ChatGPT

Open-Source Dictionary Creation Platforms

- **Lexonomy** – a MIT-licensed, web-based dictionary writing and publishing system. Users can create custom entry schemas, edit entries in an online XML editor, and publish the dictionary as a browsable site. As the project homepage notes, *"Lexonomy is a cloud-based, open-source system for writing and publishing dictionaries"* [1] . It requires no setup or coding experience, making it ideal for small indigenous lexicons.
- **FreeDict** – an open collection of bilingual dictionaries in TEI XML format (e.g. Khasi–English). The FreeDict project provides *"free (open source) dictionary databases, to be used both by humans and machines"* [2] . According to its site, FreeDict now offers 140+ dictionaries in ~45 languages, all editable under free licenses [3] [2] . These TEI-encoded resources (and conversion tools) can be adapted or imported into new dictionary backends or search interfaces.
- **OpenLexicon** – an open content project (CC BY-SA) providing scripts and web apps to access lexical databases. Its repository contains a "directory of lexical databases" plus *"scripts to download and query lexical databases"* and GUI apps to browse them [4] . In practice this means one can plug in any TEI/XML dictionary and generate an HTML table or concordancer interface (for example, as shown in the OpenLexicon web demo [4]). Although not MIT-licensed code (it uses CC BY-SA), the scripts and UI can inspire similar tools.
- **Terminology Databases (Terminologue)** – while not strictly for general dictionaries, Terminologue (open source, DCU 2019) is a cloud-based terminology management system created by the Lexonomy team [5] [6] . It shows how specialized glossaries (e.g. technical vocabularies) can be managed. Its code and design might be repurposed for a Khasi dictionary backend.

NLP Libraries for Low-Resource/Indigenous Languages

- **Indic NLP Library** – a Python toolkit for Indian languages. Its goal is *"to build Python-based libraries for common text processing and NLP in Indian languages"* [7] . It includes modules for normalization, tokenization, transliteration and more. While Khasi is Austroasiatic (not Indic), many pipeline tools (unicode normalization, tokenizers, transliteration) could be adapted to Khasi using this library as a model.
- **fastText (Facebook Research)** – an efficient library for word embeddings and text classification (MIT license). fastText provides pretrained word vectors for 157 languages [8] , and supports training on custom data. A Khasi corpus could be used to train a fastText model, yielding vector representations for Khasi words to power semantic search or clustering. (fastText also has tools for language identification and can generate embeddings for out-of-vocabulary words.)
- **spaCy** – a popular industrial-strength NLP library (MIT license) with pretrained pipelines. spaCy *"currently supports tokenization and training for 70+ languages"*, with models for tagging, parsing and NER [9] . While Khasi is not yet in spaCy's model set, its open framework and training recipe could be extended. For example, one could train new spacy pipelines on Khasi corpora (once available) to provide part-of-speech tagging or lemmatization.
- **Stanza (Stanford NLP)** – a Python toolkit for many languages (Apache license). Stanza's neural models cover *"70 (human) languages"*, and the framework makes it easy to train new language models [10] . Like spaCy, it requires training data, but its tools for dependency parsing, tokenization,

etc. are well-suited to low-resource adaptation. Building a Khasi model in Stanza could assist in cleaning and analyzing Khasi text (which in turn would help with dictionary example sentences).

- **Masakhane** – an open research community and code collection focused on African (mostly low-resource) languages. Its GitHub org contains many projects (machine translation, QA, evaluation) on dozens of languages [11] [12] . Although African-focused, Masakhane's approach (crowdsourced model training, shared datasets) is a model for low-resource NLP. The tooling (e.g. Hugging Face training scripts, data pipelines) and community framework could be adapted for Khasi.
- **NLLB-Serve (Meta's NLLB-200)** – an Apache-2.0 web service providing access to Meta's 200-language translation models [13] . NLLB-Serve offers a REST API and web UI for translation. Even if direct Khasi translation isn't available in NLLB, its codebase demonstrates how to serve large multilingual models. More generally, NLLB is a state-of-the-art open model for low-resource MT, and its GitHub (and derivatives like *Open-NLLB*) can be used to bootstrap Khasi translation or semantic similarity tasks.

Indigenous Lexical Data Projects and Digitization

- **Khasi-OCR (udaycruise2903/khasi-ocr)** – an open GitHub project to train Tesseract OCR models for Khasi script [14] . It uses LSTM training on printed Khasi text (e.g. books, dictionaries). This project shows how to convert scanned Khasi dictionaries or texts into editable Unicode. Although it has no formal license file, its code and training data (listed openly) could be reused or improved. Effective OCR is a first step toward digitizing legacy Khasi dictionaries or manuscripts.
- **Wiktionary Extraction (DBpedia/DBnary)** – projects like **DBpedia's Wiktionary RDF extraction** provide tools to pull structured data from Wiktionary. DBpedia's framework "extract[s] semantic lexical resources (an ontology about language use) from Wiktionary," including languages, parts of speech, definitions, synonyms and translations [15] . In parallel, the **DBnary** project offers an OntoLex/Lemon-based pipeline to harvest multilingual dictionary data from Wiktionary and publish it as linked data [16] . Both are open-source efforts (Scala and Java code respectively) that illustrate how a crowd-sourced, wiki-based dictionary can be parsed and indexed. In a Khasi context, one could use similar techniques (or these tools themselves) to extract any Khasi entries from Wiktionary or other collaborative lexica.
- **Wiktionary Data Dumps (Wiktextract)** – while not a single project, tools like *wiktextract* (a Python library) can parse Wiktionary XML dumps and extract word senses. Using such tools on the English and any existing Khasi Wiktionary could yield usable dictionaries. (No direct citation, but many codebases on GitHub do exactly this.)
- **Freedict Datasets** – as noted above, FreeDict's Khasi–English dictionary is already in TEI XML (FD format) and could serve as a base. The FreeDict GitHub shows a `kha-eng` directory with TEI source, licensed under GPL/BSD. This existing bilingual data can be machine-translated or reversed to seed a Khasi–Khasi lexicon (for example by using its English glosses to generate Khasi definitions, or vice versa).

Web-Based Dictionary/Glossary Projects

- **Kamusi Project** – an international collaborative dictionary project. The (now Swiss-based) Kamusi Project aims to build *"dictionaries and other language resources for every language,"* freely available online [17] . It is volunteer-driven and open in spirit. For Khasi, adopting Kamusi's platform or community model could be fruitful: local speakers can register, contribute words and definitions, and

interlink glossaries. (The Kamusi codebase itself is not public, but its design points the way to a crowdsourced dictionary portal.)

- **Wikimedia Tools (Wikidata Lexemes)** – Wikidata's lexeme section is an open lexicographic database where any language's words and definitions can be added. Although not a "project" with a standalone repo, it is an open platform (GPLv3 for Wikibase) that supports cross-lingual links. A Khasi dictionary could be built on Wikidata Lexemes or hosted via a wiki (e.g. a Khasi Wiktionary page). Data from Wikidata can then be accessed via SPARQL or its REST APIs.
- **Lexicala / Gloss APIs (open alternatives)** – While commercial (Babylon/Lexicala) APIs exist, open alternatives like WordNet's REST APIs or open WordNet projects (e.g. Open Multilingual WordNet) can inspire a Khasi lexical API. Tools such as the Python *NLTK* include open WordNet data, and services like *OpenTaal* provide open lexical resources. In summary, one could repurpose a simple dictionary API (e.g. using Flask+Solr/SQLite) as in many GitHub examples.

AI & Semantic Search for Language Documentation

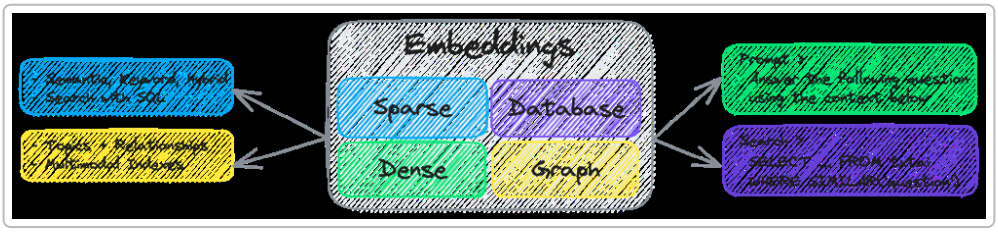

Figure: Open-source semantic search frameworks (e.g. txtai) use vector embeddings (sparse/dense/graph indexes) to enable meaning-based lookup. Modern AI frameworks can enable powerful dictionary features. For example, **txtai** (Apache 2.0) is *"an all-in-one AI framework for semantic search, LLM orchestration and language model workflows."* It builds a unified "embeddings database" combining sparse, dense and graph indexes to power search and RAG pipelines [18]. In practice, txtai can index dictionary entries as high-dimensional vectors: users could then query in Khasi and retrieve semantically similar entries, example sentences or synonyms even if exact keywords differ. Another example is **Deepset Haystack** (Apache 2.0), a production LLM framework. Haystack can orchestrate various embedding models and vector DBs to perform retrieval-augmented generation and semantic question-answering [19] [20]. A Khasi dictionary could use such a pipeline to answer queries (e.g. *"show me Khasi words related to* 'ka sait'"*) or to auto-generate example usages via an LLM.

- **Transformers/Multilingual Models** – Open-source LLMs like Hugging Face's multilingual BERT, XLM-RoBERTa or BLOOM can be fine-tuned on Khasi text for language tasks. Meta's NLLB (discussed above) provides state-of-art translation; similarly, language models like mBERT (Apache 2.0) include many low-resource languages and could be trained to embed Khasi words and definitions into a shared semantic space. These embeddings can power clustering of dictionary senses or detect synonyms.
- **Knowledge Bases & Linked Data** – Initiatives like OntoLex-Lemon and LMF define graph-based dictionary data formats. Converting Khasi dictionary content into OWL/RDF (as done by the DBpedia/DBnary projects) would allow linkage with other languages and semantic queries (via SPARQL). Tools for building semantic lexicons (e.g. Poolparty, Github *lexdict*) could be used once Khasi entries are encoded. For example, the DBnary Java toolkit automatically extracts multilingual lexica from Wiktionary to RDF [16]; a similar pipeline could be built for Khasi content.

- **Example Projects** – Other relevant open projects include *OpenMultilingualWordnet* (crowdsourced WordNets) and *ConceptNet* (graph of common knowledge). Though not Khasi-specific, they demonstrate how lexical semantics can be represented. Embedding these into a dictionary app (for definition auto-completion or semantic search) is an active area of AI-driven lexicography.

Sources: Lexonomy dictionary platform [1]; FreeDict open bilingual corpora [3] [2]; Indic NLP Library tools [7]; fastText multilingual embeddings [8]; spaCy/Stanza multilingual NLP [9] [10]; NLLB translation API [13]; txtai semantic search framework [18]; DBpedia/DBnary Wiktionary extraction [15] [16]; Kamusi collaborative dictionary [17]; Khasi OCR project [14]. Each is open-source (MIT/Apache/GPL) and actively maintained (2022–2025).

---

[1] GitHub - elexis-eu/lexonomy: A cloud-based, open-source system for writing and publishing dictionaries.
https://github.com/elexis-eu/lexonomy

[2] GitHub - freedict/fd-dictionaries: hand-written dictionaries from the FreeDict project
https://github.com/freedict/fd-dictionaries

[3] Home — FreeDict
https://freedict.org/

[4] GitHub - chrplr/openlexicon: Access to lexical databases
https://github.com/chrplr/openlexicon

[5] Making lexicography truly digital: the road to DMLex - OASIS Open
https://www.oasis-open.org/2025/01/30/the-road-to-dmlex/

[6] Terminologue
https://www.terminologue.org/

[7] GitHub - anoopkunchukuttan/indic_nlp_library: Resources and tools for Indian language Natural Language Processing
https://github.com/anoopkunchukuttan/indic_nlp_library

[8] GitHub - facebookresearch/fastText: Library for fast text representation and classification.
https://github.com/facebookresearch/fastText

[9] GitHub - explosion/spaCy: Industrial-strength Natural Language Processing (NLP) in Python
https://github.com/explosion/spaCy

[10] Overview - Stanza
https://stanfordnlp.github.io/stanza/

[11] [12] Masakhane · GitHub
https://github.com/masakhane-io

[13] GitHub - thammegowda/nllb-serve: Meta's "No Language Left Behind" models served as web app and REST API
https://github.com/thammegowda/nllb-serve

[14] GitHub - udaycruise2903/khasi-ocr: khasi-ocr is a project to create OCR model for khasi language using tesseract-ocr by LSTM layer training.
https://github.com/udaycruise2903/khasi-ocr

15 Wiktionary RDF extraction | DBpedia
https://downloads.dbpedia.org/wiki-archive/wiktionary-rdf-extraction.html

16 Wiktionary Support – Project Summary
http://kaiko.getalp.org/static/reports/3.1.21/summary.html

17 Kamusi project - Wikipedia
https://en.wikipedia.org/wiki/Kamusi_project

18 GitHub - neuml/txtai: All-in-one open-source AI framework for semantic search, LLM orchestration and language model workflows
https://github.com/neuml/txtai

19 20 GitHub - deepset-ai/haystack: AI orchestration framework to build customizable, production-ready LLM applications. Connect components (models, vector DBs, file converters) to pipelines or agents that can interact with your data. With advanced retrieval methods, it's best suited for building RAG, question answering, semantic search or conversational agent chatbots.
https://github.com/deepset-ai/haystack