Findings of the VarDial Evaluation Campaign 2025: The NorSID Shared Task on Norwegian Slot, Intent and Dialect Identification

Yves Scherrer
Language Technology Group
University of Oslo, Norway
yves.scherrer@ifi.uio.no

Rob van der Goot
Computer Science
IT University of Copenhagen
robv@itu.dk

Petter Mæhlum
Language Technology Group
University of Oslo, Norway
pettemae@ifi.uio.no

Abstract

The VarDial Evaluation Campaign 2025 was organized as part of the twelfth workshop on Natural Language Processing for Similar Languages, Varieties and Dialects (VarDial), colocated with COLING 2025. It consisted of one shared task with three subtasks: intent detection, slot filling and dialect identification for Norwegian dialects. This report presents the results of this shared task. Four participating teams have submitted systems with very high performance ($>$ 97% accuracy) for intent detection, whereas slot detection and dialect identification showed to be much more challenging, with respectively span-F1 scores up to 89%, and weighted dialect F1 scores of 84%.

1 Introduction

The workshop series on *NLP for Similar Languages, Varieties and Dialects* (VarDial), now at its twelfth edition, has traditionally hosted an evaluation campaign with shared tasks on various topics such as language and dialect identification, commonsense reasoning, question answering, and cross-lingual tagging and parsing. The shared tasks have featured many languages and dialects from different families and data from various sources, genres, and domains (Chifu et al., 2024; Aepli et al., 2023, 2022; Chakravarthi et al., 2021; Gaman et al., 2020; Zampieri et al., 2019, 2018, 2017; Malmasi et al., 2016; Zampieri et al., 2015, 2014).

The VarDial Evaluation Campaign 2025 consisted of the NorSID shared task, which focused on slot filling, intent detection and dialect identification for Norwegian dialectal data. As digital assistants are becoming more widespread, it is important that they can support a wide variety of language varieties. Where other work has focused on supporting a wider range of languages (e.g. Xu et al., 2020; FitzGerald et al., 2023), we instead focus on dialects, which has shown to be challenging for slot and intent detection systems (van der Goot et al., 2021a; Aepli et al., 2023; Winkler et al., 2024).

The NorSID shared task included three subtasks: slot filling, intent detection, and dialect classification. Each participating team was allowed to send in three submissions per subtask. It was not mandatory for the participants to provide systems for all tasks; they had the option to only take part in a specific subtask.

2 Related Work

NLP for dialects and language varieties has been a long-standing research topic, and the VarDial workshop series has contributed substantially to its popularity. Nevertheless, although important advances have been made in recent years thanks to neural architectures and large language models, engaging with linguistic variation remains one of the crucial open research questions within NLP. Several surveys summarize the state-of-the-art in NLP for dialects: Zampieri et al. (2020) summarizes the various research directions in NLP for dialects that were explored in earlier VarDial editions and introduces the reader to key issues in dialectology and sociolinguistics. Joshi et al. (2024) provides an updated perspective on NLP for dialects.

A large number of previous VarDial shared tasks focused on **language identification**, either for national varieties of pluricentric languages, or for dialects and closely related languages. The former includes tasks of discriminating between British and American English (DSL, Chifu et al., 2024; Aepli et al., 2023; Malmasi et al., 2016; Zampieri et al., 2015, 2014), or between French spoken in Belgium, Canada, France and Switzerland (FDI, Chifu et al., 2024; Aepli et al., 2022), to name but a few. The latter includes the identification of various Swiss German dialects (GDI, Zampieri et al., 2018, 2017), or of the different regional languages spoken in Italy (ITDI, Aepli et al., 2023). The di-

Proceedings of the Twelfth Workshop on NLP for Similar Languages, Varieties and Dialects, pages 1–8
January 19, 2025.

alect identification subtask of this year's NorSID task falls in the latter category. An overview of the history of language identification and its challenges can be found in Jauhiainen et al. (2019).

The two other subtasks of NorSID focus on **intent classification and slot filling** for task-oriented dialog systems, a task also sometimes referred to as **spoken language understanding**. Three recent surveys provide excellent introductions to the topic: Louvan and Magnini (2020) and Weld et al. (2022) focus mainly on methods, whereas Larson and Leach (2022) survey the available datasets. Aspects of dialectal variation and cross-lingual transfer between closely related varieties have been discussed in the SID4LR shared task at VarDial 2023 (Aepli et al., 2023), which focused on South Tyrolian and Swiss German dialects as well as Neapolitan, a language closely related to Italian.

3 Data

The data used in the NorSID shared task is taken from the NoMusic corpus, which is the Norwegian extension of the xSID dataset. We present these resources below. Table 1 provides an overview of the dataset sizes.

xSID The multilingual xSID dataset was introduced by van der Goot et al. (2021a). It consists of prompts for digital assistants taken from the English Snips (Coucke et al., 2018) and cross-lingual Facebook (Schuster et al., 2019) datasets, which were manually translated and re-annotated into 13 language varieties. xSID continues to be updated with additional languages: two languages (Neapolitan and Swiss German) were added in the context of the SID4LR shared task at VarDial 2023 (Aepli et al., 2023), and two languages (Bavarian German and Lithuanian) by Winkler et al. (2024).

The data in xSID is partitioned into 43,605 sentences for training, 300 for development and 500 for testing. The native English data is translated into the other languages, automatically in the case of the training set, and by humans in the case of the development and test sets.

NoMusic Since xSID currently does not cover Norwegian, the NoMusic corpus project (Mæhlum and Scherrer, 2024) was started to fill this gap. It complements xSID with several Norwegian versions, taking into account the prevalence of dialects (and dialect writing) in Norway. NoMusic contains translations of the English xSID development and test sets both into standard Norwegian Bokmål and into the dialects of ten native speakers of Norwegian who regularly write in these dialects.

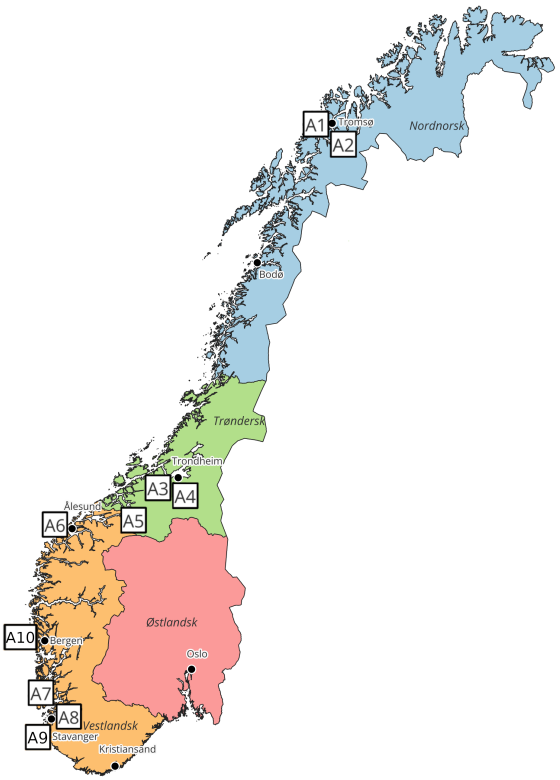

Figure 1: Map of Norway with the origins of the ten dialect translators (*A1* to *A10*). The colors represent the four major dialect areas.

Figure 1 shows the origins of the dialect speakers. 2 translators write in Northern dialects (*N*, blue on the map), 3 translators write in Central Norwegian dialects (*T* for *Trøndersk*, green) and 5 translators write in Western dialects (*V* for *Vestnorsk*, orange). None of the translators write in an Eastern dialect (red on the map), but it is common in this area to write in standard Bokmål. Therefore, the Bokmål translation can be viewed to some extent as representative of the writing traditions in Eastern Norway.

The NorSID training data As there was no training data for any Norwegian varieties, we followed the procedure from van der Goot et al. (2021a) to generate training data from the original xSID English training data using machine translation and annotation transfer. The machine translation model was trained on the Norwegian OpenSubtitles data[1],

[1] https://object.pouta.csc.fi/OPUS-OpenSubtitles/v2018/moses/en-no.txt.zip,

```
# id = 33/8
# text = Kor varmt skal det ver i dag?
# intent = weather/find
# dialect = V
1   Kor     weather/find   O
2   varmt   weather/find   B-weather/attribute
3   skal    weather/find   O
4   det     weather/find   O
5   ver     weather/find   O
6   i       weather/find   B-datetime
7   dag     weather/find   I-datetime
8   ?       weather/find   O
```

Figure 2: Example sentence with sentence-level annotation (`intent`, `dialect`) and token-level slot annotation (*i dag* of type `datetime`). The `id` field tells that it is sentence 38 from translator *A8*. It was translated from the English sentence *How warm will it be today?*

as it was the largest open parallel data based on transcribed speech. We used the FairSeq toolkit v0.9.0 with default hyperparameters, matching the original xSID setup, and relied on the attention weights for transferring the slot labels, which were afterwards automatically corrected to valid BIO sequences (i.e. first I becomes a B, and if there is a label mismatch in the span, the B-label is used). It should be noted that the automatic mapping of the slot labels led to some incorrect labeling in the target language. We also noted that the machine translation quality was relatively poor overall with a BLEU score of 18.46 (sacreBLEU on word-segmented texts). The machine-translated training set is only available in Norwegian Bokmål, not in any of the dialects covered by NoMusic (nor in the other written Norwegian norm, Nynorsk).

The NorSID development and test sets For the purpose of the shared task, we concatenated and shuffled all eleven versions of the NoMusic data, keeping intact the division into development and test sets. Furthermore, we annotated each prompt with the dialect label (N, T, V, or B for Bokmål). An example is shown in Figure 2. In the development set, we also provide a unique sentence identifier (`33/8` in the example) that determines the content (all sentences with number `33` have the same meaning) and the translator (all sentences with `/8` were produced by translator A8).

The blind test set provided to the participants consisted of the `# text` line and the first two columns of the tokenized format.

[1] http://www.opensubtitles.org/

Split	Sentences	Unique	B	N	T	V
Train	43,605	33,408	1 (MT)	–	–	–
Dev	3,300	2,736	1	2	3	5
Test	5,500	4,477	1	2	3	5

Table 1: Overview of the data used in the NorSID shared task. *Sentences* refers to the total number of sentences per split, *Unique* to the number of unique lower-cased sentences. *B, N, T, V* lists the number of translations into the four varieties (Bokmål, Nordnorsk, Trøndersk, Vestnorsk, respectively).

Team	Slots	Intents	Dialect	Reference
HiTZ	✓	✓	✓	Bengoetxea et al. (2025)
MaiNLP	✓	✓		Blaschke et al. (2025)
LTG	✓	✓		Midtgaard et al. (2025)
CUFE		✓	✓	Ibrahim (2025)

Table 2: The teams that participated in the VarDial Evaluation Campaign 2025.

Evaluation We used the standard evaluation metrics for the three tasks, namely the span F1 score for slots, accuracy for intents, and weighted F1 score for dialect classification.

The English source data in xSID is characterized by a considerable number of duplicates, and the number of duplicates further increased whenever several dialect translators produced the same translation (see Table 1). For the slot and intent evaluation, we did not perform any duplicate removal to maintain comparability with other results reported on this dataset. In contrast, the dialect identification evaluation is based on *unique lowercased sentences*, each of which is associated with a set of labels. The F1 score is computed in the same way as in multi-label classification tasks (e.g. Chifu et al., 2024).

4 Participants and Approaches

Four teams participated in the shared task (see Table 2). The organizers provided baselines for the three subtasks.

Baseline: For the slot and intent detection subtasks, the baseline we provided is the same as in the original xSID paper, trained on the English data, with an updated version of MaChAmp[2] (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

[2] https://machamp-nlp.github.io/

for intent classification.

For dialect identification, we used the same baseline model as in the ITDI shared task (Aepli et al., 2023): a Support Vector Machine (SVM) classifier with TF-IDF-weighted features of character 1-to-4-grams. The model was trained on the development set using the `scikit-learn` toolkit (Pedregosa et al., 2011).

HiTZ: Team HiTZ (Bengoetxea et al., 2025) was the only one to address all three subtasks. For slot and intent detection, they compared various combinations of the xSID training data and found that English data alone performed best overall, followed by all Germanic languages except Norwegian (i.e., English, German, Dutch and Danish). They also confirmed that multi-task modelling outperformed a single-task setup.

For dialect identification, Team HiTZ collected four additional datasets of non-Standard Norwegian and silver-labeled them using geolocation metadata and linguistic features. On the modelling side, they experimented with both encoder models (fine-tuning) and decoder models (few-shot prompting and supervised fine-tuning). In the end, one of the simplest setups consisting of the NorBERT3 encoder model fine-tuned on the provided development set (i.e., without the additionally collected data) yielded the best results.

MaiNLP: Team MaiNLP (Blaschke et al., 2025) tried to improve performance for slot and intent detection with a variety of methods: varying the training data, injecting character-level noise, training on auxiliary tasks, and combining layers of models fine-tuned on different datasets. They found that injecting character-level noise is an efficient method for improving performance, training on auxiliary tasks did not lead to substantial improvements, and replacing layers of a model fine-tuned on English SID data with layers from a model fine-tuned on the provided development set could lead to substantial performance improvements.

LTG: Team LTG (Midtgaard et al., 2025) investigated potential improvements of the automatically translated training data. They improve the alignment of the slot labels with simAlign (Jalili Sabet et al., 2020) and some heuristics, which leads to substantial performance improvements. They also use an LLM [3] for translating the training data to achieve a higher quality, but this did not lead to better performance. Finally, they map annotation from the MASSIVE dataset (FitzGerald et al., 2023) to the xSID label set, and show that training on these leads to higher performance.[4]

Team	Slots (F1)	Intents (Acc.)	Dialect (w-F1)
Baseline	64.36	84.15	77.42
HiTZ	85.37	97.69	**84.17**
MaiNLP	85.57	97.64	—
LTG	**89.27**	**98.02**	—
CUFE	—	94.38	79.64

Table 3: Highest results for each participating team for intent classification (accuracy), slot detection (Span-F1 score), and dialect identification (weighted F1).

CUFE: Team CUFE (Ibrahim, 2025) fine-tuned three BERT models (mBERT, NB-BERT and NorBERT) for the intent detection and dialect identification tasks. They only used the provided development set for fine-tuning and found that the multilingual mBERT model outperformed the Norwegian-specific models.

5 Results

We evaluated the submitted systems according to accuracy for intents, according to the span F1 score for slots (where both span and label must match exactly), and according to weighted F1 score for dialect identification.[5] Table 3 summarizes the results by showing the highest scores of each team.

For **slot detection**, all participants outperform the baseline by a large margin. Detailed results (Table 4) show that most submissions performed best on the Bokmål data, followed by Trøndersk, Vestnorsk and Nordnorsk. All participating teams found that using the original English training data in a cross-lingual transfer setting worked best, and that adding the (machine-translated) Bokmål training data led to significant drops. The participants' efforts to improve the quality of the slot annotations were largely unsuccessful (Midtgaard et al., 2025).

For **intent classification**, the baseline was also outperformed by a large margin by all participants. The range of scores show that this task is close to

[3] https://huggingface.co/norallm/normistral-7b-warm

[4] Note that training on additional SID-annotated Norwegian was not allowed in the official runs. The run including MASSIVE was submitted outside of the competition.

[5] The data, evaluation scripts and detailed results are available on Github: https://github.com/ltgoslo/NoMusic/tree/main/NorSID

Submission	B	N	T	V	all
LTG 3	90.94	87.19	**89.69**	89.49	**89.27**
LTG 2	89.92	**87.89**	89.27	**89.62**	89.25
MaiNLP 2	90.11	79.66	85.18	87.17	85.57
HiTZ 1	**91.09**	79.00	85.48	86.61	85.37
MaiNLP 1	85.60	82.66	82.99	84.11	83.68
MaiNLP 3	84.37	79.25	81.68	84.01	82.57
LTG 1	84.74	80.09	80.96	83.30	82.22
HiTZ 3	71.15	60.98	66.22	68.18	66.64
Baseline	71.49	60.68	63.23	65.05	64.36
HiTZ 2	56.74	51.94	56.69	56.25	55.66
LTG 4*	91.84	87.56	89.00	89.82	89.38

Table 4: Results (span-F1) for slots. * trained on additional Norwegian labeled data, excluded from the main ranking.

Submission	B	N	T	V	all
LTG 3	98.00	97.20	98.27	**98.20**	**98.02**
LTG 1	**98.20**	97.20	**98.33**	97.84	97.89
LTG 2	**98.20**	**97.30**	98.13	97.84	97.85
HiTZ 2	**98.20**	97.10	97.60	97.88	97.69
MaiNLP 3	97.80	96.90	98.00	97.68	97.64
MaiNLP 2	97.60	96.20	97.67	97.16	97.16
HiTZ 3	97.80	95.40	97.80	97.24	97.11
HiTZ 1	97.40	95.40	96.93	96.04	96.29
CUFE 1	96.40	93.30	95.80	93.56	94.38
MaiNLP 1	92.80	92.60	93.40	94.00	93.47
Baseline	86.40	82.60	83.33	84.80	84.15
LTG 4*	97.80	96.70	97.73	97.20	97.31

Table 5: Results (accuracy) for intents. * trained on additional Norwegian labeled data, excluded from the main ranking.

being solved, even without any annotated training data in the target language (cf. Bengoetxea et al., 2025). The detailed results in Table 5 show that the performances on the different dialects are often similar within single submissions (i.e. systems). The Northern varieties are slightly more challenging than the other dialects, but for all variants there are several systems which perform > 97%. It is also noteworthy that the additional labeled Norwegian MASSIVE dataset provided by team LTG (Midtgaard et al., 2025) did not yield any improvements for intent detection (and only marginal ones for slot filling).

For **dialect identification**, all participating systems outperform the baseline. Generally, the systems struggle most with identifying Bokmål and Nordnorsk, the two varieties with least data (1 and 2 translators, respectively). In the light of these results, data augmentation techniques targeting these two varieties specifically appear as the most promising way forward. This should not prove too difficult for Bokmål, which is standardized and therefore not particularly low-resourced.

Submission	B	N	T	V	all
HiTZ 2	**75.40**	**78.44**	**85.95**	**87.45**	**84.17**
HiTZ 3	74.91	77.50	84.29	87.08	83.32
HiTZ 1	74.10	75.72	83.97	86.61	82.71
CUFE 1	68.93	73.38	80.26	84.14	79.64
Baseline	57.38	73.46	77.76	82.59	77.42

Table 6: Results (weighted-F1) for dialects.

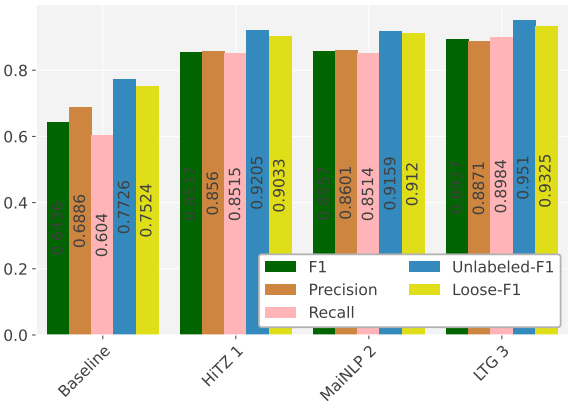

Figure 3: Performance metrics for slots.

6 Analysis

Returning to the slot filling subtask, Figure 3 shows multiple metrics for the best submission of each team over the whole test set (all dialects). Precision is higher compared to recall for most participants, except LTG. We also report unlabeled F1, where we only check if the label boundaries match and ignore the label, and loose F1 which allows for partial matches. The unlabeled F1 is substantially higher for all teams, showing that finding the right label is still an unsolved issue. The loose F1 is always lower than the unlabeled F1, but still substantially higher than the strict span F1, showing that also finding the exact boundaries of a span is still challenging.

Furthermore, we looked into the most commonly confused intent pairs. All teams have the same top-2 confusion pairs, namely: *SearchScreeningEvent–SearchCreativeWork* and *cancel_reminder–cancel_alarm* (gold–predicted). Upon inspection, almost all mistakes in these categories are on the same instances. For example, the

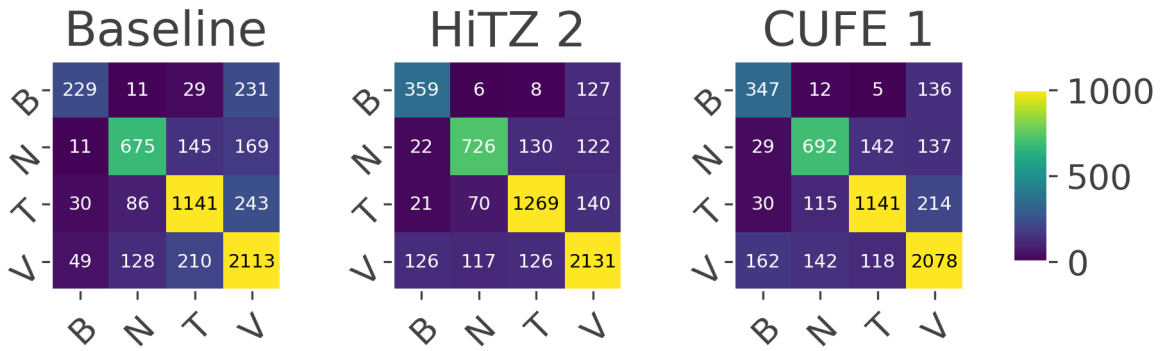

Figure 4: Confusion matrices for dialect classification.

translations of the sentence "I want to see Outcast", e.g. "Eg vil se Outcast." and "Æ vil se Outcast" are predicted as *SearchCreativeWork* by all teams, but the more precise label *SearchScreeningEvent* was annotated. We also found two erroneous annotations for the *cancel_reminder* gold label, which clearly described alarms. Other common mistakes included the prediction of *set_alarm* where *cancel_alarm* was annotated, and the prediction of *PlayMusic* where the true intent was *SearchScreeningEvent* (likely triggered by to the word 'play').

The confusion matrices for dialect classification (Figure 4) show one clear tendency, namely that the Northern (N) and Central (T) dialects are rarely confused with Bokmål (B), whereas confusions between the Western (V) dialects and Bokmål is much more common. In fact, the highest numbers of sentence-level overlap with Bokmål are observed with some of the Western dialect writers. The models also struggle delimiting the three dialect areas (N, T, V), with significant confusion between the non-adjacent areas N and V. In comparison with the baseline, the submitted systems improve mainly by better distinguishing between T and V.

7 Conclusion

This paper presented an overview of the NorSID shared task organized as part of the VarDial Evaluation Campaign 2025.

The analysis of the results presented above suggests that intent detection is largely a solved task, where most of the remaining errors can be attributed to ambiguous labels. On the other hand, the other two subtasks still show room for improvement. The submitted slot filling models struggle with finding the correct slot boundaries and assigning the correct slot labels. Since most submitted models were trained without significant amounts of Norwegian training data, the training signal may not have been strong enough to address the first issue. It is also expected that some inconsistencies have remained in the NoMusic dataset as a result of the translation and annotation.

Regarding dialect classification, the most standardized variant (Bokmål) obtains the poorest scores, most likely due to the low amount of training data provided. More generally, it remains to be investigated to what extent the four major dialect areas (based on traditional dialectological research) represent the most useful partition of our data; in particular, the five translators of the Western dialect area cover a relatively wide area where significant internal variation is expected. Finally, it would be interesting to see what levels of dialect identification performance could be achieved by humans.

Both the slot filling and dialect identification subtasks proved rather challenging, which opens up opportunities for future evaluation campaigns.

Acknowledgements

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

The NoMusic project was granted funding from the TekstHub initiative at the University of Oslo. We thank the many annotators who have contributed to this project, and in particular Marthe Midtgaard for help with the annotation transfer.

References

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

1–13, Gyeongju, Republic of Korea. Association for Computational Linguistics.

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

Jaione Bengoetxea, Mikel Zubillaga, Ekhi Azurmendi, Maite Heredia, Julen Etxaniz, Markel Ferro, and Jeremy Barnes. 2025. HiTZ at VarDial 2025 NorSID: Overcoming data scarcity with language transfer and automatic data annotation. In *Proceedings of the Twelfth Workshop on NLP for Similar Languages, Varieties, and Dialects (VarDial 2025)*, Abu Dhabi, UAE. International Committee on Computational Linguistics.

Verena Blaschke, Felicia Körner, and Barbara Plank. 2025. Add noise, tasks, or layers? MaiNLP at the VarDial 2025 shared task on Norwegian dialectal slot and intent detection. In *Proceedings of the Twelfth Workshop on NLP for Similar Languages, Varieties, and Dialects (VarDial 2025)*, Abu Dhabi, UAE. International Committee on Computational Linguistics.

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

Adrian-Gabriel Chifu, Goran Glavaš, Radu Tudor Ionescu, Nikola Ljubešić, Aleksandra Miletić, Filip Miletić, Yves Scherrer, and Ivan Vulić. 2024. VarDial evaluation campaign 2024: Commonsense reasoning in dialects and multi-label similar language identification. In *Proceedings of the Eleventh Workshop on NLP for Similar Languages, Varieties, and Dialects (VarDial 2024)*, pages 1–15, Mexico City, Mexico. 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*.

Jack FitzGerald, Christopher Hench, Charith Peris, Scott Mackie, Kay Rottmann, Ana Sanchez, Aaron Nash, Liam Urbach, Vishesh Kakarala, Richa Singh, Swetha Ranganath, Laurie Crist, Misha Britan, Wouter Leeuwis, Gokhan Tur, and Prem Natarajan. 2023. MASSIVE: A 1M-example multilingual natural language understanding dataset with 51 typologically-diverse languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4277–4302, Toronto, Canada. 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).

Michael Ibrahim. 2025. CUFE@VarDial 2025 NorSID: Multilingual BERT for Norwegian dialect identification and intent detection. In *Proceedings of the Twelfth Workshop on NLP for Similar Languages, Varieties, and Dialects (VarDial 2025)*, Abu Dhabi, UAE. International Committee on Computational Linguistics.

Masoud Jalili Sabet, Philipp Dufter, François Yvon, and Hinrich Schütze. 2020. SimAlign: High quality word alignments without parallel training data using static and contextualized embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1627–1643, Online. 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.

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

Stefan Larson and Kevin Leach. 2022. A survey of intent classification and slot-filling datasets for task-oriented dialog. *Preprint*, arXiv:2207.13211.

Samuel Louvan and Bernardo Magnini. 2020. Recent neural methods on slot filling and intent classification for task-oriented dialogue systems: A survey. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 480–496, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Petter Mæhlum and Yves Scherrer. 2024. NoMusic - the Norwegian multi-dialectal slot and intent detection corpus. In *Proceedings of the Eleventh Workshop on NLP for Similar Languages, Varieties, and Dialects (VarDial 2024)*, pages 107–116, Mexico City, Mexico. Association for Computational Linguistics.

Shervin Malmasi, Marcos Zampieri, Nikola Ljubešić, Preslav Nakov, Ahmed Ali, and Jörg Tiedemann. 2016. Discriminating between similar languages and Arabic dialect identification: A report on the third

DSL shared task. In *Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial3)*, pages 1–14, Osaka, Japan. The COLING 2016 Organizing Committee.

Marthe Midtgaard, Petter Mæhlum, and Yves Scherrer. 2025. The LTG submission to the NorSID slot and intent detection shared task: More and better training data for slot and intent detection. In *Proceedings of the Twelfth Workshop on NLP for Similar Languages, Varieties, and Dialects (VarDial 2025)*, Abu Dhabi, UAE. International Committee on 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.

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.

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.

Henry Weld, Xiaoqi Huang, Siqu Long, Josiah Poon, and Soyeon Caren Han. 2022. A survey of joint intent detection and slot filling models in natural language understanding. *ACM Comput. Surv.*, 55(8).

Miriam Winkler, Virginija Juozapaityte, Rob van der Goot, and Barbara Plank. 2024. Slot and intent detection resources for Bavarian and Lithuanian: Assessing translations vs natural queries to digital assistants. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 14898–14915, Torino, Italia. ELRA and ICCL.

Weijia Xu, Batool Haider, and Saab Mansour. 2020. End-to-end slot alignment and recognition for cross-lingual NLU. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5052–5063, Online. 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, Preslav Nakov, and Yves Scherrer. 2020. Natural language processing for similar languages, varieties, and dialects: A survey. *Natural Language Engineering*, 26(6):595–612.

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.