**1. AI4Bharat IndicNLP Corpora**

* Large-scale Indic language corpora suitable for OCR text and raw multilingual text.
* GitHub Repository & Download:
  + <https://github.com/AI4Bharat/indicnlp_corpus>
* Website with various Indic NLP resources from AI4Bharat:
  + <https://indicnlp.ai4bharat.org/pages/indicnlp-corpus/>

**2. MILPaC: Multilingual Indian Legal Parallel Corpus**

* High-quality legal text parallel corpora in English + 9 Indian languages.
* GitHub Repository (includes download and usage info):
  + <https://github.com/Law-AI/MILPaC>
* Hugging Face model for legal translation (optional for training/inference):
  + <https://huggingface.co/law-ai/InLegalTrans-En2Indic-1B>

**3. Hugging Face Indian Legal NER Datasets**

* Annotated Named Entity Recognition datasets for Indian legal texts.
* Dataset Links:
  + InLegalNER dataset: <https://huggingface.co/datasets/opennyaiorg/InLegalNER>
  + SpaCy-compatible NER model: <https://huggingface.co/opennyaiorg/en_legal_ner_trf>

**4. Anuvaad Parallel Corpus (Legal + Administrative)**

* Parallel corpora for Indic languages including legal, governance, official texts.
* GitHub Repository for parallel corpus:
  + <https://github.com/project-anuvaad/anuvaad-parallel-corpus>
* Anuvaad Corpus Tools for crawling and building datasets:
  + <https://github.com/project-anuvaad/anuvaad-corpus-tools>
* More info on the Anuvaad project:
  + [https://anuvaad.sunbird.org](https://anuvaad.sunbird.org/)

**5. OPUS Open Parallel Corpus (Indic-English subsets)**

* Additional parallel corpora that can be used for cross-lingual alignment.
* Website to download OPUS corpora:
  + <https://opus.nlpl.eu/>

**6. NLTK Corpora (Stopwords, Gazetteers, Names lists)**

* Built-in within NLTK Python package for stopwords and lexical resources for Hindi, English, and other Indian languages.
* NLTK official:
  + <https://www.nltk.org/nltk_data/>