ChatGPT

Khasi Language NLP Resources

Khasi (ISO 639-3 `kha`) is an Austroasiatic (Khasi–Palaungic) language of Meghalaya, India, spoken by about 1.05 million people [1]. It uses the Latin script (with a few digraphs) [2]. Public resources for Khasi NLP are scarce, but recent efforts have produced open models, corpora and lexicons. Below we summarize key **models, datasets and tools** (with licenses and usage) that support Khasi tasks like translation, dictionary building and linguistic processing.

Pretrained and Multilingual Models

- **Hugging Face MarianMT models (Eng→Kha)** – Two English→Khasi translation models are available from user *Bapynshngain*. The earlier `MarianMT-en-kha` (77.5M parameters) was fine-tuned from Helsinki's opus-mt-en-ROMANCE model [3]; its model and code are public (license ECL-2.0) [4]. An improved model `Bapyn-En-Kha` (also 77.5M) uses a larger (~40K sentence) parallel corpus and is MIT-licensed [5]. Both can be used via Transformers for text-to-text translation (examples on the model pages). These models achieve basic translation quality (work-in-progress) and are intended for low-resource transfer learning [3] [6].

- **Parallel Corpus** – The `English-Khasi-Parallel-Corpus` dataset (MIT license) on Hugging Face contains roughly 37.7K aligned English–Khasi sentence pairs [7]. It is the training data for the above models and can be used for NMT or other tasks. (Note: Access requires agreeing to HF terms. A CSV sample is ~8.65 MB [8].)

- **Other Multilingual Models** – No monolingual BERT or GPT exists for Khasi. Generic multilingual models (mBERT, XLM-R, MuRIL) do *not* explicitly include Khasi. For example, Google's MuRIL covers 17 Indic languages (Assamese, Hindi, etc.) but not Khasi [9]. Some researchers generate Khasi word embeddings via contextual models [10], but these are not released as standalone models. In practice, one can apply general multilingual tokenizers or language-agnostic LLMs (e.g. GPT-4) with caution, but performance on Khasi is unverified.

Corpora and Text Data

- **Khasi Annotated Corpus (Tham 2020)** – Medari Tham compiled a POS-tagged Khasi corpus of ~94.6K tokens (4,386 sentences, 5,465 word types) using the BIS tagset [11]. This corpus (not freely downloadable, but described in literature) was used to train POS taggers and parsers [10] [11]. It demonstrates linguistic properties: Khasi is analytic with *no inflectional morphology* [12] (prefixes rather than suffixes mark grammatical categories). Tham's demo site (medaritham.pythonanywhere.com) shows live POS-tagging and shallow parsing for input Khasi sentences [13].

- **English–Khasi Parallel Texts** – In addition to the HF dataset above, researchers have collected bilingual texts from stories, songs and other translated documents [14]. Hujon *et al.* (2024) report

building a “sizable” parallel corpus for NMT experiments [15]. Part of these data (~40K sentence pairs) came from NIT Silchar and the Tatoeba project, augmented by manual translations [6]. While much of this material is unpublished, it indicates that an English–Khasi corpus on the order of 10^4–10^5 sentence pairs exists for training MT systems [7] [6].

- **Dictionary/Gloss Corpora** – The FreeDict project provides downloadable bilingual dictionaries including Khasi. For example, *FreeDict Khasi–English* (2280 headwords) and *Khasi–German* (995 headwords) are available [16]. These are in dictd or XML format and presumably CC-licensed or public domain (as FreeDict aims to be fully free). They can serve as lexicon resources or seed data for building full dictionaries.

- **Monolingual Text** – There is no large published monolingual corpus (e.g. Wikipedia) for Khasi. Some news sites and literature exist (e.g. *Sal Ka Por*, *Shillong Times* in Khasi) which could be crawled. The IIT Guwahati CLST center aims to create text/speech databases for NE minority languages [17]. For now, practitioners must rely on small text collections or generate synthetic text via translation models for tasks like language modeling.

Lexicons and Dictionaries

- **Khasi–English Dictionary** – The FreeDict *Khasi–English* dictionary (v0.2.2) is open-source: it covers 2280 Khasi headwords with English glosses [16]. Likewise, *Khasi–German* exists. These can be downloaded (see FreeDict Downloads) and used in dictionary applications (GoldenDict, dictd, etc.).

- **Khasi–Khasi Dictionary (Lexonomy Project)** – A community project (GitHub: *khasi-dienshonhi*) is digitizing Iarington Kharkongor’s 1968 Khasi–Khasi dictionary using OCR and Lexonomy [18] [19]. They provide scripts to clean OCR output and import entries into Lexonomy (an open-source dictionary platform [20]). The result is a structured Khasi–Khasi lexicon (with part-of-speech and Khasi definitions) hosted locally. The underlying text appears based on a 1968 edition held at the Shillong library [18]. While still being proofread, this Lexonomy-based dictionary will allow collaborative editing and export in XML/JSON once complete.

- **Other Lexical Resources** – Traditional print dictionaries (Singh 1906, etc.) have been scanned by the Digital Library of India. For example, Kharkongor’s 1968 dictionary metadata is online [18]. The ASJP database lists a core 40-item wordlist for Khasi (useful for linguists) [1]. No WordNet or morphological analyzer for Khasi is available, reflecting the language’s analytic grammar and the paucity of NLP tooling.

NLP Tools and Libraries

- **Part-of-Speech Tagging / Parsing** – Tham (2020) developed hybrid POS taggers (HMM+CRF) and a shallow parser for Khasi, achieving ~92–95% accuracy [10] [12]. While the code isn’t officially released, the online demo (see above) shows that tokenization and tagging work on arbitrary Khasi input. These tools rely on features like capitalization and affix patterns (reflecting Khasi’s lack of inflection) [12]. In practice, one could also train a generic tagger using the 95K-token annotated corpus (if obtained under license) or by transfer from related languages.

- **Tokenization and Scripts** – Khasi text is written in Latin script, so standard Unicode Latin tokenization applies. Off-the-shelf whitespace/punctuation tokenizers (e.g. in NLTK, SpaCy's `xx_ent_wiki_sm`, HuggingFace tokenizers) will segment Khasi reasonably. The Bible or news text in Khasi show that words are separated by spaces, with digraphs like **ch**, **ng**, **sh** treated as two letters. There are no special script normalization issues beyond usual Unicode NFC normalization.

- **Transliteration / Unicode** – No transliteration tool is needed for Khasi (it already uses Latin script). All letters are in Unicode's Basic Latin block. (Khasi-specific letters like **'ng** or diacritics are not distinct codepoints.) One should ensure Khasi text uses consistent fonts (no Guwahati-style scripts) and normalize quotes/apostrophes, but no specialized lib is required.

- **Language Identification / LID** – Khasi is not covered by many LID datasets. Projects like NLLB-200 include "LID for 200 languages including 27 Indic languages" [21], but Assam's official languages list (Odia, Meitei, Nepali, Sanskrit) does not explicitly list Khasi. Bhashini's Language ID APIs will likely add Khasi soon (after the recent MoU). For now, a custom LID model could be trained on the limited text available, or one can use any 200-language LID model with Asia-focused data (it may confuse Khasi with related languages without training).

- **Morphology / Stemming** – Khasi has minimal inflection (no noun/verb conjugation) [12]. Its morphology is primarily derivational (prefixes/suffixes for aspect/negation etc.). There is no published stemming or lemmatization tool; a rule-based approach (e.g. stripping common suffixes) might help, but is application-specific. Most NLP pipelines for Khasi simply treat words as atomic tokens.

Government and Academic Initiatives

- **Bhashini (Digital India)** – In April 2025 the Meghalaya government signed an MoU to integrate Khasi (and Garo) into the national BHASHINI platform [22]. BHASHINI (MeitY's language technology initiative) aims to provide APIs for translation, speech, OCR, etc. This means Khasi could soon have official machine translation services, speech recognizers and OCR via the Unified Language Contribution API (ULCA). The MoU explicitly cites "maturing the Khasi language" on the digital map [22].

- **IIT Guwahati CLST** – The Centre for Linguistic Science & Technology (CLST) at IIT Guwahati is building resources for NE languages (including Khasi) [17]. A 2022 MEITY-funded project (BHASHINI/NLTM initiative) will develop **speech-based tools** (keyword-spotting, speech recognition) for healthcare information in Khasi (as well as Hindi, Assamese, Bengali, Bodo, Manipuri, Mizo, Nagamese, Nepali) [23]. CLST will also create text/speech corpora for Khasi and other minority languages [17]. These efforts signal government support and should produce open datasets and models (e.g. ASR training data) in coming years.

- **Local Community Efforts** – The Khasi Literature Society and Khasi Sahitya Academy help standardize and publish Khasi material (not NLP tools per se, but important for language planning). Independent researchers (e.g. at St. Anthony's College Shillong and IIIT Manipur) have published surveys and MT studies [14] [15]. The GitHub and Hugging Face projects mentioned above reflect grassroots contributions from students and enthusiasts. For dictionary-building, the use of open

tools like Tesseract OCR and Lexonomy (as in the khasi-dienshonhi project [19]) demonstrates community-driven digitization efforts.

Usage Notes and Licenses

- **Licenses** – The Bapynshngain HF models and dataset are MIT/ECL licensed [4] [5] (i.e. permissive). FreeDict data is public domain or CC0-like. The khasi-dienshonhi code is on GitHub (no license indicated, but the source texts are likely public domain). Lexonomy is open-source (GPL) but running a lexicon doesn't impose this on content. Users should check individual licenses (e.g. OU models vs MIT) when redistributing.

- **Data Formats** – The parallel corpus is CSV/UTF-8 (English, Khasi columns). FreeDict uses XML/dictd format. The khasi-dienshonhi exports XML (compatible with Lexonomy's DTD). Standard text files (UTF-8) suffice for monolingual content. Tools: any Python/Unix text processing (e.g. NLTK, spaCy, Pandas) can handle Khasi if given UTF-8 input.

- **Practical Integration** – To build a Khasi dictionary or MT system, one might combine resources: e.g. use the FreeDict and khasi-dienshonhi glosses as lexical entries; train a MarianMT model on English–Khasi data [3] [6]; use Google Translate API as a stopgap (unofficially demonstrated [24]); and deploy with NFEs (Named Entity transliteration not needed). For preservation, scanning more historical texts (as in the GitHub project) and adding them to Freedict or Lexonomy will be valuable.

Summary: In summary, Khasi benefits from a handful of open resources: the **Tham corpus** (POS-tagged text) [10] [11], the **Bapynshngain parallel dataset** (~37K sentences, MIT license) [7], and associated **MarianMT models** for translation [4] [5]. **Dictionaries** like FreeDict's Khasi–English (2280 entries) [16] and a crowdsourced Khasi–Khasi lexicon (Lexonomy) exist. Academic groups (IITG CLST, IIIT) and government initiatives (Bhashini) are actively developing Khasi NLP technologies. These tools and data provide a foundation for building dictionaries, translators and language-preservation systems for Khasi.

Sources: Most information is drawn from published papers, open catalogs and project pages, including Tham (2020) demo [10], the AI4Bharat Indic NLP Catalog [11], Hugging Face model/dataset cards [4] [5] [7] [25], FreeDict downloads [16], GitHub projects [18] [19], and news reports on Bhashini/CLST projects [22] [23]. These cover available open-source and community resources for Khasi NLP.

---

1 The ASJP Database - Wordlist Khasi
https://asjp.clld.org/languages/KHASI

2 Scripts and Languages
https://www.unicode.org/cldr/charts/45/supplemental/scripts_and_languages.html

3 4 Bapynshngain/MarianMT-en-kha · Hugging Face
https://huggingface.co/Bapynshngain/MarianMT-en-kha

5 6 7 Bapynshngain/Bapyn-En-Kha · Hugging Face
https://huggingface.co/Bapynshngain/Bapyn-En-Kha

8 25 Bapynshngain/English-Khasi-Parallel-Corpus at main
https://huggingface.co/datasets/Bapynshngain/English-Khasi-Parallel-Corpus/tree/main

9 google/muril-base-cased - Hugging Face
https://huggingface.co/google/muril-base-cased

10 12 13 NLP Tools for Khasi, a low resource language
https://aclanthology.org/2020.icon-demos.10.pdf

11 21 GitHub - AI4Bharat/indicnlp_catalog: A collaborative catalog of NLP resources for Indic languages
https://github.com/AI4Bharat/indicnlp_catalog

14 (PDF) Existing English to Khasi Translated Documents for Parallel Corpora Development : A Survey
https://www.researchgate.net/publication/328904018_Existing_English_to_Khasi_Translated_Documents_for_Parallel_Corpora_Development_A_Survey

15 Neural machine translation systems for English to Khasi: A case study of an Austroasiatic language | OpenReview
https://openreview.net/forum?id=2um5SfYsxQ

16 Downloads — FreeDict
https://freedict.org/downloads/

17 Indian Institute of Technology Guwahati : भारतीय प्रौद्योगिकी संस्थान गुवाहाटी
https://www.iitg.ac.in/iitg_academic?aca=centre-for-linguistic-science-and-technology

18 19 20 GitHub - udaycruise2903/khasi-dienshonhi: Documentation and scripts of khasi-khasi Dictionary Digitalisation project
https://github.com/udaycruise2903/khasi-dienshonhi

22 Garo Khasi Languages BHASHINI: Meghalaya govt signs MoU for integration of Garo, Khasi languages into BHASHINI, ET Government
https://government.economictimes.indiatimes.com/news/digital-india/meghalaya-govt-signs-mou-for-integration-of-garo-khasi-languages-into-bhashini/120728343

23 IIT Guwahati Researchers developing 'Speech Technologies for North Eastern Languages' – KRC TIMES
https://www.krctimes.com/stories/iit-guwahati-researchers-developing-speech-technologies-for-north-eastern-languages/

24 FREE Khasi to English Translation with EXAMPLES - Instant Khasi Translation
https://www.easyhindityping.com/khasi-to-english-translation