various data combinations by merging the augmented data blocks described in Table 3 with the original training datasets. These carefully designed data settings are then employed to conduct task-specific fine-tuning or instruction tuning on the selected base models, enabling us to evaluate the impact of different data configurations and therefore, select the suitable ones for the test-set evaluation phase.

method will transform the above example in a 'casue-specific' question as follows:

premise: My friend awoke.
choice1: I poured water on my sleeping friend.✓
choice2: My friend snored. ×

→ We utilize a publicly available English COPA-style synthetic dataset generated by GPT-4 (Achiam et al., 2023), as introduced by Whitehouse et al. (2023). To expand the coverage of this synthetic data to other languages, we translate the English examples using the NLLB-6B machine translation model (Team et al., 2022) to all the four Dialect-Copa standard languages: Croatian, Macedonian, Serbian and Slovenian.

→ The Dialect-Copa dataset does not provide any training or validation data for the Chakavian dialect. To overcome this limitation, we compile a set of Croatian to Chakavian conversion rules and corresponding examples from online language community forums (uni). In addition to these rules, we also gather a few Croatian to Chakavian lyrics translations (lyr). We then prompt the Claude-3 language model (ant) with these rules and examples, instructing it to translate the Croatian sentences from the Dialect-Copa training set into their Chakavian equivalents. Through this process, we create a synthetic Chakavian training set in the style of Dialect-Copa, which we refer to as [lang]-claude. Here is an example with ground truth Croatian to Chakavian translation (words identically translated to the available gold translations are bolded):

→ Croatian (source): Djevojka je pronašla kukca u žitaricama. Izgubila je apetit.

→ Chakavian (gold-translation): Mlada je **našla** neko blago va žitaricah. Je **zgubila tiek**.

→ Chakavian (claude-translation): Divojka je **našla** buba u žitarican. **Zgubila** je **tiek**.

We observe that only a small number of words, specifically three in this instance, are correctly translated from Croatian to Chakavian. Despite the limited accuracy of the translation, this synthetic translated dataset enables us to train and evaluate models on the Chakavian dialect, despite the absence of original training data for this specific dialect. The detailed report on the dialect conversion rules and the Claude-3 prompt template used for generating the synthetic Chakavian dataset can be found in Appendix A.

Table 3 provides a comprehensive overview of the data augmentation techniques employed and the languages covered by each approach.

Phase 3: Data Selection Following the data augmentation process, we create various data combinations by merging the augmented data with the original training datasets. Table 4 provides a comprehensive overview of the various training data combination settings we employ, along with their respective descriptions and the specific data sources included in each combination. These combinations are designed to investigate the impact of different data characteristics on the performance of our models. For instance, the otrsl setting combines all original, transliterated, reverse-augmented, and synthetic data while excluding any data written in the Cyrillic script. The rationale behind this combination is to assess whether our Latin-only Dialect-Copa test set benefits from the absence of script variations in the training data. Additionally, we introduce a language-agnostic data combination denoted as $\texttt{otrsl}_{\texttt{mix}}$, in which we perform cross-lingual modifications by ensuring that the premise, choice1, and choice2 for each example are presented in different languages. This combination allows us to evaluate the models' ability to handle language-agnostic reasoning.

Phase 4: Prompt Design Encoder-based models can be fine-tuned using any of the data settings created in the previous steps. However, to perform few-shot prompting or instruction tuning with generative language models (LLMs), we need to design prompt-based instructions. During our preliminary experiments, we observed that using 4-shot same-class prompting (i.e., providing 4 cause examples for a cause-based question) yields slightly better results compared to combining 2 cause and 2 effect examples in the prompt. Specifically, this approach led to a 4.9% improvement on the English validation set. So we opted for 4-shot same-class prompting to perform inference. Here, these 4-shot same-class examples are randomly drawn from the training set of the target dialect. When the training set is unavailable for a specific dialect, a closely related language is used (eg. Croatian training set for Chakavian examples).

The following prompt template is used for inference and instruction tuning of the Aya-101 model:

```
Instruction: Given the premise, {premise}, What is the correct
{question} {'before'/'after'} this?
A: {choice1}
B: {choice2}
Correct {question}: {correct_answer}
```

By designing the prompt in this manner, we provide the model with a clear instruction, the premise, and the two answer choices. The model is then expected to select the correct answer based on the given question type (cause or effect). This template is employed both during inference and instruction tuning of the Aya-101 model to ensure consistency and optimize performance on the Dialect-Copa dataset.

Phase 5: Task-Specific Tuning We employ two distinct approaches for task-specific tuning of our selected models. The first approach, known as full model fine-tuning, involves updating all the weights of the model during the training process. We apply this method to the BERTić model, fine-tuning it for 5-10 epochs on the Dialect-Copa dataset. However, for mid-size models like Aya-101, full fine-tuning may be unnecessarily computationally expensive, especially considering the limited amount of training data available. To address this concern, we use LoRA (Low-Rank Adaptation) adapter tuning (Hu et al., 2022) which is a more parameter-efficient tuning approach. LoRA introduces a small number of trainable parameters in the form of low-rank matrices, which are inserted between the layers of the pre-trained model. Note that this draws from a long history of efficient adaption using dedicated units (Houlsby et al., 2019; Pfeiffer et al., 2020; Faisal and Anastasopoulos, 2022). During training, only these newly introduced parameters are updated, while the original model weights remain frozen. This approach significantly reduces the number of trainable parameters, making it more suitable for fine-tuning on smaller datasets. By employing LoRA adapter tuning, we can effectively adapt the Aya-101 model to the Dialect-Copa dataset without the need for full model fine-tuning, thereby striking a balance between performance and computational efficiency.

4 Results and Discussion

In this section, we present and discuss the results of our experiments on the Dialect-Copa dataset.

4.1 Validation Set Insights

Table 5 summarizes the key takeaways from our incremental experiments conducted on the validation dataset using BERTić and Aya-101. First, we observe that combining datasets from multiple languages boosts performance on the Croatian subset, as opposed to training on a single language. This finding motivates us to utilize all available training data and prepare different data combinations, as described in Table 4. Second, we find that increasing the data quantity through various data augmentation techniques (Table 3) primarily improves performance for most languages and low-resource dialects. Furthermore, discarding instances written in the Cyrillic script can boost performance for certain languages and dialects (e.g., Croatian, Cerkno, and Serbian), while hurting others.

We also explore cross-lingual mix-and-match strategies, but we do not find any conclusive patterns indicating that this approach consistently makes the model more language-agnostic, as it helps in some cases while hindering performance in others. Additionally, we experiment with discarding English examples and upsampling specific language groups (e.g., Serbian and Torlak examples for the Torlak dialect), which leads to slight performance improvements for the Torlak dialect.

Notably, we observe that full fine-tuning of the comparatively smaller, non-instruction-tuned, but language-specific BERTić model cannot surpass the performance of the multilingual, instruction-tuned Aya-101 model. Finally, we apply the same data combinations to perform instruction tuning on the Aya-101 model and observe an overall performance boost. However, our experiments with different numbers of training epochs (5 and 10) yield inconclusive findings.

4.2 Test Set Insights

Team-specific ranking Table 6 presents a comparison of the best-performing submissions from different teams on the Dialect-Copa test set. We categorize the submissions into two groups: Category 1 includes teams that utilize closed-source model weights, while Category 2 consists of teams that rely on open-source model weights. Our submissions belong to the latter category. We observe that the closed-source GPT-4 model achieves the best overall performance. Team JSI employs GPT-4 with a 10-shot prompting approach, where they provide the first 10 test instances without revealing

Base Model	Setting	en	hr	mk	sl	sl-cer	sr	sr-tor	Avg. (acc)
Takeaway 1: Combining all training data helps									
BERTić	Finetune (hr)	-	64	-	-	-	-	-	-
BERTić	`o`	0.65	0.67	0.55	0.67	0.49	0.66	0.59	0.61
Takeaway 2: Data augmentation helps for low-resource languages in most cases									
BERTić	`otrslc`	0.46	0.7	0.69	0.67	0.42	0.68	0.65	0.61
Takeaway 3: Script choice makes a difference (Using only Latin script performs better on Latin script evaluation)									
BERTić	`otrsl`	0.51	0.77	0.64	0.64	0.59	0.72	0.64	0.65
Takeaway 4: Cross-lingual mix-and-match effect: inconclusive									
BERTić	$\texttt{otrsl}_{\texttt{mix}}$	0.56	0.76	0.68	0.57	0.52	0.66	0.63	0.63
Takeaway 5: Upsampling certain language groups might help targeted evaluation in some cases									
BERTić	$\texttt{otrsl}_{\texttt{sr-tor}}$	0.55	0.74	0.63	0.61	0.48	0.64	0.66	0.62
BERTić	$\texttt{otrsl}_{\texttt{sl-cer}}$	0.54	0.68	0.65	0.57	0.58	0.66	0.64	0.62
Takeaway 6: A multilingually-instruction-tuned 13B parameter model performs significantly better									
Aya-101	4-shot	0.83	0.77	0.75	0.76	0.62	0.81	0.73	0.75
Takeaway 7: Further task-specific instruction tuning helps even more									
Aya-101	`otrsl`	0.86	0.79	0.81	0.91	0.7	0.82	0.77	0.81

Takeaway 7: Training for 10 epochs instead of 5: inconclusive

setting	epochs	en	hr	mk	sl	sl-cer	sr	sr-tor	mean	max count
`otrsl`	10	0.52	0.74	0.65	0.56	0.61	0.67	0.59	0.62	5
	5	0.50	0.72	0.62	0.57	0.58	0.65	0.61	0.61	2
$\texttt{otrsl}_{\texttt{mk-hr-ckm}}$	10	0.49	0.76	0.66	0.62	0.51	0.67	0.70	0.63	1
	5	0.48	0.76	0.68	0.64	0.52	0.71	0.67	0.64	4
$\texttt{otrsl}_{\texttt{sl-cer}}$	10	0.56	0.64	0.65	0.62	0.58	0.65	0.60	0.62	5
	5	0.54	0.68	0.65	0.57	0.58	0.66	0.64	0.62	5
$\texttt{otrslc}_{\texttt{sr-tor}}$	10	0.50	0.70	0.65	0.55	0.48	0.62	0.68	0.60	4
	5	0.49	0.69	0.63	0.60	0.51	0.62	0.70	0.61	4

Table 5: Takeaways from incremental experiments performed on the Dialect-Copa validation dataset. The best language-specific scores for each setting are underlined (Takeaway 7).

Team	Base Model	System Description	sl-cer	hr-ckm	sr-tor	Avg. (acc)
Closed Source Model Weights						
JSI	GPT-4	10-shot with first 10 test instances (without answer)	**0.734**	**0.890**	**0.974**	**0.866**
UNIRI	GPT-4	RAG implementation; Chakavian and Cerkno lexical dictionary; Reasoning instruction and self referral grading task	0.708	0.764	-	-
UNIRI	GPT-4	0-shot iterative prompt	0.664	0.774	0.894	0.777
Open Source Model Weights						
GmuNLP	Aya-101	4-shot prompting	0.694	0.756	0.840	**0.763**
GmuNLP	Aya-101	LORA adapter tuning on $\texttt{otrsl}_{\texttt{hr-ckm}}$ $\rightarrow$ 4-shot prompting	**0.700**	0.750	0.824	0.758
GmuNLP	Aya-101	LORA adapter tuning on `otrsl` $\rightarrow$ 4-shot prompting	0.682	**0.760**	0.824	0.755
GmuNLP	Aya-101	LORA adapter tuning on $\texttt{otrsl}_{\texttt{hr-ckm}}$ $\rightarrow$ 4-shot prompting	0.660	0.742	**0.848**	0.750
WueNLP	Mixtral	LORA adapter tuning on standard variety of target dialect	0.556	0.606	0.738	0.633
CLaC	XLM-R	Fine-tuning XLM-RoBERTa base for multiple choice QA task	0.564	0.522	0.570	0.552

Table 6: Performance comparison of different submissions on Dialect-Copa test set.

base model	setting	epoch	sl-cer	hr-ckm	sr-tor	Avg. (acc)
Aya-101	4-shot	-	0.694	**0.756**	0.840	**0.763**
Aya-101	$\texttt{otrslc}_{\texttt{mix-mk-hr-ckm}}$	5	0.690	**0.756**	0.836	0.761
Aya-101	$\texttt{otrsl}_{\texttt{mix-hr-ckm}}$	5	**0.700**	0.750	0.824	0.758
Aya-101	`otrsl`	5	0.682	**0.760**	0.824	0.755
Aya-101	$\texttt{otrsl}_{\texttt{mk-hr-ckm}}$	5	0.660	0.742	0.848	0.750
Aya-101	$\texttt{otrsl}_{\texttt{sl-cer}}$	5	0.686	0.718	0.836	0.747
BERTić	$\texttt{otrsl}_{\texttt{hr-ckm}}$	10	0.572	0.626	**0.722**	**0.640**
BERTić	$\texttt{otrsl}_{\texttt{mk-hr-ckm}}$	5	**0.582**	**0.634**	0.682	0.633
BERTić	$\texttt{otrslc}_{\texttt{mix-testset}}$	5	0.576	0.622	0.692	0.630
BERTić	`otrsl`	10	0.540	0.622	0.700	0.621

Table 7: GMUNLP system submissions for test-set evaluation. The best dialect-specific scores for each base-model type are bolded.

the answers. Interestingly, even the 0-shot prompting using GPT-4 (by team UNRI) outperforms all submissions in Category 2 using open-source models. Among the Category 2 submissions, GMUNLP (our submission) achieves the highest performance on all varieties. The base Aya-101 model with 4-shot prompting yields the best average score across all languages. However, LoRA adapter tuning on different data combinations results in language-specific best scores.

GMUNLP submission Table 7 presents the results of our selected 10 system submissions. We observe that the best performance achieved by the BERTić model on the `otrsl` setting is 62%, which is approximately 17% lower compared to the `otrcl`-tuned Aya-101 model. When comparing language-specific results, we find that the Torlak (sr-tor) dialect is the easiest to predict for both the Aya-101 and BERTi'c models, while the Cerkno dialect (sl-cer) proves to be the most challenging to learn.

Interestingly, upsampling the Cerkno dialect-related data ($\texttt{otrsl}_{\texttt{sl-cer}}$-tuned) does not yield the best score for the Cerkno test-set. Instead, upsampling the Chakavian dialect-related data using the $\texttt{otrsl}_{\texttt{mk-hr-ckm}}$ setting leads to better scores on the Cerkno test set. This observation holds true for both the Aya-101 and BERTić base models, indicating that leveraging data from more closely related languages does not always provide the most significant benefit. However, all varieties here are from the same language group and it is not unnatural that oversampling other languages or dialects can have a positive impact as well (given the brittleness of the fine-tuning process). We believe this phenomenon warrants further investigation to gain a deeper understanding of the complex interplay between language-relatedness and task-specific model performance.

5 Conclusion

In this study, we explored the impact of data augmentation techniques on fine-tuning multilingual language models for improving common sense reasoning in dialectal variations. Our experiments encompassed a range of language models, from smaller to mid-sized architectures, to investigate their adaptability to dialectal nuances. The observed variations in performance and the upper limits achieved by different models reflect the diverse ways in which language models handle and adapt to dialectal variations. The insights gained from this work may contribute to the development of more robust and adaptable language models that can handle the challenges posed by dialectal variations. Future work can explore advanced data augmentation techniques, investigate the impact of domain-specific knowledge integration, and develop novel architectures tailored to the unique characteristics of dialects.

Acknowledgements

This work has been generously supported by the National Science Foundation under grants IIS-2125466 and IIS-2327143. We are thankful to the anonymous reviewers and area chairs for their constructive feedback. This project was supported by resources provided by the Office of Research Computing at George Mason University (https://orc.gmu.edu) and funded in part by grants from the National Science Foundation (Awards Number 1625039 and 2018631).

References

Claude — anthropic.com. https://www.anthropic.com/claude. [Accessed 25-03-2024].

Croatian-chakavian features -UniLang — forum.unilang.org. https://forum.unilang.org/viewtopic.php?t=14771. [Accessed 28-03-2024].

Elitni Odredi - Ljubavi Moja lyrics + Croatian (Chakavian dialect) translation — lyricstranslate.com. https://lyricstranslate.com/en/ljubavi-moja-jubavi-moja.html. [Accessed 28-03-2024].

OpenAI Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mo Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Benjamin Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Sim'on Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Raphael Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Lukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Hendrik Kirchner, Jamie Ryan Kiros, Matthew Knight, Daniel Kokotajlo, Lukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Adeola Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel P. Mossing, Tong Mu, Mira Murati, Oleg Murk, David M'ely, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Ouyang Long, Cullen O'Keefe, Jakub W. Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alexandre Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Pondé de Oliveira Pinto, Michael Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario D. Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin D. Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas A. Tezak, Madeleine Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cer'on Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2023. Gpt-4 technical report.

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.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Tech-*

nologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Fahim Faisal, Orevaoghene Ahia, Aarohi Srivastava, Kabir Ahuja, David Chiang, Yulia Tsvetkov, and Antonios Anastasopoulos. 2024. Dialectbench: A nlp benchmark for dialects, varieties, and closely-related languages.

Fahim Faisal and Antonios Anastasopoulos. 2022. Phylogeny-inspired adaptation of multilingual models to new languages. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 434–452, Online only. Association for Computational Linguistics.

Andrew S. Gordon, Zornitsa Kozareva, and Melissa Roemmele. 2011. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *AAAI Spring Symposium: Logical Formalizations of Commonsense Reasoning*.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2790–2799. PMLR.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Peiqin Lin, Shaoxiong Ji, Jörg Tiedemann, André F. T. Martins, and Hinrich Schütze. 2024. Mala-500: Massive language adaptation of large language models.

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, Kiyv, Ukraine. Association for Computational Linguistics.

Nikola Ljubešić, Nada Galant, Sonja Benčina, Jaka Čibej, Stefan Milosavljević, Peter Rupnik, and Taja Kuzman. 2024. DIALECT-COPA: Extending the standard translations of the COPA causal commonsense reasoning dataset to south slavic dialects. In *Eleventh Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2024)*, Mexico City, Mexico. Association for Computational Linguistics.

Jonas Pfeiffer, Andreas Rücklé, Clifton Poth, Aishwarya Kamath, Ivan Vulić, Sebastian Ruder, Kyunghyun Cho, and Iryna Gurevych. 2020. Adapterhub: A framework for adapting transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020): Systems Demonstrations*, pages 46–54, Online. Association for Computational Linguistics.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Chenxi Whitehouse, Monojit Choudhury, and Alham Aji. 2023. LLM-powered data augmentation for enhanced cross-lingual performance. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 671–686, Singapore. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Ahmet Üstün, Viraat Aryabumi, Zheng-Xin Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, Shayne Longpre, Niklas Muennighoff, Marzieh Fadaee, Julia Kreutzer, and Sara Hooker. 2024. Aya model: An instruction finetuned open-access multilingual language model.

A Croatian to Chakavian conversion rules

Here we report the collected Croatian to Chakavian conversion rules and their corresponding examples (uni), the lyrics translations (lyr) and the Claude-3 prompt we used to generate the Cahakavian synthetic translations.

Conversion rules

```
Rule: m's at the end of words become n's
Examples:
Ja sam = Ja san
osam = osan
s ženom = s ženon
vidim = vidin

Rule: đ becomes j
Examples:
mlađi=>mlaji
među=>meju

Rule: The sounds ć and đ actually do not
exist separately from č and dž

Rule: genitive plural in the feminine and
neuter takes a zero ending:
Examples: žena = žen, sela = sel

Rule: dative/instrumental plurals are
also somewhat different.

Rule: Masculine and neuter nouns can have
an alternate ending.
Examples:
it can be gradovima or gradoviman

Rule: For feminine nouns, it's shortened.
Examples:
ženama=>ženan

Rule:-ao endings are shortened to just -a
Examples:
išao je => iša je
rekao sam => reka san

Rule:-io endings change to -ija
Examples:
govorio je => govorija je
vidio sam => vidija san

Rule:ča, aside from meaning "what" can
also be used as a particle meaning "away"
or "out".
Examples:
gremo ča (let's get out of here)

Rule: some of the third person plural
forms can often be extended into longer
-u ending forms.
Examples:
govore -> govoru -> govoridu
rade -> radu -> radidu
pišu -> pišedu

Rule: Infinitive is shortened .
There is no I at the end.
In some speeches there is neither T/Ć at
the end
Examples:
bit =>bi
pivat =>piva
znat => zna
plivat => pliva
ronit => roni

Rule: change lj=>j
Examples:
zaljubiti se=>zajubit se
ljubav =>jubav
ljuska =>juska
ljudi=>judi

Rule: Change O => E in some cases
Example:
nekoga-nikega
svakoga-svakega
tomu-temu
toga-tega
bijeloga-bilega
jednoga-jenega
jednomu-jenemu

Rule: In standard croatian third person
plural has ending in some verbs E.
In chakavian it is always U
Examples:
vide =>vidu
hoće =>hoću
stoje =>stoju
```

```
stave =>stavu
motre =>motru
leže=>ležu

Rule: Change H=>V
Examples:
kruh =>kruv
kuhati =>kuvati
suh =>suv
gluh =>gluv[/b]
```

Lyrics translations We collected Croatian to Chakavian lyrics translations from the `lyricstranslate.com` site (lyr).

Claude-3 prompt We use the following prompt consisting the the above mentioned conversion rules, examples and lyrics translations:

```
<conversion_file.txt>

This file contains Croatian to Chakavian
dialect conversion grammar rules with
examples.

Now here are some Croatian sentences and
it's parallel Chakavian sentences:

<Croatian_lyrics.txt>
<Chakavian_lyrics.txt>.

Given these resources, I want you to translate
the following Croatian sentences to Chakavian
dialect.

<Croatian_training_sentences.txt>
```

JSI and WüNLP at the DIALECT-COPA Shared Task: In-Context Learning From Just a Few Dialectal Examples Gets You Quite Far

Nikola Ljubešić[1,2], Taja Kuzman[1], Peter Rupnik[1],
Ivan Vulić[3], Fabian David Schmidt[4], and Goran Glavaš[4]
[1]Dept. of Knowledge Technologies, Jožef Stefan Institute
[2]University of Ljubljana
[3]Language Technology Lab, University of Cambridge
[4]Center for AI and Data Science (CAIDAS), University of Würzburg
{nikola.ljubesic, taja.kuzman, peter.rupnik}@ijs.si
{fabian.schmidt, goran.glavas}@uni-wuerzburg.de, iv250@cam.ac.uk

Abstract

The paper presents the JSI and WüNLP systems submitted to the DIALECT-COPA shared task on causal commonsense reasoning in dialectal texts. Jointly, we compare LLM-based zero-shot and few-shot in-context inference (JSI team), and task-specific few-shot fine-tuning, in English and respective standard language, with zero-shot cross-lingual transfer (ZS-XLT) to the test dialects (WüNLP team). Given the very strong zero-shot and especially few-shot in-context learning (ICL) performance, we further investigate whether task semantics, or language/dialect semantics explain the strong performance, showing that a significant part of the improvement indeed stems from learning the language or dialect semantics from the in-context examples, with only a minor contribution from understanding the nature of the task. The higher importance of the dialect semantics to the task semantics is further shown by the finding that the in-context learning with only a few dialectal instances achieves comparable results to the supervised fine-tuning approach on hundreds of instances in standard language.

1 Introduction

Causal commonsense reasoning is an important aspect of natural language understanding (NLU) abilities of the large language models (LLMs); their performance on such tasks probes the extent to which the LLMs have acquired commonsense and world knowledge. Choice Of Plausible Alternatives (COPA) dataset (Roemmele et al., 2011) has de facto been the standard evaluation benchmark for causal commonsense reasoning for over a decade.[1] Like on most other NLU tasks, state-of-the-art LLMs exhibit impressive performance on the English COPA dataset (Chowdhery et al., 2023; Zhong et al., 2022). LLMs, unlike their smaller encoder-based predecessors (e.g., BERT, RoBERTa), also offer spectacular COPA performance for other languages (Ponti et al., 2020; Žagar and Robnik-Šikonja, 2022; Shi et al., 2023), including South Slavic languages, both with Latin and Cyrillic scripts, reaching accuracy levels between 94% and 97%[2]. Though LLMs excel on high-resource and moderately resourced standard languages, their utility for commonsense reasoning in truly low-resource languages (Senel et al., 2024) and especially dialects (Joshi et al., 2024) has been much less scrutinized. In the DIALECT-COPA shared task of the VarDial Evaluation Campaign 2024 (Chifu et al., 2024), COPA is extended to geographically very localized dialects (i.e., micro- or nano-dialects) of South Slavic languages that are very rarely present in texts online, and thus could not have been (except perhaps in minimal traces) present in the pretraining corpora of LLMs.

In this work, we focus on benchmarking decoder-style LLMs in the DIALECT-COPA task, covering a variety of closed-source and open-source LLMs in zero-shot and few-shot in-context learning (ICL) inference setups. Subsequently, we select the best-performing open-source model during in-context learning (Mixtral Instruct) and fine-tune it for the task in the standard supervised fashion – assuming a somewhat larger training dataset – with training instances either in English or in the respective standard language of the target dialect (e.g., Slovenian for the Cerkno dialect).

[1]Inter alia, the COPA dataset is included to the selection of tasks in the well-known benchmark for general-purpose natural language understanding SuperGLUE (Wang et al., 2019).

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

Proceedings of the Eleventh Workshop on NLP for Similar Languages, Varieties, and Dialects (VarDial 2024), pages 209–219
June 20, 2024

We make use of all the development data provided inside the shared task, namely the translations of the COPA dataset (Roemmele et al., 2011) into the standard Slovenian, Croatian, Serbian and Macedonian languages, as well as the translations available for two out of three dialects, namely the Cerkno and the Torlak dialects (Ljubešić et al., 2024). While we have access to both training and development portions of COPA datasets for other languages and dialects, the Chakavian dialect is a surprise dialect: findings from the other two dialects thus steered decisions for Chakavian too.

To sum up, we evaluate the LLMs in the following scenarios: **1)** *zero-shot inference* where the model is presented with the task description in English and needs to provide an answer to the COPA instances in the South Slavic dialects; **2)** *few-shot in-context learning* (ICL) where the prompt is extended with additional examples from the respective COPA dataset; and **3)** *fine-tuning zero-shot cross-lingual transfer* (ZS-XLT) (Lauscher et al., 2020; Schmidt et al., 2022), in which an LLM is (in a parameter-efficient manner) fine-tuned on training data in English or a standard South Slavic language (Slovenian, Serbian, and Croatian, respectively) and then used to make predictions in the corresponding target dialect (Cerkno, Torlak, and Chakavian, respectively).

The ICL variants in general, with few target dialect instances in the context, exhibit a significantly improved performance in comparison to zero-shot performance. Comparing ICL to fune-tuning zero-shot cross-lingual transfer, we observe a comparable performance.

Following the finding of significant improvements through just a few target dialect examples, we investigate the source of these few-shot ICL performance gains. We find that the exposure to the dialect itself through the few in-context instances is key, as opposed to exposure to the COPA task itself.

2 Multi-Parallel COPA Datasets

Our work focuses on the Choice Of Plausible Alternatives (COPA) dataset, originally published in English (Roemmele et al., 2011), and its translation-based derivatives in a selection of South Slavic languages and dialects. All COPA datasets have the same set of instances, and they differ only in the language variety in which the instances are written. The COPA dataset consists of 1,000 examples, split into 400 training, 100 development and 500 test instances. Each instance consists of three sentences: a statement (*premise*) and two possible *effects* or *causes* (alternatives) for the statement, e.g., a premise *All my socks were in the laundry* is coupled with two *effect* choices: *I wore sandals* (correct/plausible) and *I wore boots* (incorrect).

We evaluate the models on 'standard language' and dialectal versions of a selection of South Slavic languages. More precisely, we use the following COPA datasets for three South Slavic dialects – the Slovenian Cerkno dialect (COPA-SL-CER), the Croatian Chakavian dialect (COPA-HR-CKM), and the Torlak dialect of Serbian (COPA-SR-TOR) (Ljubešić et al., 2024). The models' performance on the dialectal datasets is compared with their performance on the datasets in the standard South Slavic language that is closest to them, namely Slovenian (COPA-SL) (Žagar et al., 2020), Croatian (COPA-HR) (Ljubešić, 2021), Serbian (COPA-SR) (Ljubešić et al., 2022b) and Macedonian (COPA-MK) (Ljubešić et al., 2022a). All the datasets were translated from the English COPA dataset (Roemmele et al., 2011) following the XCOPA translation and adaptation methodology (Ponti et al., 2020), except for Slovenian which was translated as part of the Slovenian SuperGLUE benchmark (Žagar and Robnik-Šikonja, 2022). Torlak, Serbian and Macedonian datasets are written in Cyrillic and all other in Latin script.

The COPA datasets for the standard South Slavic languages and English are openly available, whereas the dialectal COPA datasets have been introduced in the DIALECT-COPA shared task, part of the VarDial Evaluation Campaign 2024 (Chifu et al., 2024) and are currently only partly available: as part of the shared task, the training and development portions were made publicly available for all languages (Ljubešić et al., 2024);[3] the test splits have been made available only to the shared task participants. Inside the shared task, no training and development data were given for the Chakavian dialect, to enable estimation and analysis of models' performance "in the wild" for a new (truly low-resource) dialect.

3 Models in Evaluation

In this work, we extend the prior experiments that focused on the use of LLMs for the task (Wi-

[3]The training and development splits can be accessed at the CLARIN.SI repository: http://hdl.handle.net/11356/1766.

bowo et al., 2023; Ljubešić et al., 2024) by (i) evaluating a larger number of open- and closed-source instruction-tuned generative LLMs, and by (ii) widening investigation from the basic zero-shot scenarios to few-shot in-context learning and cross-lingual transfer of supervised fine-tuning. In this section, we outline all the models, with links to the models provided in Appendix A.

GPT-3.5 Turbo and GPT-4 are closed-source models provided by OpenAI through their payable API (OpenAI, 2023a,b). We use the versions `gpt-3.5-turbo-0125` and `gpt-4-0125-preview` through the chat completion endpoint, with temperature set to 0. The models are said to be trained on massive multilingual web text collections; however, the details on pretraining data, as well as the details of the training procedure and model architecture are not publicly known.

Mistral 7B Instruct is an open-source model provided by Mistral AI (Jiang et al., 2023). We experiment with two 7B model variants, `Mistral-7B-Instruct-v0.1` and `Mistral-7B-Instruct-v0.2`, where the main difference is that v0.2 extends the context size from 8k to 32k input tokens. The details on the pretraining data have not been made available.

Mixtral 8×7B Instruct is another open-source model from Mistral AI (Jiang et al., 2024). We use the `Mixtral-8x7B-Instruct-v0.1` variant. The main difference between Mistral and Mixtral is the introduction of a sparse mixture-of-experts network in Mixtral, where 8 feed-forward blocks are added to each layer. For each token, two blocks are selected to process it. As a consequence, despite having 47B parameters in total, only 13B active parameters are used for each token. Furthermore, it is stated that Mixtral was pretrained on much larger quantities of multilingual data than Mistral. The context size is 32K tokens.

mT0-XXL is an open-source model developed by the BigScience academic initiative (Muennighoff et al., 2023). We use the `mT0-XXL` variant which has 13 billion parameters. The model is a fine-tuned version of the multilingual mT5 model (Xue et al., 2021), which was pretrained on a sample from the mC4 dataset covering 101 languages.

Aya 101 is an open-source model developed by Cohere For AI (Üstün et al., 2024). We use the `aya-101` variant with 13B parameters. As mT0 above, it is an instruction-tuned version of mT5 (Xue et al., 2021), relying on a multilingual dataset that covers 101 languages.

Gemma 7B It is an open-source model provided by Google (Mesnard et al., 2024). It is a lightweight 7B version of Google's closed-source Gemini model family (Anil et al., 2023), and it was trained primarily on English data.

Falcon-7B-Instruct is an open-source 7B model developed by the Technology Innovation Institute (Almazrouei et al., 2023). It is an instruction-tuned version of the Falcon-7B language model which was pretrained on English and French data.

Llama-2-7B-Chat is a 7B open-source model from Meta (Touvron et al., 2023), with the context size of 4000 tokens, intended primarily for English.

In sum, the coverage of evaluated models is extensive, where the models vary in their availability (open-sourced versus 'black-box' commercial models), size, as well as their pretraining data. For instance, while mT0 and Aya 101 were pretrained on massively multilingual datasets, other models are primarily built for English only, such as Gemma and Llama-2-Chat. Further, while most models have 7B parameters, Mixtral 8×7B Instruct, mT0 and Aya 101 have 13B parameters.

To maximize the comparability between the results of the models, we provide them all with identical prompts (available in Appendix B). We ran all our experiments on a single A100 40GB.[4]

4 Results and Discussion

We now delve into the main experiments, covering zero-shot and different 10-shot ICL scenarios, followed by ablations on the importance of learning 'language/dialect semantics' versus 'task semantics/structure' in ICL. Finally, we report experiments with supervised fine-tuning.[5]

4.1 Zero-Shot Inference

Table 1 summarizes the results of zero-shot inference with LLMs on the training portions of the datasets (400 examples), with models listed in decreasing order of performance on standard language datasets (column STD), that is, Slovenian (sl), Croatian (hr), Serbian (sr) and Macedonian

[4]Due to this, we relied on an 8-bit quantization for Mistral models and a 4-bit quantization for Mixtral models.

[5]While the data for Serbian, Macedonian and Torlak are available both in the Latin and in the Cyrillic script, we report only the results on Serbian and Torlak Latin data and Macedonian Cyrillic data; these options yielded higher absolute scores across the models.

(mk). The ranking of the models based on the dialectal performance (column DIA), i.e., on Cerkno (sl-cer) and Torlak (sr-tor), is similar.

While the model ranking is relatively similar relative on both standard and dialect varieties, all models expectedly perform substantially worse on dialectal datasets. For instance, the best-performing system, GPT-4, drops 36.5 accuracy points (from 96 to 59.5) between Slovenian (sl) and its Cerkno dialect (sl-cer), and from 95.8 to 76 on average. Such drops are observed for all the other models as well (e.g., mT0 as the best-performing open-source model has 14 points lower average accuracy on DIA compared to STD).

Overall, GPT-4 outperforms the open-model competition by a wide margin, with mT0 as the closest follower (10 accuracy points difference on DIA). Expectedly, Mixtral performs much better than its smaller Mistral 7B Instruct counterparts. Two systems that perform worse than expected are Aya 101, which closely follows the design of an earlier mT0 model, and Gemma 7B It. Finally, Falcon-7B-Instruct and Llama-2-7B-Chat perform worse than the random baseline of 50% due to their inability to follow instructions, frequently providing answers in which neither of the two alternatives is chosen. This might stem from their limited multilingual capabilities, as outlined in Section 3.

Other Observations. It is worth noting that models generally tend to exhibit similar performance across the standard language variety: there are no large or consistent differences in performance on Slovenian, Croatian, Serbian and Macedonian, despite the fact that these languages are not equally resourced (e.g., Slovenian is by far the most resourced of the four, whereas Macedonian is the least resourced (Terčon and Ljubešić, 2023)).

In contrast, models' performance across the two dialects is vastly different. The Cerkno dialect seems to be much more challenging for all models than the Torlak dialect. This, we believe, stems from the fact that Torlak is significantly closer to the standard Serbian and Macedonian than Cerkno is to standard Slovenian (Ljubešić et al., 2024).

4.2 Few-Shot In-Context Learning

We next perform in-context learning (ICL) only over the models that performed above the random baseline in the zero-shot evaluation. First, we note that mT0 and Aya 101, both based on mT5, actually experienced performance decrease when moving from zero-shot to few-shot ICL scenarios. We speculate that this might be a consequence of limited context size and encoder capacity, which might be incapable of encoding a longer prompt. We thus present only the results where models show gains moving from zero-shot to ICL scenarios.

In our preliminary experiments, we varied the number of few-shot examples from the development set provided to the models. The results show consistent improvements as the number of shots increases up to 10, followed by minor and negligible gains with 20 instead of 10 shots. For that reason, we report the results in the 10-shot scenario. An example of a prompt is provided in Appendix B. An overview of results with zero-shot (Section 4.1) versus 10-shot prompting scenarios is provided in Table 2.

The main finding is that ICL, for the models with sufficient context sizes where ICL works as expected, offers substantial performance benefits both for the standard languages (column STD) and for the target dialects (column DIA). Interestingly, the largest absolute gains from 10-shot ICL are observed for the most difficult, Cerkno dialect: performance of GPT-4 rises from 60% to 74% in accuracy.

The observed gains with ICL thus open up the following question – where do the gains come from? Is it the adaptation to the task and its structure, or is it rather the adaptation to the target language and dialect and a better understanding of it? We discuss this next in the prompt ablation tests.

Prompt Ablation Experiments. We aim to discriminate between the contributions of learning the 'task semantics' versus learning the 'language/dialect semantics' by performing two experiments: **1)** in the `list` experiment we add to the initial zero-shot prompt only lists of sentences of the target language, and **2)** in the `task` experiment the structure of the task is added to the initial prompt by providing instances from the COPA dataset but without any answer. As before, we use the development dataset instances for few-shot prompts. With `list` we ablate the task definition, while with `task` we ablate the information on the answer, but still provide information on task itself.

The results are given in Table 3. The main finding is that the substantial part of the total improvement comes from the language/dialect semantics, represented by the `list` results. An answer to the task, missing in the `task` scenario, but included in

Model	STD	DIA	sl	sl-cer	hr	sr	sr-tor	mk
gpt-4-0125-preview	0.958	0.760	0.960	0.595	0.960	0.968	0.925	0.943
mT0-xxl	0.798	0.660	0.787	0.540	0.738	0.765	0.713	0.838
gpt-3.5-turbo-0125	0.799	0.646	0.802	0.547	0.820	0.830	0.745	0.745
aya-101	0.710	0.610	0.728	0.530	0.645	0.665	0.623	0.720
Mixtral-8x7B-Instruct-v0.1	0.691	0.521	0.682	0.405	0.705	0.713	0.637	0.665
gemma-7b-it	0.599	0.546	0.593	0.522	0.570	0.618	0.552	0.605
Mistral-7B-Instruct-v0.2	0.524	0.396	0.515	0.285	0.542	0.537	0.487	0.497
Mistral-7B-Instruct-v0.1	0.510	0.495	0.507	0.487	0.507	0.515	0.500	0.502
falcon-7b-instruct	0.432	0.442	0.500	0.485	0.463	0.458	0.510	0.357
llama-2-7b-chat	0.114	0.032	0.175	0.020	0.152	0.145	0.090	0.035

Table 1: Zero-shot results, with additional averages reported over the three standard languages (STD) and the three dialects (DIA). Results are reported in accuracy scores.

Model	# shots	STD	DIA	sl	sl-cer	hr	sr	sr-tor	mk
Mistral-7B-Instruct-v0.2	0	0.524	0.396	0.515	0.285	0.542	0.537	0.487	0.497
Mistral-7B-Instruct-v0.2	10	0.734	0.570	0.718	0.507	0.757	0.752	0.632	0.708
Mixtral-8x7B-Instruct-v0.1	0	0.691	0.521	0.682	0.405	0.705	0.713	0.637	0.665
Mixtral-8x7B-Instruct-v0.1	10	0.780	0.624	0.802	0.5	0.818	0.795	0.748	0.703
gpt-3.5-turbo-0125	0	0.799	0.646	0.802	0.547	0.820	0.830	0.745	0.745
gpt-3.5-turbo-0125	10	0.828	0.666	0.845	0.53	0.84	0.858	0.802	0.77
gpt-4-0125-preview	0	0.958	0.760	0.960	0.595	0.960	0.968	0.925	0.943
gpt-4-0125-preview	10	0.984	0.853	0.98	0.738	0.988	0.99	0.968	0.978

Table 2: Zero- and ten-shot results in terms of accuracy across models that improve with few-shot prompting. Averages for datasets in standard languages (STD), i.e., Slovenian, Croatian, Serbian and Macedonian, and dialectal datasets (DIA), i.e., Cerkno and Torlak, are given.

Variant	STD	DIA	sl	sl-cer	hr	sr	sr-tor	mk
zero-shot	0.691	0.521	0.682	0.405	0.705	0.713	0.637	0.665
10-shot	0.780	0.624	0.802	0.5	0.818	0.795	0.748	0.703
list	0.745	0.607	0.74	0.515	0.775	0.757	0.698	0.708
task	0.786	0.619	0.818	0.492	0.805	0.802	0.745	0.72

Table 3: Results over the ablated 10-shot examples on the Mixtral 8×7B Instruct model, either to the level of a list of sentences (list) or tasks without any answer given (task), compared to the previous results of zero-shot and 10-shot experiments. We additionally provide averages over standard languages (STD) and dialects (DIA).

the `10-shot` scenario, seems to be almost irrelevant for ICL. However, the remaining gap between the `list` and the `task` rows in Table 3 indicates that providing examples of the task, although without the answer, is still beneficial.

These results shed important light on *why* in-context learning offers substantial gains both on standard languages and on dialects. However, there is another angle, specific to this shared task, that these results open up. Namely, both the `list`- and the `task`- transformed prompts do not require the correct answer to be known as part of in-context examples; they can therefore be run even on the Chakavian dialect, for which no training and development data were available in the shared task. Interestingly, omitting an answer even yields minor gains on the datasets in standard languages, and just a minor drop in performance on the dialectal datasets.

4.3 Fine-Tuning and ZS-XLT

The WüNLP team next investigates zero-shot cross-lingual transfer (ZS-XLT) with an LLM fine-tuned on English training data or the training data in the corresponding standard language (e.g., for Chakavian as target, we train on the instances from the training portion of Croatian COPA). Following the JSI team's zero-shot inference and few-shot ICL results, we opt to tune Mixtral 8×7B Instruct as the best-performing open-source LLM in their ICL experiments. We fine-tune the model generatively, using the prompt below, and constraining the output vocabulary to “1”, “2” (we minimize the standard negative log likelihood loss):

```
'Premise: "{PREMISE}"
 Question: "{QUESTION}"
 Choice 1: "{CHOICE1}"
 Choice 2: "{CHOICE2}"
 Answer: '
```

Since we are running supervised fine-tuning, we chose to prepend the task description to the prompt.[6] We carry out fine-tuning in a parameter-efficient manner, using quantized (4-bit) low-rank adaptation (Q-LoRA) (Hu et al., 2021; Dettmers et al., 2024), optimizing the LoRA matrices with AdamW (Loshchilov and Hutter, 2018) (learning rate 10^{-5} with linear decay, no warmup). We train on the whole training set (400 instances) in batches of 32 instances, for 10 epochs, checkpointing the model after every update.

Although the DIALECT-COPA shared task offers validation portions in target languages/dialects, one should note that, following Schmidt et al. (2022, 2023b), using target language development set for model selection violates true zero-shot cross-lingual transfer: the labeled target language validation instances would, in fact, be better used as training data (Schmidt et al., 2023b). Because of this, we report results for two model variants: (1) training on English instances (`en`) vs. instances of the corresponding standard language (`x`) $\times$ (2) selecting the last checkpoint (`last`) of the training run (true ZS-XLT) vs. selecting the model checkpoint that has the best performance on the target language validation set (`val`, violates true ZS-XLT). These four variants are named as: MixtralLoRA-{`en`,`x`}-{`last`,`val`}. Table 4 summarizes the performance for all four variants on the validation data of standard South Slavic languages as well as target dialects. The final official shared task results for all four variants (runs), on the test portions of target dialects, are reported in the next section.

4.4 Results on Test Data

We present the official test data results of both teams in Figure 1. The runs from WüNLP comprise fine-tuning Mixtral 8×7B Instruct either on the English or the standard data across two model selection scenarios, as described in Section 4.3. Similar to the results during the development phase (Table 4), there is no strong difference between the variants: the averages are almost identical. However, comparing this set of results to the zero-shot approach with Mixtral 8×7B Instruct, we observe positive impact of fine-tuning, even if fine-tuning was conducted on English or standard language data.

The `list` and the `task` approaches in the 10-shot ICL scenario with Mixtral 8×7B Instruct, conducted by JSI, improve over the zero-shot scenario, arriving roughly to the level of the WüNLP fine-tuning results.

The best results of the two teams, as in the shared task overall, are obtained, not surprisingly, with the GPT-4-based take on zero-shot inference, and even more on the two approaches to 10-shot ICL without having the correct answers at hand. While zero-shot prompting already improves over any of Mixtral results on each of the three dialects, achieving an average result of 75% accuracy, the model excels further once 10 examples of the language are provided for ICL, even only as examples of the dialect in question, with the average result rising to 83%. Describing the nature of the task combined with the 10 shots, but without the correct answer, yields an additional gain, resulting in an average accuracy of 87%.

Interestingly, the 'harder' the dialect, the more is gained by just submitting exemplary sentences of the dialect during in-context learning, with a much more significant jump from zero-shot scenario (`gpt4-zero`) to the scenario with a list of sentences in the dialect added (`gpt4-list`) on the Cerkno dialect (considered a 'hard dialect') than on the Torlak dialect (considered an 'easy dialect'). We see similar further gains moving from the scenario with the list of sentences in the dialect (`gpt4-list`) to the scenario where examples of the task are added (`gpt4-task`).

5 Conclusion

In this work, we benchmark three mainstream approaches for using LLMs for causal commonsense reasoning in three South Slavic dialects: (1) zero-shot inference with LLMs, (2) few-shot in-context learning, and (3) supervised fine-tuning and zero-shot cross-lingual transfer. We find that, for the same LLM, both few-shot ICL and cross-lingual transfer with supervised fine-tuning (with training instances in English or in the standard language of the target dialect) expectedly outperform zero-shot inference with LLMs. Somewhat surprisingly, few-shot ICL with as few as 10 in-dialect instances tends to perform comparably to fine-tuning based

[6] Recent work indeed suggests that, unlike in zero-shot inference and few-shot ICL, task description prompts have limited effect on performance in supervised fine-tuning (Li et al., 2023).

Variant	STD	DIA	sl	sl-cer	hr	sr	sr-tor	mk
MixtralLoRA-en-last	0.815	0.615	0.82	0.52	0.82	0.87	0.71	0.75
MixtralLoRA-en-val	0.82	0.645	0.82	0.57	0.82	0.87	0.72	0.77
MixtralLoRA-x-last	0.825	0.675	0.80	0.57	0.83	0.89	0.78	0.78
MixtralLoRA-x-val	0.833	0.69	0.82	0.60	0.84	0.89	0.78	0.78

Table 4: Fine-tuning zero-shot cross-lingual transfer results (ZS-XLT) on the *validation* data: fine-tuning Mixtral-Instruct 8x7B with Q-LoRA, either on English training data (en) or the training portion of the standard language corresponding to the target dialect (x); For each of the two models (*en* vs. *x*) we report the performance of the `last` checkpoint as well as the checkpoint that yields the best validation performance. We additionally provide averages over standard datasets (STD) and dialectal datasets (DIA).

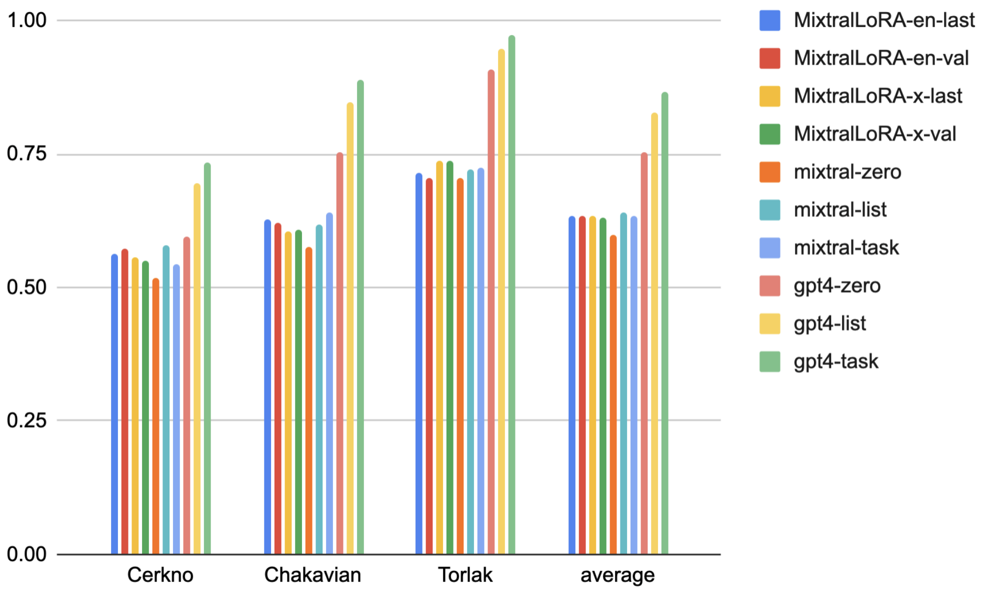


Figure 1: Test data results

on 400 instances in standard languages that are related to the corresponding dialect. Further inspection reveals that the LLMs leverage the few provided in-dialect instances to improve their understanding of the target dialect, rather than to learn the task and its structure. Future work will investigate further recent strategies for improving performance of LLMs for low-resource languages and in cross-lingual transfer, including, *inter alia*, checkpoint averaging in fine-tuning (Schmidt et al., 2023a) and supervised in-context learning (Li et al., 2023).

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) funded by the Slovenian Research and Innovation Agency, and withing the research programme "Language resources and technologies for Slovene" (P6-0411), both funded by the Slovenian Research and Innovation Agency (ARIS).

6 Limitations

One of the limitations of the presented paper is the use of closed-source models. While we decided to include them in the analyses to be able to obtain an insight into how well the open-source models perform in comparison to the closed-source models, we should note that we have limited insights to the architecture of these models and that the reproducibility of these results might be hindered by updates to the models that might not be communicated openly.

References

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Merouane Debbah, Etienne Goffinet, Daniel Heslow, Julien Launay, Quentin Malartic, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. 2023. Falcon-40B: an open large language model with state-of-the-art performance.

Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

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.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2024. Qlora: Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems*, 36.

Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2021. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

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*.

Aditya Joshi, Raj Dabre, Diptesh Kanojia, Zhuang Li, Haolan Zhan, Gholamreza Haffari, and Doris Dippold. 2024. Natural language processing for dialects of a language: A survey. *arXiv preprint arXiv:2401.05632*.

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020. From zero to hero: On the limitations of zero-shot language transfer with multilingual transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4483–4499.

Chengzu Li, Han Zhou, Goran Glavaš, Anna Korhonen, and Ivan Vulić. 2023. On task performance and model calibration with supervised and self-ensembled in-context learning. *arXiv preprint arXiv:2312.13772*.

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, 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, Nada Galant, Sonja Benčina, Jaka Čibej, Stefan Milosavljević, Peter Rupnik, and Taja Kuzman. 2024. DIALECT-COPA: Extending the standard translations of the COPA causal commonsense reasoning dataset to south slavic dialects. In *Eleventh Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2024)*, Mexico City, Mexico. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2018. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, and et al. 2024. Gemma.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, et al. 2023. Crosslingual generalization through multitask finetuning. In *The 61st Annual Meeting Of The Association For Computational Linguistics*.

OpenAI. 2023a. ChatGPT General FAQ. `https://help.openai.com/en/articles/6783457-chatgpt-general-faq`. Accessed: March 3, 2023.

OpenAI. 2023b. 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*.

Fabian David Schmidt, Ivan Vulić, and Goran Glavaš. 2022. Don't stop fine-tuning: On training regimes for few-shot cross-lingual transfer with multilingual language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10725–10742.

Fabian David Schmidt, Ivan Vulić, and Goran Glavaš. 2023a. Free lunch: Robust cross-lingual transfer via model checkpoint averaging. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5712–5730, Toronto, Canada. Association for Computational Linguistics.

Fabian David Schmidt, Ivan Vulić, and Goran Glavaš. 2023b. One for all & all for one: Bypassing hyperparameter tuning with model averaging for cross-lingual transfer. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Lütfi Kerem Senel, Benedikt Ebing, Konul Baghirova, Hinrich Schuetze, and Goran Glavaš. 2024. Kardeş-NLU: Transfer to low-resource languages with big brother's help – a benchmark and evaluation for Turkic languages. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1672–1688, St. Julian's, Malta. Association for Computational Linguistics.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. 2023. Language models are multilingual chain-of-thought reasoners. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Luka Terčon and Nikola Ljubešić. 2023. CLASSLA-Stanza: The Next Step for Linguistic Processing of South Slavic Languages.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Ahmet Üstün, Viraat Aryabumi, Zheng-Xin Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, et al. 2024. Aya model: An instruction finetuned open-access multilingual language model. *arXiv preprint arXiv:2402.07827*.

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*.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mt5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498.

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 Overview of Models

The models in evaluation (see Section 3) along with the links to access them are available in Table 5.

B Overview of Prompts

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:

Model	Link
gpt-3.5-turbo-0125	–
gpt-4-0125-preview	–
Mistral-7B-Instruct-v0.1	https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.1
Mistral-7B-Instruct-v0.2	https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.2
Mixtral-8x7B-Instruct-v0.1	https://huggingface.co/mistralai/Mixtral-8x7B-Instruct-v0.1
mT0-xxl	https://huggingface.co/bigscience/mt0-xxl
aya-101	https://huggingface.co/CohereForAI/aya-101
gemma-7b-it	https://huggingface.co/google/gemma-7b-it
falcon-7b-instruct	https://huggingface.co/tiiuae/falcon-7b-instruct
llama-2-7b-chat	https://huggingface.co/meta-llama/Llama-2-7b-chat-hf

Table 5: Models in evaluation along with their huggingface.co links.

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:

List prompt The ten-shot prompt, but omitting the structure of the task in the examples, and rather giving just samples of the language the task will be in.

You will be given a task. The task definition is in English, but the task itself is in another language. Here are some samples of the language the task is in:
Sample 1:
"Muški je otpra špino."
"Školjka ot zahoda se je napunila z oduon."
"Oda je počela teć z mlaznici."
Sample 2:
"Mlada je našla neko blago va žitaricah."
"Nalila je mlieko va škudelico."
"Je zgubila tiek."
Sample 3:
...
Sample 10:
"Šlovek je čuda popi na fešte."
"Ta drugi dan ga je bolela glava."
"Ta drugi dan mu je kapa nuos."
Now to your task! 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".
Premise: "Moje tielo je hitalo hlat na travo."
Question: "cause"
Hypothesis 1: "Sunce je hodilo van."
Hypothesis 2: "Trava je bila pokošena."
Answer:

Task prompt The ten-shot prompt, but without an answer provided. 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 without a solution:
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."
Example 2:
Premise: "Mlada je našla neko blago va žitaricah."

Question: "effect"
Hypothesis 1: "Nalila je mlieko va škudelico."
Hypothesis 2: "Je zgubila tiek."
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."
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:

Incorporating Dialect Understanding Into LLM Using RAG and Prompt Engineering Techniques for Causal Commonsense Reasoning

Benedikt Perak
University of Rijeka,
Faculty of Humanities
and Social Sciences,
AIRI Center
bperak@uniri.hr

Slobodan Beliga
University of Rijeka,
Faculty of Informatics and
Digital Technologies,
AIRI Center
sbeliga@uniri.hr

Ana Meštrović
University of Rijeka,
Faculty of Informatics and
Digital Technologies,
AIRI Center
amestrovic@uniri.hr

Abstract

The choice of plausible alternatives (COPA) task requires selecting the most plausible outcome from two choices based on understanding the causal relationships presented in a given text. This paper outlines several approaches and model adaptation strategies to the VarDial 2024 DIALECT-COPA shared task, focusing on causal commonsense reasoning in South-Slavic dialects. We utilize and evaluate the GPT-4 model in combination with various prompts engineering and the Retrieval-Augmented Generation (RAG) technique. Initially, we test and compare the performance of GPT-4 with simple and advanced prompts on the COPA task across three dialects: Cerkno, Chakavian and Torlak. Next, we enhance prompts using the RAG technique specifically for the Chakavian and Cerkno dialect. This involves creating an extended Chakavian-English and Cerkno-Slovene lexical dictionary and integrating it into the prompts. Our findings indicate that the most complex approach, which combines an advanced prompt with an injected dictionary, yields the highest performance on the DIALECT-COPA task.

1 Introduction

The choice of plausible alternatives task, introduced by Roemmele et al. in 2011 (Roemmele et al., 2011), presents a scenario where the model is required to comprehend cause-effect relationships from provided input and select the most plausible outcome from two choices. This task encapsulates various linguistic aspects, including lexical semantics, syntactic structure, and world knowledge, making it a challenging assessment for NLP systems (Ponti et al., 2020). It is an important benchmark in the domain of NLP, specifically designed to evaluate models' abilities in causal reasoning and inference making.

In recent years, the advancement of neural network architectures has led to the creation of a wide range of neural text representation models (Babić et al., 2020). Particularly transformer-based models like BERT (Devlin et al., 2018) and GPT (Radford et al., 2019), has significantly impacted the landscape of NLP tasks. Models pre-trained on extensive text corpora are commonly referred to as Large Language Models (LLMs). Through this process, they capture linguistic patterns and semantic connections, enabling them to perform better than traditional language models in a range of NLP tasks. Thus, LLMs have demonstrated remarkable performance across tasks such as text classification (Sun et al., 2019; Balkus and Yan, 2022), sentiment analysis (Babić et al., 2021; Beliga et al., 2021), paraphrase detection (Vrbanec and Meštrović, 2023), information spreading prediction (Meštrović et al., 2022), machine translation (Zhu et al., 2020; Yang et al., 2020), metaphor generation (Tong et al., 2024), question answering (Wang et al., 2019), etc.

Evaluating the performance of LLMs in tasks like COPA remains important to assess their true capabilities in causal reasoning inference. Furthermore, while LLMs exhibit strong performance in languages with abundant resources, such as English, their effectiveness decreases for small(er) dialects of non-major and low-resource languages (Kantharuban et al., 2023). Thus, it is important to explore novel approaches and model adaptation strategies that may enhance LLMs' abilities to comprehend dialects.

This paper aims to study language understanding in micro-dialects for moderately-resourced South-Slavic languages. Precisely, we experiment with the COPA task in Cerkno dialect (Slovenia), Chakavian dialect (Croatia) and Torlak dialect (Serbia). In addition, we evaluated some of our approaches to COPA task on datasets in several languages: English, Slovenian, Croatian, Serbian, and Macedonian language.

In this work, our focus is on examining and dif-

Proceedings of the Eleventh Workshop on NLP for Similar Languages, Varieties, and Dialects (VarDial 2024), pages 220–229
June 20, 2024 ©2024 Association for Computational Linguistics

ferentiating various methods specifically designed for the DIALECT-COPA task, with a particular emphasis on causal commonsense reasoning. The essence of this task lies in its ability to test models on understanding causal relationships within given scenarios, a crucial component of commonsense reasoning. For instance, consider the premise: "The girl found a bug in her cereal." In this case, the DIALECT-COPA task presents two possible effects: 1) "She poured milk in the bowl," and 2) "She lost her appetite." The challenge for the model is to deduce the most plausible effect of the initial event. While the first option is a neutral action that could occur in any context of preparing cereal, the second option directly relates to the discovery of a bug in the cereal, which would naturally lead to a loss of appetite. This task not only assesses the model's ability to infer logical consequences from specific events but also its capacity to navigate and understand nuanced human reactions, thereby evaluating its grasp of causal commonsense reasoning within varied contexts.

We consider and examine four distinct model adaptation strategies of utilising GPT-4 in the DIALECT-COPA task as follows.

1. Simple Prompt Engineering for GPT-4: This initial method involves straightforward simple prompt engineering with GPT-4, employing an iterative, 0-shot framework. It serves as our baseline, testing the model's innate ability to understand and reason about the given dialectal inputs without prior examples.

2. Advanced Prompt Engineering for GPT-4: Building on the first approach, we introduce advanced prompt engineering techniques that incorporate explicit reasoning instructions. This method enhances the model's capacity for logical deduction and causal inference, aimed at improving its performance on the task.

3. Simple Prompt Engineering with Retrieval-Augmented Generation (RAG): The third strategy extends the simple prompt engineering approach by integrating the RAG technique. This implementation includes an expanded lexical database, featuring 11,000 Chakavian and 4,000 Cerkno lexical entries, to facilitate deeper understanding and generation capabilities in these dialects.

4. Advanced Prompt Engineering with RAG for GPT-4: Our most sophisticated approach combines advanced prompt engineering, reasoning instructions, and RAG (see Fig. 1). This comprehensive method leverages the expanded dictionaries — comprising 11,000 Chakavian and 4,000 Cerkno lexical items — and integrates reasoning instructions to optimize the model's performance on the DIALECT-COPA task by enhancing its reasoning capabilities and dialect understanding.

Augmenting GPT models with dialect dictionaries represents a strategic enhancement aimed at enhancing LLMs' efficacy in dialect-rich linguistic environments. This augmentation approach, by integrating dialect-specific lexical resources into the model's pre-training regimen, is designed to elevate the model's comprehension and operational performance across varied linguistic landscapes. Such a strategy not only promises improvements in understanding diverse dialects but also champions the cause of inclusivity and accuracy in natural language processing applications.

To complement this augmentation strategy, we incorporate a novel self-referral grading task. This mechanism serves as an internal feedback loop, enabling the model to assess its own performance and adapt more effectively to the nuances of different dialects. This framework showcases a pioneering blend of dialect augmentation and self-evaluation capabilities, setting new approach for adaptability and accuracy in processing dialectical variations within LLMs.

2 Experiment Setup

2.1 DIALECT-COPA Shared Task

The DIALECT-COPA task, a component of the VarDial 2024 shared tasks, presents a unique challenge in understanding causal relationships within scenarios expressed in various South-Slavic dialects (Ljubešić et al., 2024; Chifu et al., 2024). The task involves selecting the most plausible outcome from two options provided, based on the comprehension of causal connections embedded within the context of these dialectal variations. This task aims to explore and evaluate models' capabilities in reasoning across different South-Slavic languages and dialects, emphasizing the nuanced linguistic differences and their impact on reasoning abilities.

Within the task, participants are presented with datasets containing sentences written in South-Slavic dialects, along with adequate annotations of sentences designed to probe the understanding of causal relationships within these linguistic contexts. The datasets encompass a diverse range of dialects, namely Cerkno, Chakavian, and Torlak. This broad representation of dialects enables a comprehensive

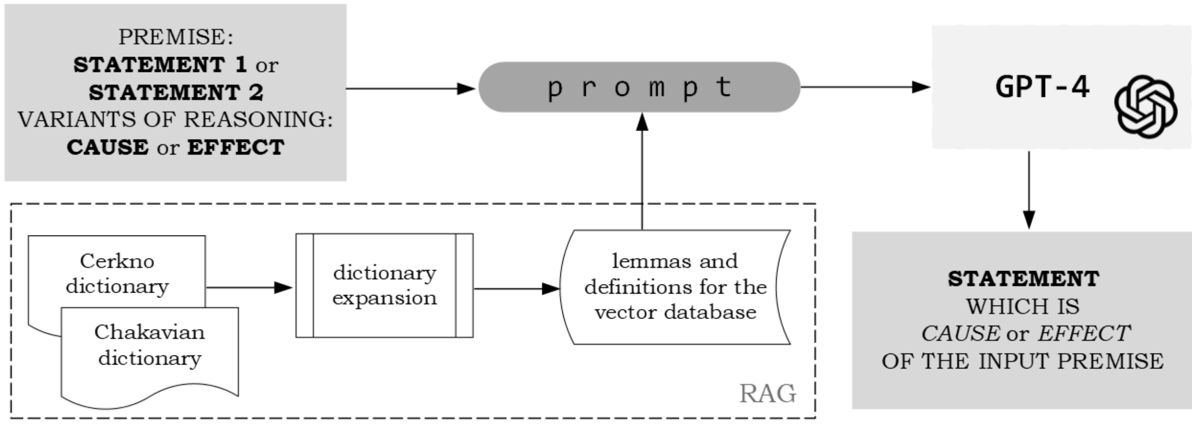


Figure 1: Framework whick inlude advanced prompt engineering and Retrieval Augmented Generation using external data incorporated in dialect dictionaries.

examination of how language variations influence the interpretation and identification of causal connections within the given scenarios. In addition, the organizers have provided several datasets in various languages, such as English, Slovenian, Croatian, Serbian, and Macedonian, annotated specifically for the COPA task. These datasets can be utilized for training and validating the proposed models, enhancing the robustness and versatility of the solutions developed.

To address the complexities of reasoning within dialectal contexts, participants are required to analyze the provided sentences and discern the underlying causal relationships. This task challenges participants to navigate the linguistic intricacies inherent in South-Slavic dialects, such as variations in vocabulary, syntax, and grammatical structures, which may impact the interpretation of causal connections. By engaging with dialectal variations, participants have the opportunity to explore the intersection of language diversity and causal reasoning, thereby contributing to a deeper understanding of how linguistic differences shape cognitive processes.

Taking the example: *The girl found a bug in her cereal.* is translated in Croatian (Hr): *Djevojka je pronašla kukca u žitaricama.*, in Croatian dialect Chakavian (Hr-Ckm) as: *Mlada je našla neko blago va žitaricah.* with options: 1) *Nalila je mlijeko u zdjelicu.* (Hr) and *Nalila je mlieko va škudelico.* (Hr-Ckm) (She poured milk in the bowl.), or 2) *Izgubila je apetit.* (Hr) *Je zgubila tiek.* (Hr-Ckm) (She lost her appetite.) In this analysis, the primary challenge revolves around the semantic ambiguity inherent in language translation, exemplified by the Croatian noun *kukac* and its Chakavian dialect translation *blago*. The term "blago" in Standard Croatian predominantly connotes "treasure" and, to a lesser extent, "cattle". It is notably perplexing to encounter its use in the Chakavian dialect where it assumes the meaning "bug".

This example highlights the broader linguistic issue of polysemy—where a single word has multiple meanings depending on context—which poses significant challenges for accurate translation and understanding. Such ambiguities are especially pronounced in dialects and regional languages, which may not be sufficiently documented, thereby complicating the task of linguistic models like Large Language Models (LLMs) in processing and interpreting such data accurately. For instance, with no data implying the meaning "bug" the reasoning is guided in the direction of misleading translation *The girl found a treasure in her cereal.* with no meaningful connotation to the either of two causal options, which often leads to phenomenon referred as AI semantic hallucination. This situation underscores the necessity for enhanced data inclusivity and sophisticated contextual analysis capabilities in language processing technologies.

2.2 Datsets

We use several available datasets annotated for the COPA task: (i) COPA-HR for Croatian language (Ljubešić, 2021), (ii) COPA-SR for Serbian language (Ljubešić et al., 2022b) and (iii) COPA-MK for Macedonian language (Ljubešić et al., 2022a). In addition, we also used data sets for the corresponding dialects, Chakavian from Croatian, Cerkno from Slovenian and Torlak from Serbian, as well as a data set for the English language.

All datasets use the Latin alphabet, but for the Macedonian and Serbian languages, as well as for the Torlak dialect, we used a dataset written in Cyrillic. The statistics of the distribution of examples into sets for training, validation and testing are shown in the Table 1.

Table 1: Datasets distribution for languages and South-Slavic dialects.

Lang./Dialect	Train	Val.	Test
English	400	100	-
Slovenian	400	100	-
Croatian	400	100	-
Serbian	400	100	-
Macedonian	400	100	-
Cerkno Dialect	400	100	500
Chakavian Dialect	400	100	500
Torlak Dialect	400	100	500

2.3 Prompt Engineering

Prompt engineering has emerged as a valuable technique for improving the effectiveness of LLMs (Reynolds and McDonell, 2021). This method leverages task-specific instructions to expand model capabilities without altering the core model parameters.

In this research, we explore various strategies to formulate prompts to Generative Pre-trained Transformer 4 (GPT-4) (OpenAI et al., 2023) for the COPA task. In the first step, we experiment with the 0-shot learning approach defining a simple prompt and iterating it across all examples in the datasets, for all languages. Here is the prompt used for generating responses in the simple version.

```
This is a reasoning task
with statement in {language} language.
Given the PREMISE = {premise}
please determine {question} of the PREMISE,
choosing between two statements
offered in choices:
CHOICE1 = {choice1}
or
CHOICE2 = {choice2}
Format response in JSON format
with keys:
index:{index}, label: integer(range: 0-1).
```

Next, we pursue the 0-shot learning approach further by providing more comprehensive instructions that incorporates reasoning on the causal relationships within the given sentences. We extend the previous prompt (simple prompt) to advanced prompt with a following instruction:

```
This is statement in {language}.
Based on the PREMISE = {premise}
please determine what could be
the plausible {"cause"|"effect"}
of the PREMISE,
choosing between two choices:
CHOICE1 = {choice1}
or
CHOICE2 = {choice2}
Reason about this problem step by step.
```

Again, we iterate the advanced prompt across all examples in the datasets, for all languages.

In this approach we were able to analyze the reasoning process in the LLM, as well as the elements guiding the erroneous decision in the system. For instance, in the example with the girl and the bug in her cereal, the reasoning in English is correct:

"Finding a bug in one's cereal is generally an unpleasant and unexpected experience that could lead to a loss of appetite. The presence of the bug would likely overshadow the normal routine of adding milk to the cereal, as the discovery would be off-putting and could cause disgust. Therefore, it is more plausible that the effect of finding a bug in her cereal would be that she lost her appetite rather than her continuing with the meal preparation by pouring milk into the bowl."

However, the Chakavian reasoning explanation is plagued with wrong translation of nouns *bug* (Chak. blago) and *appetite* (Chak. tiek). The GPT4 reasoning is as follows: "The premise 'Mlada je našla neko blago va žitaricah' translates to 'The young woman found some treasure in the cereals'. Neither of the choices directly relate to finding treasure in cereals. However, CHOICE_0 'Nalila je mlieko va škudelico' translates to 'She poured milk into a bowl', which could be indirectly related to cereals as one might pour milk into a bowl of cereal. CHOICE_1 'Je zgubila tiek' does not seem to be related to the premise as it translates to 'She lost a thread', which has no apparent connection to finding treasure or cereals. Therefore, CHOICE_0 is selected as it has a weak indirect connection to the premise.".

Consequently, without a comprehensive resolution of the lexical concepts embedded within the sentence, the effectiveness of the reasoning prompt in facilitating accurate comprehension and analysis is significantly diminished. This highlights the critical need for precise semantic interpretation to ensure that cognitive processing mechanisms can effectively engage with and extrapolate meaningful insights from the textual content presented.

2.4 Retrieval-Augmented Generation

In the next step, we propose an approach that combines prompts using GPT-4 and RAG techniques for dialect processing. RAG is a general-purpose fine-tuning technique which combines pre-trained parametric and non-parametric memory for language generation (Lewis et al., 2020). This approach has shown promise through the integration of knowledge from external databases, resulting in improved accuracy and credibility of the models, especially for tasks requiring substantial domain knowledge (Gao et al., 2023). It also facilitates ongoing knowledge updates and the incorporation of domain-specific information.

Within the framework of the DIALECT-COPA task, we utilized an external knowledge base, specifically, a dictionary of the Chakavian dialect and a dictionary of the Cerkno dialect (we could not find a freely available version of the Torlak dialect dictionary). By including dialect dictionaries as external knowledge into the model we enhance our model's understanding of this linguistic variety.

The cornerstone of the Chakavian dictionary dictionary is the work of Cvjetana Miletić (Miletić, 2019), which catalogues approximately 10,800 words, predominantly from the Kastav region in Croatia. In addition, an online version of the Cerkno dictionary called Ana Mičken's zbíerka crklajnskih besít[1], was used as a source of external knowledge for the Cerkno dialect. This dictionary contains about 4000 lemmas with corresponding descriptions and some linguistic examples of usage.

In dialectal dictionaries, in addition to lemmas and corresponding definitions, we also found determinants about the meaning and grammatical category of words. However, we expanded the dictionaries in order to offer as much external knowledge as possible to the model. Thus, we expanded the dictionaries to have the following determinants: lemmas in original dialectal form, lemmas translated into English, examples of use, examples of use translated into English, definitions, expanded definitions and suggestions of use offered by GPT.

We thus proceeded with broadening definitions and examples of usage within the Chakavian and Cerkno contexts, as well as facilitating automatic translations into English. The prompt used for expanding the dictionary examples was structured as follows:

```
Given a dictionary entry "{line}",
expand the definition and provide usage
examples in {language_dict},
with the source lexeme in {language_source}.
Format the response in JSON, including:
'definition': A string containing
the expanded definition,
'GPT_suggestion_of_use':
Two examples of usage in {language_source},
'source_nd_translated':
The source lexeme translated into English,
'example_of_use_translated':
Usage examples translated into English.
```

This methodology was designed to enrich the dictionary and augment GPT-4's ability to comprehend and utilize terms and expressions unique to the Chakavian dialect.

Dictionaries containing linguistic data from the Croatian Chakavian and Slovenian Cerkno dialect were processed using a Retrieval-Augmented Generation (RAG) approach. The process began by transforming data into a CSV format for ease of manipulation and analysis. Following data importation, specialized Recursive Character Text Splitter tool in Langchain library was employed to segment the text into manageable parts, enhancing the handling and vectorization of the data, using the chunk size 200 characters with 50 characters overlap.

The segmented data was then vectorized using an embedding function that facilitated the creation of a persistent vector store in a Chroma database. This vector store serves as a retriever, enabling the efficient retrieval of vectorized text segments based on their semantic content. This setup was designed for integrating the enriched dictionaries with advanced language models, thereby allowing for more contextually aware processing and generation of text based on the Chakavian and Cerkno dialect, respectively.

Enriched dictionaries were used as context in conjunction with the first two previously described prompt strategies (simple prompt engineering strategy and advanced prompt engineering strategy) by inserting the whole enriched dictionary into the prompt.

```
This is a reasoning task
```

[1] https://www2.arnes.si/ supmrazp/zbirkacb.htm

```
with statement in {language} language.
Using aditional knowledge
about the {language} language
from the provided dictionary in the
{context}, answer the following
Question:{question}
```

By introducing retrieval augmentation approach, we further evaluated the model's performance on the DIALECT-COPA task.

In addition to the augmentation approach, we introduced a self-referential grading aspect into the prompt. This mechanism establishes an internal feedback loop, empowering the model to autonomously evaluate its own performance. By integrating this self-referral aspect, the proposed approach provides a deeper insight into the reasoning procedures within the DIALECT-COPA task. For this purpose, we define a specific prompt in which we instruct the model to translate the premise and choices into the English language using the dialect dictionary. And then we instruct the model to provide a reasoning for the choice it made. The prompt is formulated as follows:

```
Translate the premise and choices using
the dictionary knowledge in the {context},
especially if the connection of the
premises and choices is apparently weak,
or you do not understand the phrase or
word and the choices.
Provide a reasoning for the choice
you made.
Assess your certainty in a range 0-1.
Format response in JSON format with
keys: index:{index},
label: integer(range: 0-1),
reasoning: str,
certainty: float (range:0-1).
```

The inclusion of a self-referential grading mechanism called *certainty* within the prompt is using the model's autonomous evaluative capabilities with a potential to elevate its ability to introspect and rationalize decision-making processes. Although we did not proceed with the feedback mechanisms based on these values, this approach can be further developed as a framework for the development of more self-aware and adaptive language processing systems in future linguistic research.

3 Results

The initial measurement for English, Macedonian, Serbian, Slovenian, and Croatian was performed on the training data[2] to serve as an orientation and provide an initial insight into the success of the GPT-4 0-shot approach. The performance of the dialects (Chakavian, Cerkno and Torlak) was also measured on available training datasets. Thus, the baseline performance of the GPT-4 model was obtained for all languages in their standard variant, as well as the baseline values for 3 dialects.

The left part of Table 2 shows the model accuracy results for the GPT-4 0-shot with simple prompt (left) and GPT-4 0-shot with advanced prompt framework (right). The results are shown in 3 columns in both cases. The first column contains the measured values for the classification task where the question is *cause*, the second column shows the accuracy when the question is *effect*, and the third column presents a question-independent accuracy (i.e. accuracy on average).

As expected, in the 0-shot - simple prompt variant, but also in the 0-shot - advanced prompt variant demanding iterative reasoning, the model for the English language achieves the highest accuracy (96% and 98.3% on average, respectively). This is expected given the large amount of English language data available to train the model in contrast to the availability of data in other languages. Unexpected, the success of the model in Slovenian and English is identical in the 0-shot - simple prompt variant.

If we consider the 0-shot - simple prompt cases in more detail (left part of the Table 2), two observed facts are interesting. First, the other languages are not far behind the results for English and Slovenian. English, and Slovenian reaches an accuracy on average 96%, while all other languages are below that, but none are below 91%. They are behind by a small number of percentage points. Second, the differences in results with respect to the examination of *cause* or *effect* do not oscillate drastically. The differences are up to 4 percentage points in all cases, except for the Croatian and Serbian, where the deviation is approximately 6 or 8%. Third, also expected, the results for dialects are significantly worse than the results of the same languages in standard varieties. 4% worse for Serbian, 18.3% for Croatian, and even 31.7% for the case of Slovenian.

Unexpectedly, slightly worse results are achieved in the case where the baseline for GPT-4

[2] training and validation data are available at https://github.com/clarinsi/dialect-copa/?tab=readme-ov-file

Language	Dialect	Alphabet	GPT-4 0-shot - simple			GPT-4 0-shot - advanced			$\overline{\Delta}$
			Cause	Effect	Avg	Cause	Effect	Avg	
ENG	-	Latin	**.98**	**.941**	**.96**	**.985**	**.98**	**.983**	**+.0225**
MCD	-	Cyrillic	.934	.896	.915	.894	.931	.913	-.0025
SRB	-	Cyrillic	.965	.881	.923	.939	.936	.938	**+.015**
SRB	Torlak	Cyrillic	.894	.871	.883	.828	.881	.855	-.0275
SLO	-	Latin	**.98**	**.941**	**.96**	.904	.921	.913	-.0475
SLO	Cerkno	Latin	.621	.663	.643	.581	.604	.593	-.05
CRO	-	Latin	.97	.916	.943	.965	.901	.933	-.01
CRO	Chakavian	Latin	.748	.772	.76	.677	.629	.653	-.1075

Table 2: Achieved results in terms of accuracy for GPT-4 0-shot - simple promt (left) and GPT-4 0-shot - advanced prompt (right) in terms of accuracy representing the baselines for English, Macedonian, Serbian, Slovenian and Croatian as well as Torlak, Cerkno and Chakavian dialects. The delta column ($\overline{\Delta}$) indicates the average change in accuracy, and refers to the average difference in accuracy achieved by the andvanced and simple prompts.

0-shot are prompted with more sophisticated prompt (right part of the Table 2) against plain 0-shot. This is the case for Macedonian (-0.2%), Croatian (-1%), and Slovenian (-4.7%). However, the results are favorable for Serbian (+1.5% improvement) and English (+2.3% improvement) when sophisticated reasoning is prompted. Although small, a positive change in the performance of the advanced prompt relative to the simple prompt is evident for English (+.0225) and Serbian (+.015) in the last column of the Table 2, where $\overline{\Delta}$ indicates the average accuracy difference between the advanced and simple prompts. A positive value indicates the advantage of advanced prompts over simple prompts.

The preliminary results and findings from this study provide initial evidence that further investigation into sophisticated causal reasoning is warranted, given its modest improvements in accuracy. This suggests potential benefits that could enhance understanding and application in related fields. Moreover, it gave us a strong indication that the model needs an external source of dialect knowledge and the need to apply the RAG paradigm to augment the model's ability to more efficiently handle a dialect it was not initially trained for.

Therefore, in Table 3 experimental results based on advanced prompt engineering which includes iterative reasoning and the RAG approach were presented only for dialects. Test data prepared for the DIALECT-COPA shared task were used (Chifu et al., 2024; Ljubešić et al., 2024) .

Results in terms of accuracy for upgrading the basic 0-shot technique with an iterative reasoning prompt (GPT-4 0-shot-advanced) did not yield fruitful results. The success of such an approach for all dialects recorded a negative effect, namely -5.6% for Cerkno, -8.8% for Torlak and -11% for the Chakavian dialect.

Approach	Cer.	Cha.	Tor.
`GPT-4 0-shot-simple`	.664	**.774**	**.894**
`GPT-4 0-shot-advanc.`	.608	.664	.806
`GPT-4-augm-simple`	.688	.76	-
`GPT-4-augm-advanc.`	**.708**	.764	-

Table 3: Results in terms of accuracy for 4 different approaches: (1) GPT-4 0-shot - simple promt, (2) GPT-4 0-shot - advanced prompt, (3) GPT-4 augmented with external knowledge for Cerkno and Chakavian dialect, and (4) GPT-4 augmented with external knowledge for Cerkno and Chakavian dialect and advanced prompt engineering.

Augmentation of the model with external knowledge (GPT-4-augm-simple approach) from the dialect dictionaries for Chakavian and Cerkno managed to improve the accuracy of Cerkno to 68.8%, which is 2.4% better than the initial 0-shot. However, this was not the case for the Chakavian dialect. One possible reason for this outcome is the incompleteness of the dictionary and lack of polysemous definitions. For example, there is no relation of the Chakavian word *blago* to suggested English equivalent *bug*. Therefore, one way to improve this process would require a better dictionary resources, or even a different way of representing the polysemous nature of lexical networks, such as using lexical graphs (Ban Kirigin et al., 2022; Perak and

Kirigin, 2023).

Finally, with RAG and prompt engineering that requires sophisticated reasoning (GPT-4-augm-advanced), progress is made in both dialects. With such an approach, Chakavian still lags behind the basic 0-shot approach, but Cerkno reaches 70.8%, which is 4.4% more successful than the initial 0-shot.

Experimental findings suggest that the advanced prompt in our study tends to underperform when dealing with languages and dialects that are less familiar to the GPT-4 model. This observation is further substantiated by the statistical measure of Pearson's correlation, which explores the relationship between the simple prompt and the delta (Δ)[3]. The correlations for all languages are negative. The most pronounced values are observed for English with statistically significant correlations ($r = -.7655, p = 3.1040 \times 10^{-78}, \alpha = .05$) and Serbian ($r = -.6202, p = 6.9432 \times 10^{-44}, \alpha = .05$), which is expected given their positive average delta values (refer to the rightmost column in Table 2). Correlation values are also negative for all other languages, but significantly lower, and extremely low for dialects (although for Chakavian and Cerkno are not statistically significant).

4 Conclusion

In this paper, we propose and compare several model adaptation strategies for DIALECT-COPA task. We combine prompt engineering and RAG techniques to enhance the dialect understanding of GPT-4 model.

Research indicated that sophisticated causal reasoning has slight advantages in accuracy, suggesting the need for further investigation. The model requires external dialect knowledge and the application of the RAG paradigm for more efficient handling of dialects for which it was not initially trained. The enhancement of the basic 0-shot method with iterative reasoning did not produce significant results, negatively impacting Cerkno and Chakavian dialects. The model's augmentation with external knowledge from Chakavian and Cerkno dialect dictionaries increased Cerkno's accuracy to 68.8%, a 2.4% improvement over the initial 0-shot. Lastly, employing RAG and prompt engineering that demands complex reasoning led to improvements in both dialects. Although Chakavian still trails the basic 0-shot method, Cerkno achieves an accuracy of 70.8%, marking a 4.4% improvement over the initial 0-shot.

The top achievements of our UNIRI team, as detailed in this paper, hold the second position overall when oposed with the outcomes reported by other teams in the DIALECT-COPA Task on Causal Commonsense Reasoning, a part of the VarDial Evaluation Campaign 2024. Specifically, for the Cerkno dialect, we achieved the second-best result using a simple prompt, with an accuracy of .774. In the case of the Chacavian dialect, we obtained the third-best result by combining an advanced prompt with reasoning and the RAG approach, achieving an accuracy of .708.

The main contribution of this paper is the proposed approach, which utilizes prompt engineering alongside the RAG technique. In this method, RAG facilitates enhancements by integrating Chakavian and Cerkno dictionaries into the prompt. During testing, we showcased that prompts augmented with RAG outperform those without RAG on the Cerkno test dataset. To the best of our knowledge, this represents the first attempt to integrate dialect dictionaries into RAG with the objective of addressing the COPA task focused on causal commonsense reasoning in South-Slavic dialects.

Future work will be concentrated on extending these methods to encompass other dialects and tasks akin to COPA. This includes exploring the adaptation of similar approaches to additional South-Slavic dialects and extending the application to a broader spectrum of tasks requiring nuanced linguistic understanding and reasoning abilities.

Acknowledgments

This work has been fully supported by the University of Rijeka projects uniri-mladi-drustv-23-33, uniri-drustv-23-95, uniri-class-1.

References

Karlo Babić, Sanda Martinčić-Ipšić, and Ana Meštrović. 2020. Survey of neural text representation models. *Information*, 11(11):511.

Karlo Babić, Milan Petrović, Slobodan Beliga, Sanda Martinčić-Ipšić, Mihaela Matešić, and Ana Meštrović. 2021. Characterisation of covid-19-related tweets in the croatian language: framework based on the cro-cov-csebert model. *Applied Sciences*, 11(21):10442.

[3]Delta (Δ) represents an absolute value that is calculated as the difference in accuracy achieved by a advanced and simple prompt.

Salvador V Balkus and Donghui Yan. 2022. Improving short text classification with augmented data using gpt-3. *Natural Language Engineering*, pages 1–30.

Tajana Ban Kirigin, Sanda Bujačić Babić, and Benedikt Perak. 2022. Graph-based taxonomic semantic class labeling. *Future Internet*, 14(12):383.

Slobodan Beliga, Sanda Martinčić-Ipšić, Mihaela Matešić, Irena Petrijevčanin Vuksanović, and Ana Meštrović. 2021. Infoveillance of the croatian online media during the covid-19 pandemic: one-year longitudinal study using natural language processing. *JMIR public health and surveillance*, 7(12):e31540.

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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.

Anjali Kantharuban, Ivan Vulić, and Anna Korhonen. 2023. Quantifying the dialect gap and its correlates across languages. *arXiv preprint arXiv:2310.15135*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Nikola Ljubešić. 2021. Choice of plausible alternatives dataset in croatian copa-hr.

Nikola Ljubešić, Mirjana Starović, Taja Kuzman, and Tanja Samardžić. 2022a. Choice of plausible alternatives dataset in serbian copa-mk.

Nikola Ljubešić, Mirjana Starović, Taja Kuzman, and Tanja Samardžić. 2022b. Choice of plausible alternatives dataset in serbian copa-sr.

Nikola Ljubešić, Nada Galant, Sonja Benčina, Jaka Čibej, Stefan Milosavljević, Peter Rupnik, and Taja Kuzman. 2024. DIALECT-COPA: Extending the standard translations of the COPA causal commonsense reasoning dataset to south slavic dialects. In *Eleventh Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2024)*, Mexico City, Mexico. Association for Computational Linguistics.

Ana Meštrović, Milan Petrović, and Slobodan Beliga. 2022. Retweet prediction based on heterogeneous data sources: the combination of text and multilayer network features. *Applied Sciences*, 12(21):11216.

Cvjetana Miletić. 2019. *Slovar kastavskog govora [Dictionary of the Kastav Speech]*. Udruga Čakavski senjali, Kastav.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2023. Gpt-4 technical report. Computer Science 2303.08774, arXiv.

Benedikt Perak and Tajana Ban Kirigin. 2023. Construction grammar conceptual network: Coordination-based graph method for semantic association analysis. *Natural Language Engineering*, 29(3):584–614.

Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. Xcopa: A multilingual dataset for causal commonsense reasoning. *arXiv preprint arXiv:2005.00333*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Laria Reynolds and Kyle McDonell. 2021. Prompt programming for large language models: Beyond the few-shot paradigm. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–7.

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*.

Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2019. How to fine-tune bert for text classification? In *Chinese computational linguistics: 18th China national conference, CCL 2019, Kunming, China, October 18–20, 2019, proceedings 18*, pages 194–206. Springer.

Xiaoyu Tong, Rochelle Choenni, Martha Lewis, and Ekaterina Shutova. 2024. Metaphor understanding challenge dataset for llms. *arXiv preprint arXiv:2403.11810*.

Tedo Vrbanec and Ana Meštrović. 2023. Comparison study of unsupervised paraphrase detection: Deep learning—the key for semantic similarity detection. *Expert systems*, 40(9):e13386.

Zhiguo Wang, Patrick Ng, Xiaofei Ma, Ramesh Nallapati, and Bing Xiang. 2019. Multi-passage bert: A globally normalized bert model for open-domain question answering. *arXiv preprint arXiv:1908.08167*.

Jiacheng Yang, Mingxuan Wang, Hao Zhou, Chengqi Zhao, Weinan Zhang, Yong Yu, and Lei Li. 2020. Towards making the most of bert in neural machine translation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 9378–9385.

Jinhua Zhu, Yingce Xia, Lijun Wu, Di He, Tao Qin, Wengang Zhou, Houqiang Li, and Tie-Yan Liu. 2020. Incorporating bert into neural machine translation. *arXiv preprint arXiv:2002.06823*.

One-shot Prompt for Language Variety Identification

Nat Gillin
nat.gillin@gmail.com

Abstract

We present a one-shot prompting approach to multi-class classification for similar language identification with an off-the-shelf pre-trained large language model that is not particularly trained or tuned for the language identification task. Without post-training or fine-tuning the model, we simply include one example per class when prompting the model and surprisingly the model is able to generate the language and locale labels accordingly.

1 Introduction

Recent works validated the idea of using language models generation performs well in classification task (Li et al., 2018a; Thant and Nwet, 2020; Hadar and Shmueli, 2021) and generation models can also perform competitively as zero-shot text-classifiers (Yin et al., 2019; Meng et al., 2022; Sun et al., 2023; Wang et al., 2023). Particular to language identification, Gillin (2022) have trained an encoder-decoder model for the French Cross-Domain Dialect Identification (FDI) dataset (Găman et al., 2023) for the VarDial 2022 shared task (Aepli et al., 2022a).[1]

Previously one might find it appealing to train or fine-tune a model to achieve state-of-the-art NLP models for specific tasks, but recent advancement in large language models and prompt-based solutions have made us think,

> *What if we just prompt a popular LLM and make it work like a classifier without tuning it?*

To test the idea of just prompting a pre-trained model for language identification, we evaluated the approach on the *English*, *French*, *Portuguese* and *Spanish* subset of the DSL Multi-label classification of similar languages (DSL-ML) shared task at VarDial 2024 (Chifu et al., 2024).[2] A few of example inputs and outputs of the DSL-ML test data are as follows:

`[IN]:` It took a lifetime, three trips to the moon and the downfall of communism to make it happen...

`[OUT]:` EN-GB,EN-US

`[IN]:` ...as an artist, there is no shortage of colour in my life.

`[OUT]:` EN-GB

`[IN]:` ...the annual pop culture event bringing colorful cosplayers, entertainment aficionados and comic book lovers together under one roof...

`[OUT]:` EN-US

The English varieties contains 3 classes, `EN-US`, `EN-GB` or both `EN-GB,EN-US`. The Portuguese and Spanish varieties also have 3 classes. Respectively, `PT-BR`, `PT-PT` and `PT-BR,PT-PT` for Portuguese from Brazil, Portugal or both and `ES-AR`, `ES-ES` and `ES-AR,ES-ES` for Spanish from Argentina, Spain or both.

For the French varieties, the single label classes comprises the Belgium, Canada, Switzerland and France, viz. `FR-BE`, `FR-CA`, `FR-CH` and `FR-FR`. And the combinations of multi-labels may come from either of the labels, e.g. `FR-CA,FR-CH,FR-FR` to represent texts that could be in classified as either Canadian, Swiss and French varieties. Also, we note that the input text from the French varieties subtask contains masked named-entities represented by the `$NE$` tokens.

[1] The general idea is to generate language labels as how a language model will generate the next token/word in natural text (Li et al., 2018b,c).

[2] https://sites.google.com/view/vardial-2022/shared-tasks

Proceedings of the Eleventh Workshop on NLP for Similar Languages, Varieties, and Dialects (VarDial 2024), pages 230–234
June 20, 2024 ©2024 Association for Computational Linguistics

2 TL;DR (Experimental Setup)

We use the Mistral instruct model with 7 billion parameters (Mistral-7B) (Jiang et al., 2023) for all our experiments .[3]

Off the shelf, we did not *post-train*, i.e. continue training the language model generation with raw monolingual texts, nor *fine-tune* the model with language identification datasets.

Without using the training data, we only selected one example per class from each language family from the development dataset provided by the DSL-ML shared task organizers. These examples were used as one-shot prompt and prepended to texts in the test sets.

For example, given an example from each class in the English development set from Section 1 and a input text from the test set:

> **`[IN]:`** Conducting an amateur orchestra and performing with it as a soloist are parts of the learning process for young professionals.

We process the above to put them in the format that the Mistral-7B model expects, e.g.

```
<s>[INST] It took a lifetime... [/INST]
EN-GB,EN-US</s>
<s>[INST] ...as an artist... [/INST]
EN-GB</s>
<s>[INST] ...the annual pop culture... [/INST]
EN-US</s>
[INST] Conducting an amateur orchestra... [/INST]
```

And we expected the model to generate `EN-US`, `EN-GB` or `EN-GB,EN-US` as a continuation to the examples and input sentence we entered. We will refer to this as *one-shot prompting* for the rest of the paper.[4] We repeated the *one-shot prompting* approach for the French, Portuguese and Spanish test sets (Zampieri et al., 2024, 2023; Găman et al., 2023; Bernier-colborne et al., 2023).

2.1 One-shot Prompting with Instructions

Additionally, for the English variety test set (Tan et al., 2014a), we experimented with a instruction prompt where we prepend the following instructions before the examples and the test instance, aka. *instructed one-shot prompting*.

```
Label the following text as (i) EN-US if it's
in United States English or (ii) EN-GB if it's
in United Kingdom English or (iii) EN-US,EN-GB
if it can be both in United States or
United Kingdom. <s>[INST] ...[/INST]... </s>...
[INST] Conducting an amateur orchestra... [/INST]
```

3 Results

Lang	Train	Dev	Test
EN	75.1	74.8	74.5
ES	21.2	20.6	21.3
PT	20.0	20.6	18.5
FR	-	15.6	12.9

Table 1: Weighted Averaged F1 Score of One-shot Prompting

Table 1 presents the weighted F1 scores of the one-shot prompting without instructions. In addition to the test set scores, we report the performance of the results of classifying the training (*Train*) and development (*Dev*) of the one-shot prompting approach.

We note that these numbers for the test set F1 scores differ from the ones reported in the official shared task findings papers (Chifu et al., 2024) since we didn't do any special label processing to compute partial matches for multi-class true labels before computing the weighted F1-score with `sklearn`.[5]

Split	One-shot	Prompt-shot
Train	75.1	69.9
Dev	74.8	68.7
Test	74.5	74.8

Table 2: Results of English Variety Classification between One-shot Prompting without (`One-shot`) vs with Instructions (`Prompt-shot`)

Table 2 reports the results of the English variety classification with and without the pre-example instruction prompt as described in Section 2.1. The one-shot prompts with instructions consistently performs worse on the training and development sets as compared to the one-shot prompting without instructions. However, one-shot prompting performs almost equally on F1-scores on the test sets with or without instructions.

[3] https://huggingface.co/mistralai/Mistral-7B-v0.1

[4] We acknowledge that the terminology of "*-shot" has not been defined formally in previous literature, e.g. https://datascience.stackexchange.com/q/120637/122. In this case, we refer to *one-shot* as giving the model one example per class as context before requiring it to infer the label given the test instance.

[5] https://scikit-learn.org/stable/modules/generated/sklearn.metrics.precision_recall_fscore_support.html

4 Related Works

As language coverage of identification systems increases (Jauhiainen et al., 2019; Agarwal et al., 2023; Burchell et al., 2023), language identification between similar languages, dialects and national varieties remains an active and challenging task in NLP (Tiedemann and Ljubešić, 2012; Gaman et al., 2020; Bouamor et al., 2019; Chakravarthi et al., 2021; Aepli et al., 2022b, 2023).

Early studies on language varieties classification created annotations through proxy signals such using the top-level domain of the text source's website as the locale label (Tan et al., 2014b). However, datasets with locale labels created through proxy signals are often unreliable since there might be no linguistics marker that distinguish one language variety to another language variety (Zampieri et al., 2014; Ács et al., 2015; Goutte et al., 2016).

Zampieri et al. (2023) and Bernier-colborne et al. (2023) redefined the language variety identification task as a multi-label task instead of assigning only a single language variety to each text.

5 Conclusion

By prompting the Mistral-7B model, which was not particularly known to be trained on language identification, we were able to make it classify language varieties to some extent. However, like many large language models, it is largely English-centric and we observed that the English variety classification performance far exceeds the French, Portuguese or Spanish varieties classification task. While a language model '*open source*' its model parameters, the lack of transparency in what goes into training the model makes its usage a grey-box probing exercise.[6]

[6] https://gist.github.com/alvations/af68bd50d4e59d4e74f3632d9ce44e7c (Tan, 2023)

References

Judit Ács, László Grad-Gyenge, and Thiago Bruno Rodrigues de Rezende Oliveira. 2015. A two-level classifier for discriminating similar languages. In *Proceedings of the Joint Workshop on Language Technology for Closely Related Languages, Varieties and Dialects*, pages 73–77, Hissar, Bulgaria. Association for Computational Linguistics.

Noëmi Aepli, Antonios Anastasopoulos, Adrian Chifu, William Domingues, Fahim Faisal, Mihaela Găman, Radu Tudor Ionescu, and Yves Scherrer. 2022a. Findings of the VarDial evaluation campaign 2022. In *Proceedings of the Ninth Workshop on NLP for Similar Languages, Varieties and Dialects*, Gyeongju, Republic of Korea. International Committee on Computational Linguistics (ICCL).

Noëmi Aepli, Antonios Anastasopoulos, Adrian-Gabriel Chifu, William Domingues, Fahim Faisal, Mihaela Gaman, Radu Tudor Ionescu, and Yves Scherrer. 2022b. Findings of the VarDial evaluation campaign 2022. In *Proceedings of the Ninth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 1–13, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Noëmi Aepli, Çağrı Çöltekin, Rob Van Der Goot, Tommi Jauhiainen, Mourhaf Kazzaz, Nikola Ljubešić, Kai North, Barbara Plank, Yves Scherrer, and Marcos Zampieri. 2023. Findings of the VarDial evaluation campaign 2023. In *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*, pages 251–261, Dubrovnik, Croatia. Association for Computational Linguistics.

Milind Agarwal, Md Mahfuz Ibn Alam, and Antonios Anastasopoulos. 2023. LIMIT: Language identification, misidentification, and translation using hierarchical models in 350+ languages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14496–14519, Singapore. Association for Computational Linguistics.

Gabriel Bernier-colborne, Cyril Goutte, and Serge Leger. 2023. Dialect and variant identification as a multi-label classification task: A proposal based on near-duplicate analysis. In *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*, pages 142–151, Dubrovnik, Croatia. Association for Computational Linguistics.

Houda Bouamor, Sabit Hassan, and Nizar Habash. 2019. The MADAR shared task on Arabic fine-grained dialect identification. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 199–207, Florence, Italy. Association for Computational Linguistics.

Laurie Burchell, Alexandra Birch, Nikolay Bogoychev, and Kenneth Heafield. 2023. An open dataset and model for language identification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 865–879, Toronto, Canada. Association for Computational Linguistics.

Bharathi Raja Chakravarthi, Gaman Mihaela, Radu Tudor Ionescu, Heidi Jauhiainen, Tommi Jauhiainen, Krister Lindén, Nikola Ljubešić, Niko Partanen, Ruba Priyadharshini, Christoph Purschke, Eswari Rajagopal, Yves Scherrer, and Marcos Zampieri. 2021. Findings of the VarDial evaluation campaign 2021. In *Proceedings of the Eighth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 1–11, Kiyv, Ukraine. Association for Computational Linguistics.

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. Association for Computational Linguistics.

Mihaela Gaman, Dirk Hovy, Radu Tudor Ionescu, Heidi Jauhiainen, Tommi Jauhiainen, Krister Lindén, Nikola Ljubešić, Niko Partanen, Christoph Purschke, Yves Scherrer, and Marcos Zampieri. 2020. A report on the VarDial evaluation campaign 2020. In *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 1–14, Barcelona, Spain (Online). International Committee on Computational Linguistics (ICCL).

Nat Gillin. 2022. Is encoder-decoder transformer the shiny hammer? In *Proceedings of the Ninth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 80–85, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Cyril Goutte, Serge Léger, Shervin Malmasi, and Marcos Zampieri. 2016. Discriminating similar languages: Evaluations and explorations. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1800–1807, Portorož, Slovenia. European Language Resources Association (ELRA).

Mihaela Găman, Adrian-Gabriel Chifu, William Domingues, and Radu Tudor Ionescu. 2023. Fredo: A large corpus for french cross-domain dialect identification. *Procedia Computer Science*, 225:366–373. 27th International Conference on Knowledge Based and Intelligent Information and Engineering Sytems (KES 2023).

Yonatan Hadar and Erez Shmueli. 2021. Categorizing items with short and noisy descriptions using ensembled transferred embeddings. *arXiv preprint arXiv:2110.11431*.

Tommi Jauhiainen, Marco Lui, Marcos Zampieri, Timothy Baldwin, and Krister Lindén. 2019. Automatic language identification in texts: a survey. *J. Artif. Int. Res.*, 65(1):675–682.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Maggie Yundi Li, Stanley Kok, and Liling Tan. 2018a. Don't classify, translate: Multi-level e-commerce product categorization via machine translation.

Maggie Yundi Li, Stanley Kok, and Liling Tan. 2018b. Don't classify, translate: Multi-level e-commerce product categorization via machine translation. *CoRR*, abs/1812.05774.

Maggie Yundi Li, Liling Tan, Stanley Kok, and Ewa Szymanska. 2018c. Unconstrained product categorization with sequence-to-sequence models. In *Proceedings of the Workshop on eCommerce (co-located with SIGIR)*, pages 1–6.

Yu Meng, Jiaxin Huang, Yu Zhang, and Jiawei Han. 2022. Generating training data with language models: Towards zero-shot language understanding. In *Advances in Neural Information Processing Systems*.

Xiaofei Sun, Xiaoya Li, Jiwei Li, Fei Wu, Shangwei Guo, Tianwei Zhang, and Guoyin Wang. 2023. Text classification via large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8990–9005, Singapore. Association for Computational Linguistics.

Liling Tan. 2023. Transparent, opaque and translucent open source llms. *alvations.com*.

Liling Tan, Marcos Zampieri, Nikola Ljubešic, and Jörg Tiedemann. 2014a. Merging comparable data sources for the discrimination of similar languages: The dsl corpus collection. In *Proceedings of the 7th Workshop on Building and Using Comparable Corpora (BUCC)*, pages 11–15, Reykjavik, Iceland.

Liling Tan, Marcos Zampieri, Nikola Ljubešić, and Jörg Tiedemann. 2014b. Merging comparable data sources for the discrimination of similar languages: The dsl corpus collection. In *Proceedings of the 7th Workshop on Building and Using Comparable Corpora (BUCC)*, pages 6–10. Workshop on Building and Using Comparable Corpora (BUCC) ; Conference date: 27-05-2014 Through 27-05-2014.

Khin Yee Mon Thant and Khin Thandar Nwet. 2020. Comparison of supervised machine learning models for categorizing e-commerce product titles in myanmar text. In *2020 International Conference on Advanced Information Technologies (ICAIT)*, pages 194–199. IEEE.

Jörg Tiedemann and Nikola Ljubešić. 2012. Efficient discrimination between closely related languages. In *Proceedings of COLING 2012*, pages 2619–2634, Mumbai, India. The COLING 2012 Organizing Committee.

Zhiqiang Wang, Yiran Pang, and Yanbin Lin. 2023. Large language models are zero-shot text classifiers.

Wenpeng Yin, Jamaal Hay, and Dan Roth. 2019. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3914–3923, Hong Kong, China. Association for Computational Linguistics.

Marcos Zampieri, Kai North, Tommi Jauhiainen, Mariano Felice, Neha Kumari, Nishant Nair, and Yash Bangera. 2023. Language variety identification with true labels.

Marcos Zampieri, Kai North, Tommi Jauhiainen, Mariano Felice, Neha Kumari, Nishant Nair, and Yash Bangera. 2024. Language variety identification with true labels. In *Proceedings of LREC-COLING*.

Marcos Zampieri, Liling Tan, Nikola Ljubešić, and Jörg Tiedemann. 2014. A report on the DSL shared task 2014. In *Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects*, pages 58–67, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.

A Appendix

All generations from the Mistral models used to produce the results from Table 1 and 2 can be found on `https://huggingface.co/collections/alvations/jelly-shots-662f2661e4a1f7302a85488a`

Improving Multi-label Classification of Similar Languages by Semantics-Aware Word Embeddings

The Quyen Ngo[1], Thi Anh Phuong Nguyen[2], My Linh Ha[1], Thi Minh Huyen Nguyen[1], Phuong Le-Hong[1]*
[1]Vietnam National University, Hanoi, Vietnam
[2]Institute of Information Technology, Vietnam Academy of Science and Technology, Hanoi, Vietnam
Correspondence: *phuonglh@vnu.edu.vn*

Abstract

The VLP team participated in the DSL-ML shared task of the VarDial 2024 workshop which aims to distinguish texts in similar languages. This paper presents our approach to solving the problem and discusses our experimental and official results. We propose to integrate semantics-aware word embeddings which are learned from ConceptNet into a bidirectional long short-term memory network. This approach achieves good performance – our system is ranked in the top two or three of the best performing teams for the task.

1 Introduction

Discriminating between similar languages (e.g., Croatian and Serbian) and language varieties (e.g., Brazilian and European Portuguese) has been a popular research topic related to the study of diatopic language variation from a computational perspective (Aepli et al., 2023). In the DSL-ML shared tasks of The Eleventh Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial) 2024 (Chifu et al., 2024), participating teams are expected to provide multi-label annotations for the instances of datasets from five different macro-languages and with different types of multi-label annotations, including BCMS (Bosnian, Croatian, Montenegrin, Serbian) (Rupnik et al., 2023; Miletić and Miletić, 2024), EN (American and British English), ES (Argentinian and Peninsular Spanish), Portuguese (Brazilian and European Portuguese) (Zampieri et al., 2024), and FR (Belgian, Canadian, French and Swiss French) (Găman et al., 2023; Tan et al., 2014; Bernier-Colborne et al., 2023). Participating systems are evaluated on macro-average F1 for each test set, and aggregated over the five test sets.

This paper presents an approach to improving the performance of our participating system in this shared task. The main idea of our approach is the integration of semantic word embeddings which are learned from the ConceptNet knowledge graph into a recurrent neural network model. The ConceptNet word embeddings are readily available for multiple languages that are concerned with this task.

The paper is structured as follows. Section 2 describes our method. Section 3 presents and discusses empirical results on the development datasets and on the private test set as announced by the shared task organizers. Section 4 concludes the paper and outlines several possible directions for future work.

2 Methods

In this shared task, participants are expected to provide multi-label annotations for the test set instances. There are two tracks. In the closed track, systems may only use the labeled training data provided for the task. The use of pre-trained models is allowed as long as they are not specifically pre-trained or fine-tuned on language identification tasks. In the open track, systems may use any data and pre-trained models, except the prohibited datasets listed in the language description. Our system is essentially in the closed track since we do not use any external training data and the ConceptNet embeddings are not specifically pre-trained or fine-tuned on any language identification task.

We aim to develop a method which does not utilize pre-trained models for this task. Thus, we use bidirectional long short-term memory networks (BiLSTMs) for learning text representation.

2.1 Bidirectional LSTM Model

Let $\mathbf{x}_j$ be the embedding of token w_j and $\mathrm{RNN}_\theta(\mathbf{x})$ be an abstraction of a LSTM that processes the sequence of vectors $\mathbf{x} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n]$, then output for $\mathbf{x}_j$ is defined as $\vec{v}_j := \mathrm{RNN}^l_\theta(\mathbf{x}_j) \oplus \mathrm{RNN}^r_\theta(\mathbf{x}_j)$. We consider multi-layer BiLSTMs where the output $\vec{v}^k_j$ of the k-th layer is fed as input

Proceedings of the Eleventh Workshop on NLP for Similar Languages, Varieties, and Dialects (VarDial 2024), pages 235–240
June 20, 2024

to the $(k+1)$-th layer. In our experiments, each token embedding $\mathbf{x}_j$ is either *initialized randomly* or is a *static pre-trained word embedding* provided by ConceptNet Numberbatch as presented in the next subsection.

For decoding, we use a fully-connected feed-forward network which is fed the output of the last BiLSTM. The output is simply computed by a soft-max layer as common in multiway classification:

$$P(y_j|\vec{v}_j) = \text{softmax}(W\vec{v}_j + \vec{b}),$$

where W and b are parameter matrices. The overall network architecture that we use is as follows:

$$\texttt{EmbeddingLayer}(w) \to \texttt{stacked BiLSTM}(h) \to \texttt{Dense}(d,\ \texttt{ReLu}) \to \texttt{Dense(softmax)},$$

where the hyperparameters w, h and d are the word embedding size, the recurrent hidden size and the dense hidden size, which are tuned on the development datasets for the best performance.

2.2 ConceptNet Numberbatch

ConceptNet is a freely-available semantic network, designed to help computers understand the meanings of words that people use (Speer et al., 2017)[1]. Figure 1 illustrates an excerpt of the concept of ConceptNet. It has been used to create word embeddings – representations of word meanings as vectors, similar to word2vec (Mikolov et al., 2013), GloVe (Pennington et al., 2014), or fastText (Bojanowski et al., 2016). These word embeddings are free, multilingual, aligned across languages, and designed to avoid representing harmful stereotypes. Their performance at word similarity, within and across languages, was shown to be state of the art at SemEval 2017 (Speer and Lowry-Duda, 2017).

ConceptNet Numberbatch is a set of semantic vectors which are trained on ConceptNet that can be used directly as a representation of word meanings. These embeddings benefit from the fact that they have semi-structured, common sense knowledge from ConceptNet, giving them a way to learn about words that isn't just observing them in context. Unlike most embeddings, ConceptNet Numberbatch is multilingual from the ground up. Words in different languages share a common semantic space, and that semantic space is informed by all of the languages. These appealing properties of ConceptNet embeddings make them suitable for multilingual processing tasks which deal with lexical semantics. Discrete structures of ConceptNet have been recently exploited to improve natural language inference (Le-Hong and Cambria, 2023) and dependency parsing (Le-Hong and Cambria, 2024). In this work, we demonstrate that ConceptNet Numberbatch is also helpful in the problem of similar languages classification.

[1] `https://conceptnet.io/`

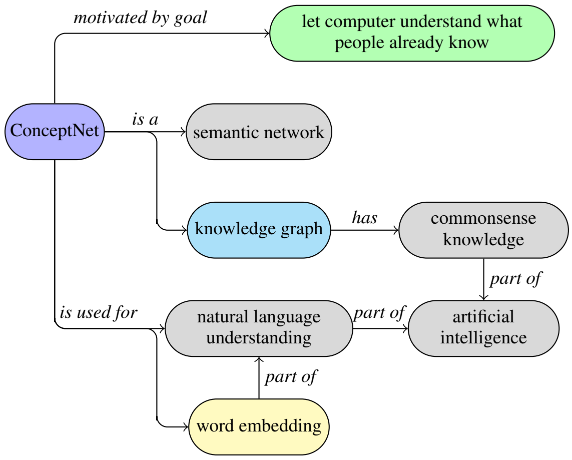


Figure 1: An illustration of ConceptNet in graph.

Language	Training	Dev.	Test
BCMS	368	122	123
EN	2,097	599	300
ES	3,467	989	495
FR	340,363	17,090	12,000
PT	3,467	991	495

Table 1: Statistics of the datasets used in the shared task.

3 Results

3.1 Datasets

Some statistics of the five datasets of the DSL-ML-2024 shared task are given in Table 1. Some observations about the datasets are as follows.

First, the BCMS dataset contains texts in Bosnian, Croatian, Montenegrin, Serbian. There are no multi-label samples in the training split of this dataset but multi-label samples are present in the development and test splits. The size of this dataset is quite small but its sample text is often very long[2]. These properties make supervised models less accurate. Second, while the English and Portuguese datasets are of the same size, the French training dataset is about 100 times larger. This makes the training of French models much more time consuming.

[2] The longest training sample has 159,440 characters.

3.2 Experimental Settings

We carry out two experiments. In the first experiment, the model is applied on randomly initialized word embeddings which are fine-tuned on the training set. This experiment allows us to estimate the performance that a purely supervised learning system can achieve. In the second experiment, the same model is applied on the ConceptNet embeddings. This experiment investigates the advantage of using semantics-aware embeddings in detecting similar languages in a multilingual context. All the models have the same training objective, which is to set the score of the correct language label above the scores of incorrect ones. We use the common cross-entropy loss to minimize the objective function over the training data. This correlates with maximizing the number of correct predictions in the predicted outputs. Note that we consider each target label as atomic; for example, "*EN-GB,EN-US*" is considered a single label instead of two labels "*EN-GB*" and "*EN-US*"[3].

The ConceptNet Numberbatch word embeddings are freely available for download from the ConceptNet open data project[4]; we use the 19.08 version, *numberbatch-en-19.08.txt.gz* for English and *numberbatch-19.08.txt.gz* for multilingual word vectors, each also has 300 dimensions.

Since the model is trained in an end-to-end fashion, the gradients of the entire network, including the embedding matrices for tokens with respect to the sum of the losses are computed using the backpropagation algorithm. We perform multiple training epochs, using early stopping – the training process is stopped when the accuracy does not increase after three consecutive epochs on the development dataset. The maximal sequence length of each sentence is set to 40 tokens[5]. The models are all trained by the Adam optimizer (Kingma and Ba, 2015) with a learning rate of 5×10^{-5}. The batch size is set to 32[6]. On each dataset, we run a set of experiments with a different number of hidden units in each recurrent layer or in the dense layer (cf. subsection 3.4). Each experiment is run five times, its results are averaged for reporting.

[3]We have not tried any multi-label classification method in this task; the problem is considered multi-class classification.

[4]ConceptNet Numberbatch: `https://github.com/commonsense/conceptnet-numberbatch`

[5]This threshold is validated on the training split of the English dataset where 84.07% of samples are $\leq$ 40 tokens.

[6]All models are implemented in the Scala programming language using the BigDL library.

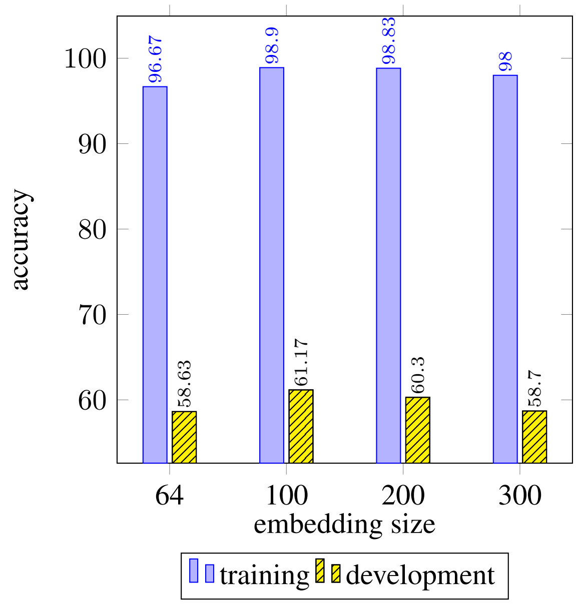


Figure 2: Maximal performance of the `LSTM-r` model with respect to the word embedding size and the recurrent size on the English training and development sets.

3.3 Evaluation Metrics

The organizers of this task provide an evaluation script and a baseline system which uses tf-idf-weighted character-level and word-level n-gram features in a linear SVM classifier[7]. The official scoring script provides per-class F1-scores, weighted and macro-averaged F1-scores. However, during the development stage, we did not use this script to evaluate our models; we used the common accuracy score on the training set and development set when validating the models. Despite of not being a good score for evaluating imbalanced datasets, this metric is found to be effective in model tuning.

3.4 Results

In this subsection, we first present the performance of our models on the development datasets. We then report the performance on the test datasets. We carried out the same experiments for all the languages. For brevity, we report only the process on the English dataset.

In the simple `LSTM-r` model where the word embeddings are initialized randomly, we vary the word embedding size w in the range $[64, 100, 200, 300]$ and the recurrent size h in the range $[100, 200, 300]$. The dense hidden size is fixed at 32 heuristically. Figure 2 shows the accuracy of this model on the

[7]`https://github.com/yvesscherrer/DSL-ML-2024/`

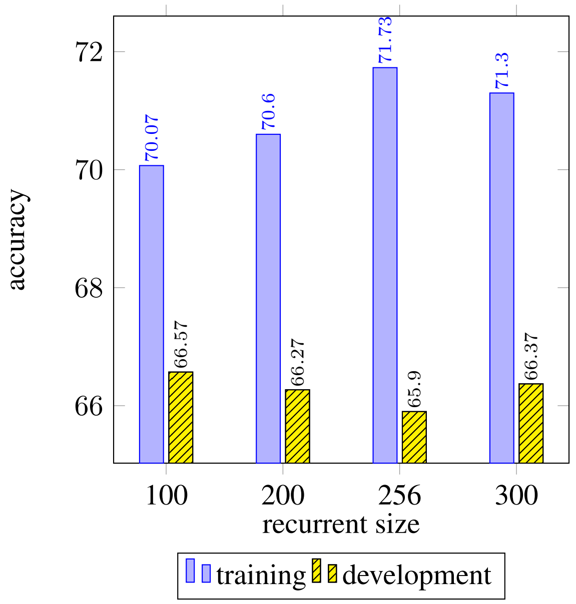


Figure 3: Performance of the `LSTM-c` model with respect to the recurrent size on the English training and development sets. The ConceptNet word embedding size is 300.

training and development splits. This model has a peak performance when $w = 100$ (and $h = 300$, not shown in the figure), having an accuracy of 61.17% on the development split. It seems that the model overfits the training data, which has a relatively small size.

In the enriched `LSTM-c` model where the words embeddings are ConceptNet embeddings, we vary the recurrent size h in the range $[100, 200, 256, 300]$. As above, the dense hidden size is 32. Figure 3 shows its accuracy. This model is not able to achieve a high accuracy on the training set but its development accuracy is significantly better than the `LSTM-c` model. This is maybe due to our choice of freezing the embedding layer, that is, the ConceptNet embeddings are not fine-tuned during training. The best accuracy of `LSTM-c` on the development dataset is 66.57%, which is 5.4% of absolute points better than that of `LSTM-r`.

Table 2 presents the official results of our `LSTM-r` and `LSTM-c` models on the test datasets of all the languages (Chifu et al., 2024). The ConceptNet embeddings are not available for BCMS languages, there is thus only one submission for this dataset.

As shown in these results, the ConceptNet embeddings help improve the accuracy of the English and French datasets by about 1.2% for English and 0.2% for French. However, they are not helpful for the Spanish and Portuguese datasets. It is surprising that the `LSTM-c` model is significantly worse than the `LSTM-r` model on the Portuguese datasets with a gap of about 10% of macro F1. It is possible due to a technical problem of our system during the training stage for this model. We plan to investigate further on this problem once the gold labels of the test datasets are available for additional analysis.

Language	M. F1	W. F1	EM	VLP
BCMS	27.22	36.97	00.00	1
EN	76.98	77.64	16.67	2
	75.88	76.30	26.67	1
ES	75.39	76.06	45.51	1
	74.14	74.36	42.31	2
FR	25.96	25.96	–	2
	25.74	25.74	–	1
PT	66.36	69.13	13.56	1
	56.58	62.01	00.00	2

Table 2: Official results of our systems on the test datasets as announced by the organizers. M. F1, W. F1 and EM is the macro F1, the weighted F1 and the exact match score, respectively. The VLP column is the submission index where number 1 indicates the `LSTM-r` model and number 2 indicates the `LSTM-c` model.

4 Conclusion

In this paper, we have presented a recurrent neural network model for tackling the problem of distinguishing similar languages. Our method utilizes semantics-aware ConceptNet embeddings for four languages. Despite its simplicity, the proposed model achieves relatively good results.

We are currently using the simple multi-class classification approach for this task. We plan to apply specific methods of multi-label classification for the task in a future work.

Recent works have shown that learning to classify texts can be beneficial by unsupervised representation learning methods such as contrastive learning (Su et al., 2022). The goal of contrastive learning is to learn a representation of text such that similar instances are close together in the representation space, while dissimilar instances are far apart. A combination of similarity embeddings learned by contrastive learning and semantics embeddings learned from knowledge graphs such as WordNet and ConceptNet can be helpful for this task.

Finally, in the last few years, pre-trained large language models such as XLM-R (Conneau et al., 2020), GPT (Brown et al., 2020) and LLaMa (Touvron et al., 2023) are making new waves in the field of natural language processing due to their emergent ability and generalizability. We have investigated using a pre-trained XLM-R model for this shared task but initial results are mediocre compared to our proposed approach. However, a more throughout inquiry of using large language models is necessary before a firm conclusion about their usefulness can be drawn.

Acknowledgments

This study is supported by Vingroup Innovation Foundation (VINIF) in project code VINIF.2020.DA14.

References

Noëmi Aepli, Çağrı Çöltekin, Rob Van Der Goot, Tommi Jauhiainen, Mourhaf Kazzaz, Nikola Ljubešić, Kai North, Barbara Plank, Yves Scherrer, and Marcos Zampieri. 2023. Findings of the VarDial evaluation campaign 2023. In *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*, pages 251–261, Dubrovnik, Croatia.

Gabriel Bernier-Colborne, Cyril Goutte, and Serge Leger. 2023. Dialect and variant identification as a multi-label classification task: A proposal based on near-duplicate analysis. In *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*, pages 142–151, Dubrovnik, Croatia. Association for Computational Linguistics.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

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.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Mihaela Găman, Adrian-Gabriel Chifu, William Domingues, and Radu Tudor Ionescu. 2023. FreCDo: A large corpus for French cross-domain dialect identification. *Procedia Computer Science*, 225:366–373.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations, ICLR 2015*, pages 1–15, San Diego, CA, USA.

Phuong Le-Hong and Erik Cambria. 2023. A semantics-aware approach for multilingual natural language inference. *Language Resources and Evaluation*, 57(2):611–639.

Phuong Le-Hong and Erik Cambria. 2024. Integrating graph embedding and neural models for improving transition-based dependency parsing. *Neural Computing and Applications*, 36(6):2999–3016.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *Proceedings of Workshop at ICLR*, pages 1–12, Scottsdale, Arizona, USA.

Aleksandra Miletić and Filip Miletić. 2024. A gold standard with silver linings: Scaling up annotation for distinguishing Bosnian, Croatian, Montenegrin and Serbian. In *Proceedings of the 4th Workshop on Human Evaluation of NLP Systems*, Turin, Italy. European Language Resources Association.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of EMNLP*, pages 1532–1543, Doha, Qatar.

Peter Rupnik, Taja Kuzman, and Nikola Ljubešić. 2023. BENCHić-lang: A benchmark for discriminating between Bosnian, Croatian, Montenegrin and Serbian. In *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*, pages 113–120, Dubrovnik, Croatia. Association for Computational Linguistics.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. ConceptNet 5.5: An open multilingual graph of general knowledge. In *Proceedings of AAAI*, pages 4444–4451.

Robyn Speer and Joanna Lowry-Duda. 2017. ConceptNet at SemEval-2017 task 2: Extending word embeddings with multilingual relational knowledge. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 85–89, Vancouver, Canada.

Xi'ao Su, Ran Wang, and Xinyu Dai. 2022. Contrastive learning-enhanced nearest neighbor mechanism for multi-label text classification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 672–679, Dublin, Ireland. Association for Computational Linguistics.

Liling Tan, Marcos Zampieri, Nikola Ljubešic, and Jörg Tiedemann. 2014. Merging comparable data sources for the discrimination of similar languages: The dsl corpus collection. In *Proceedings of the 7th Workshop on Building and Using Comparable Corpora (BUCC)*, pages 11–15, Reykjavik, Iceland.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Marcos Zampieri, Kai North, Tommi Jauhiainen, Mariano Felice, Neha Kumari, Nishant Nair, and Yash Bangera. 2024. Language variety identification with true labels. In *Proceedings of LREC-COLING*.

Brandeis at VarDial 2024 DSL-ML Shared Task: Multilingual Models, Simple Baselines and Data Augmentation

Jonne Sälevä and **Chester Palen-Michel**
Mitchom School of Computer Science
Brandeis University
{jonnesaleva, cpalenmichel}@brandeis.edu

Abstract

This paper describes the Brandeis University submission to VarDial 2024 DSL-ML Shared Task on multilabel classification for discriminating between similar languages. Our submission consists of three entries per language to the closed track, where no additional data was permitted. Our approach involves a set of simple non-neural baselines using logistic regression, random forests and support vector machines. We follow this by experimenting with finetuning multilingual BERT, either on a single language or all the languages concatenated together. In addition to benchmarking the model architectures against one another on the development set, we perform extensive hyperparameter tuning, which is afforded by the small size of the training data. Our experiments on the development set suggest that finetuned mBERT systems significantly benefit most languages compared to the baseline. However, on the test set, our results indicate that simple models based on scikit-learn can perform surprisingly well and even outperform pretrained language models, as we see with BCMS. Our submissions achieve the best performance on all languages as reported by the organizers. Except for Spanish and French, our non-neural baseline also ranks in the top 3 for all other languages.

1 Introduction

Language identification (LID) is the task of determining which language a piece of text is written in (Jauhiainen et al., 2019). While robust LID software already exists (e.g. Google's CLD3[1]), there are still several unsolved problems that plague current state-of-the-art LID models. One of the most pressing issues is lack of proper language coverage, which recent work has fortunately started to address as more data becomes available for more languages (e.g. Adebara et al., 2022; Burchell et al., 2023a; Kargaran et al., 2023).

Despite these promising developments, detection of lower-resourced languages, variants, and dialects still poses problems for modern NLP. The lack of resources also generally correlates with poor quality of the resources that are available which can lead to, for instance, datasets with unusually short sentences which may make the task difficult (Baldwin and Lui, 2010). To make matters worse, low-resource language variants tend to also be deceptively similar to other languages or dialects which makes differentiating between them accurately all the more challenging (Jauhiainen et al., 2019).

In the last ten years, the NLP for Similar Languages, Varieties, and Dialects workshop (VarDial) has emerged as the principal venue for discussion around these problems (e.g. Aepli et al., 2023, 2022; Chakravarthi et al., 2021). The workshop also features an annual shared task on discriminating between similar languages (DSL). The first VarDial DSL shared task DSL was organized with the purpose of better understanding the difficulties faced by state-of-the-art systems when differentiating between similar languages and varieties (Zampieri et al., 2014). Since then, multiple DSL shared tasks have been organized, leading to the development of a robust research community (Zampieri et al., 2014, 2015; Malmasi et al., 2016; Zampieri et al., 2017).

In the most recent VarDial DSL shared task, annotated datasets were added (Aepli et al., 2023). In the current iteration of the task, the labels were treated as a multi-label classification problem as proposed in Bernier-colborne et al. (2023).

In this paper, we describe our submission to the most recent VarDial shared task. For our submission, we experimented with simple non-neural baselines using scikit-learn, extensive hyperparameter tuning, data augmentation, and concatenating the

[1] https://github.com/google/cld3

Proceedings of the Eleventh Workshop on NLP for Similar Languages, Varieties, and Dialects (VarDial 2024), pages 241–251
June 20, 2024 ©2024 Association for Computational Linguistics

Language	Split	Total Documents	Mean Sentences per doc	Mean Tokens per doc
EN	train	2,097	1.5	38.3
EN	dev	599	1.4	34.8
EN	test	300	1.4	35.3
BCMS	train	368	428.7	6,540.3
BCMS	dev	122	429.0	6,672.8
BCMS	test	123	465.1	6,999.1
FR	train	340,363	9.0	80.4
FR	dev	17,090	7.7	78.4
FR	test	12,000	12.0	96.7
PT	train	3,467	1.8	44.3
PT	dev	991	1.8	44.0
PT	test	495	1.8	43.7
ES	train	3,467	1.9	58.7
ES	dev	989	1.9	58.6
ES	test	495	1.9	60.2

Table 1: Counts of documents, average sentences per document, and tokens per document for each dataset.

datasets in an attempt at enhancing multilingual transfer. Ultimately, we found the best performing models for all languages tended to be fine-tuned mBERT variants (Devlin et al., 2018), except BCMS whose best performing model was a non-neural random forest model implemented in `scikit-learn` (Pedregosa et al., 2011).

2 Task Description

The shared task (Chifu et al., 2024) consisted of distinguishing between different varieties of a macro-language. There were 5 macro-language groups in the shared task. Some datasets differ notably in the size of a single classification instance, which we refer to as documents. In Table 1, the number of total documents for each of the splits is shown along with the mean sentences and tokens per document. The tokens and sentences are obtained by using the `spaCy` library and the `*_core_small` models for each language. For BCMS, we used the Croatian model, since it was the only language explicitly supported by `spaCy`. It can be seen that the French dataset is much larger than the others and that the BCMS dataset contains much longer documents in terms of sentences and tokens than any of the other datasets.

Data Sources The English, Spanish, and Portuguese data is from DSL-TL (Zampieri et al., 2024), which is manually annotated labels from the Discriminating Similar Languages Corpus Collection (DSLCC) (Tan et al., 2014). The French data partially comes from FreCDo (Găman et al., 2023) and DSLCC. French is also the only language whose dataset has named entities masked out. The Bosnian, Croatian, Montenegrin, and Serbian (BCMS) data comes from BENCHić-lang (Rupnik et al., 2023) and Twitter HBS 1.0 (Ljubešić and Rupnik, 2022) as well as Miletić and Miletić (2024). Given that much of the BCMS data is derived from Twitter, it is fairly different than the other datasets in terms of content. Details regarding the origins of the datasets and how they were annotated are summarized in Table 2.

3 System Descriptions

We made three submissions for the closed track. The three submissions consisted of our best performing models for scikit learn based classifiers, our best performing models using fine-tuning of mBERT, and a fine-tuned mBERT model using the concatenation of all datasets.

3.1 Run 1: `scikit-learn` Baselines

For Run 1, we submitted our best model from testing a series of scikit-learn classifiers: logistic regression models, linear-kernel SVMs and random forest models. For all models, we used bag-of-n-grams-style features where the n-grams were defined over (a) space-separated tokens (`analyzer=word`) or (b) characters (`analyzer=char`). In addition to integer counts (`CountVectorizer`), we also experimented with real-valued tf-idf weights (`TfidfVectorizer`) as an alternative representation. To prevent overfitting, we did not consider n-grams beyond $n = 2$. The full set of hyperparameters is shown in Table 3. The best performing configurations can be found in Table 4.

3.2 Run 2: Per-language mBERT Models

For our second run, we experimented with fine-tuning multilingual BERT (Devlin et al., 2018) independently on each language. We used `bert-base-multilingual-cased` for each submission[2]. The multilingual BERT model is pre-trained on masked language modeling and next sentence prediction. All macro-languages are included in mBERTs pre-training data. While the documentation of mBERT is less clear about variants of the macro-languages are included, for BCMS, individual languages are listed. All BCMS languages are

[2]`https://huggingface.co/google-bert/bert-base-multilingual-cased`

Lang.	Original data	Varieties	Train	Dev / Test	Annotation	Entities
English	DSL-TL	British English American English	Multi-label	Multi-label	Manually	Present
Spanish	DSL-TL	Castillian Spanish Argentinian	Multi-label	Multi-label	Manually	Present
Portuguese	DSL-TL	Brazilian, Portugal	Multi-label	Multi-label	Manually	Present
French	FreCDo, DSLCC	Canadian, Belgian Metropolitan French, Swiss	Multi-label	Multi-label	Automatically	Masked
BCMS	BENCHić-lang / Twitter HBS 1.0	Bosnian, Serbian, Montenegrin, Croatian	Single-label	Multi-label	Manually	Present

Table 2: Description of datasets included in the shared task.

Hyperparameter	Values
Architecture	Random forest, log. reg., SVM
Mode	multilabel, multiclass
Feature type	count, tf-idf
n-gram level	word, char
n-grams range	unigrams, bigrams, both
Solver	newton-cg, lbfgs, liblinear, sag, saga
Regularizer (C)	0.001, 0.01, 0.1, 1, 10, 100
Class weight	unadjusted, balanced
Max. iterations	off, 5000
Max. features	off, sqrt
No. of estimators	50, 100
Max. depth	30, 50

Table 3: Hyperparameter values used in non-neural `scikit-learn` experiments (Run 1).

represented in mBERTs pre-training data except for Montenegrin. We experimented with different hyperparameters for fine-tuning; the full set of values used can be seen in Table 5.

We adapt mBERT to multi-label classification by using a linear layer for classification, applying a sigmoid function to the logits and setting a threshold of 0.5 for the label to be included in the output. At inference time, if no output label meets the threshold, we relax the threshold to ensure each example is labeled first to .25, then .05. If after relaxing the threshold no label is assigned, we assign the most common label for the dataset.

Because the BCMS dataset had particularly longer documents with multiple sentences, we segmented each example first into sentences using spaCy (Honnibal et al., 2020). We then trained a model to predict on independent sentences. For inference we segment the documents first and classify each of their sentences. We then obtain final labels for the document by including labels that occur over a threshold of a proportion of the composite sentences. The threshold was set at 0.2 by adjusting to the development set.

All hyperparameters were tuned using an exhaustive grid search through all possible options. The hyperparameter configurations we experimented with for Run 2 can be found in Table 5.

3.3 Run 3: Finetuning All Languages at Once

For Run 3, we submitted mBERT fine-tuned on the concatenation of all the datasets. As we had already performed extensive hyperparameter tuning for Run 2, we opted to re-use well-performing hyperparameters from prior mBERT training runs for Run 2. Specifically, we used a learning rate of 2.0E-5, a batch size of 64, and 3 epochs to train the model with the concatenated dataset. We used a naive concatenation for this run and did not weight or sample the combined dataset in any special way. The motivation for this run is that it would provide a single model capable of distinguishing between similar languages for multiple macro-languages. As we discuss further in Section 5, this combined single model works decently well for most languages, but performs very poorly on the BCMS data.

4 Additional Experiments

In addition to the submitted systems, we conducted other experiments. These additional experiments included exploring data augmentation and segmentation of BCMS documents. Ultimately the BCMS segmentation was used for Run 2, but the data augmentation approaches did not appear to be useful enough to be included any of our submitted systems.

4.1 Segmenting BCMS

Noticing that performance was lower on BCMS and that the dataset had a much higher proportion of sentences per document compared with the datasets of other macro-languages, we compared

Language	BCMS	English	Spanish	Portuguese	French
Model	Random forest	Log. Reg. (OvR)	Random forest	SVM (OvR)	SVC (OvR)
Text features					
Count type	tf-idf	tf-idf	tf-idf	tf-idf	count
n-gram level	word	word	word	word	char
n-gram range	unigrams	unigrams	unigrams	both	bigrams
Common hyperparameters					
Solver	-	sag	-	-	-
Regularization (C)	-	10	-	10	100
Max iterations	-	100	-	5000	5000
Random forest params					
Bootstrap	False	-	False	-	-
Class weight	balanced	-	-	-	-
Max depth	50	-	50	-	-
Max features	-	-	sqrt	-	-
No. of estimators	50	-	100	-	-
F1 (macro)	71.33	79.75	82.99	72.01	55.00

Table 4: Best hyperparameters for scikit-learn models as computed on the development set.

Language	Batch Size	Learning Rate	Epochs
EN	16	2.0E-05	3
BCMS	16	2.0E-05	3
FR	16	2.0E-05	3
ES	64	3.0E-05	3
PT	16	2.0E-05	3

Table 5: Hyperparameters for individual mBERT models submission (Run 2).

	Orig. BCMS	Segmented BCMS
Macro F1	20.67	**72.2**
Weighted avg. F1	47.73	**79.8**

Table 6: Comparison of mBERT model on original BCMS dataset with segmented data.

performance from segmenting and not segmenting the data first. When segmenting the data into sentences, we used spaCy (Honnibal et al., 2020) with the Croatian model for all BCMS languages. In order to map back to the original examples, we label the example with any label that shows up in more than 20% of the composite sentences.

The results of this experiment are shown in Table 6. When applying segmentation and the strategy of classifying on each sentence individually, we saw a large gain of more than 50 points of macro F1 when segmenting first and then recombining.

4.2 Data Augmentation

Since some of the datasets had only a few thousand samples, we explored data augmentation as a way to obtain additional samples while still using only the datasets available for the closed track. Because the French and BCMS datasets contained hundreds of thousands of training sentences, we focused our data augmentation experiments on English, Spanish, and Portuguese. We attempted two simple data augmentation strategies.

First, since very simple word replacements have been shown to help model robustness (Wei and Zou, 2019; Kolomiyets et al., 2011) we tried naively splitting documents in half and recombined these half sentences with other half sentences of the same labels. The pieces from each sentence must have the same label. An example of this process is shown in Figure 1, where the label is EN-GB for all sentences in the example.

Second, similar to Zhang et al. (2022) or Andreas (2020), we attempted to replace segments based on spans from dependency trees with spans from other documents with the same labels. For the syntactic span augmentation, we use spaCy to get a dependency parse of each sentence. We then take a node and replace its children token span with another token span from a node of the same part of speech and parent dependency relation from a randomly sampled sentence with the same label. An example can be seen in Figure 2. In Figure 2, the label is EN-US for each sentence.

Unfortunately, neither of these approaches ended up providing a significant performance increase when evaluating on the development set.

We compare the naive augmentation, tree-based

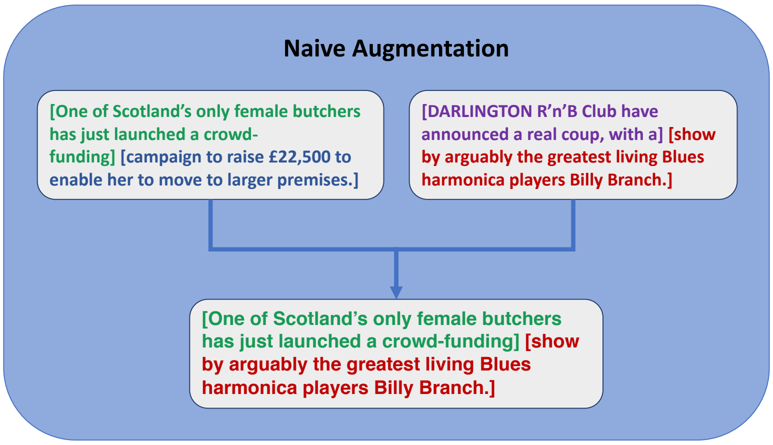


Figure 1: Naive augmentation approach.

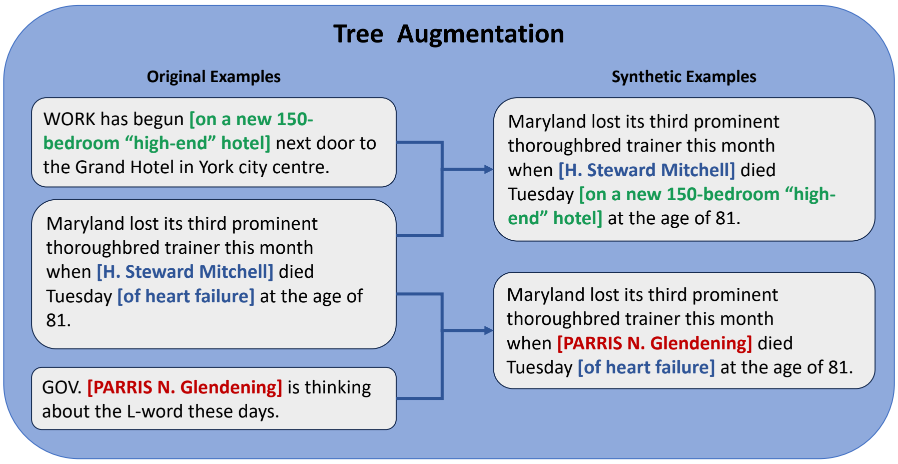


Figure 2: Tree augmentation approach.

Augmentation Strategy	EN	ES	PT
No Augmentation	**84.18**	**82.36**	74.45
Naive Aug.	82.47	82.09	**76.05**
Tree Aug.	81.8	81.19	73.69

Table 7: Results from data augmentation experiments. Scores are Macro-F1.

	EN	ES	FR	BCMS	PT
Run 1	79.75	74.49	54.26	69.32	72.01
Run 2	83.49	**83.50**	**96.58**	**72.20**	75.20
Run 3	**84.67**	82.75	68.40	20.67	**76.01**

Table 8: Macro F1 scores on the development set for each of our submissions on each language group.

augmentation, and no augmentation in Table 7 and find the macro-average F1 for each language is lower with the augmentations except for Portuguese. Since the Portuguese performance was only .04 higher than the concatenation model (run 3) and only seemed to benefit Portuguese, we decided not to submit any of the data augmentation approaches as part of our final submission.

5 Results

Based on performance on the development set as seen in Table 8, we expected Run 2 to perform best for Spanish, French, and BCMS and Run 3 to perform best for English and Portuguese.

Results from each submission are reported in Table 9. Run 3, the concatenated dataset with mBERT, does perform best for English and Por-

Language	Run	F1 (m.)	F1 (w.)
BCMS	Run 1: scikit-learn	**76.20**	**84.28**
BCMS	Run 2: mBERT	71.90	75.61
BCMS	Run 3: mBERT-all	19.85	45.30
EN	Run 1: scikit-learn	80.60	80.78
EN	Run 2: mBERT	85.27	85.56
EN	Run 3: mBERT-all	**85.48**	**85.62**
ES	Run 1: scikit-learn	74.59	75.31
ES	Run 2: mBERT	**82.27**	**82.68**
ES	Run 3: mBERT-all	82.09	82.31
PT	Run 1: scikit-learn	72.36	75.49
PT	Run 2: mBERT	71.40	74.10
PT	Run 3: mBERT-all	**75.21**	**77.71**
FR	Run 1: scikit-learn	27.03	27.03
FR	Run 2: mBERT	26.53	26.53
FR	Run 3: mBERT-all	**38.51**	**38.51**

Table 9: Test set results for all submitted runs. F1 (m.) and F1 (w.) refer to macro-F1 and weighted F1.

tuguese. However, for Run 1, Random Forest performed better on the test set for BCMS than mBERT-based models. Additionally, for Run 3, the concatenated dataset with mBERT, outperformed for French instead of Run 2 as seen on the development dataset.

To better understand the results, we created confusion matrices for our submitted runs for each dataset. Figure 3 shows the confusion matrix for Run 1 and 4 for Run 2. A confusion matrix for Run 3 is included in Appendix A.

Class imbalance appears to be a challenge, especially for BCMS and French. For Run 3, all predictions were for Serbian. Run 2 appears most capable for BCMS of making predictions that are ambiguous but still at least partially correct. Run 1 clearly performs well on BCMS, but seems to struggle with French class imbalance. For French, class imbalance seems to affect Run 1 the most with all varieties being mistaken for Metropolitan French at a higher rate than other runs. Run 3 appears to do better at correctly classifying Belgian and Swiss French.

For English, Run 2 predicts British English more often. All runs appear to struggle with ambiguous examples in English and Portuguese. It appears models are better able to correctly predict ambiguous examples in Spanish than in other macrolanguages.

6 Discussion and Conclusion

In this paper, we presented the Brandeis submissions to the VarDial 2024 DSL-ML Shared Task. We conclude by discussing some relevant aspects of our findings.

Baselines Perform Remarkably Well Somewhat contrary to our initial expectations, `scikit-learn`-based models seemed to perform well on both the development and test sets for many languages. On English, Portuguese and BCMS, the non-neural baselines underperformed mBERT by less than 4 macro-F1 points which is remarkable given the drastically smaller size of the baselines. This suggests that simple baselines may carry more utility than initially anticipated.

Further, the baseline performance on the test set shows stronger evidence of their utility. On French, Portuguese and BCMS, the baselines even outperform mBERT. While the differences in test set macro-F1 are less than 1 point in for both Portuguese and French, on BCMS the best baseline outperforms mBERT by more than 4.3 F1 points.

While this is a positive sign, we find the trend reversal somewhat perplexing. Since other trends, such as the universally low performance of Run 3 on BCMS, are replicated on both the test and development set, it stands to reason that this may not entirely be an issue of domain mismatch. Instead, we hypothesize that this may have to do with inherent noisiness in the kinds of low-resource data the shared task deals with.

Concatenation of Fine-Tuning Languages Contrary to the findings of Baldwin and Lui (2010), who showed that language identification becomes more difficult as the number of languages increases, we find that performance does not degrade significantly even after we increase the number of output labels from 2-4 per macrolanguage (independent mBERT models) to 14 (mBERT finetuned on all languages). One exception to this is BCMS, where mBERT-all underperforms even the official baseline. We hypothesize that with such a comparatively small number of languages (with other models like Burchell et al. (2023b) handling more than 200), increasing the number of languages to be classified does not degrade performance when the number of samples is comparable between languages. We speculate that BCMS languages may have underperformed with the concatenated model because there were drastically fewer examples. The majority class for BCMS is Serbian, and the minority classes are especially under-represented.

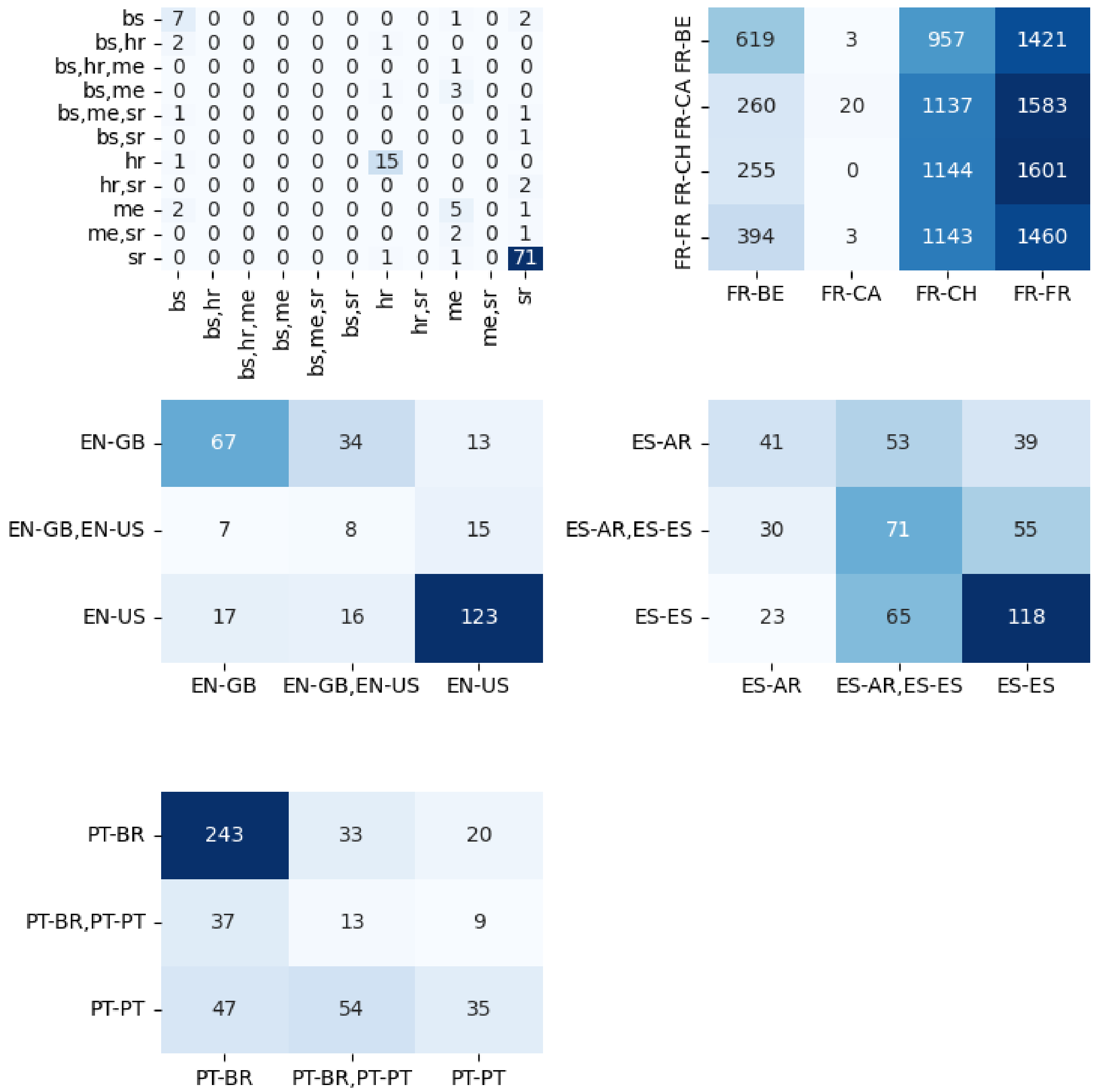


Figure 3: Confusion matrices for Run 1 on the test set. Correct labels are the x-axis and predicted are on the y-axis.

Simple Data Augmentation Does Not Help Much. We did not see improvement from fairly simple data augmentation approaches. It is possible that the models for discriminating similar models mostly rely on small spans of tokens that are already well represented in the original data. It is plausible that changing mixing spans of tokens into different contexts does not make much of a difference if those spans are already well weighted features and do not highly depend on what context they occur in. In future work, it may be worth attempting to better identify which spans are more informative features and experiment with data augmentation approaches that focus on the portion of the text that is most helpful in distinguishing the language variety.

Acknowledgements

The authors wish to thank the organizers of the VarDial 2024 DSL-ML Shared Task and the workshop in general. The authors also wish to thank Christopher Allison for providing technical support with the Brandeis Student Compute Cluster.

References

Ife Adebara, AbdelRahim Elmadany, Muhammad Abdul-Mageed, and Alcides Alcoba Inciarte. 2022. Afrolid: A neural language identification tool for african languages.

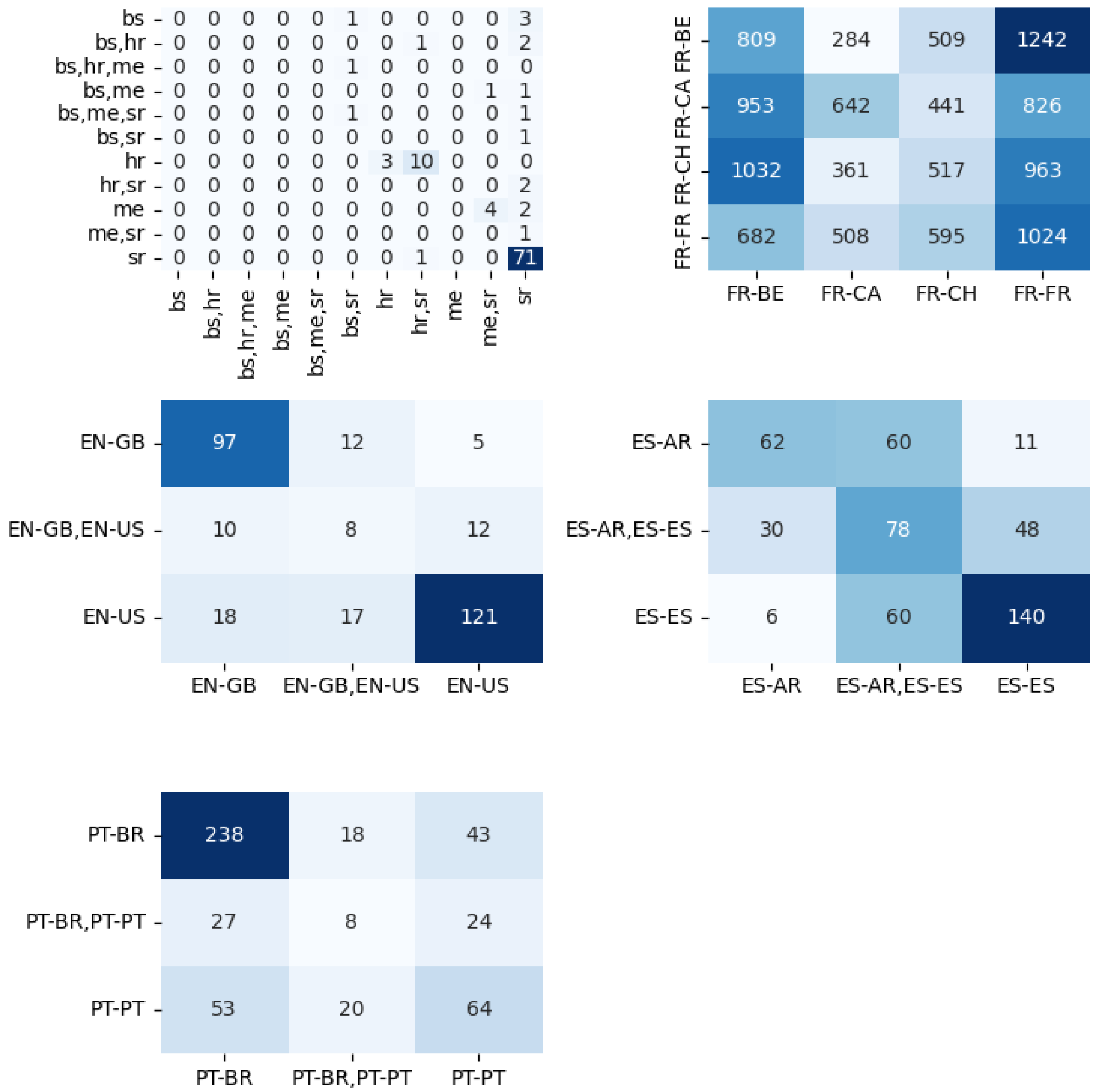


Figure 4: Confusion matrices for Run 2 on the test set. Correct labels are the x-axis and predicted are on the y-axis.

Noëmi Aepli, Antonios Anastasopoulos, Adrian-Gabriel Chifu, William Domingues, Fahim Faisal, Mihaela Gaman, Radu Tudor Ionescu, and Yves Scherrer. 2022. Findings of the VarDial evaluation campaign 2022. In *Proceedings of the Ninth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 1–13, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Noëmi Aepli, Çağrı Çöltekin, Rob Van Der Goot, Tommi Jauhiainen, Mourhaf Kazzaz, Nikola Ljubešić, Kai North, Barbara Plank, Yves Scherrer, and Marcos Zampieri. 2023. Findings of the VarDial evaluation campaign 2023. In *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*, pages 251–261, Dubrovnik, Croatia. Association for Computational Linguistics.

Jacob Andreas. 2020. Good-enough compositional data augmentation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7556–7566, Online. Association for Computational Linguistics.

Timothy Baldwin and Marco Lui. 2010. Language identification: The long and the short of the matter. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 229–237, Los Angeles, California. Association for Computational Linguistics.

Gabriel Bernier-colborne, Cyril Goutte, and Serge Leger. 2023. Dialect and variant identification as a multi-label classification task: A proposal based on near-duplicate analysis. In *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial*

2023), pages 142–151, Dubrovnik, Croatia. Association for Computational Linguistics.

Laurie Burchell, Alexandra Birch, Nikolay Bogoychev, and Kenneth Heafield. 2023a. An open dataset and model for language identification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 865–879, Toronto, Canada. Association for Computational Linguistics.

Laurie Burchell, Alexandra Birch, Nikolay Bogoychev, and Kenneth Heafield. 2023b. An open dataset and model for language identification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 865–879, Toronto, Canada. Association for Computational Linguistics.

Bharathi Raja Chakravarthi, Gaman Mihaela, Radu Tudor Ionescu, Heidi Jauhiainen, Tommi Jauhiainen, Krister Lindén, Nikola Ljubešić, Niko Partanen, Ruba Priyadharshini, Christoph Purschke, Eswari Rajagopal, Yves Scherrer, and Marcos Zampieri. 2021. Findings of the VarDial evaluation campaign 2021. In *Proceedings of the Eighth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 1–11, Kiyv, Ukraine. Association for Computational Linguistics.

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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Mihaela Găman, Adrian-Gabriel Chifu, William Domingues, and Radu Tudor Ionescu. 2023. FreCDo: A large corpus for French cross-domain dialect identification. *Procedia Computer Science*, 225:366–373.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength natural language processing in python.

Tommi Jauhiainen, Marco Lui, Marcos Zampieri, Timothy Baldwin, and Krister Lindén. 2019. Automatic language identification in texts: A survey. *Journal of Artificial Intelligence Research*, 65:675–782.

Amir Hossein Kargaran, Ayyoob Imani, François Yvon, and Hinrich Schütze. 2023. Glotlid: Language identification for low-resource languages.

Oleksandr Kolomiyets, Steven Bethard, and Marie-Francine Moens. 2011. Model-portability experiments for textual temporal analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 271–276, Portland, Oregon, USA. Association for Computational Linguistics.

Nikola Ljubešić and Peter Rupnik. 2022. The news dataset for discriminating between bosnian, croatian and serbian SETimes.HBS 1.0. Slovenian language resource repository CLARIN.SI.

Shervin Malmasi, Marcos Zampieri, Nikola Ljubešić, Preslav Nakov, Ahmed Ali, and Jörg Tiedemann. 2016. Discriminating between similar languages and Arabic dialect identification: A report on the third DSL shared task. In *Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial3)*, pages 1–14, Osaka, Japan. The COLING 2016 Organizing Committee.

Aleksandra Miletić and Filip Miletić. 2024. A gold standard with silver linings: Scaling up annotation for distinguishing Bosnian, Croatian, Montenegrin and Serbian. In *Proceedings of the 4th Workshop on Human Evaluation of NLP Systems*, Turin, Italy. European Language Resources Association.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Peter Rupnik, Taja Kuzman, and Nikola Ljubešić. 2023. BENCHić-lang: A benchmark for discriminating between Bosnian, Croatian, Montenegrin and Serbian. In *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*, pages 113–120, Dubrovnik, Croatia. Association for Computational Linguistics.

Liling Tan, Marcos Zampieri, Nikola Ljubešic, and Jörg Tiedemann. 2014. Merging comparable data sources for the discrimination of similar languages: The dsl corpus collection. In *Proceedings of the 7th Workshop on Building and Using Comparable Corpora (BUCC)*, pages 11–15, Reykjavik, Iceland.

Jason Wei and Kai Zou. 2019. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6382–6388, Hong Kong, China. Association for Computational Linguistics.

Marcos Zampieri, Shervin Malmasi, Nikola Ljubešić, Preslav Nakov, Ahmed Ali, Jörg Tiedemann, Yves Scherrer, and Noëmi Aepli. 2017. Findings of the VarDial evaluation campaign 2017. In *Proceedings of the Fourth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial)*, pages 1–15, Valencia, Spain. Association for Computational Linguistics.

Marcos Zampieri, Kai North, Tommi Jauhiainen, Mariano Felice, Neha Kumari, Nishant Nair, and Yash Bangera. 2024. Language variety identification with true labels. In *Proceedings of LREC-COLING*.

Marcos Zampieri, Liling Tan, Nikola Ljubešić, and Jörg Tiedemann. 2014. A report on the DSL shared task 2014. In *Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects*, pages 58–67, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.

Marcos Zampieri, Liling Tan, Nikola Ljubešić, Jörg Tiedemann, and Preslav Nakov. 2015. Overview of the DSL shared task 2015. In *Proceedings of the Joint Workshop on Language Technology for Closely Related Languages, Varieties and Dialects*, pages 1–9, Hissar, Bulgaria. Association for Computational Linguistics.

Le Zhang, Zichao Yang, and Diyi Yang. 2022. TreeMix: Compositional constituency-based data augmentation for natural language understanding. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5243–5258, Seattle, United States. Association for Computational Linguistics.

A Run 3 Confusion Matrix

Figure 5 shows the confusion matrix for Run 3. Run 3 performs poorly on the BCMS dataset and only predicts Serbian for all examples. For French, Run 3 appears to do worse at predicting Metropolitan French, but better at Swiss and Belgian than Run 2.

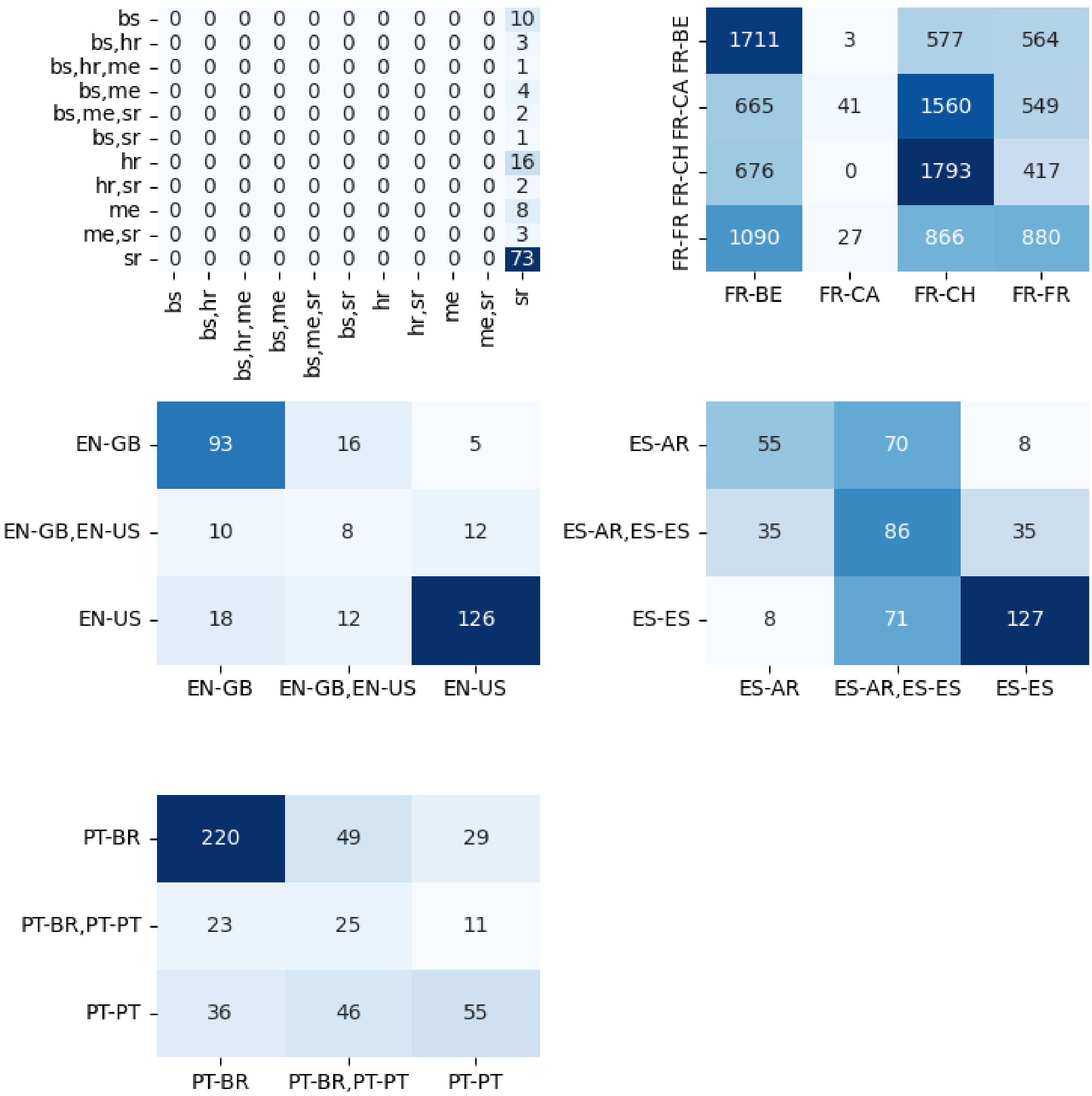


Figure 5: Confusion matrices for Run 3 on the test set. Correct labels are the x-axis and predicted are on the y-axis.

Author Index