Low-Resource Languages & Scaling

Annotation Beyond High-Resource Settings

Jin Zhao

Brandeis University

April 27, 2025

Today's Agenda

1. The low-resource challenge
2. Finding annotators for rare languages
3. Cross-lingual transfer approaches
4. Quality assurance at scale
5. Case studies: MasakhaNER, AmericasNLU
6. Scaling annotation projects

Project: Final presentations May 4

The Low-Resource Problem

Most of the world's languages lack NLP resources

Statistics:

- 7,000+ languages worldwide
- NLP resources exist for $\sim$100
- Well-resourced: $\sim$20 languages

Consequences:

- Billions of speakers excluded from NLP benefits
- AI systems perpetuate linguistic inequality
- Cultural knowledge encoded in languages is lost

What Makes a Language "Low-Resource"?

Resource scarcity across dimensions:

- **Data:** Limited digital text corpora
- **Tools:** No tokenizers, taggers, parsers
- **Annotators:** Hard to find qualified speakers
- **Evaluation:** No standard benchmarks
- **Research:** Little prior NLP work

Examples:

- Many African languages (Yoruba, Igbo, Swahili)
- Indigenous American languages (Quechua, Nahuatl)
- Southeast Asian languages (Khmer, Lao)

Annotator Recruitment Challenges

Finding qualified annotators is difficult

Issues:

1. **Pool size:** Fewer speakers = fewer annotators
2. **Literacy:** Some languages have low written literacy
3. **Digital access:** Annotators may lack internet/devices
4. **Location:** Speakers concentrated in specific regions
5. **Payment:** International payments can be difficult

Crowdsourcing platforms:

- MTurk has very limited low-resource coverage
- Need specialized recruitment strategies

Annotator Recruitment Strategies

Where to find annotators:

1. **University partnerships:**
   - Linguistics departments
   - Area studies programs
   - International student groups
2. **Community organizations:**
   - Cultural associations
   - Religious communities
   - Diaspora groups
3. **Local collaborators:**
   - In-country researchers
   - NGOs working in region
   - Local universities

Native Speaker Requirements

Quality depends on annotator qualifications

Minimum requirements:

- Native or near-native fluency
- Written literacy in the language
- Familiarity with local varieties

Verification methods:

- Self-reported proficiency
- Language background questionnaire
- Qualification tasks in target language
- Validation by other native speakers

Challenge: Dialectal variation may require specific regional speakers

Cross-Lingual Transfer

Leveraging high-resource languages

Approach:

1. Train model on high-resource language
2. Transfer to low-resource target
3. Fine-tune with small amount of target data

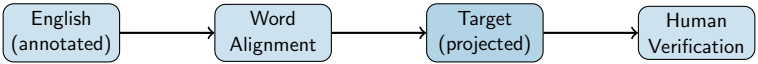
How it relates to annotation:

- Need less annotated data in target language
- Annotation effort focused on validation
- Can start with model predictions

Tools: mBERT, XLM-R, BLOOM

Projection-Based Annotation

Transfer annotations through parallel text



Annotation role:

- Correct projection errors
- Validate label quality
- Handle non-parallel structures

Active Learning for Low-Resource

Maximize value of each annotation

Strategy:

1. Train initial model on small seed set
2. Identify most informative examples
3. Request annotation only for those
4. Iterate until performance plateaus

Benefits for low-resource:

- Reduces annotation volume needed
- Focuses expert annotator time
- Achieves good performance with less data

Quality Assurance at Scale

Maintaining quality as you scale up

Challenges:

- More annotators = more variability
- Harder to maintain consistent training
- Quality monitoring becomes complex

Solutions:

1. Tiered training program
2. Regular calibration sessions
3. Automated quality checks
4. Gold standard embedded in batches
5. Per-annotator performance tracking

Scaling Annotation Infrastructure

Managing large annotation projects

Infrastructure needs:

- Assignment management system
- Progress tracking dashboard
- Communication channels
- Payment processing
- Data storage and backup

Tools:

- Label Studio (self-hosted)
- Argilla (with team features)
- Custom platforms (for very large scale)

Case Study: MasakhaNER

NER for African Languages

Project scope:

- 10 African languages initially (now 20+)
- Named Entity Recognition task
- Community-driven annotation

Key practices:

- Native speaker annotators only
- Adapted guidelines for local contexts
- Community ownership of data
- Open release for research

Impact: Major benchmark for African NLP

Case Study: AmericasNLU

NLU for Indigenous Languages of the Americas

Languages:

- Quechua, Guarani, Aymara, Nahuatl, etc.
- 10 languages in initial release

Approach:

- Translation and adaptation of NLI datasets
- Local university partnerships
- Native speaker verification

Lessons:

- Cultural adaptation matters (not just translation)
- Community involvement increases buy-in
- Local expertise is essential

Ethical Considerations

Responsible low-resource annotation

Key principles:

1. **Community consent:** Involve speakers in decisions
2. **Benefit sharing:** Results should help community
3. **Data sovereignty:** Community controls their data
4. **Fair compensation:** Pay fair wages, not exploitative
5. **Attribution:** Credit annotator contributions

Avoid:

- Extractive research (take data, give nothing back)
- Helicopter science (no local collaboration)

Building Sustainable Projects

Beyond one-time annotation

Sustainability strategies:

1. Train local researchers
2. Build lasting partnerships
3. Open-source tools and data
4. Document processes for replication
5. Create maintainable infrastructure

Goal: Enable community to continue work independently

Practical Tips

For low-resource annotation projects:

1. Start small (pilot with 1 language first)
2. Invest in annotator training
3. Adapt guidelines to local context
4. Build relationships with communities
5. Plan for long-term sustainability
6. Document everything for reproducibility

Resources:

- Masakhane community
- AmericasNLP workshop
- ACL SIGTYP (typologically diverse NLP)

Next Class: Best Practices & Future

Lecture 26 (Apr 29): Best Practices & Future Directions

Topics:

- Course summary
- Best practices checklist
- Emerging trends in annotation
- Future of human annotation
- Career paths in annotation

Project: Final presentations May 4

Key Takeaways

1. **Most languages** lack NLP resources
2. **Annotator recruitment** requires creative strategies
3. **Cross-lingual transfer** reduces annotation needs
4. **Quality at scale** requires systematic processes
5. **Community involvement** is ethically essential
6. **Sustainability** should be planned from the start

Questions?

Office Hours: Wednesdays 1-3pm, Volen 109

jinzhao@brandeis.edu