Language beyond the Standard

NLP for Low-Resource Varieties

Sina Ahmadi

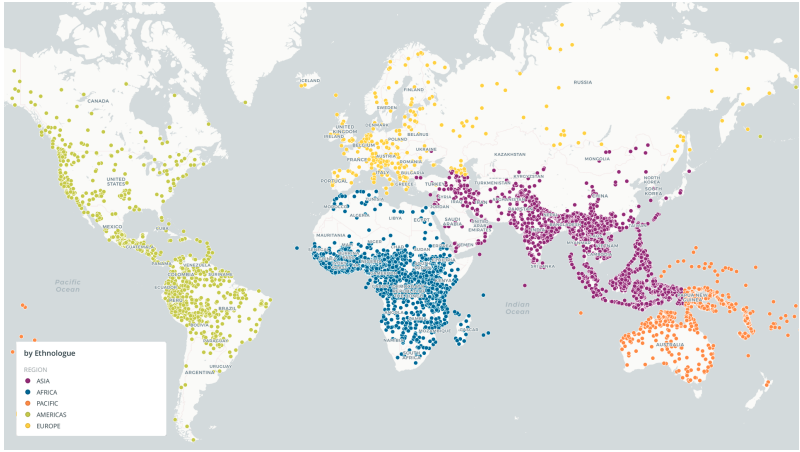
ZurichNLP - ETH AI Center

January 19, 2026

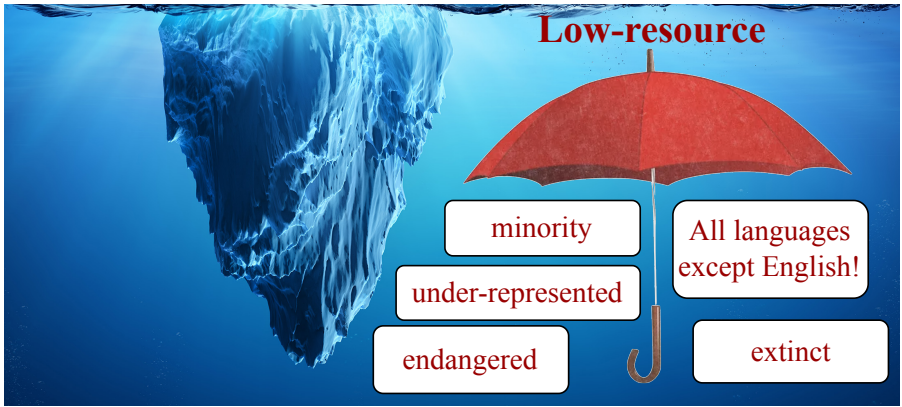
Universität Zürich[UZH]

Background

- More than 7,000 "languages" are spoken today (Ethnologue, 2026)
- Not all languages have equal status (see `https://endangeredlanguages.com`)



Background: Low-Resourced NLP



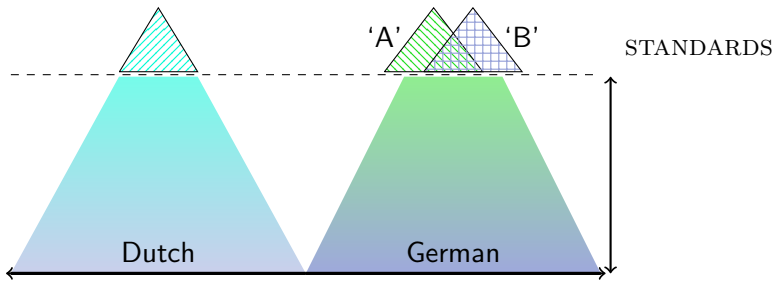
- **98%** of languages around the globe are low-resourced!
- Any language can be considered low-resourced depending on **domain** and **task**.

Background: Languages are not monoliths

Language and Speech Technology (LST) operates on a monolithic assumption, privileging standardized written forms and neglecting intra-language diversity.

Background

Language and Speech Technology (LST) operates on a monolithic assumption, privileging standardized written forms and neglecting intra-language diversity.



What about all the 'varieties'?

Background

- Many studies have gone beyond the monolithic concept of a language, as for the varieties of English (Ziems et al., 2023)
- LST for dialects and varieties is challenging (Zampieri et al., 2020)
- Non-standard varieties are limited to spoken language (*almost*)
- Differences in written language: orthographic supremacy (Lew, 2012)
- Lexical variations: more than 30 words for "hedgehog" in Kurdish!
- Loanwords and terminologies ("*char*" vs. "*voiture*")
- ... and this is where we face our greatest challenge

The Holy Grail: Data

Manual Curation

- Manually-curated resources as for Arabic vernaculars (Bouamor et al., 2018) or Swiss German (Dogan-Schönberger et al., 2021)
- Atlases as for Italian (De Mareüil et al., 2021)
- Crowd-sourcing: 180,000+ utterances across 6 dialects (>100 hours) from movies and series (Ahmadi et al., 2024, LREC)

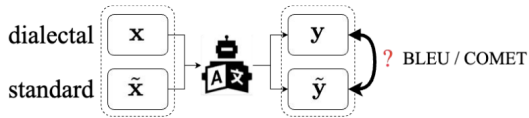
Variety	Sentence
Standard	*Wir leben im Zeitalter der Technik.*
Bern	*mir läbä im zitauter vor technik.*
Graubünden	*miar leben im ziitalter dr technik.*
St. Gallen	*mir lebed im ziitalter de technik.*
Wallis	*mir läbu im zitalter der technik.*
Zürich	*mir läbu im zitalter der technik.*

CODET (Alam et al., 2024a, EACL)

- Extract contrastive data from previous studies (Italian, Basque, Swiss German)
- Re-purpose contrastive data from other sources
- Create new contrastive data (Bengali, Central Kurdish)
- Benchmark dialects using machine translation (MT) models
- Quantify discrepancies across varieties

Languages/Varieties	**# Sents**	**# Varieties**
Arabic Vernaculars	12,000	25
Farsi Varieties	3071	2
Malay-Indonesian	3071	2
Swahili	1919	2
Tigrinya Varieties	3071	2
Italian Varieties	792	439
Swiss German Varieties	118	368
Basque Varieties	370	39
Bengali Varieties	200	5
Central Kurdish Varieties	300	4
Griko	163	1

CODET: Benchmarking

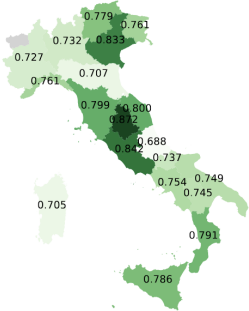


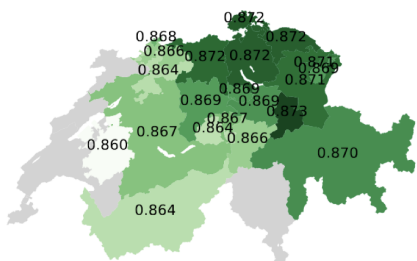
- We evaluate CODET in the X → English direction using four different-sized NLLB-200 in two setups:
- **With reference:** Compare dialectal output against standard reference translation
- **Without reference:** Treat non-standard sentences as adversarial or non-native noisy inputs

⇒ **A robust MT system should produce the same output for dialectal inputs regardless of variation.**

CODET: Benchmarking

- MT systems excel at handling standard variants
- As dialectal variation deviates further from the standard, the quality of translations decreases
- Possible causes: 1) spelling variation, 2) inadequate representation



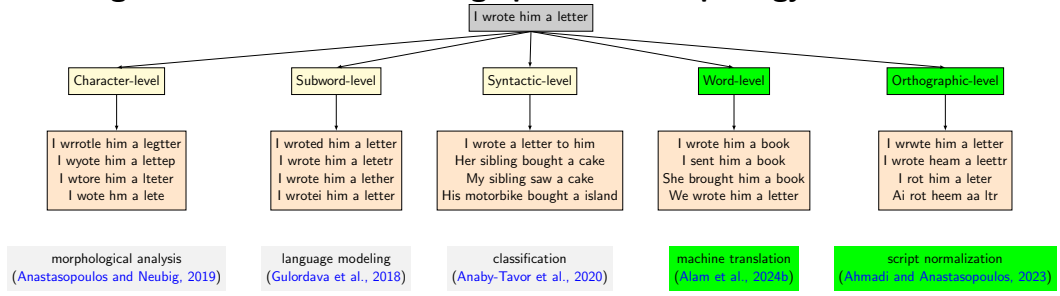


Gamification

How to collect dialectal data more efficiently? $\Rightarrow$ Dia-Lingle (Sun et al., 2025, ACL)

- Gamify dialectal data collection
- Challenge the player to outsmart an oracle
- Optimize using active learning
- Collect through feedback learning
- Supports five languages (Swiss German, Romansh, Kurdish, Japanese, Korean)
- Play Dia-Lingle: `https://dia-lingle.ivia.ch/`

Synthetic Data Augmentation

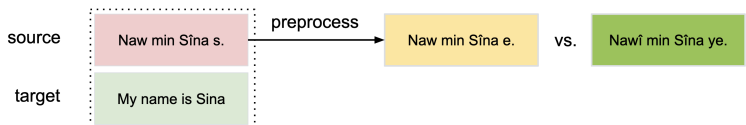


* Examples inspired by Şahin's *To Augment or Not to Augment* (Şahin, 2022)

Synthetic Data Augmentation: Morphology

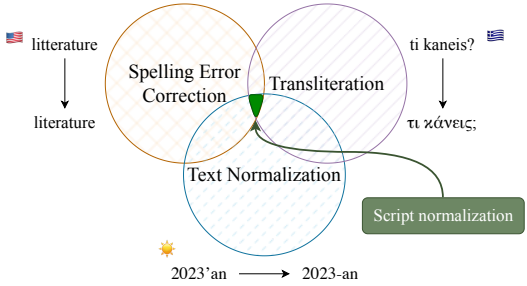
Using rules, convert sentences from a dialect to the standard (** synthetic sentences)

- Learn and apply morphosyntactic variation
- Map vocabulary
- Replace terminology



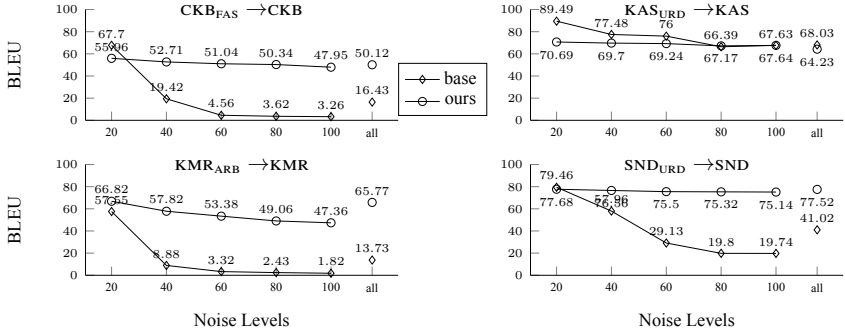
Synthetic Data Augmentation: Orthography

- "*Non-standard varieties are limited to spoken language*" → Unconventional Writing
  - *ana raye7 el gam3a el sa3a 3 el 3asr.* (Arabic in Latin script, aka Arabizi)
  - *Twra ti na sou pw* (Greek in Latin, aka Greeklish)
  - *mer6 pr tn mess pr mn anif* (French SMS language)
- Script Normalization: normalization of a text written in an unconventional script based on the conventional script and orthography



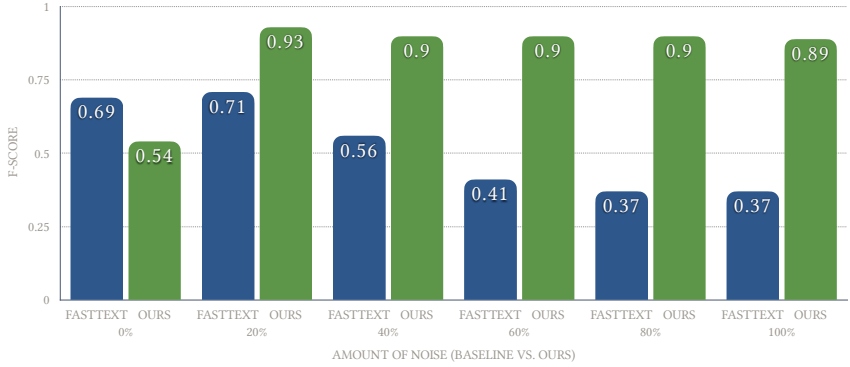
Script Normalization (Ahmadi and Anastasopoulos, 2023, ACL 2023)

- Baseline: a naive "copy" system
- Ours: trained models on different levels of noise
- **Our models dramatically improve over the baseline**
- Including when evaluated on real data
- The more similar the scripts, the more difficult the normalization!



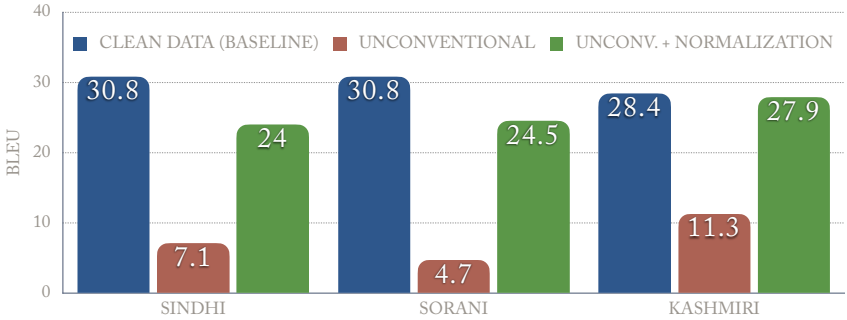
Script Normalization (Ahmadi and Anastasopoulos, 2023, ACL 2023)

1. **Language identification (LID)**
   - Compare LID with and without normalization
   - Terrible performance by any existing model
   - Models trained on normalized datasets improve the F-scores
   - Closely-related languages (scripts) are confused!



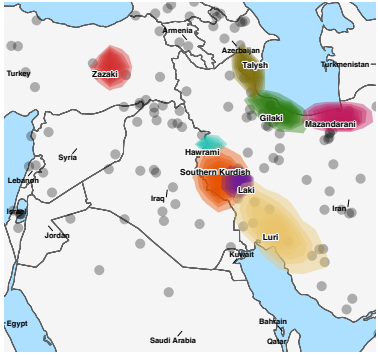
Script Normalization: Extrinsic Experiments

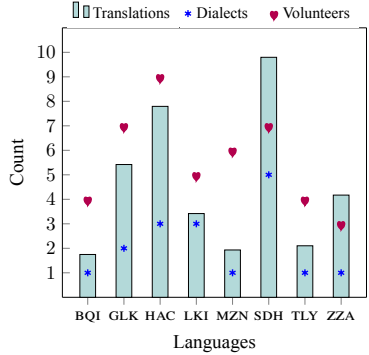
1. Language identification (LID)
2. **Machine Translation (MT)**
   - Evaluate MT with and without normalization
   - Terrible performance on noisy data (NLLB as baseline)
   - Models trained on normalized datasets improve the F-scores



Participatory Research

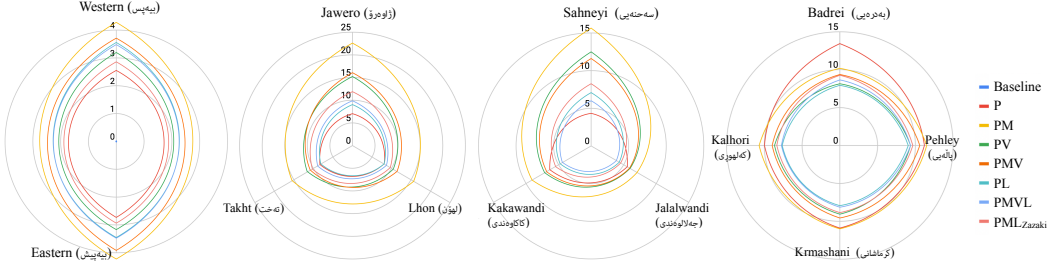
- Mobilize communities for data collection (and research) (Nekoto et al., 2020)
- Parallel corpora for under-represented Middle Eastern languages: PARME (Ahmadi et al., 2025, ACL)
- 45 contributors translated 36,384 sentences into eight severely under-resourced ME languages (18-23M speakers)
- No standard variety ⇒ 18 varieties written in seven orthographies!





Participatory Research

- Considerable performance disparities
- Not intrinsic translation difficulty
- Varying degrees of representation in data
- Linguistic proximity to the source material



Takeaways

- Creative data collection methods can bridge resource gaps
- There are significant performance discrepancies across different varieties
- Dialect-aware NLP should account for orthographic, lexical, and morphosyntactic variations
- LLMs? $\rightarrow$ *"significant challenges persist in dialect identification, generation, and translation"* (Mousi et al., 2025)
- More robust metrics to not penalize spelling variation
- Still a lot of room for improvement
- **The path forward: moving from discrete modeling toward continuous variation modeling**

Thanks!

Contact: sina.ahmadi@uzh.ch

Language beyond the Standard

Questions?

References (I)

Ahmadi, S. and Anastasopoulos, A. (2023). Script normalization for unconventional writing of under-resourced languages in bilingual communities. In Rogers, A., Boyd-Graber, J., and Okazaki, N., editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14466–14487, Toronto, Canada. Association for Computational Linguistics.

Ahmadi, S., Jaff, D., Alam, M. M. I., and Anastasopoulos, A. (2024). Language and speech technology for Central Kurdish varieties. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10034–10045.

Ahmadi, S., Sennrich, R., Karami, E., Marani, A., Fekrazad, P., Baghban, G. A., Hadi, H., Heidari, S., Dogan, M., Asadi, P., Bashir, D., Ghodrati, M. A., Amini, K., Ashourinezhad, Z., Baladi, M., Ezzati, F., Ghasemifar, A., Hosseinpour, D., Abbaszadeh, B., Hassanpour, A., Hamaamin, B. J., Hama, S. K., Mousavi, A., Hussein, S. N., Nejadgholi, I., Ölmez, M., Osmanpour, H., Ramezani, R. R., Aziz, A. S., Salehi, A., Yadegari, M., Yadegari, K., and Roodsari, S. Z. (2025). PARME: parallel corpora for low-resourced Middle Eastern languages. In Che, W., Nabende, J., Shutova, E., and Pilehvar, M. T., editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 30032–30053. Association for Computational Linguistics.

Alam, M. M. I., Ahmadi, S., and Anastasopoulos, A. (2024a). CODET: A benchmark for contrastive dialectal evaluation of machine translation. In Graham, Y. and Purver, M., editors, *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1790–1859, St. Julian's, Malta. Association for Computational Linguistics.

Alam, M. M. I., Ahmadi, S., and Anastasopoulos, A. (2024b). A morphologically-aware dictionary-based data augmentation technique for machine translation of under-represented languages. *arXiv preprint arXiv:2402.01939*.

Anaby-Tavor, A., Carmeli, B., Goldbraich, E., Kantor, A., Kour, G., Shlomov, S., Tepper, N., and Zwerdling, N. (2020). Do not have enough data? deep learning to the rescue! In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 7383–7390.

Anastasopoulos, A. and Neubig, G. (2019). Pushing the limits of low-resource morphological inflection. *arXiv preprint arXiv:1908.05838*.

Bouamor, H., Habash, N., Salameh, M., Zaghouani, W., Rambow, O., Abdulrahim, D., Obeid, O., Khalifa, S., Eryani, F., Erdmann, A., and Oflazer, K. (2018). The MADAR Arabic dialect corpus and lexicon. In Calzolari, N., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Hasida, K., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S., and Tokunaga, T., editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

De Mareüil, P. B., Bilinski, E., Vernier, F., De Iacovo, V., and Romano, A. (2021). For a mapping of the languages/dialects of italy and regional varieties of italian. *New Ways of Analyzing Dialectal Variation*, pages 267–288.

References (II)

Dogan-Schönberger, P., Mäder, J., and Hofmann, T. (2021). SwissDial: Parallel multidialectal corpus of spoken Swiss German. *arXiv preprint arXiv:2103.11401*.

Ethnologue (2026). Ethnologue: Languages of the world.

Gulordava, K., Aina, L., and Boleda, G. (2018). How to represent a word and predict it, too: Improving tied architectures for language modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing; 2018 Oct 31-Nov 4; Brussels, Belgium. Stroudsburg: Association for Computational Linguistics; 2018. p. 2936–41.* ACL (Association for Computational Linguistics).

Lew, R. (2012). *How can we make electronic dictionaries more effective?* Oxford University Press.

Mousi, B., Durrani, N., Ahmad, F., Hasan, M. A., Hasanain, M., Kabbani, T., Dalvi, F., Chowdhury, S. A., and Alam, F. (2025). Aradice: Benchmarks for dialectal and cultural capabilities in LLMs. In Rambow, O., Wanner, L., Apidianaki, M., Al-Khalifa, H., Eugenio, B. D., and Schockaert, S., editors, *Proceedings of the 31st International Conference on Computational Linguistics, COLING 2025, Abu Dhabi, UAE, January 19-24, 2025*, pages 4186–4218. Association for Computational Linguistics.

Nekoto, W., Marivate, V., Matsila, T., Fasubaa, T. E., Fagbohungbe, T., Akinola, S. O., Muhammad, S. H., Kabenamualu, S. K., Osei, S., Sackey, F., Niyongabo, R. A., Macharm, R., Ogayo, P., Ahia, O., Berhe, M. M., Adeyemi, M., Mokgesi-Selinga, M., Okegbemi, L., Martinus, L., Tajudeen, K., Degila, K., Ogueji, K., Siminyu, K., Kreutzer, J., Webster, J., Ali, J. T., Abbott, J. Z., Orife, I., Ezeani, I., Dangana, I. A., Kamper, H., Elsahar, H., Duru, G., Kioko, G., Murhabazi, E., Biljon, E. V., Whitenack, D., Onyefuluchi, C., Emezue, C. C., Dossou, B. F. P., Sibanda, B. K., Bassey, B. I., Olabiyi, A., Ramkilowan, A., Öktem, A., Akinfaderin, A., and Bashir, A. (2020). Participatory research for low-resourced machine translation: A case study in african languages. In Cohn, T., He, Y., and Liu, Y., editors, *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020*, volume EMNLP 2020 of *Findings of ACL*, pages 2144–2160. Association for Computational Linguistics.

Şahin, G. G. (2022). To augment or not to augment? A comparative study on text augmentation techniques for low-resource NLP. volume 48, pages 5–42. MIT Press One Broadway, 12th Floor, Cambridge, Massachusetts 02142, USA ....

Sun, J., Sevastjanova, R., Ahmadi, S., Sennrich, R., and El-Assady, M. (2025). Dia-lingle: A gamified interface for dialectal data collection. In Mishra, P., Muresan, S., and Yu, T., editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 148–158, Vienna, Austria. Association for Computational Linguistics.

Zampieri, M., Nakov, P., and Scherrer, Y. (2020). Natural language processing for similar languages, varieties, and dialects: A survey. *Natural Language Engineering*, 26(6):595–612.

Ziems, C., Held, W., Yang, J., Dhamala, J., Gupta, R., and Yang, D. (2023). Multi-value: A framework for cross-dialectal English NLP. In Rogers, A., Boyd-Graber, J. L., and Okazaki, N., editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 744–768. Association for Computational Linguistics.