Findings of the VarDial Evaluation Campaign 2023

Noëmi Aepli[1], Çağrı Çöltekin[2], Rob van der Goot[3], Tommi Jauhiainen[4]
Mourhaf Kazzaz[2], Nikola Ljubešić[5,6], Kai North[7], Barbara Plank[8]
Yves Scherrer[4], Marcos Zampieri[7]

[1]University of Zurich, [2]University of Tübingen, [3]IT University of Copenhagen,
[4]University of Helsinki, [5]Jožef Stefan Institute, [6]University of Zagreb,
[7]George Mason University, [8]LMU Munich

Abstract

This report presents the results of the shared tasks organized as part of the VarDial Evaluation Campaign 2023. The campaign is part of the tenth workshop on Natural Language Processing (NLP) for Similar Languages, Varieties and Dialects (VarDial), co-located with EACL 2023. Three separate shared tasks were included this year: Slot and intent detection for low-resource language varieties (SID4LR), Discriminating Between Similar Languages – True Labels (DSL-TL), and Discriminating Between Similar Languages – Speech (DSL-S). All three tasks were organized for the first time this year.

1 Introduction

The workshop series on *NLP for Similar Languages, Varieties and Dialects* (VarDial), traditionally co-located with international conferences, has reached its tenth 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., 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 2023, we offered three shared tasks which we present in this paper:

- **SID4LR:** Slot and intent detection for low-resource language varieties[1]
- **DSL-TL:** Discriminating Between Similar Languages – True Labels[2]
- **DSL-S:** Discriminating Between Similar Languages – Speech[3]

DSL-TL and DSL-S continue the long line of language and dialect identification (Jauhiainen et al., 2019) shared tasks at VarDial, whereas the SID4LR features a task novel to the evaluation campaigns.

This overview paper is structured as follows: in Section 2, we briefly introduce the three shared tasks. Section 3 presents the teams that submitted systems to the shared tasks. Each task is then discussed in detail, focusing on the data, the participants' approaches, and the obtained results. Section 4 is dedicated to SID4LR, Section 5 to DSL-TL, and Section 6 to DSL-S.

2 Shared Tasks at VarDial 2023

The evaluation campaign took place in January – February 2023. Due to the ACL placing the workshop at the EACL conference in early May, the schedule from the shared tasks' first announcement to completion was relatively tight. The call for participation in the shared tasks was first published in early January, the training data sets for the shared tasks were released on January 23rd, and the results were due to be submitted on February 27th.[4]

2.1 SID for Low-resource Language Varieties (SID4LR)

The SID4LR shared task focused on Slot and Intent Detection (SID) for digital assistant data in three low-resource language varieties: Swiss German (GSW) from the city of Bern, South Tyrolean (DE-ST), and Neapolitan (NAP). Intent detection is the task of automatically classifying the intent of an utterance and slot detection aims at finding the relevant (labeled) span. Figure 1 illustrates these two tasks with an example. The objective of this shared

[1] Task organizers: Noëmi Aepli, Rob van der Goot, Barbara Plank, Yves Scherrer.

[2] Task organizers: Marcos Zampieri, Kai North, Tommi Jauhiainen.

[3] Task organizers: Çağrı Çöltekin, Mourhaf Kazzaz, Tommi Jauhiainen, Nikola Ljubešić.

[4] https://sites.google.com/view/vardial-2023/shared-tasks

Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023), pages 251–261
May 5, 2023 ©2023 Association for Computational Linguistics

English (EN)	Remind me to go to the dentist next Monday
Italian (IT)	Ricordami di andare dal dentista lunedì prossimo
Neapolitan (NAP)	**Ricuordam' 'e 'i addo dentista lunnerì prossimo**
German (DE)	Erinnere mich am nächsten Montag zum Zahnarzt zu gehen
Swiss German (GSW)	**Du mi dra erinnere nöchscht Mänti zum Proffumech zga**
South Tyrolean (DE-ST)	**Erinner mi in negschtn Muntig zin Zohnorzt zu gian**

Figure 1: Example of the SID tasks. The **three target languages (NAP, GSW, DE-ST)** are in bold, the corresponding high-resource languages (DE and IT) and the translation (EN) are included for comparison. The *slot* annotations are coloured: datetime and reminder/todo. The *intent* for this sentence is `reminder/set_reminder`.

task is to address the following question: *How can we best do zero-shot transfer to low-resource language varieties without standard orthography?*

The xSID-0.4 corpus[5], which includes data from both Snips (Coucke et al., 2018) and Facebook (Schuster et al., 2019), constitutes the training data, providing labeled information for slot and intent detection in 13 different languages. The original training data is in English, but we also provided automatic translations of the training data into German, Italian, and other languages. These translations are obtained with the Fairseq library (Ott et al., 2019), using spoken data for training (more details in van der Goot et al. (2021a)). Bleu scores (Papineni et al., 2002) were 25.93 and 44.73 for respectively German and Italian. Slot label annotations were transferred using the attention weights. Participants were allowed to use other data to train on as long as it was not annotated for SID in the target languages. Specifically, the following resources were allowed:

1. annotated data from other (related and unrelated) languages in the xSID-0.4 corpus;
2. raw text data from the target languages, if available (e.g., Wikipedia, web crawls);
3. pre-trained language models containing data from the target languages.

It was not mandatory for the participants to provide systems for all tasks and languages; they had the option to only take part in a specific subset. We used the standard evaluation metrics for these tasks, namely the span F1 score for slots and accuracy for intents.

2.2 Discriminating Between Similar Languages – True Labels (DSL-TL)

Discriminating between similar languages (e.g., Croatian and Serbian) and national language varieties (e.g., Brazilian and European Portuguese) has been a popular topic at VarDial since its first edition. The DSL shared tasks organized from 2014 to 2017 (Zampieri et al., 2017; Malmasi et al., 2016; Zampieri et al., 2015, 2014) have addressed this issue by providing participants with the DSL Corpus Collection (DSLCC) (Tan et al., 2014), a collection of journalistic texts containing texts written in groups of similar languages (e.g., Indonesian and Malay) and language varieties (e.g., Brazilian and European Portuguese).[6] The DSLCC was compiled assuming each instance's gold label is determined by where the text is retrieved from. While this is a straightforward and primarily accurate practical assumption, previous research (Goutte et al., 2016) has shown the limitations of this problem formulation as some texts may present no linguistic marker that allows systems or native speakers to discriminate between two very similar languages or language varieties.

At VarDial 2023, we tackle this important limitation by introducing the DSL True Labels (DSL-TL) shared task. DSL-TL provided participants with the DSL-TL dataset (Zampieri et al., 2023), the first human-annotated language variety identification dataset where the sentences can belong to several varieties simultaneously. The DSL-TL dataset contains newspaper texts annotated by multiple native speakers of the included language and language varieties, namely English (American and British varieties), Portuguese (Brazilian and European varieties), and Spanish (Argentinian and Peninsular varieties). More details on the DSL-TL shared task and dataset are presented in Section 5.

[5] https://bitbucket.org/robvanderg/sid4lr

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

Team	SID4LR	DSL-TL	DSL-S	System Description Paper
UBC	✓			Kwon et al. (2023)
Notre Dame	✓			Srivastava and Chiang (2023)
VaidyaKane		✓		Vaidya and Kane (2023)
ssl		✓		Hohl and Shim (2023)
UnibucNLP		✓		Gaman (2023)
SATLab		✓		

Table 1: The teams that participated in the VarDial Evaluation Campaign 2023.

2.3 Discriminating Between Similar Languages – Speech (DSL-S)

In the DSL-S 2023 shared task, participants were using the training, and the development sets from the Mozilla Common Voice (CV, Ardila et al., 2020) to develop a language identifier for speech.[7] The nine languages selected for the task come from four different subgroups of Indo-European or Uralic language families (Swedish, Norwegian Nynorsk, Danish, Finnish, Estonian, Moksha, Erzya, Russian, and Ukrainian).

The 9-way classification task was divided into two separate tracks. Only the training and development data from the CV dataset were allowed in the closed track, and no other data were to be used. This prohibition included systems and models trained (unsupervised or supervised) on any other data. On the open track, the use of any openly available (available to any possible shared task participant) datasets and models not including or trained on the Mozilla Common Voice test set was allowed. The evaluation measure used was the Macro F1 score over the nine languages.

3 Participating Teams

A total of six teams submitted runs to the SID4LR and DSL-TL tasks. Two teams registered for the DSL-S shared task, but neither provided any submissions. In Table 1, we list the teams that participated in the shared tasks, including references to the system description papers, which are published as parts of the VarDial workshop proceedings. Detailed information about the submissions is included in the task-specific sections below.

4 SID for Low-resource Language Varieties

4.1 Dataset

The xSID-0.4 corpus[8] makes up the training data and provides labeled information for slot and intent detection in 13 different languages. The xSID dataset consists of sentences from the English Snips (Coucke et al., 2018) and cross-lingual Facebook (Schuster et al., 2019) datasets, which were manually translated into 12 other languages (van der Goot et al., 2021a). There are 43,605 sentences in the English training data. The evaluation data contains 500 test sentences and 300 validation sentences per language. For the test data, we took the existing South Tyrolean (DE-ST) part of xSID (van der Goot et al., 2021a) and two novel translations created for this shared task: Bernese Swiss German (GSW) and Neapolitan (NAP). The new translations were done by native speakers of the two language varieties. They translated directly from English without seeing the Italian or German source sentences. The translations were then processed and annotated by the shared task organizers (who have passive knowledge of the two language varieties). The two steps were done according to the guidelines from the original paper by van der Goot et al. (2021a).

4.2 Participants and Approaches

UBC: Team UBC (Kwon et al., 2023) participated in both subtasks: slot and intent detection. They used several multilingual Transformer-based language models, including mBERT, XLM-R, SBERT, LaBSE, LASER, and mT0. Furthermore, they experimented with a variety of settings to improve performance: varying the source languages, combining different language models, data augmentation via paraphrasing and machine trans-

[7] Further information available at: https://dsl-s.github.io.

[8] https://bitbucket.org/robvanderg/sid4lr

lation, and pre-training on the target languages. For the latter, they made use of additional external data from various sources for all three target languages for the training.

Notre Dame: Team Notre Dame (Srivastava and Chiang, 2023) submitted a research paper to the VarDial workshop, within which they also described their participation in the intent detection subtask. The team applied zero-shot methods, i.e., they did not use any data from the target language in the training process. They fine-tuned monolingual language models[9] with noise-induced data. The noising technique they applied is similar to that of Aepli and Sennrich (2022) with three main differences: they 1) add an additional noise type: *swapping* between adjacent letters; 2) they employ higher levels of noise and include multiple copies of the fine-tuning data; and 3) remove the step of continued pre-training to avoid using any target language data.

Baseline: The baseline we provided is the same as in the original xSID paper, trained on the English data, with an updated version of MaChAmp (van der Goot et al., 2021b). The model uses an mBERT encoder and a separate decoder head for each task, one for slot detection (with a CRF layer) and one for intent classification.

4.3 Results

We evaluated the submitted systems according to accuracy for intents and according to the span F1 score for slots (where both span and label must match exactly). Table 2 contains the scores.

For intent classification, the winner for all three languages is the team Notre Dame. Both teams beat the baseline by a large margin. All systems reached the highest scores on DE-ST and the lowest scores on GSW, but both participating teams managed to significantly close the gaps between the languages compared to the baseline.

For slot detection, the UBC team outperformed the baseline for DE-ST and GSW but not for NAP. Again, GSW turned out to be the most difficult language variety of the three. We must note, however, that the UBC submission contained a large amount of ill-formed slots. Between 13% (DE-ST, NAP) and 28% (GSW) of predicted slots start with an `I-` label instead of `B-`; the evaluation script simply ignores such slots. Furthermore, a small number of predicted spans have inconsistent labels (e.g., `I-datetime` immediately followed by `I-location`). This suggests that the model architecture chosen by the UBC team was not appropriate for span labeling tasks and that a different architecture could have led to further improvements compared to the baseline. The baseline system, which uses a CRF prediction layer, did not produce any such inconsistencies.

		Baseline	UBC	Notre Dame
Intent detection	**DE-ST**	0.6160	0.8940	**0.9420**
	GSW	0.4720	0.8160	**0.8860**
	NAP	0.5900	0.8540	**0.8900**
Slot detection	**DE-ST**	0.4288	**0.4692**	–
	GSW	0.2530	**0.2899**	–
	NAP	**0.4457**	0.4215	–

Table 2: Results for intent classification (accuracy) and slot detection (Span-F1 score). UBC submitted several models for intent detection, and here we report their best-performing system for each language.

4.4 Summary

The UBC submissions are based on a pre-trained multilingual language model (mT0), which was fine-tuned on the 12 languages of the xSID dataset. Among these languages are Italian and German, but all training sets except the English one have been produced by machine translation. This setup worked better than using only the related languages of xSID (IT and DE) or only English. Also, further data augmentation with paraphrasing and machine translation did not have any positive effect. These findings suggest that task-specific knowledge is more important than having access to linguistic material in the target languages (or even in related high-resource languages).

The Notre Dame participation provides a somewhat contrasting result. They start with a monolingual BERT model of the related high-resource language (IT or DE) and use fine-tuning to make the model more robust to character-level noise. The possibility of including unrelated languages was not explored here.

The contributions proposed by the participants are thus largely complementary, and it would be interesting to see if their combination leads to further improvements on the task. For instance, task-specific fine-tuning (using all of the xSID data)

[9] German BERT: https://huggingface.co/dbmdz/bert-base-german-uncased and Italian BERT: https://huggingface.co/dbmdz/bert-base-italian-uncased

could be combined with language-specific fine-tuning (based on the noise induction task) and complemented with the baseline's CRF architecture to provide consistent slot labels.

A striking finding of this shared task are the poor results on Swiss German compared to the other two low-resource varieties, Neapolitan and South-Tyrolean German. This may be due to the particular Swiss German dialect used in this dataset and/or to some translator-specific preferences or biases. Further analysis will be required to fully explain these differences.

5 Discriminating Between Similar Languages – True Labels

The DSL-TL shared task contained two tracks:

- **Track 1 – Three-way Classification:** In this track, systems were evaluated with respect to the prediction of all three labels for each language, namely the variety-specific labels (e.g., PT-PT or PT-BR) and the common label (e.g., PT).

- **Track 2 – Binary Classification:** In this track, systems were scored only on the variety-specific labels (e.g., EN-GB, EN-US).

In addition to the two tracks mentioned above, we provided participants with the option of using external data sources (open submission) or only the DSL-TL dataset (closed submission).

5.1 Dataset

Data DSL-TL contains 12,900 instances split between three languages and six national language varieties, as shown in Table 3. Instances in the DSL-TL are short extracts (1 to 3 sentences long) from newspaper articles randomly sampled from two sources (Zellers et al., 2019; Tan et al., 2014). Considering the source's ground truth label, the DSL-TL creators randomly selected 2,500 instances for each Portuguese and Spanish variety and 1,500 instances for each English variety.

Annotation DSL-TL was annotated using crowdsourcing through Amazon Mechanical Turk (AMT).[10] The annotation task was restricted to annotators based on the six national language variety countries, namely Argentina, Brazil, Portugal, Spain, United Kingdom, and the United States. The annotators were asked to label each instance with what they believed to be the most representative variety label, namely European (pt-PT) or Brazilian Portuguese (pt-BR), Castilian (es-ES) or Argentine Spanish (es-AR), and British (en-GB) or American English (en-US). The label distributions are shown in Table 3. The annotators were presented with three choices: (1) language variety A, (2) language variety B, or (3) both or neither for cases in which no clear language variety marker (either linguistic or named entity) was present in the text. The annotator agreement calculations and filtering carried out after the annotation stage are described in detail in the dataset description paper (Zampieri et al., 2023). Finally, the instances in DSL-TL have been split into training, development, and testing partitions, as shown in Table 4.

5.2 Participants and Approaches

Four teams provided submissions to the shared task.

VaidyaKane: All submissions from the team VaidyaKane used a pre-trained multilingual XLM-RoBERTa fine-tuned to language identification[11] to classify the language of the sentence (Conneau et al., 2020b). After the initial language identification, they experimented with several language-specific BERT models to identify the exact variety. Their best submission on track one used "bert-base-uncased"[12] for English (Devlin et al., 2019), "bertin-project/bertin-roberta-base-spanish"[13] for Spanish (la Rosa et al., 2022), and "neuralmind/bert-base-portuguese-cased"[14] for Portuguese (Souza et al., 2020). On track two, the models for Spanish and Portuguese were the same, but "roberta-base"[15] was used for English (Liu et al., 2019).

ssl: Team ssl submitted one submission to each of the four track combinations. For the closed tracks, they trained an SVM classifier using TF-IDF weighted character n-grams from one to four and word n-grams from one to two. On the open

[10] https://www.mturk.com/

[11] https://huggingface.co/papluca/xlm-roberta-base-language-detection

[12] https://huggingface.co/bert-base-uncased

[13] https://huggingface.co/bertin-project/bertin-roberta-base-spanish

[14] https://huggingface.co/neuralmind/bert-base-portuguese-cased

[15] https://huggingface.co/roberta-base

Language	**Variety A**	**Variety B**	**Both/Neither**	**Total**
Portuguese	1,317 (pt-PT)	3,023 (pt-BR)	613 (pt)	4,953
Spanish	2,131 (es-ES)	1,211 (es-AR)	1,605 (es)	4,947
English	1,081 (en-GB)	1,540 (en-US)	379 (en)	3,000
Total				**12,900**

Table 3: DSL-TL's class splits and the total number of instances.

Variety	**Train**	**Dev**	**Test**	**Total**
Portuguese	3,467	991	495	4,953
Spanish	3,467	985	495	4,947
English	2,097	603	300	3,000
Total				12,900

Table 4: DSL-TL's train, dev, and test splits are 70/20/10% of the total number of instances, respectively.

tracks, they also used names of people obtained from Wikidata (Vrandečić and Krötzsch, 2014).

UnibucNLP: On track one, the UnibucNLP team submitted a run using an XGBoost stacking ensemble (Chen and Guestrin, 2016). The classifier stack for the ensemble consisted of one SVM and one KRR classifier. For track two, the stack classifiers were the same, but Logistic Regression was used for the stacking ensemble.

SATLab: On both tracks, the SATLab team used a Logistic Regression classifier from the LIBLinear package with character n-grams from one to five weighted by BM25 and L2 normalization. The n-grams had to appear in at least two different sentences in the training data. The system was very similar to the one used by Bestgen (2021) in the Dravidian Language Identification (DLI) shared task in 2021 (Chakravarthi et al., 2021).

5.3 Results

Tables 5 to 8 show the recall, precision, and F1 scores for the baselines and best submissions for all track combinations.

Rank	**Model**	**R**	**P**	**F1**
	baseline-mBERT	0.5490	0.5450	0.5400
	baseline-XLM-R	0.5280	0.5490	0.5360
1	run-3-UnibucNLP	0.5291	0.5542	0.5318
	baseline-NB	0.5090	0.5090	0.5030
2	run-1-SATLab	0.4987	0.4896	0.4905
3	run-1-ssl	0.4978	0.4734	0.4817

Table 5: The macro average scores of the best run for each team on **closed track 1**.

Rank	**Model**	**R**	**P**	**F1**
	baseline-ANB	0.8200	0.7990	0.7990
	baseline-NB	0.8110	0.7920	0.7940
	baseline-XLM-R	0.7830	0.7820	0.7800
1	run-1-ssl	0.7521	0.7885	0.7604
	baseline-mBERT	0.7600	0.7530	0.7550
2	run-2-SATLab	0.7520	0.7430	0.7452
3	run-1-UnibucNLP	0.6502	0.7756	0.6935

Table 6: The macro average scores of the best run for each team on **closed track 2**.

Rank	**Model**	**R**	**P**	**F1**
1	run-3-VaidyaKa	0.5962	0.5866	0.5854
2	run-1-ssl	0.4937	0.5068	0.4889

Table 7: The macro average scores of the best run for **open track 1**.

Rank	**Model**	**R**	**P**	**F1**
1	run-1-VaidyaKa	0.8705	0.8523	0.8561
	baseline-NB	0.8200	0.8030	0.8030
2	run-1-ssl	0.7647	0.7951	0.7729

Table 8: The macro average scores of the best run for each team on **open track 2**.

Team UnibucNLP (Gaman, 2023) achieved the first place out of nine submissions on the closed version of track one. Their XGBoost stacking ensemble attained an F1 score of 0.5318. The results were still slightly worse than the multilingual BERT[16] (mBERT) (Devlin et al., 2019) and the XLM-RoBERTa[17] (XLM-R) (Liu et al., 2019) baselines. All other submissions achieved slightly worse F1 scores. In the second place, team SATLab's logistic regressor obtained an F1 score of 0.4905. In third place, team ssl's SVM produced an F1 score of 0.4817. The similarity between the top three F1 scores shows that automatically differentiating between similar language varieties is a challenging task, especially when taking into consideration neutral labels (EN, ES, or PT), as well as only using the provided data.

[16] mBERT: `https://huggingface.co/bert-base-multilingual-cased`

[17] XLM-R: `https://huggingface.co/xlm-roberta-base`

Team ssl (Hohl and Shim, 2023) achieved the best performance out of ten submissions on the closed version of track two. Their SVM was able to more effectively differentiate between six labels that did not include the aforementioned neutral labels (en-GB, en-US, es-AR, es-ES, pt-PT, or pt-BR). They achieved an F1 score of 0.7604. Their results were closely followed by the performance of SATLab's logistic regressor, having attained an F1 score of 0.7452, and UnibucNLP's XGBoost stacking ensemble with an F1 score of 0.6935. All submissions were clearly behind the adaptive and traditional Naive Bayes baselines, which were identical to the systems winning the Identification of Languages and Dialects of Italy (ITDI) shared task in 2022 (Jauhiainen et al., 2022a; Aepli et al., 2022). SVMs are well-known to perform well when there is a clear distinction between class boundaries. This likely explains why team ssl's SVM has outperformed UnibucNLP's ensemble since neutral labels that contained features of both classes were no longer considered.

Team VaidyaKane's (Vaidya and Kane, 2023) submission to the open version of track 1 outperformed all other open and closed submissions for this track. Their two-stage transformer-based model achieved an F1 score of 0.5854. Team ssl was the only other team to submit predictions for open tracks 1 and 2. Their open submission for track 1 achieved an F1 score of 0.4889 which surpassed that of their closed submission for this track. The use of additional data was, therefore, found to improve overall performances.

Team VaidyaKane produced the highest F1 score on the open version of track 2. They achieved an F1 score of 0.8561, which was greater than all other open and closed submissions for either track. Team ssl also saw a further improvement in their SVM's model performance when using additional data for track 2. Their SVM model produced an F1 score of 0.7729, which was superior to their closed-track submission. These performances show that the use of additional data is beneficial and further proves that the classification of language varieties is an easier task than the classification of language varieties with neutral labels.

5.4 Summary

The DSL-TL shared task introduced a novel problem formulation in language variety identification. The new human-annotated dataset with the presence of the 'both or neither' class represent a new way of looking at the problem. Given the similarity between language varieties, we believe this new problem formulation constitutes a fairer way of evaluating language identification systems, albeit rather challenging in terms of performance as demonstrated in this shared task.

6 Discriminating Between Similar Languages – Speech

6.1 Dataset

The DSL-S shared task uses Mozilla Common Voice data (version 12 released in Dec 2022) in 9 languages from two language families. The data comes from volunteers reading a pre-selected set of sentences in each language. The audio is recorded through a web-based interface. For training and development sets, we follow the training and development set of the source data. Even though the test data used in this task comes from the Common Voice test data for the nine languages, we do not use the entire test set of the CV release but sample 100 audio files for each language. There is no overlap of sentences and speakers between the data sets. Table 9 presents the test set's statistics. The total amount of unpacked speech data is around 15 gigabytes. The data includes severe class imbalance, as well as substantial differences in the number of speakers. Generalization from a small number of speakers is a known challenge in similar speech data sets, including earlier VarDial evaluation campaigns.[18] The CV data set makes this task further challenging since the variety of speakers in the test set is much larger than the training and the development sets.

Similar to the earlier VarDial shared tasks with audio data (Zampieri et al., 2017, 2018, 2019), we provided 400-dimensional i-vector and 512-dimensional x-vector features, both extracted using Kaldi (Povey et al., 2011). Unlike earlier tasks, however, the raw audio data was also available to the potential participants.

6.2 Participants and Approaches

Two teams registered for the shared task, but neither provided any submissions. In this section, we briefly introduce the baselines we provided. For the closed track, we provided a linear SVM baseline with x-vectors features (Snyder et al., 2018). The

[18] See Jauhiainen et al. (2018) and Wu et al. (2019) for earlier approaches to this problem.

	Train			Dev			Test		
	n	spk	duration	n	spk	duration	n	spk	duration
DA	2734	3	3:17:38	2105	10	2:50:46	100	48	0:07:50
ET	3137	221	5:49:04	2638	167	4:57:54	100	88	0:11:12
FI	2121	3	2:43:47	1651	13	1:59:23	100	63	0:07:46
MDF	173	2	0:15:39	54	1	0:04:39	100	7	0:08:40
MYV	1241	2	1:58:26	239	1	0:22:55	100	9	0:09:07
NO	314	3	0:22:43	168	4	0:13:28	100	18	0:07:35
RU	26043	252	37:16:50	10153	394	15:23:17	100	98	0:09:15
SV	7421	22	8:11:54	5012	73	5:32:33	100	89	0:07:24
UK	15749	28	18:38:31	8085	103	10:58:25	100	28	0:08:22

Table 9: Number of instances (n), number of speakers (spk) and total duration (hour:minute:seconds) for each split of the DSL-S shared task. The speaker numbers are approximated based on client id detection by CV.

System	**P**	**R**	**F1**
SVM + x-vectors	0.0914	0.1189	0.0876
XLS-R	0.6736	0.5953	0.5856
XLS-R + NB	0.7331	0.7167	0.7031

Table 10: Baseline scores of the DSL-S shared task.

SVM baseline was implemented using scikit-learn (Pedregosa et al., 2011), and tuned only for the SVM margin parameter ‘C’. The open track baseline uses two baselines - the XLS-R multilingual pre-trained transformer speech model (Conneau et al., 2020a)[19] with a classification head for direct speech classification, and a multilingual speech recognition system [20] based on XLS-R (Babu et al., 2021) to transcribe the speech, and uses Naive Bayes (Jauhiainen et al., 2022a,b) to identify the language.[21]

6.3 Results

The scores for the baselines are presented in Table 10. The SVM baseline performs particularly badly on the test set (the development precision, recall, and F1 scores are 0.4088, 0.4011, 0.3777, respectively). The reason behind this is likely due to the fact that, although they were used for language identification in earlier research, the x-vectors are designed for speaker identification. Given the variability of speaker features in the test set, any classifier relying on speaker features are likely to fail. The baselines relying on pre-trained transformer models perform substantially better, with the direct speech classifier being more than 10 points behind the transcription and text classification approach. While the direct speech classification approach could be further improved through hyperparameter optimisation (currently we fine-tune for 3 epochs with a batch size of 24 and a learning rate of 1e-04) and a selection of the layer from which the features are extracted (related work suggests that lower transformer layers are more informative for discriminating between languages (Bartley et al., 2023)), these baseline results show that transcription and text classification might still be a shorter path to a reasonably performing system for discriminating between similar languages than direct speech classification.

6.4 Summary

Although we did not have any submissions for this shared task, we believe that the task includes many interesting challenges. Based only on our baseline results, identifying languages from a limited amount of data (without pre-trained speech models) seems challenging, yet this is particularly interesting for low-resource settings and for investigating differences and similarities for closely related language varieties. We hope to see more interest in the community for language/dialect identification from speech.

7 Conclusion

This paper presented an overview of the three shared tasks organized as part of the VarDial Evaluation Campaign 2023: Slot and intent detection for low-resource language varieties (SID4LR), Discriminating Between Similar Languages – True La-

[19] https://huggingface.co/facebook/wav2vec2-large-xlsr-53

[20] https://huggingface.co/voidful/wav2vec2-xlsr-multilingual-56

[21] https://github.com/tosaja/TunPRF-NADI

bels (DSL-TL), and Discriminating Between Similar Languages – Speech (DSL-S).

Acknowledgements

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

The work related to the SID4LR shared task has received funding from the Swiss National Science Foundation (project nos. 191934 and 176727) and ERC Grant 101043235. The work related to the DSL-TL and DSL-S shared tasks has received partial funding from the Academy of Finland (funding decision no. 341798). The work related to the DSL-S shared task has received funding from the Slovenian Research Agency within the research project J7-4642 and the research programme P6-0411.

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 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.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4218–4222, Marseille, France. European Language Resources Association.

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, et al. 2021. XLS-R: Self-supervised cross-lingual speech representation learning at scale. *arXiv preprint arXiv:2111.09296*.

Travis M. Bartley, Fei Jia, Krishna C. Puvvada, Samuel Kriman, and Boris Ginsburg. 2023. Accidental learners: Spoken language identification in multilingual self-supervised models.

Yves Bestgen. 2021. Optimizing a supervised classifier for a difficult language identification problem. In *Proceedings of the Eighth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 96–101, Kiyv, Ukraine. 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.

Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 785–794, New York, NY, USA. Association for Computing Machinery.

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. 2020a. Unsupervised cross-lingual representation learning for speech recognition.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020b. 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.

Alice Coucke, Alaa Saade, Adrien Ball, Théodore Bluche, Alexandre Caulier, David Leroy, Clément Doumouro, Thibault Gisselbrecht, Francesco Caltagirone, Thibaut Lavril, et al. 2018. Snips voice platform: an embedded spoken language understanding system for private-by-design voice interfaces. *arXiv preprint arXiv:1805.10190*.

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.

Mihaela Gaman. 2023. Using ensemble learning in language variety identification. In *Proceedings of the Tenth Workshop on NLP for Similar Languages, Varieties and Dialects*, Dubrovnik, Croatia. 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).

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).

Fritz Hohl and Soh-Eun Shim. 2023. Vardial in the wild: Industrial applications of lid systems for closely-related language varieties. In *Proceedings of the Tenth Workshop on NLP for Similar Languages, Varieties and Dialects*, Dubrovnik, Croatia. Association for Computational Linguistics.

Tommi Jauhiainen, Heidi Jauhiainen, and Krister Lindén. 2018. HeLI-based experiments in Swiss German dialect identification. In *Proceedings of the Fifth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2018)*, pages 254–262, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Tommi Jauhiainen, Heidi Jauhiainen, and Krister Lindén. 2022a. Italian language and dialect identification and regional French variety detection using adaptive naive Bayes. In *Proceedings of the Ninth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 119–129, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Tommi Jauhiainen, Heidi Jauhiainen, and Krister Lindén. 2022b. Optimizing naive Bayes for Arabic dialect identification. In *Proceedings of the The Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 409–414, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

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.

Sang Yun Kwon, Gagan Bhatia, ElMoatez Billah Nagoudi, Alcides Alcoba Inciarte, and Muhammad Abdul-Mageed. 2023. Sidlr: Slot and intent detection models for low-resource language varieties. In *Proceedings of the Tenth Workshop on NLP for Similar Languages, Varieties and Dialects*, Dubrovnik, Croatia. Association for Computational Linguistics.

Javier De la Rosa, Eduardo G. Ponferrada, Manu Romero, Paulo Villegas, Pablo González de Prado Salas, and María Grandury. 2022. Bertin: Efficient pre-training of a spanish language model using perplexity sampling. *Procesamiento del Lenguaje Natural*, 68(0):13–23.

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.

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.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

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.

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.

Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, and Karel Vesely. 2011. The kaldi speech recognition toolkit. In *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society.

Sebastian Schuster, Sonal Gupta, Rushin Shah, and Mike Lewis. 2019. Cross-lingual transfer learning for multilingual task oriented dialog. 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 3795–3805, Minneapolis, Minnesota. Association for Computational Linguistics.

David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur. 2018. X-Vectors: Robust DNN embeddings for speaker recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5329–5333.

Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2020. BERTimbau: pretrained BERT models for Brazilian Portuguese. In *9th Brazilian Conference on Intelligent Systems, BRACIS, Rio Grande do Sul, Brazil, October 20-23 (to appear)*.

Aarohi Srivastava and David Chiang. 2023. Fine-tuning bert with character-level noise for zero-shot transfer to dialects and closely-related languages. In *Proceedings of the Tenth Workshop on NLP for Similar Languages, Varieties and Dialects*, Dubrovnik, Croatia. Association for Computational Linguistics.

Liling Tan, Marcos Zampieri, Nikola Ljubesic, and Jörg Tiedemann. 2014. Merging comparable data sources for the discrimination of similar languages : The dsl corpus collection. In *International Conference on Language Resources and Evaluation*.

Ankit Vaidya and Aditya Kane. 2023. Two-stage pipeline for multilingual dialect detection. In *Proceedings of the Tenth Workshop on NLP for Similar Languages, Varieties and Dialects*, Dubrovnik, Croatia. Association for Computational Linguistics.

Rob van der Goot, Ibrahim Sharaf, Aizhan Imankulova, Ahmet Üstün, Marija Stepanović, Alan Ramponi, Siti Oryza Khairunnisa, Mamoru Komachi, and Barbara Plank. 2021a. 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.

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.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: A free collaborative knowledgebase. *Commun. ACM*, 57(10):78–85.

Nianheng Wu, Eric DeMattos, Kwok Him So, Pin-zhen Chen, and Çağrı Çöltekin. 2019. Language discrimination and transfer learning for similar languages: Experiments with feature combinations and adaptation. In *Proceedings of the Sixth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 54–63, Ann Arbor, Michigan. 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, 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.

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019. Defending against neural fake news. *Advances in neural information processing systems*, 32.