A Gold Standard with Silver Linings: Scaling Up Annotation for Distinguishing Bosnian, Croatian, Montenegrin and Serbian

Aleksandra Miletić[1] **Filip Miletić**[2]
[1]Department of Digital Humanities, University of Helsinki, Finland
[2]Institute for Natural Language Processing, University of Stuttgart, Germany
aleksandra.miletic@helsinki.fi filip.miletic@ims.uni-stuttgart.de

Abstract

Bosnian, Croatian, Montenegrin and Serbian are the official standard linguistic varieties in Bosnia and Herzegovina, Croatia, Montenegro, and Serbia, respectively. When these four countries were part of the former Yugoslavia, the varieties were considered to share a single linguistic standard. After the individual countries were established, the national standards emerged. Today, a central question about these varieties remains the following: How different are they from each other? How hard is it to distinguish them? While this has been addressed in NLP as part of the task on Distinguishing Between Similar Languages (DSL), little is known about human performance, making it difficult to contextualize system results. We tackle this question by reannotating the existing BCMS dataset for DSL with annotators from all target regions. We release a new gold standard, replacing the original single-annotator, single-label annotation by a multi-annotator, multi-label one, thus improving annotation reliability and explicitly coding the existence of ambiguous instances. We reassess a previously proposed DSL system on the new gold standard and establish the human upper bound on the task. Finally, we identify sources of annotation difficulties and provide linguistic insights into the BCMS dialect continuum, with multiple indicators highlighting an intermediate position of Bosnian and Montenegrin.

Keywords: BCMS, Distinguishing Between Similar Languages, human upper bound, gold standard, corpus annotation

1. Introduction

Bosnian, Croatian, Montenegrin and Serbian are the official standard linguistic varieties in their respective countries: Bosnia and Herzegovina (3.3M inhabitants), Croatia (3.9M), Montenegro (0.6M) and Serbia (6.7M) (Figure 1).[1] When the four countries were part of the former Yugoslavia, these varieties were considered to belong to the same language, which was commonly referred to as Serbo-Croatian or Croato-Serbian. After the civil wars of the 1990s and the establishment of individual countries, national linguistic standards also emerged. Thirty years later, one of the central questions about Bosnian, Croatian, Montenegrin and Serbian remains the following: How different are they from each other? In other words, how hard (or how easy) is it to distinguish between them?

One of the rare empirical studies that address this issue shows that Croatian and Serbian are situated at the opposing ends of the continuum, whereas Bosnian and Montenegrin tend to lean towards the one or the other depending on the considered linguistic feature (Ljubešić et al., 2018). Results from NLP, specifically on the task of Distinguishing Between Similar Languages (DSL) (Zampieri et al., 2014, 2017, 2015; Malmasi et al., 2016), seem to point in the same direction. In particular, Rupnik et al. (2023) introduce a four-class dataset for this task and evaluate two models. Model performance varies widely per class: it is perfect on Serbian and solid on Croatian, but the results are weaker on Bosnian, and low on Montenegrin.

However, contextualizing model performance remains difficult since the human upper bound has not been determined. Furthermore, the four-class test set used in the system evaluation cited above allows only a single label per instance. Previous research has shown that this can be insufficient for DSL since some instances contain no variety-specific markers (Goutte et al., 2016; Bernier-Colborne et al., 2023; Zampieri et al., 2023). Finally, the dataset was annotated by a single human annotator. This may be suboptimal and potentially calls into question the reliability of the annotation, and thus of the evaluation.

This paper presents the first large-scale multi-annotator study on distinguishing Bosnian, Croatian, Montenegrin and Serbian (BCMS). Our goal is twofold. First, we seek to consolidate the existing four-class dataset by scaling up the number of annotators and introducing a multi-label annotation. Second, we systematically examine how human performance aligns with previous observations on the relationship between these varieties as well as system performance on the DSL task.

Our contributions are as follows. (1) We **release a new gold standard set with multiple labels per instance**[2] for the DSL task on BCMS, drawing on multiple annotations per instance and an annota-

[1]Note that the number of inhabitants is not directly equivalent to the number of speakers of each variety.

[2]https://doi.org/10.5281/zenodo.10998042

Proceedings of the HumEval @LREC-COLING 2024, pages 36–46
Turin, Italy. May 21st, 2024.

tor population originating from all target countries. (2) We use this dataset to **reassess a previously proposed computational system**, investigating performance differences with respect to the original single-annotator, single-label test set. (3) We **establish the human upper bound** on this task and identify sources of annotation difficulties. (4) We **provide linguistic insights into the BCMS dialect continuum**, with multiple indicators highlighting an intermediate position of the varieties spoken in Bosnia and Herzegovina and Montenegro. To the best of our knowledge, this is the first perception study on the BCMS language area. Moreover, our contributions underline the validity of our methodology for experiments based on human annotation, independently of the tasks and languages at hand.

This paper is organized as follows. We first summarize related work (§ 2), present our annotation procedure (§ 3), and introduce the resulting dataset (§ 4). We then examine it from three perspectives: reassessing an existing DSL system (§ 5), analyzing human accuracy (§ 6), and comparing human and system performance (§ 7). We conclude with a summary and outlook (§ 8).

2. Related Work

Empirical research into the relationship between Bosnian, Croatian, Serbian and Montenegrin remains scarce. To address this issue, Ljubešić et al. (2018) conduct a corpus-based dialectometric study. The authors look at the geographical distribution of 16 linguistic variables on phonological, morphosyntactic and lexical levels. The results situate Croatian and Serbian at the opposing ends of the continuum, whereas Bosnian and Montenegrin tend to align with the one or the other depending on the variable. Furthermore, the variables do not necessarily have an even spread over the continuum or the same frequency. For example, the opposition between ekavian and ijekavian forms (e.g. ***de**te* in ekavian vs. ***dije**te* in ijekavian, meaning 'child') is a distinguishing feature for Serbian (the only of the four national standards based on both the ekavian and the ijekavian pronunciation); it is also by far the most frequent feature identified in the corpus by Ljubešić et al. (2018). This asymmetry can be expected to make some varieties harder to identify.

This hypothesis is corroborated by current results in the DSL task on these varieties. A DSL shared task has been organized regularly by the VarDial Workshops since 2014, and Bosnian, Croatian, Montenegrin and Serbian have been part of it from the very first iteration, albeit as a three-class problem focusing on Bosnian, Croatian and Serbian (Zampieri et al., 2014, 2015; Malmasi et al., 2016; Zampieri et al., 2017). In more recent work, Rupnik et al. (2023) introduce a novel benchmark,

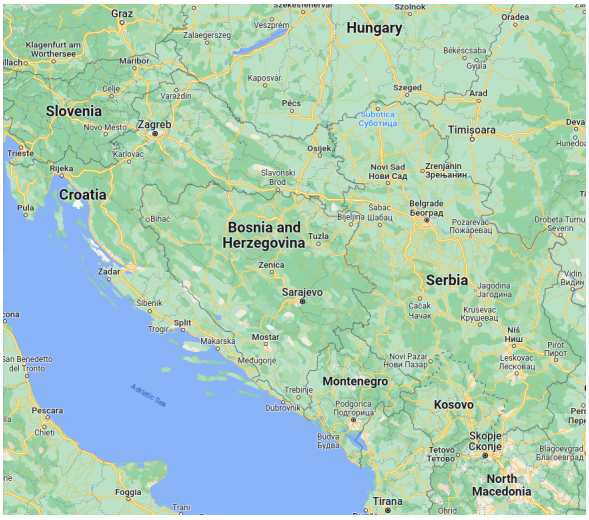

Figure 1: Bosnia and Herzegovina, Croatia, Montenegro, Serbia and neighbouring countries. Map data ©2023 GeoBasis-DE/BKG (©2009), Google.

containing two datasets: SETimes, based on newspaper texts in Bosnian, Croatian and Serbian; and a Twitter dataset containing instances in Bosnian, Croatian, Montenegrin and Serbian. Their evaluation of two DSL systems on the four-class Twitter dataset yields a global micro F1 score of 0.87 for both models, but the results vary widely per class: both models achieve 1.0 micro F1 on Serbian and 0.83 on Croatian, but the scores are somewhat lower on Bosnian (0.75–0.80) and drop significantly on Montenegrin (0.10–0.36).

Goutte et al. (2016) identify similar areas of difficulty. In their comprehensive overview of available DSL methods, the authors report that classifiers show a high degree of confusion when discriminating between Bosnian and Croatian texts. They call on six native speakers from the three countries to manually annotate the 12 most problematic instances, obtaining mean annotator accuracy of 16.6%. Some of the instances receive no correct annotations; in the gold standard, these are systematically labelled either as Croatian or as Bosnian.

These observations have contributed to a drive to redefine the DSL task. To this end, Zampieri et al. (2023) introduce a True Labels dataset for English, Spanish and Portuguese, which introduces the `both/neither` label for instances without any variety-specific markers. The dataset is annotated manually. In a similar vein, Bernier-Colborne et al. (2023) argue for framing DSL as a multi-label classification task and introduce such a dataset for four varieties of French. A model trained and evaluated on their dataset achieves an absolute gain of 0.225 on the macro F1 score on ambiguous texts.

The findings discussed above coalesce around two main points. First, the current four-class BCMS dataset would benefit from redefining the anno-

tation as multi-label. Also, to ensure annotation reliability and determine the human upper bound, the number of human annotators should be scaled up (the current version is annotated by a single annotator). Second, a systematic examination of human performance is required in order to better understand both the relationship between these four varieties and the issues faced by NLP systems.

We address these challenges as follows. We recruit 33 annotators from the four target countries and reannotate the test set from the dataset proposed by Rupnik et al. (2023). The collected annotation is used to derive a new, multi-label gold standard test set, against which we reevaluate an existing system. We measure inter-annotator agreement and determine the human upper bound on the task, thereby enabling a better contextualization of NLP system performance. Finally, we contrast human and system performance and draw conclusions about sources of difficulty and the underlying properties of the dialect continuum.

3. Annotation Process

This section describes our data collection. First, we present the original dataset on which this work is based (§ 3.1). Next, we provide details on the task definition (§ 3.2) and data preprocessing (§ 3.3). Finally, we describe the demographic structure of our annotator pool (§ 3.4).

3.1. Original Gold Standard

The original dataset was collected from the social media platform Twitter (rebranded as X in 2023) using the TweetGeo (Ljubešić et al., 2016) and TweetCaT (Ljubešić et al., 2014) tools. It has been part of the VarDial shared task on DSL since 2016 (Malmasi et al., 2016) as an out-of-domain test set for systems trained on the newspaper-based SETimes dataset. The current version of the dataset was published by Rupnik et al. (2023).

In the dataset, a single instance corresponds to the concatenation of all tweets produced by a given user. The dataset contains 614 instances (4,456,087 tokens) in total, with a strong skew towards Serbian (Table 1). Results obtained on this dataset may therefore be less reliable for the other three varieties. The instances contain 7,257 tokens on average. Occasional tweets in languages other than BCMS were not filtered out. The instances were labeled manually by a single annotator.

The dataset is split into train, dev and test in a 3:1:1 ratio. We conduct our annotation on the test set, allowing us to reevaluate a previously proposed DSL system, establish the human upper bound, and more generally assess the relevance of multi-label annotation for this task.

	Label				
Split	bs	hr	me	sr	Total
train	45	53	34	236	368
dev	15	18	11	79	123
test	15	18	11	79	123
Total	75	89	56	394	614

Table 1: Label distribution in the original gold standard across data splits. **sr** = Serbia, **hr** = Croatia, **bs** = Bosnia and Herzegovina, **me** = Montenegro.

3.2. Task Definition

The basic task in our annotation process is defined as follows: for a given instance, determine the country you think the author is from. We explicitly avoid asking the participants to identify the language of the author, since the interplay between national, ethnic and linguistic identity in this language area is complex (see e.g. Ljubešić et al., 2018). A speaker living in country A may exhibit linguistic features consistent with variety A, but self-identify as speaking variety B, C or D based on their ethnic identity. Since we are interested in the geographic spread of linguistic features independently of perceived ethnic identity, we ask for the country of origin to limit this type of bias. This is also reflective of the model we reevaluate: it was trained on top-level web domains of each country, which correspond more closely to geographic origin than to language.

Participants can provide a two-level annotation. In case of ambiguity, they are instructed to choose the country they find the most appropriate as the first-level choice, and can add multiple optional annotations as their second-level choice. This is in line with the previously discussed recent developments of VarDial DSL-TL (discriminating between similar languages – true labels) datasets for English, Spanish and Portuguese, which introduce the `neither/both` label for instances without variety-specific linguistic markers (Zampieri et al., 2023). However, the instances in these datasets are much shorter, spanning several sentences instead of hundreds of tweets per instance in our case. It is therefore much less probable to find a fully ambiguous instance in the BCMS dataset. We still include the multiple choice option, both for its linguistic relevance and to estimate annotator uncertainty.

Participants are also asked to highlight text segments on which they based their decision. They can choose between two types of segments: linguistic indicators and world knowledge. Annotation guidelines illustrate linguistic indicators with phonetic, morphological, lexical and syntactic phenomena; world knowledge pertains to country-specific named entities (TV channels, political parties, cities etc.). The guidelines explain the difference between the two types of indicators and ask for deci-

sions not to be based solely on world knowledge.

Finally, annotators are asked to mark the spot in the instance where they reached their decision. They may also report offensive content through the interface. Annotation is run using `potato` (Pei et al., 2022); a screenshot is provided in Figure 2.

3.3. Data Preparation

Unlike the original manual annotation, which was based on unaltered tweet content, we preprocess the data. We remove retweets (reposts of another user's tweet) since they are not produced by the users themselves. We also anonymize URLs and mentions in tweets by respectively replacing them with `[link]` and `@ime` (meaning 'name' in BCMS). This is done for two main reasons: to avoid priming the participants based on the content of these elements, and to improve readability for participants not familiar with Twitter. Hashtags are left unaltered, since they are often part of sentence structure. A brief description of these elements and their processing was provided in the annotation manual.

3.4. Participants

Participants were recruited through the authors' personal and professional contacts. Participants needed to be legally of age, to be native speakers of one of the four varieties, and to have spent most of their lives in one of the four countries.

Participation was not remunerated. This fact, as well as the expected duration of the task, was clearly stated both in the call for participation and the informed consent form. Prospective annotators were required to email the authors, read the task instructions and informed consent form, sign it and return it by email. Their willingness to complete this process was taken as an indicator of their motivation to participate despite the lack of remuneration. Further details are provided in the Ethics Statement (Section 10).

A total of 33 participants were recruited. A pre-annotation survey asked for participants' gender, year of birth, place of birth, current country of residence, the country in which they spent most of their lives until now, and until the age of 18.

A total of 25 participants identified as female, and 8 as male. Mean annotator age was 44.6 (SD = 12.1). In the analyses presented here, we consider the participants to come from the country in which they spent most of their lives according to the pre-annotation survey. The distribution of participants per country is given in Table 2.

Note that not all of the participants annotated the full dataset. Because participation was not remunerated, we aimed to limit the expected task duration to 1h. To this end, we split the dataset into four subsamples. The first subsample had the highest number of participants (17) and they were the most diverse. For the remaining three, most if not all participants were from Serbia.

Country	Total	S1	S2	S3	S4
Bosnia and Herz.	4	4	—	—	—
Croatia	7	6	1	—	—
Kosovo	1	—	1	—	—
Montenegro	1	1	—	—	—
Serbia	20	6	4	5	5
Total	33	17	6	5	5

Table 2: Distribution of participants by self-reported country. S1-S4: subsamples 1-4.

4. Establishing the New Gold Standard

The collected annotations were used to establish a new, multi-label gold standard. We describe how the new gold labels were determined (§ 4.1) and analyze the resulting label distribution (§ 4.2).

4.1. Resolving Annotations

Data collection ran from June to September 2023. After excluding participants who annotated less than 5 instances, the collected data contains a total of 1,098 annotations, out of which 988 were first-level annotations, and 110 were optional second-level choices. The median number of annotators per instance was 5 (min = 3, max = 17).

Inter-annotator agreement is evaluated using Krippendorff's α (Krippendorff, 1970), computed via the `Fast Krippendorff` implementation (Castro, 2017). As shown in Table 3, there are notable differences between the four subsamples, with α ranging from 0.668 on Subsample 3 to 0.893 on Subsample 1. This may be an indicator of sample difficulty, but further investigation is required to confirm this. All scores correspond to acceptable levels of agreement (Krippendorff, 2004).

We establish the new gold standard using a weighted voting strategy. The label selected as the first-level choice receives the weight of 1, and all

Subsample	α
S1	0.893
S2	0.734
S3	0.668
S4	0.768
Average	0.765

Table 3: Inter-annotator agreement measured as Krippendorff's α. $-1 \leq \alpha < 0$: inverse agreement; $\alpha = 0$: no agreement beyond chance; $0 < \alpha \leq 1$: agreement beyond chance.

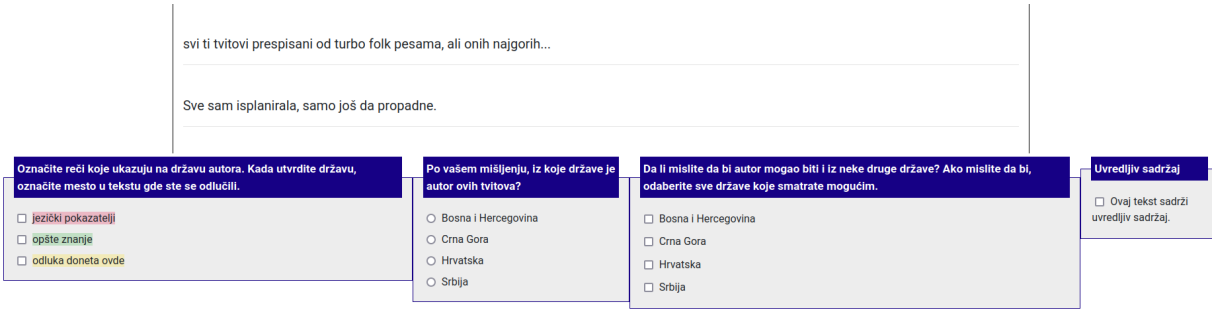

Figure 2: Annotation interface.

second-level choices receive the weight of 0.5. The votes are summed instance-level for each country and normalized by number of participants. Each country receives a final score between 0 and 1.

The gold first-level label is the one with the highest score. At this level, we do not accept multiple labels. One instance in the dataset did not receive a first-level annotation due to a tie in label scores and was excluded from the subsequent analyses. For the second-level annotation, we set a threshold at 0.2 in order to filter out labels which received low scores. In case of a tie on the second level, all labels with the second-best score are retained.

4.2. New Gold

The final label distribution in the new gold standard is given in Table 4. In the resulting annotation, 25 instances (20.3% of the dataset) have more than one label. For the instances that carry two labels, all combinations of countries are instantiated, except for the one combining Croatia and Montenegro. Note, however, that one instance in the dataset carries all four labels. This is also the only instance that has more than two labels.

Label combo	Count
sr	70
hr	16
bs	7
hr, sr	6
bs, me	5
me, sr	5
bs, hr	4
me	4
bs, sr	4
bs, hr, me, sr	1
Total	122

Labels	1st	2nd
sr	81	5
hr	18	5
bs	13	8
me	10	9
Total	122	27

Table 4: Distribution of labels in the new gold standard. Left panel: counts for all label combinations found in the new gold. Within a combination, labels are ordered alphabetically. Right panel: counts for each label as the first- and second-level choice.

Whereas Montenegro is the least frequent first-level annotation, it is the most frequent second-level choice (on 9 instances), followed by Bosnia and Herzegovina (on 8 instances). With Serbia and Croatia receiving only 5 second-level annotations each, this may point towards an uncertainty when it comes to identifying varieties from Bosnia and Herzegovina and Montenegro. This trend is explored in more detail in Section 6.

When compared to the original gold standard annotation, first-level labels differ on three instances. Two instances originally labelled as Montenegrin were relabelled as Serbian, and one instance initially annotated as Bosnian was recoded as Montenegrin. Such a low number of differences may be perceived as wasted annotation effort. However, the value of reliable annotations should not be underestimated. Moreover, the reannotation process had another goal: establishing a multi-label gold standard. This goal was achieved and its impact is evaluated in the following section. Finally, this process also allowed us to collect rich information on how humans perform on this task, which provide valuable observations laid out in Sections 6 and 7. We consider these as the silver linings of our work on the gold standard.[3]

5. System Evaluation

We examine the effect of changes to the gold standard on evaluations of DSL models. Specifically, we reevaluate the *NB Web* model introduced by Rupnik et al. (2023), which was the most robust in their evaluation. This is a Naive Bayes classifier trained on a web-based corpus using around 800 regionally distinctive words as features.

We compute the accuracy, macro-averaged and micro-averaged F1 scores using (i) the initial gold standard test set published by model authors; (ii) our reannotated test set in the single-label version; and (iii) a permissive evaluation, where a prediction is deemed correct if it corresponds to any

[3]This is also indicative of the reliability of the original annotator.

one label included in the multi-label version of our test set. The results are presented in Table 5.

Gold standard	Acc.	F1 macro	F1 micro
initial	86.9	67.7	86.9
ours (one label)	88.5	69.0	88.5
ours (all labels)	91.0	—	—

Table 5: Reevaluation of the DSL system by Rupnik et al. (2023). For comparability, initial test set results are recalculated to account for one instance excluded after reannotation.

The reannotated test set leads to a higher assessment of performance in the single-label setup (+1.6 accuracy points). Considering any label from the multi-label set as correct yields a further improvement (+4.1 accuracy points over the initial test set). These differences are overall limited – unsurprisingly, given the previously noted similarity between the initial and reannotated test sets – but they still confirm the relevance of multi-annotator and multi-label judgments on this task.

6. Human Performance

This section presents an analysis aiming to establish sources of difficulty for human annotators. We accomplish this by looking at two main indicators: annotators' accuracy as measured against the new gold annotation (§ 6.1), and their uncertainty (§ 6.2). For the latter, we rely on two indirect indicators: the presence of secondary labels and the duration of reading before the annotation decision is reached.

6.1. Accuracy

Compared against our single-label gold standard, mean participant accuracy on this task stands at 94.3 (SD = 6.2), or 5.8 points above model performance. It ranges from 76.7 to 100.0, indicating a considerable degree of variability across speakers.

To better understand the potential sources of this variability, we consider the available demographic information. We first check the effect of age under the assumption that older speakers may be better at distinguishing the varieties due to a higher degree of exposure prior to the breakup of Yugoslavia, but we find no correlation with annotator-level accuracy ($rho = -0.01$, $p = 0.97$).

We further look into the effect of the annotators' country of origin in relation to their accuracy on individual classes (Table 6). The analysis points to some intuitive patterns: for instance, speakers from Bosnia and Herzegovina obtain higher accuracy on instances labeled as coming from their country (+6.0) or from Croatia (+2.6) compared to speakers from Serbia, who are likely more susceptible to confusing those two varieties due to their shared

	Accuracy on gold labels			
Country	bs	hr	me	sr
Bosnia & Herz.	91.7	94.7	50.0	98.6
Croatia	96.4	100.0	93.3	100.0
Kosovo	75.0	80.0	100.0	100.0
Montenegro	75.0	100.0	100.0	100.0
Serbia	85.7	92.1	69.4	96.7
Overall	88.3	94.1	74.0	97.8

Table 6: Accuracy on individual gold labels cross-tabulated with annotators' self-reported countries of origin. Note that the number of annotators per country is highly variable.

frequent features (e.g. ijekavian forms). However, other patterns are less readily interpretable.

We further assess if class-level performance differs by country of origin using the Mann-Whitney–U test.[4] We compare the accuracy of annotators on a given class for one pair of countries at a time, and find no statistically significant differences.[5] The country-level trends may therefore be related to the uneven geographical distribution of annotators, but they should nevertheless be reexamined with a larger participant pool.

That said, Table 6 clearly shows that overall human performance varies strongly across the classes. Accuracy is highest on instances labeled as coming from Serbia and Croatia – the endpoints of the BCMS continuum – as opposed to those from Montenegro and Bosnia and Herzegovina. The Wilcoxon signed-rank test indicates that annotator accuracy is significantly different for all pairs of labels except Croatia and Serbia ($p = 0.129$). Variable degrees of difficulty in determining the correct label may also be reflected by other indicators of participants' uncertainty, to which we now turn.

6.2. Uncertainty

Secondary labels. Recall that participants annotated each instance using a primary country label and, optionally, one or more secondary labels. We now look into their tendency to use secondary labels as an indirect indicator of their uncertainty. Out of 988 individual annotations, 110 (11.1%) include a secondary country label. This tendency may seem overall limited; however, secondary labels were provided for 62 out of 123 annotated instances (50.4%). Furthermore, 28 out of 33 participants (84.8%) provided a secondary annotation at least once. This indicates that less-than-certain annotation decisions are in fact prominent.

[4] For all statistical significance tests, we set alpha to 0.05. Full results with individual test statistics and p-values are provided in Appendix A.

[5] We do not extend this analysis to the Kosovo and Montenegro groups as each only has one annotator.

1st choice annotations				2nd choice labels				Time to decision			Chars to decision		
Label	Total	w\ 2nd choice		bs	hr	me	sr	med.	min	max	med.	min	max
bs	119	27	(22.7%)	—	15	10	10	1’ 32”	0’ 02”	6’ 04”	1,200	0	5,239
hr	159	16	(10.1%)	8	—	5	5	1’ 12”	0’ 01”	4’ 52”	1,028	105	5,173
me	68	19	(27.9%)	10	—	—	12	1’ 37”	0’ 01”	7’ 12”	2,204	657	9,218
sr	642	48	(7.5%)	28	10	23	—	1’ 21”	0’ 01”	5’ 35”	1,366	0	7,077
Total	988	110	(11.1%)	46	25	38	27	1’ 23”	0’ 01”	7’ 12”	1,330	0	9,218
	(a)				(b)				(c)			(d)	

Table 7: Distribution of individual annotations by choice of primary country labels. Panels from left: (a) number of annotations; (b) distribution of secondary country choices (may not sum to first country totals due to multiple choices being allowed); (c) time taken to annotate an instance; (d) character index where decision was made, indicated by highlighting tweet text. Outliers excluded in panels (c) and (d).

We further examine this trend with respect to different primary country choices under the assumption that different regional varieties are not equally easy to distinguish. The results in Table 7 show a clear distinction between annotations resulting in primary labels of Serbia or Croatia, with secondary choices present in up to 10% of cases; and those of Bosnia and Herzegovina or Montenegro, where secondary choices are two to three times more frequent. This is consistent with the intermediate position of these two countries in the regional dialect continuum (previously noted in Section 6.1).

The distribution of secondary labels varies depending on the primary country, but without clear tendencies: whatever the primary country choice, most (if not all) other countries may be considered as potential alternatives. These overlaps are striking as we would expect them to more clearly pattern with similarities between the varieties. We therefore conduct a qualitative analysis to better understand the motivations for secondary choices.

Qualitative analysis. Consider the following sample tweets (normalized to include diacritics), taken from a single instance where both primary and secondary choices hesitated between Montenegro and Bosnia and Herzegovina.

(1) Ako mi nestane interneta, umrijet ću.
If I run out of internet, I will die.

(2) Komšija pošalje poruku da mu lajkujem profilnu.
A neighbor messaged me to like his profile pic.

Example (1) includes the future tense form *umrijet ću* ‘I will die’, which is atypical for most of Serbia. It is the only dialect region where this construction would generally be realized with ekavian phonological features and fully synthetically (*umreću*). Example (2) contains the lexical item *komšija* ‘neighbor’. Its use excludes Croatia, the one dialect region where the equivalent *susjed* is predominant. This would leave the annotator with the choice between the varieties of Bosnia and Herzegovina and Montenegro, which have many more shared linguistic features. In other words, the difficulty comes from insufficiently distinctive regional linguistic features.

A different pattern is illustrated by the following tweets, taken from an instance where annotators were hesitant between Montenegro and Serbia.

(3) Današnji dan – jedva čekam sjutra.
Today – I can’t wait for tomorrow.

(4) i lep i jak
both handsome and strong

Example (3) contains the form *sjutra* ‘tomorrow’. It distinguishes Montenegro from all other varieties, which have the equivalent *sutra*. But a minority of this user’s tweets contain forms typical of varieties spoken in Serbia. Example (4) includes the ekavian variant *lep* ‘pretty, handsome’, whereas in Montenegro we would expect the ijekavian *lijep*. This can be seen as codeswitching. It is often spurious (e.g. quoted song lyrics), but codeswitched instances are not systematically flagged on Twitter. Annotation is therefore complicated by linguistic features which are sufficiently distinctive on their own, but which together point to multiple regional varieties.

Duration of reading. A final type of information on annotation difficulties comes from behavioral data: the automatically recorded amount of time spent to annotate an instance; and the character index at which the decision was made, indicated by highlighting tweet text. Distribution by primary country choice is shown in Table 7, panels (c) and (d). For each variable, we use the Mann-Whitney–U test to determine whether it differs significantly across individual pairs of labels.

Annotation duration varies depending on the chosen primary label. Annotators spend less time on instances they label as Croatian or Serbian, and more on those labeled as Bosnian or Montenegrin. The difference in median annotation duration is up to 25 seconds (Croatia vs Montenegro). These differences are statistically significant in all pairs of labels, except for those with a similar status in the dialect continuum: Croatia and Serbia, and Bosnia and Herzegovina and Montenegro.

Looking at the amount of read text, it is by far the highest when labelling an instance as coming from Montenegro – up to twice more compared to the other labels. The differences are statistically significant in all pairs of labels, except when comparing Bosnia and Herzegovina – which has the second lowest median – with Croatia and with Serbia. This is a slight reversal of the previous tendency; a potential explanation is that identifying features distinctive of Bosnia and Herzegovina requires somewhat less text, but more careful consideration, compared to those typical of Serbia.

Overall, behavioral information aligns with other indicators of annotation uncertainty: varieties at the extremes of the regional dialect continuum are easier to discriminate than those with an intermediate position. We now ask whether these trends also hold for system performance.

7. Human vs. System Performance

As previously noted (§ 6.1), mean human accuracy is noticeably higher than system performance on this task. We now compare human and system performance at a finer-grained level by contrasting their respective confusion matrices (Figure 3).

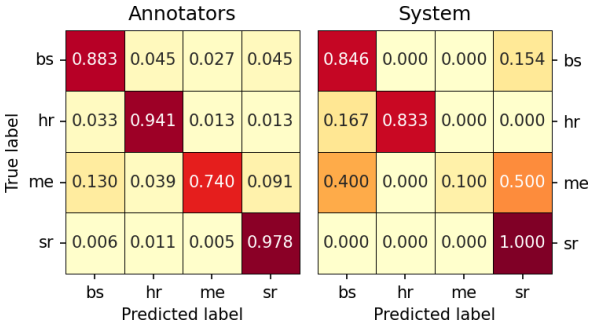

Figure 3: Confusion matrices for human and system performance. The matrix for annotators is computed on all individual annotations. The values are normalized per true label.

Both humans and the model obtain the highest results on instances labeled as coming from Serbia. The model in fact achieves perfect performance, potentially reflecting the skew in its training data.

The class with the second-highest human accuracy is Croatia, with the misclassified instances spread over all three remaining classes. The system obtains an accuracy that is over 10 points lower. Moreover, for misclassified items, it systematically falls back onto Bosnia and Herzegovina. We find a similar pattern for the class of Bosnia and Herzegovina: the system performs somewhat worse than our annotators and, unlike them, always misclassifies into the same class – in this case, Serbia.

Finally, both the annotators and the system struggle the most with instances labeled as coming from Montenegro, although to a very different extent. Our participants produce misclassifications in 26% of cases; in half of these annotations, they opt for Bosnia and Herzegovina, which again confirms the closeness of the two varieties. By contrast, the system misclassifies 90% of instances, splitting them between Bosnia and Herzegovina and Serbia.

8. Conclusions and Future Work

We have presented the first large-scale multi-annotator study on distinguishing Bosnian, Croatian, Montenegrin, and Serbian – four closely related but distinct national linguistic varieties. In order to consolidate an existing single-annotator, single-label test set for the task of Distinguishing Between Similar Languages, we scale up the number of annotators and recruit them from all target regions. This results in a multi-judgment, multi-label gold standard which allows us to analyze both system and human performance on this task.

Compared to the original test set, our reannotated version leads to a somewhat higher assessment of accuracy of an existing system (88.5, or +1.6 points, on single-label evaluation). More importantly, we establish mean human accuracy (94.3), showing that the system still lags behind it. We further identify sources of annotation difficulties using a broad range of indicators and observe consistent effects in line with the properties of the regional dialect continuum. These results may be partly due to an imbalanced geographic distribution of our annotators, but they point to important considerations which can be further validated on a larger participant sample. Specifically, instances coming from the endpoints of the dialect continuum – Croatia and Serbia – are the most accurately annotated and the easiest to judge; the reverse is true for Bosnia and Herzegovina and (especially) Montenegro, which occupy an intermediate position and have been shown to exhibit less distinctive features. Finally, a comparative error analysis shows that human misclassifications are spread across the false classes and likely explained by linguistic similarities. By contrast, the system generally falls back onto one dominant class, reflecting the label distribution in its training data.

Our results also raise questions to be explored in future work. The use of optional secondary labels in human annotation has shown that one-fifth of instances give rise to ambiguous interpretations. Formulating the DSL task as multilabel classification on these varieties would therefore more closely align model design with the perceptions of native speakers. More generally, the target varieties vary in terms of their relative annotation difficulty, with the one spoken in Montenegro proving particularly challenging. But this is also the most recently es-

tablished of the four national standards, suggesting an important role of diachronic developments. Additional annotators from as yet underrepresented countries would enable a further analysis of this and other empirically established patterns, providing novel insights into this linguistically rich region.

9. Limitations

A central aim of our study was to reannotate an existing dataset; we were therefore bound by its original class distribution. This however implies a strong skew towards data from Serbia, with Montenegro being the least frequent of the remaining three classes. This trend may have an impact on the analysis of human behavior, which could be verified through a replication study on a balanced subsample. A connected issue is the geographic skew in our annotator sample, as discussed throughout the paper. The reliability of the results is particularly affected for the Kosovo and Montenegro groups, with only one annotator each.

More generally, the annotated instances are long, with an average of over 7,000 tokens. Rather than request that annotators read all instances in their entirety – which does not seem reasonable in terms of cognitive effort – we asked them to take a decision as soon as they had seen sufficient linguistic indicators. We note that individual annotators differ with respect to the amount of text they deem necessary to read. In addition, this approach is not strictly comparable to computational models, which generally use all available text.

10. Ethics Statement

This study draws on data provided by 33 human annotators. All participants gave informed consent prior to accessing the annotation platform. The informed consent form described the task to be performed; the nature of the data to be annotated (tweets), including the risk of being exposed to potentially offensive content; the estimated duration of the task; the specific demographic information to be collected; the non-remunerated nature of participation; the right to withhold answers to any questions and to withdraw from the study at any moment; and the procedures used to anonymize and store the collected information. The participants could further freely opt into receiving the results of the study; being contacted for participation in extensions of the same study or in other similar studies; and being publicly acknowledged as participants in resulting scientific publications and dataset documentation.

We collected personal information on the participants: gender, year of birth, place of birth, and country-level residential history. We used this information to provide aggregate analyses of annotation performance and perception of different regional language varieties. Moreover, we aimed to fully respect the self-reported nature of this information. For example, in selecting their country of origin, one participant chose the option "other" and entered "Kosovo", while self-identifying as a speaker of Serbian in correspondence with the authors. We assigned this participant to the Kosovo group in line with their choice. Participant-level personal information is anonymized and securely stored. We disclose the names of a subset of participants in order to acknowledge their participation, but without linking the names to any other information. This was explicitly agreed through an opt-in procedure.

In terms of more general risks, we note that linguistic research in socially complex contexts – including areas with a history of conflict – may be instrumentalized with respect to broader societal or political issues. We stress that our research empirically examines regional patterns of language use as attested in the data we collected, without a predetermined view of the linguistic communities under study or suggestion that the observed patterns generalize to the population level.

11. Acknowledgements

We thank the anonymous reviewers for their feedback, as well as Mikko Aulamo, Yves Scherrer, and Amelie Wührl for their help in setting up the annotation platform. We are also indebted to our volunteer annotators, whose participation enabled this work: Vesna Arsenović, Katja Bilać, Bojana Damnjanović, Ljubomir Ivanović, Biljana Kaurin, Marijana Kaurin, Maida Kojić McAndrew, Irina Masnikosa, Snežana Naić, Marija Runić, Tibor Weigand, and the remaining 22 participants who wished to remain anonymous. Aleksandra Miletić was supported by Academy of Finland project number 342859. Filip Miletić was supported by DFG research grant SCHU 2580/5-1.

12. Bibliographical References

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.

Santiago Castro. 2017. Fast Krippendorff: Fast computation of Krippendorff's alpha agree-

ment measure. https://github.com/pln-fing-udelar/fast-krippendorff.

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).

Klaus Krippendorff. 1970. Estimating the reliability, systematic error and random error of interval data. *Educational and psychological measurement*, 30(1):61–70.

Klaus Krippendorff. 2004. *Content Analysis: An Introduction to Its Methodology*, second edition. SAGE publications.

Nikola Ljubešić, Darja Fišer, and Tomaž Erjavec. 2014. TweetCaT: a tool for building Twitter corpora of smaller languages. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2279–2283, Reykjavik, Iceland. European Language Resources Association (ELRA).

Nikola Ljubešić, Maja Miličević Petrović, and Tanja Samardžić. 2018. Borders and boundaries in Bosnian, Croatian, Montenegrin and Serbian: Twitter data to the rescue. *Journal of Linguistic Geography*, 6(2):100–124.

Nikola Ljubešić, Tanja Samardžić, and Curdin Derungs. 2016. TweetGeo - a tool for collecting, processing and analysing geo-encoded linguistic data. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3412–3421, Osaka, Japan. The COLING 2016 Organizing Committee.

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.

Jiaxin Pei, Aparna Ananthasubramaniam, Xingyao Wang, Naitian Zhou, Apostolos Dedeloudis, Jackson Sargent, and David Jurgens. 2022. POTATO: The portable text annotation tool. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 327–337, Abu Dhabi, UAE. 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.

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. 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.

A. Full statistical results for human behavior analysis

Label	Countries of origin		U	p
bs	Croatia	Bosnia & Herz.	15.5	0.778
	Serbia	Bosnia & Herz.	32.0	0.502
	Serbia	Croatia	50.0	0.197
hr	Croatia	Bosnia & Herz.	17.5	0.257
	Serbia	Bosnia & Herz.	37.5	0.845
	Serbia	Croatia	49.0	0.117
me	Croatia	Bosnia & Herz.	23.0	0.060
	Serbia	Bosnia & Herz.	52.0	0.344
	Serbia	Croatia	44.0	0.111
sr	Croatia	Bosnia & Herz.	17.5	0.257
	Serbia	Bosnia & Herz.	35.5	0.700
	Serbia	Croatia	49.0	0.119

Table 8: Results of the Mann-Whitney–U test comparing annotator-level accuracy for a given gold label across pairs of annotators' countries of origin. The Kosovo and Montenegro groups are limited to one annotator each and are therefore not included in the analysis.

Labels		W	p
bs	hr	26.0	0.053
bs	me	25.5	0.015
bs	sr	8.0	0.003
hr	me	10.0	0.001
hr	sr	16.0	0.129
me	sr	12.0	0.001

Table 9: Results of the paired Wilcoxon signed-rank test comparing annotator-level accuracy across pairs of gold labels.

		time to annotate		character offset	
Labels		U	p	U	p
bs	hr	11001.0	0.007	3127.0	0.242
bs	me	3839.5	0.694	773.5	0.005
bs	sr	32842.5	0.046	13522.5	0.672
hr	me	6557.5	0.009	2115.0	0.000
hr	sr	53433.0	0.205	18759.5	0.041
me	sr	18419.5	0.046	4939.0	0.006

Table 10: Results of the Mann-Whitney–U test comparing instance-level behavioral information (time taken to annotate an instance; character offset at which the decision was taken) across pairs of gold labels.