Natural Language Processing for Low-Resource and Marginalized Language Varieties

Pillar I: Data

Sina Ahmadi (sina.ahmadi@uzh.ch)

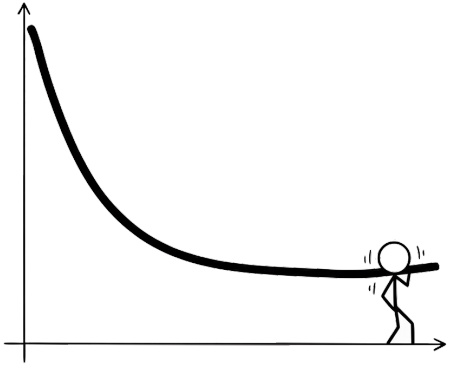
Department of Computational Linguistics

February 25, 2026

Recap

Low-Resourced Languages Lack Data

- The long tail of data
- How to tackle this?



Data Sources

- Naturally occurring sources of text data
  - News: Local news, Voice of America
  - Government documents: Governments often mandate translation
  - Wikipedia: Some Wikipedia articles are translated into many languages
  - Subtitles: Subtitles of movies and TED talks
  - Religious documents: Bible, Jehovah's Witness publications
  - Social media outlets
- Naturally occurring sources of speech data
  - Transcribed news: Spoken radio news with transcriptions
  - Audio books: Regular audio books or religious books
  - Subtitled talks/videos: TED(x) talks or YouTube videos with transcriptions
  - Manually transcribed datasets: Record speech and manually transcribe

Not much available for low-resourced languages!

Five Essential Questions

Five Essential Questions About Data

Q1. **What type of data?** Identify language resources (lexica, corpora, treebanks)

Q2. **What modality?** Text, audio, image, video

Q3. **For what purpose?** Define the downstream task

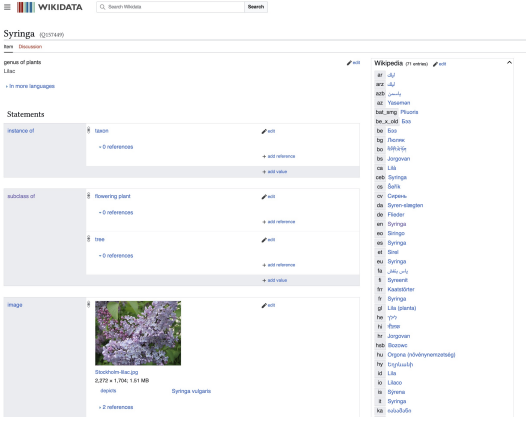
Q4. **How to create it?** Adapt your data to your task

Q5. **Where to maintain it?** Make it sustainable and accessible

Q1: What type of data?

A grammar of Tuatschin: A Sursilvan Romansh dialect (Maurer-Cecchini, 2021)

1. Dictionaries & lexica
2. Terminologies & thesauri
3. Encyclopediae
4. Corpora
5. Language descriptions
6. Knowledge graphs
7. Language models?



Q2: What Modality?

- **Text**: the default modality in NLP
- **Speech/Audio**: radio, oral traditions, community recordings
  Can bypass the text bottleneck for oral languages
- **Image**: document scans, sign language frames
- **Video**: subtitled media, sign language

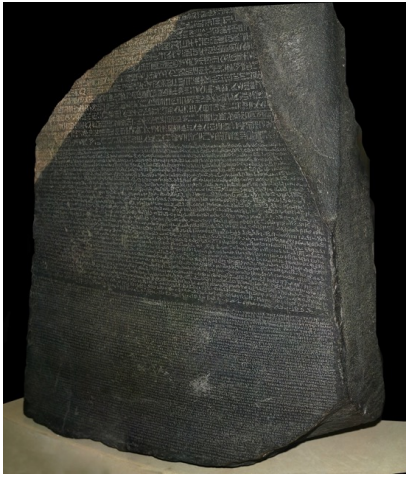
Example multimodal models:

- Vision–Text: CLIP (Radford et al., 2021)
- Sign language: SignCLIP (Jiang et al., 2024)
- Speech-to-speech: SeamlessM4T (2023) (100+ languages)

Q3: For what purpose?

Tasks and their primary data types:

- **Language modeling** → monolingual corpus
- **Machine translation** → parallel corpus (bitext in 2+ languages)
- **Speech synthesis / ASR** → speech corpus (audio + transcriptions)
- **Information extraction** → annotated corpus
- **Multimodal tasks** → aligned data across modalities



Q4. How to create it?

The naïvest approach:

- Collect data, then annotate
- This has a number of obvious shortcomings:
  - Raw data is often difficult to obtain
    - Domains where only limited text exists (law, medicine)
    - Low-resourced languages!
  - Annotation is expensive
    - Crowdworkers are cheap but unskilled, and still cost money
    - Experts are expensive and slow
  - Standard pipelines assume data abundance
    - Pre-train $\rightarrow$ fine-tune requires large corpora
    - Evaluation benchmarks may not exist

Q5. Where to maintain it?

- **Data governance for heritage and minority languages**
  - Who owns the data? Who benefits?
  - Consent, attribution, and community control
- **Decolonising language technology** (Bird, 2020)
  - Technology should serve communities, not extract from them
  - Community agency in deciding what gets digitized and how
- **Open-source matters**
  - Reproducibility, community contribution, long-term access
  - Platforms: Hugging Face, GitHub, CLARIN, ELRA

Low-Resource Approaches

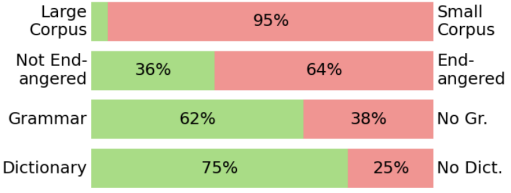
Approaches to Low-Resource NLP

Innovative approaches to

- Data collection
- Data annotation
- Data augmentation

Data Collection

- **Crowdsourcing**: Amazon Mechanical Turk, Prolific
- **Community-driven initiatives**: Mozilla Common Voice, Masakhane
- **Grammar books as data sources**
  Leverage linguistic descriptions directly in LLMs (Aycock et al., 2025)



60% of the world's languages have a grammar book; 75% have a dictionary (Zhang et al., 2024)

Data Collection: OCR/ASR Bootstrapping

Many low-resource languages have printed or spoken materials, but no digital text.

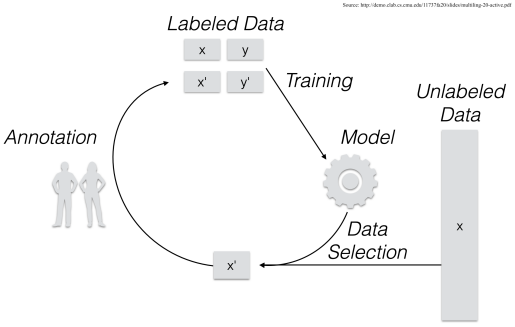
- **OCR bootstrapping**: Historical newspapers, books, government records
  - Small seed data $\rightarrow$ train initial OCR model $\rightarrow$ correct $\rightarrow$ retrain
- **ASR bootstrapping**: Radio broadcasts, oral traditions, community recordings
  - Leverage pre-trained multilingual models (Whisper, MMS) and fine-tune

Data Annotation

- **Expert annotation**
  - High quality, but expensive and slow
  - Often the only option for morphologically complex or under-documented languages
- **Crowdsourced annotation**
  - Scalable, but limited for low-resource languages (few qualified workers)
  - Quality control: inter-annotator agreement
- **Cross-lingual annotation projection**
  - Project labels from high-resource language via word alignment or translation
  - Noisy but useful as silver-standard data
- **LLM-as-a-Judge / LLM-as-annotator**
  - Use LLMs to label, rank, or filter data
  - Caveat: LLM knowledge of low-resource languages is often shallow

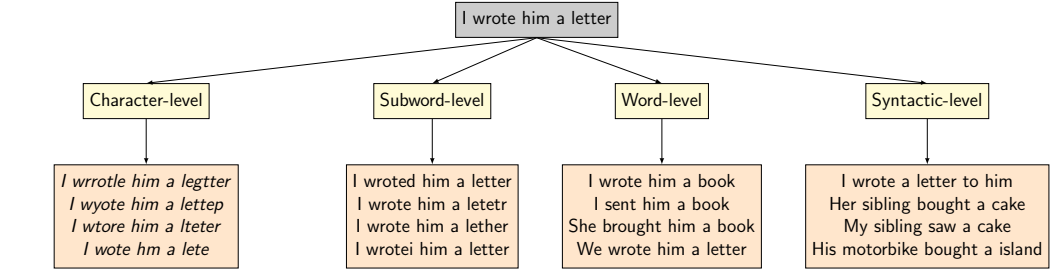
Data Annotation using Active Learning

- Incremental creation of data and model improvement
- Query a human annotator to efficiently generate $\langle X, Y \rangle$ examples from X
- Fundamental ideas:
  - **Uncertainty**: select data that are hard for current models
  - **Representativeness**: select data similar to the target distribution



Data Augmentation

Data augmentation techniques that aim to increase the sample size



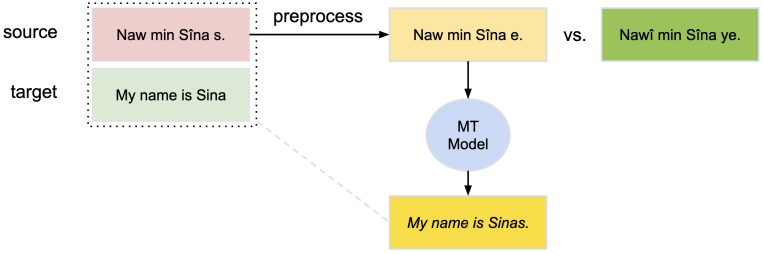
morphological analysis (Anastasopoulos and Neubig, 2019)	language modeling (Gulordava et al., 2018)	machine translation (Alam et al., 2024)	classification (Anaby-Tavor et al., 2020)

* Inspired by Şahin (2022)'s *To Augment or Not to Augment*

Data Augmentation (an example)

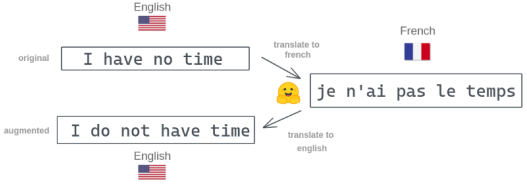
Using rules, convert sentences from a dialect to the standard (* synthetic sentences)

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



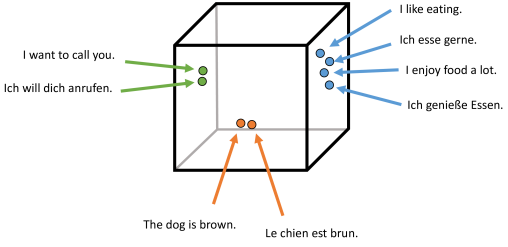
Data Augmentation: Back-Translation

- **Core idea**: use translation as a data generation tool (Sennrich et al., 2016)
- **Forward**: translate from a high-resource language into the target low-resource language
- **Back-translate**: translate the generated target text back into the source language
- The resulting parallel pairs (original + back-translated) augment your training data
- Introduces lexical and syntactic diversity while preserving meaning
- Widely used to bootstrap MT systems and improve downstream tasks

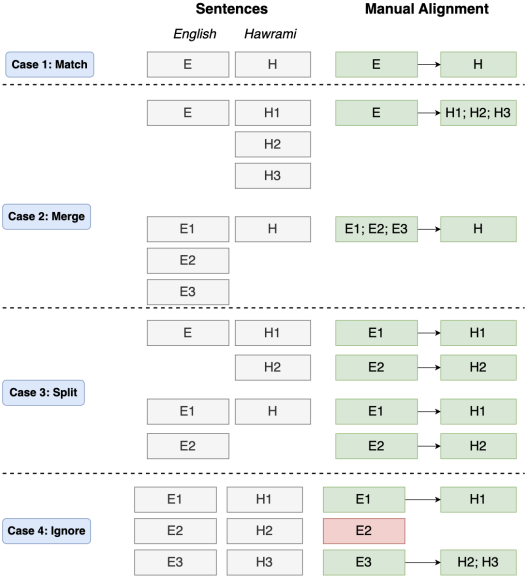


Data Augmentation: Bitext Mining

- How to find parallel sentences when you only have monolingual data in two languages?
- Encode sentences from both languages into a shared multilingual embedding space
- Sentences that are close in this space are likely translations of each other
- Critical for building MT systems from scratch

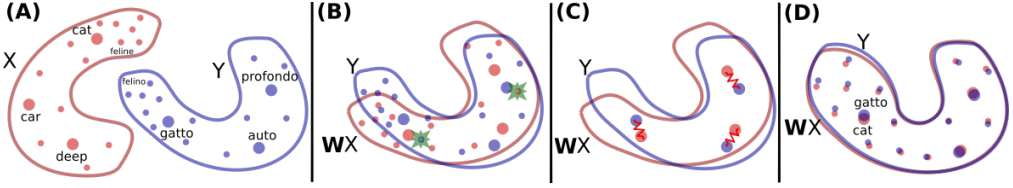


Data Augmentation: Bitext Mining (an example)



Data Augmentation: Bilingual Lexicon Induction (BLI)

- Same intuition, but at the **word level**
- Train word embeddings independently in each language
- Learn a mapping to align the two embedding spaces (aka *Procrustes alignment*)
- Nearest neighbours across spaces $\approx$ translation pairs
- Can be supervised, semi-supervised, or fully unsupervised (Lample et al., 2017)



Data Augmentation: Gamification and Feedback Learning

How to collect dialectal data more efficiently? ⇒ Dia-Lingle (Sun et al., 2025)

- Gamify dialectal data collection
- Challenge the player to outsmart an oracle
- Optimize using active learning
- Collect through feedback learning
- Play Dia-Lingle: `https://dia-lingle.ivia.ch/`

Takeaways

Takeaways

- There is no single prescription for all languages when it comes to data
- Innovative techniques are required to collect, annotate, and augment data
- Multimodal approaches can unlock resources for languages with limited written traditions
- Sustainability, ethics, and community involvement are not optional, they are foundational

Next week: How to optimize learning in frugality? $\rightarrow$ *Pillar II: Learning*

References

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

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

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

S. Aycock, D. Stap, D. Wu, C. Monz, and K. Sima'an. Can LLMs really learn to translate a low-resource language from one grammar book? In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=aMBSY2ebPw.

S. Bird. Decolonising speech and language technology. In *Proceedings of the 28th international conference on computational linguistics*, pages 3504–3519, 2020.

K. Gulordava, L. Aina, and G. Boleda. 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), 2018.

Z. Jiang, G. Sant, A. Moryossef, M. Müller, R. Sennrich, and S. Ebling. SignCLIP: Connecting text and sign language by contrastive learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9171–9193, 2024.

G. Lample, A. Conneau, L. Denoyer, and M. Ranzato. Unsupervised machine translation using monolingual corpora only. *arXiv preprint arXiv:1711.00043*, 2017.

P. Maurer-Cecchini. *A grammar of Tuatschin: A Sursilvan Romansh dialect*, volume 3. BoD–Books on Demand, 2021.

A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

G. G. Şahin. 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 . . . , 2022.

Seamless Communication. Seamless: Multilingual expressive and streaming speech translation. 2023.

R. Sennrich, B. Haddow, and A. Birch. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 86–96, 2016.

J. Sun, R. Sevastjanova, S. Ahmadi, R. Sennrich, and M. El-Assady. Dia-lingle: A gamified interface for dialectal data collection. In P. Mishra, S. Muresan, and T. Yu, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 148–158, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-253-4. doi: 10.18653/v1/2025.acl-demo.15. URL https://aclanthology.org/2025.acl-demo.15/.

B. Thompson and P. Koehn. Vecalign: Improved sentence alignment in linear time and space. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 1342–1348, 2019.

K. Zhang, Y. Choi, Z. Song, T. He, W. Y. Wang, and L. Li. Hire a linguist!: Learning endangered languages in LLMs with in-context linguistic descriptions. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15654–15669, 2024.