Natural Language Processing for Low-Resource and Marginalized Language Varieties

Introduction

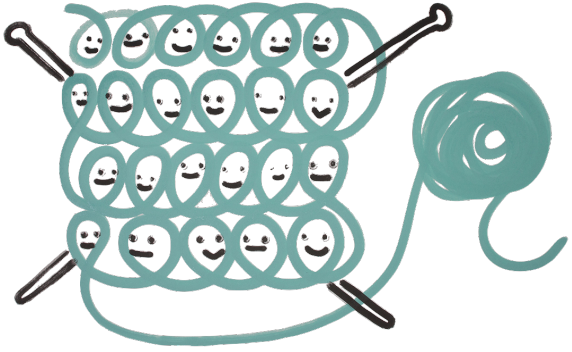
Sina Ahmadi (sina.ahmadi@uzh.ch)

Department of Computational Linguistics

February 18, 2026

Welcome

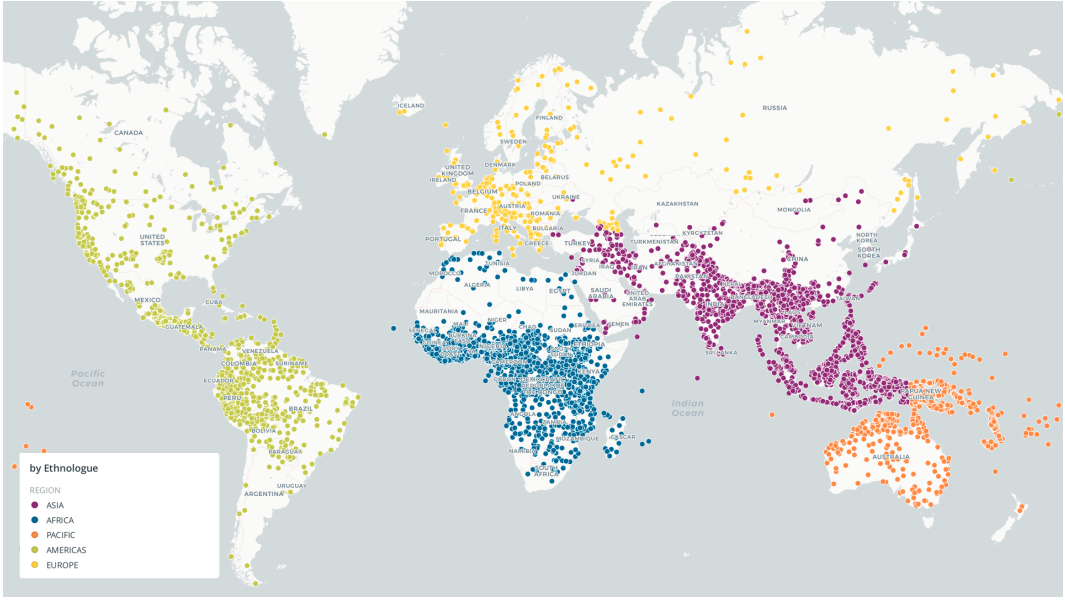
Get to Know Each Other



- Introduce yourself! What languages do you speak?
- What's your background?
- Why are you interested in low-resource NLP?

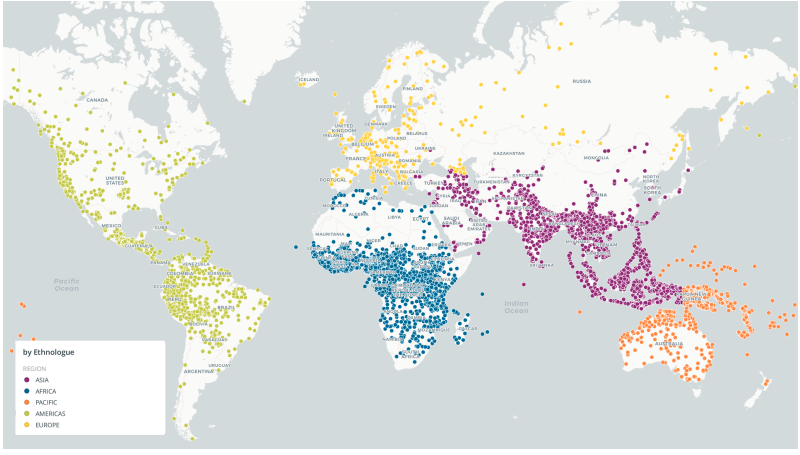
Low-resource NLP

What is Low-Resource NLP?



Low-resource NLP: Linguistic Disparity

- More than 7,000 "languages" are spoken today (Ethnologue, 2024)
- Not all languages have equal status (`https://endangeredlanguages.com`)



Linguistic Disparity

What does low-resource mean?

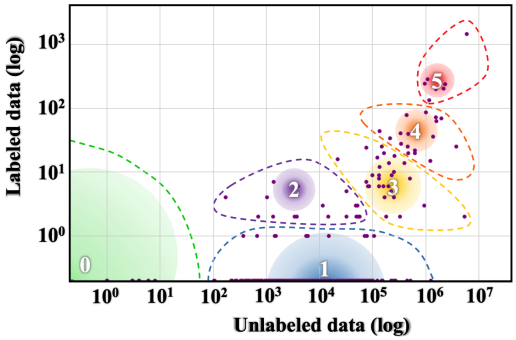
The term "low-resource" can have multiple implications:

- In machine learning: *generalizing with* **minimal or no feedback**, *to new domains or tasks facing data paucity*
  e.g., biomedical, legal and literary domains
- In NLP, *generalize to languages for which there is* **limited data**
  e.g., archaeological fragments or descriptions of the language and contexts
- **Computational efficiency**: reduce the time of inference or the size/energy footprint of a model (i.e. Green deep learning (Xu et al., 2021))
- **Linguistically sensitive supervision**: reduce the risk of harmful biases or misunderstandings
  e.g., CIDAR: Culturally Relevant Instruction Dataset For Arabic (Alyafeai et al., 2024)

Every language can be low-resource!

- Most languages around the globe are *somehow* low-resourced!
- Any language can be considered low-resourced depending on **domain** and **task**.
- Examples:
  - English: high-resource for news, low-resource for 16th-century dialects
  - German: high-resource for standard written text, low-resource for Swiss German dialects
  - Medical domain: even high-resource languages lack annotated clinical data
- Low-resource $\neq$ only small languages. It's about **what resources exist for what purpose**.

Low-Resource $\neq$ Small Languages



The power-law distribution of language resources by Joshi et al. (2020)

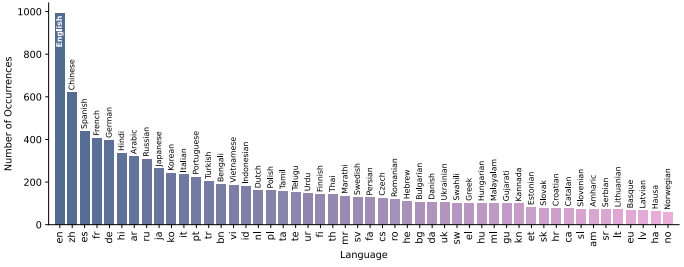
- **0 - The Left-Behinds:** Virtually no resources. Impossible to lift up digitally.
- **1 - The Scraping-Bys:** Some unlabeled data, almost no labeled datasets.
- **2 - The Hopefuls:** Small set of labeled datasets. NLP tools possible in a few years.
- **3 - The Rising Stars:** Strong web presence, but insufficient labeled data.
- **4 - The Underdogs:** Large unlabeled data, moderate labeled data. High potential.
- **5 - The Winners:** Dominant online presence, massive resources.

NLP for all languages

- Inequality of information and representation affects how we understand
- Facilitate communication between humans
- Decreasing the digital divide
- Mitigating cross-cultural biases
- Deploying language technology for underrepresented languages, dialects, minorities
- Societal impact and *prestige* (Crystal, 2002)
- Understanding cross-linguistic differences
- Enhance interaction between humans and machines
  e.g., speech recognition/synthesis, question answering, dialogue
- Analyze and comprehend language
  e.g., syntactic analysis, text classification, entity/relation recognition/linking

Challenges

- The long tail of data
- Lack of domain expertise for data annotation
- Algorithmic bias toward high-resource languages
- Evaluation gaps & benchmark limitations



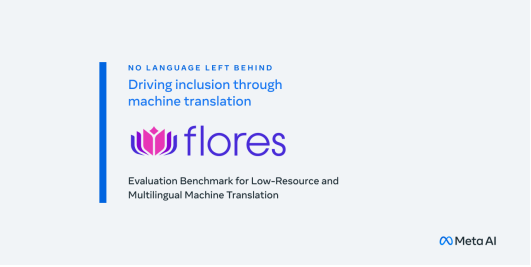
Distribution of languages in multilingual benchmarks (Wu et al., 2025)

Lack of data is not the cause but a consequence

- Language status: How cool is your language?
- Standardization and written tradition
- Language infrastructure:
  - Standard orthography/spelling
  - Character encoding (e.g., Unicode)
  - Fonts and rendering technologies
  - Input methods

Expand to More Languages

- CommonVoice (`https://commonvoice.mozilla.org`)
- Masakhane for African languages (`https://www.masakhane.io`)
- Increasing number of multilingual LLMs (still challenging: (Zhang et al., 2023))
- Meta's Massively Multilingual Speech project aiming ASR for +1000 languages
- Multilingual benchmarks such as FLORES-200 for MT
- Special Interest Groups (SIGs) affiliated to the Association for Computational Linguistics (ACL)



Approaches & Framework

A Framework for Low-Resource NLP

Pillar I: Data

- Resource creation
- Annotation strategies
- Data augmentation
- Quality & coverage

Pillar II: Learning

- Transfer learning
- Cross-lingual models
- Zero- & few-shot methods
- Leveraging high-resource languages

Pillar III: Evaluation

- Metrics & benchmarks
- Human judgments
- Culturally aware evaluation
- Beyond standard assumptions

Any robust approach to low-resource NLP must address all three pillars.

Pillar I: Data

How do we obtain and curate data for low-resource languages?

Resource Creation

- Naturally occurring sources (news, subtitles, religious texts)
- Community-driven initiatives
- Crawling & scraping digital content

Annotation

- Expert vs. crowdsourced annotation
- Cross-lingual annotation projection
- LLM-as-a-Judge

Data Augmentation

- Character-, word-, and syntactic-level perturbations
- Back-translation
- Synthetic data generation

Quality & Coverage

- Domain and register diversity
- Dialect and variety representation
- Data contamination & bias

Pillar II: Learning

Learn a model to map an input X into an output Y:

Input X	Output Y	Task
Text	Text in another language	Machine Translation
Text	Response	Dialog
Speech	Transcript	Speech Recognition
Text	Linguistic Structure	Language Analysis

- **Supervised Learning**: Paired data $\langle X, Y \rangle$, source data X, target data Y
- **Unsupervised Learning**: Find hidden patterns in unlabeled data X without Y
- **Semi-supervised Learning**: X with a limited number of labels

Pillar III: Evaluation

Do our metrics actually measure what we think they measure?

- **Surface-level metrics** (e.g., BLEU) rely on exact or near-exact string matching
- **Trained metrics** (e.g., COMET) inherit biases from their training data
- **Benchmarks** are often translated from English
  - Translated benchmarks correlate less with human judgments (0.47) than localized ones (0.68) (Wu et al., 2025)
  - Cultural and pragmatic nuances are lost in translation

Variety	Source Sentence	MT Output (NLLB)
Standard	*Wir leben im Zeitalter der Technik.*	We live in the age of technology.
Bern	*mir läbä im zitauter vor technik.*	I'm a little overwhelmed by the technique.
Graubünden	*miar leben im ziitalter dr technik.*	I think we're living in the age of technology.
St. Gallen	*mir lebed im ziitalter de technik.*	I lived in the age of technology.
Wallis	*mir läbu im zitalter der technik.*	I was in the age of technology.
Zürich	*mir läbu im zitalter der technik.*	I was in the age of technology.

NLLB performance on Swiss German varieties (Dogan-Schönberger et al., 2021).

Course Organization

Schedule

#	Date	Topic	Type
1	18 Feb	Introduction	Lecture
2	25 Feb	Pillar I: Data	Lecture
3	04 Mar	Pillar II: Learning	Lecture
4	11 Mar	Pillar III: Evaluation	Lecture
5	18 Mar	Theme 1: Data Augmentation & Synthetic Data	Seminar
6	25 Mar	Theme 2: Cross-Lingual Transfer & Multilingualism	Seminar
7	01 Apr	Theme 3: Dialects & Non-Standard Varieties	Seminar
–	08 Apr	*No Class (Easter Break)*	–
8	15 Apr	Theme 4: Script, Orthography & Normalization	Seminar
9	22 Apr	Theme 5: Code-Switching & Hybrid Identities	Seminar
10	29 Apr	Theme 6: Morphologically Rich LR Languages	Seminar
11	06 May	Theme 7: Low-Resource Speech & Multimodal NLP	Seminar
12	13 May	Theme 8: Human-Centric NLP & Ethical Deployment	Seminar
13	20 May	Theme 9: Efficiency & Small Language Models	Seminar
14	27 May	Final Showcase: Technical Interventions	Presentations

Attendance

- Attendance is expected, especially during seminar sessions
- If you must miss a class, notify the instructor in advance
- Think of attendance as a collegial responsibility; your peers are counting on you

Assessment

- Paper presentation: 50%
  - Study a paper carefully and present it in the class
  - Assigning papers automatically or by yourselves?
  - More details later
- Final project: 50%
  - Problem Identification: Select a language variety and a task (Week 5)
  - Solution Proposal: A brief outline of the proposed methodology (Week 8)
  - Implementation & Presentation: Delivering the final solution and presenting the results (Week 14)

Academic Integrity

AI Tools Policy

- ✓ **Permitted:** Brainstorming and editing
- ! **Must cite:** Any machine-generated text (dedicated section)
- Students bear full responsibility for accuracy

Why This Policy?

- You need to **understand** the concepts, not just generate text
- In low-resource NLP, you often can't trust AI outputs
- Critical thinking is essential

Full details available in CL UZH Academic Rules

Selected Readings

- *Data Quality Issues in Multilingual Speech Datasets: The Need for Sociolinguistic Awareness and Proactive Language Planning* (Lau et al., 2025)
- *Approaches for NLP in low-resource scenarios* (Hedderich et al., 2020)
- *Rule-Based Language Technology* (Hurskainen et al., 2023)
- *Natural Language Processing RELIES on Linguistics* (Opitz et al., 2024)
- *Big AI is accelerating the metacrisis: What can we do?* (Bird, 2025)

References

Z. Alyafeai, K. Almubarak, A. Ashraf, D. Alnuhait, S. Alshahrani, G. A. Abdulrahman, G. Ahmed, Q. Gawah, Z. Saleh, M. Ghaleb, et al. CIDAR: Culturally Relevant Instruction Dataset For Arabic. *arXiv preprint arXiv:2402.03177*, 2024.

S. Bird. Big AI is accelerating the metacrisis: What can we do? *arXiv preprint arXiv:2512.24863*, 2025.

D. Crystal. *Language death*. Cambridge University Press, 2002.

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

Ethnologue. Ethnologue: Languages of the world, 2024.

M. A. Hedderich, L. Lange, H. Adel, J. Strötgen, and D. Klakow. A survey on recent approaches for natural language processing in low-resource scenarios. *arXiv preprint arXiv:2010.12309*, 2020.

A. Hurskainen, K. Koskenniemi, T. Pirinen, A. Ranta, I. Listenmaa, F. A. Pirinen, E. Axelson, S. Hardwick, K. Lindén, S. N. Moshagen, et al. Rule-based language technology. 2023.

P. Joshi, S. Santy, A. Budhiraja, K. Bali, and M. Choudhury. The state and fate of linguistic diversity and inclusion in the NLP world. In D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.560. URL https://aclanthology.org/2020.acl-main.560/.

M. Lau, Q. Chen, Y. Fang, T. Xu, T. Chen, and P. Golik. Data quality issues in multilingual speech datasets: The need for sociolinguistic awareness and proactive language planning. In W. Che, J. Nabende, E. Shutova, and M. T. Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7466–7492, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.370. URL https://aclanthology.org/2025.acl-long.370/.

J. Opitz, S. Wein, and N. Schneider. Natural Language Processing RELIES on Linguistics. *arXiv preprint arXiv:2405.05966*, 2024.

M. Wu, W. Wang, S. Liu, H. Yin, X. Wang, Y. Zhao, C. Lyu, L. Wang, W. Luo, and K. Zhang. The bitter lesson learned from 2,000+ multilingual benchmarks. *arXiv preprint arXiv:2504.15521*, 2025.

J. Xu, W. Zhou, Z. Fu, H. Zhou, and L. Li. A survey on green deep learning. *arXiv preprint arXiv:2111.05193*, 2021.

X. Zhang, S. Li, B. Hauer, N. Shi, and G. Kondrak. Don't trust ChatGPT when your question is not in English: A study of multilingual abilities and types of LLMs. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7915–7927, 2023.