Day 7: Language Resources for Urdu & Pashto NLP
Libraries, Corpora & Open Datasets

Sami Uddin Shinwari
Subject Specialist – IT | NLP Researcher
MS Data Science (FAST NU Peshawar)
BS Computer Science (UET Peshawar)

December 23, 2025

Day 7 Overview

- Why language resources matter
- Urdu NLP datasets and libraries
- Pashto NLP datasets and tools
- Challenges in low-resource languages

Why Language Resources Matter

- Models depend on data quality
- Low-resource languages suffer from data scarcity
- Most NLP progress starts with corpora

Core Idea:

Better data often beats better models

Urdu NLP Resources

Corpora & Datasets

- EMILLE Urdu Corpus
- CRULP Urdu Treebank
- Urdu text datasets (Kaggle)

Libraries & Tools

- UrduHack (Python)
- Stanza (limited support)
- Indic NLP Library

Urdu Data Example

Sentence:

یہ اردو زبان کی ایک مثال ہے

Challenges:

- Tokenization ambiguity
- No capitalization
- Morphological variation

Pashto NLP Resources

Corpora & Datasets

- Leipzig Pashto Corpus
- BBC / VOA Pashto datasets
- Universal Dependencies (limited)

Models & Tools

- FastText Pashto embeddings
- HuggingFace multilingual models

Pashto Data Example

Sentence:

دا پښتو ژبې یوه بیلګه ده

Challenges:

- Dialectal variation
- Sparse annotated data
- Script normalization issues

English vs Urdu & Pashto Resources

- **English**
  - Massive labeled datasets
  - Mature NLP libraries
  - Strong pretrained models
- **Urdu & Pashto**
  - Data scarcity
  - Manual preprocessing
  - Research-driven datasets

Open Challenges

- Lack of standard benchmarks
- Inconsistent annotations
- Limited domain diversity
- Dependency on multilingual models

Key Takeaways (Day 7)

- Language resources are the backbone of NLP
- Urdu & Pashto need more open datasets
- Dataset creation is high-impact research
- Community contribution is essential

#Day7 #UrduNLP #PashtoNLP #LowResourceLanguages

References

- Hardie, A. (2003). Urdu corpus development. Lancaster University.
- Rahman et al. (2023). Pashto NLP challenges. IJNLC.
- Leipzig Corpora Collection.
- HuggingFace Model Hub.