Welcome Note
VarDial 2025

VarDial Workshop and Shared Task Organizers

Abu Dhabi, UAE, January 2025

The VarDial Workshop Series

1. VarDial 2014 at COLING in Dublin, Ireland
2. LT4VarDial 2015 at RANLP in Hissar, Bulgaria
3. VarDial 2016 at COLING in Osaka, Japan
4. VarDial 2017 at EACL in Valencia, Spain
5. VarDial 2018 at COLING in Santa Fe, United States
6. VarDial 2019 at NAACL in Minneapolis, United States
7. VarDial 2020 at COLING virtually in Barcelona, Spain
8. VarDial 2021 at EACL virtually in Kiev, Ukraine
9. VarDial 2022 at COLING in Gyeongju, Korea (and online)
10. VarDial 2023 at EACL in Dubrovnik, Croatia (and online)
11. VarDial 2024 at NAACL in Mexico City, Mexico (and online)
12. VarDial 2025 at COLING in Abu Dhabi, UAE

Schedule

9:00 – 9:30 — Opening and Findings of the Evaluation Campaign
9:30 – 9:50 — Shared Task Participants Poster Boosters
9:50 – 10:15 — Oral presentation 1
10:15 – 10:30 — Poster boosters I
10:30 – 11:00 — Coffee break
11:00 – 12:00 — Invited talk: Fajri Koto
12:00 – 12:30 — Poster boosters II
12:30 – 14:00 — Lunch break
14:00 – 15:00 — Poster session
15:00 – 15:30 — Oral presentation 2
15:30 – 16:00 — Coffee break
16:00 – 17:15 — Oral presentations 3–5
17:15 – 17:30 — Closing remarks

VarDial 2025 Organizers

Workshop organizers:

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

Organizers of the NorSID shared task:

- Yves Scherrer
- Rob van der Goot
- Petter Mæhlum

Overview of the VarDial 2025 Evaluation Campaign

The NorSID Shared Task:

Norwegian Slot, Intent and Dialect Identification

Slot and Intent Detection

Add VarDial to the calendar for tomorrow 9 AM

Slot and Intent Detection

Add VarDial to the calendar for tomorrow 9 AM

Intent: AddCalendar

Slot and Intent Detection

Add VarDial to the calendar for tomorrow 9 AM

Intent: AddCalendar

Slots: *Add* $[VarDial]_{Event}$ *to the calendar for* $[tomorrow\ 9\ AM]_{Datetime}$

xSID (2021)

ar	أود أن أرى مواعيد عرض فيلم Silly Movie 2.0 في دار السينما
da	Jeg vil gerne se spilletiderne for Silly Movie 2.0 i biografen
de	Ich würde gerne den Vorstellungsbeginn für Silly Movie 2.0 im Kino sehen
de-st	I mecht es Programm fir Silly Movie 2.0 in Film Haus sechn
en	I'd like to see the showtimes for Silly Movie 2.0 at the movie house
id	Saya ingin melihat jam tayang untuk Silly Movie 2.0 di gedung bioskop
it	Mi piacerebbe vedere gli orari degli spettacoli per Silly Movie 2.0 al cinema
ja	Silly Movie 2.0 の上映時間を見せて。映画館
kk	Мен Silly Movie 2.0 бағдарламасының кинотеатрда көрсетілім уақытын
nl	Ik wil graag de speeltijden van Silly Movie 2.0 in het filmhuis zien
sr	Želela bih da vidim raspored prikazivanja za Silly Movie 2.0 u bioskopu
tr	Silly Movie 2.0'ın sinema salonundaki seanslarını görmek istiyorum
zh	我想看 Silly Movie 2.0 在影院 的放映

Van der Goot, R., et al.: *From Masked Language Modeling to Translation: Non-English Auxiliary Tasks Improve Zero-shot Spoken Language Understanding*. Proceedings of NAACL 2021.

SID4LR (Shared Task at VarDial 2023)

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

Aepli, N., et al.: *Findings of the VarDial Evaluation Campaign 2023*. Proceedings of VarDial 2023.

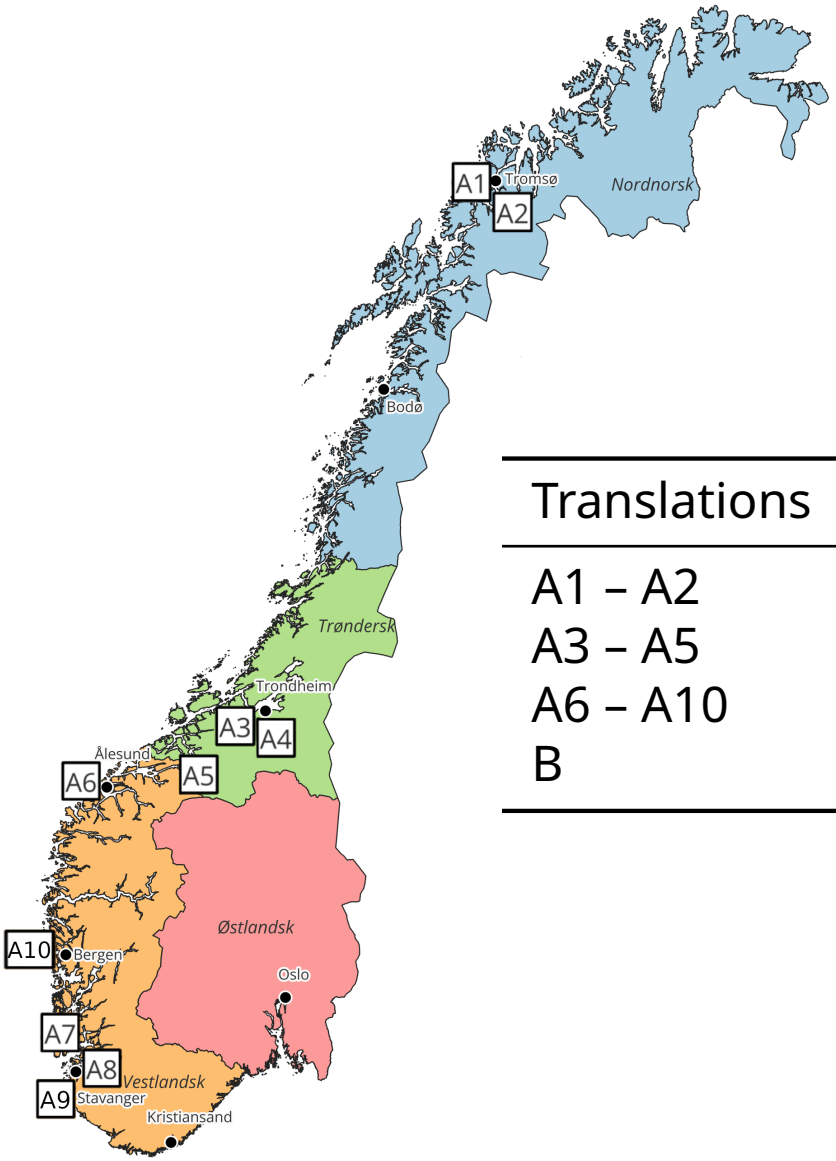
NoMusic (VarDial 2024)

Extension of the xSID data to 11 Norwegian varieties:
Standard Bokmål + 10 dialects (A1 – A10):

English	Set a reminder to go to the grocery store later
Bokmål	Sett på en påminnelse om å gå i butikken etterpå
A1	Minn mæ på at æ skal dra på butikken seinere.
A2	Sett enn påminnelse om å fære tel butikken seinar.
A3	Sett en alarm for å da te matbutikken seinere
A4	Sett en påminnelse om å gå te matbutikken seinar
A5	Sett en påminnelse for å gå t butikken seinar
A6	Sett en påminnelse om å stikke på butikken seinere.
A7	Sett på en påminnelse om å gå t butikken seinare
A8	Lag ein påminnelse om å gå på butikken seinere
A9	Sett ein påminnelse for å dra te matbutikken seinåre
A10	Sett på en påminnelse for å gå på butikken senere.

Mæhlum, P. & Scherrer, Y.: *NoMusic – The Norwegian Multi-Dialectal Slot and Intent Detection Corpus.* Proceedings of VarDial 2024.

NoMusic (VarDial 2024)



Translations	Dialect region	Label
A1 – A2	North Norwegian / Nordnorsk	N
A3 – A5	Central Norwegian / Trøndersk	T
A6 – A10	West Norwegian / Vestnorsk	V
B	Bokmål	B

The NorSID Shared Task

Three subtasks:

- Slot identification
  - BIO span annotation task (40 slot types)
  - Metric: span F1 score
- Intent identification
  - Text classification task (18 intent labels)
  - Metric: accuracy
- Dialect identification
  - Text classification task (4 labels)
  - Metric: weighted F1 score on deduplicated data

The NorSID Shared Task

Three subtasks:

- Slot identification
  - BIO span annotation task (40 slot types)
  - Metric: span F1 score
- Intent identification
  - Text classification task (18 intent labels)
  - Metric: accuracy
- Dialect identification
  - Text classification task (4 labels)
  - Metric: weighted F1 score on deduplicated data

Four participating teams:

Team	Slots	Intents	Dialects
HiTZ	✓	✓	✓
MaiNLP	✓	✓	
LTG	✓	✓	
CUFE		✓	✓

Data

Dataset	Size	S	I	D
Manually annotated English xSID **training** set	43k	✓	✓	–
Machine-translated xSID training sets (12 languages, e.g. German, Dutch, Danish, Norwegian Bokmål)	43k	✓	✓	–

Data

Dataset	Size	S	I	D
Manually annotated English xSID **training** set	43k	✓	✓	–
Machine-translated xSID training sets (12 languages, e.g. German, Dutch, Danish, Norwegian Bokmål)	43k	✓	✓	–
NorDial (dialectal tweets, not annotated)		–	–	(✓)
NordicTweetStream (geotagged tweets, not necessarily dialectal)		–	–	(✓)
Nordic Dialect Corpus + LIA corpus (dialectological transcriptions, different genre)		–	–	(✓)

Data

Dataset	Size	S	I	D
Manually annotated English xSID **training** set	43k	✓	✓	–
Machine-translated xSID training sets (12 languages, e.g. German, Dutch, Danish, Norwegian Bokmål)	43k	✓	✓	–
NorDial (dialectal tweets, not annotated)		–	–	(✓)
NordicTweetStream (geotagged tweets, not necessarily dialectal)		–	–	(✓)
Nordic Dialect Corpus + LIA corpus (dialectological transcriptions, different genre)		–	–	(✓)
Concatenation of the 11 NoMusic **validation** sets (allowed for **training**)	3300	✓	✓	✓

Data

Dataset	Size	S	I	D
Manually annotated English xSID **training** set	43k	✓	✓	–
Machine-translated xSID training sets (12 languages, e.g. German, Dutch, Danish, Norwegian Bokmål)	43k	✓	✓	–
NorDial (dialectal tweets, not annotated)		–	–	(✓)
NordicTweetStream (geotagged tweets, not necessarily dialectal)		–	–	(✓)
Nordic Dialect Corpus + LIA corpus (dialectological transcriptions, different genre)		–	–	(✓)
Concatenation of the 11 NoMusic **validation** sets (allowed for **training**)	3300	✓	✓	✓
Concatenation of the 11 NoMusic **test** sets	5500	✓	✓	✓

Slots and Intents – Participants and Approaches

- **Baseline:** Multi-task mBERT fine-tuned on English xSID training data

Slots and Intents – Participants and Approaches

- **Baseline:** Multi-task mBERT fine-tuned on English xSID training data
- Multi-task $>$ single-task models (HiTZ)
- Norwegian/Scandinavian $\approx$ multilingual base models (HiTZ, MaiNLP, LTG, CUFE)
- Norwegian $>$ English training data for intents (HiTZ, MaiNLP, LTG)
- English $>$ Norwegian training data for slots (HiTZ, MaiNLP, LTG)

Slots and Intents – Participants and Approaches

- **Baseline:** Multi-task mBERT fine-tuned on English xSID training data
- Multi-task $>$ single-task models (HiTZ)
- Norwegian/Scandinavian $\approx$ multilingual base models (HiTZ, MaiNLP, LTG, CUFE)
- Norwegian $>$ English training data for intents (HiTZ, MaiNLP, LTG)
- English $>$ Norwegian training data for slots (HiTZ, MaiNLP, LTG)
- Improved label projection and translation of the Norwegian training data 👎 (LTG)

Slots and Intents – Participants and Approaches

- **Baseline:** Multi-task mBERT fine-tuned on English xSID training data
- Multi-task $>$ single-task models (HiTZ)
- Norwegian/Scandinavian $\approx$ multilingual base models (HiTZ, MaiNLP, LTG, CUFE)
- Norwegian $>$ English training data for intents (HiTZ, MaiNLP, LTG)
- English $>$ Norwegian training data for slots (HiTZ, MaiNLP, LTG)
- Improved label projection and translation of the Norwegian training data 👎 (LTG)
- Noise injection to simulate spelling and dialectal variation 👍 (MaiNLP)
- Training on auxiliary tasks (NER, POS, Dep, DID) 👎 (MaiNLP)
- Combining layers of models fine-tuned on different datasets 👍 (MaiNLP)

Slots and Intents – Results

Intents (accuracy %):

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

- Results are close together
- Similar errors across teams
- Subtask is close to being solved

Slots and Intents – Results

Intents (accuracy %):

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

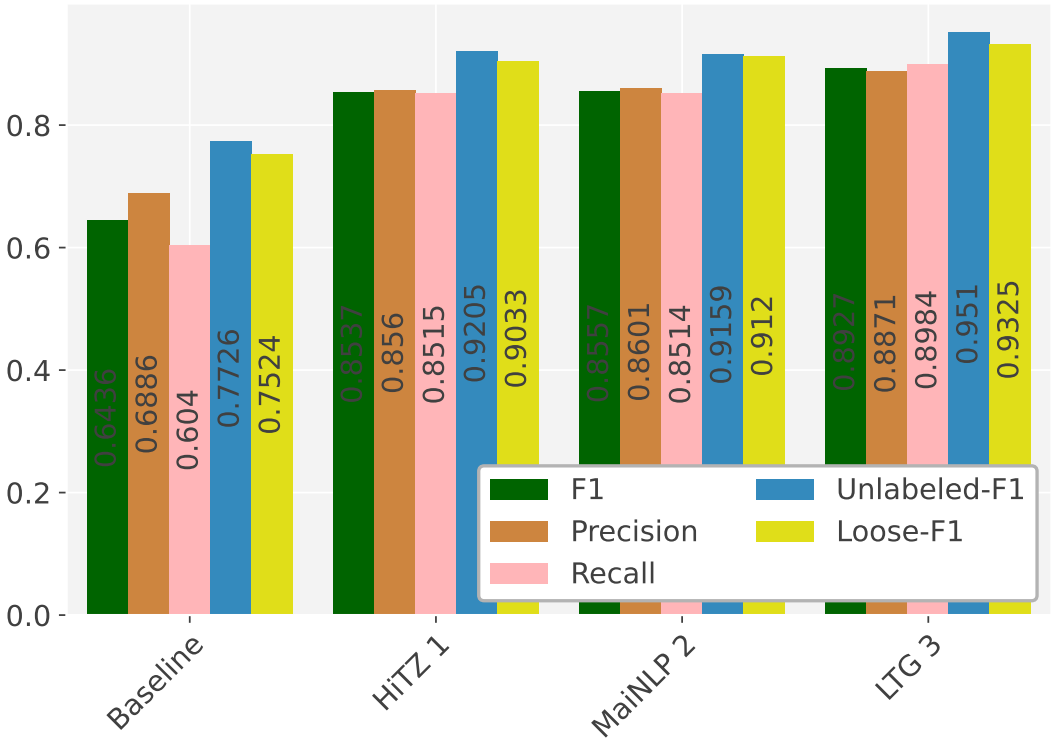
- Results are close together
- Similar errors across teams
- Subtask is close to being solved

Slots (span F1-score %):

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

- Northern dialects seem most difficult
- LTG 4* includes Norwegian MASSIVE training dataset (not allowed by ST guidelines)

Slots – Results



- Precision $>$ recall (except LTG)
- Unlabeled F1 $>$ labeled F1 (difficulties finding the correct label)
- Loose F1 $>$ strict F1 (difficulties finding the exact span boundaries)

Dialects – Participants and Approaches

- **Baseline:** SVM classifier with TF-IDF-weighted features of character 1-to-4-grams, trained on validation set

Dialects – Participants and Approaches

- **Baseline:** SVM classifier with TF-IDF-weighted features of character 1-to-4-grams, trained on validation set
- Multilingual BERT $>$ Norwegian BERT (CUFE)
- Encoder models with fine-tuning $>$ decoder models with few-shot prompting or supervised fine-tuning (HiTZ)

Dialects – Participants and Approaches

- **Baseline:** SVM classifier with TF-IDF-weighted features of character 1-to-4-grams, trained on validation set
- Multilingual BERT $>$ Norwegian BERT (CUFE)
- Encoder models with fine-tuning $>$ decoder models with few-shot prompting or supervised fine-tuning (HiTZ)
- Include additional silver-labeled datasets 👎 (HiTZ)

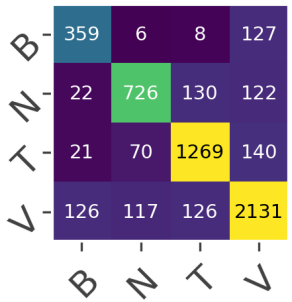
Dialects – Results

Weighted-average F1-score %:

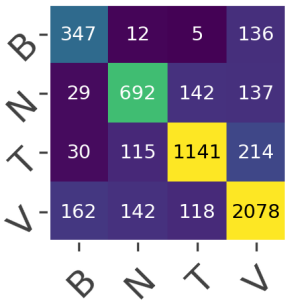
Submission	B	N	T	V	all
HiTZ 2	**75.40**	**78.44**	**85.95**	**87.45**	**84.17**
HiTZ 3	74.91	77.50	84.29	87.08	83.32
HiTZ 1	74.10	75.72	83.97	86.61	82.71
CUFE 1	68.93	73.38	80.26	84.14	79.64
Baseline	57.38	73.46	77.76	82.59	77.42

- Systems struggle most with identifying Bokmål and Nordnorsk, the two varieties with least data (1 and 2 translators, respectively)
- Confusions between the Western (V) dialects and Bokmål are most common
- Also significant confusion between the non-adjacent dialect areas N and V

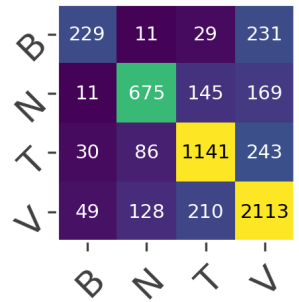
HiTZ 2



CUFE 1



Baseline



Takeaways

1. Intent identification is mostly solved, whereas slot and dialect identification show room for improvement:
   - Insufficient high-quality Norwegian training data
   - Inconsistencies in annotations
   - Unbalanced data distribution across the four dialect areas

Takeaways

1. Intent identification is mostly solved, whereas slot and dialect identification show room for improvement:
   - Insufficient high-quality Norwegian training data
   - Inconsistencies in annotations
   - Unbalanced data distribution across the four dialect areas
2. Slot and intent scores for Norwegian are substantially higher than for other low-resource and dialect scenarios (SID4LR, Bavarian)
   - Is Norwegian dialect writing closer to standard?
   - Is there better cross-lingual transfer from English?

Takeaways

1. Intent identification is mostly solved, whereas slot and dialect identification show room for improvement:
   - Insufficient high-quality Norwegian training data
   - Inconsistencies in annotations
   - Unbalanced data distribution across the four dialect areas
2. Slot and intent scores for Norwegian are substantially higher than for other low-resource and dialect scenarios (SID4LR, Bavarian)
   - Is Norwegian dialect writing closer to standard?
   - Is there better cross-lingual transfer from English?
3. What kind of variation does the Norwegian data actually contain?
   - Is individual speaker variation (punctuation, word choices, translationese, …) more salient than dialectal variation?

Welcome Note
VarDial 2025

VarDial Workshop and Shared Task Organizers

Abu Dhabi, UAE, January 2025

Announcements

Tutorial 6: Connecting Ideas in Lower-Resource Scenarios: NLP for National Varieties, Creoles, and Other Low-Resource Scenarios

Organizers: Aditya Joshi, Diptesh Kanojia, Heather Lent, Hour Kaing and Haiyue Song

Time: Tomorrow Monday, 09:00 - 17:30

Location: Conference Hall B (C)

Announcements

Tutorial 6: Connecting Ideas in Lower-Resource Scenarios: NLP for National Varieties, Creoles, and Other Low-Resource Scenarios

Organizers: Aditya Joshi, Diptesh Kanojia, Heather Lent, Hour Kaing and Haiyue Song

Time: Tomorrow Monday, 09:00 - 17:30

Location: Conference Hall B (C)

I have a 3-year post-doc opening at the University of Oslo!

Deadline: 6 April