Background

Amchigale Konkani, a dialect of Konkani spoken by approximately 2 million people along the west coast of Karnataka, is at risk of extinction. Unlike Goan Konkani, which has a digital presence and is supported by Google Translate, Amchigale Konkani has no such resources. One of the key challenges is that Amchigale Konkani lacks its own script and is traditionally written using either Kannada or Devanagari (Marathi) script.

However, Amchigale Konkani (hereafter simply called Konkani) follows the same grammatical structure as other Indian languages like Hindi and Marathi. It also has words that are derived from Sanskrit. We could potentially use this to our benefit to recreate sentences in Konkani.

Since the language is not well-documented or widely available on the internet, its usage is declining with each generation. There are no comprehensive dictionaries, grammar references, or structured learning materials, making it difficult for younger generations to learn and use the language.

With advancements in AI—specifically Large Language Models (LLMs), speech-to-text, and Optical Character Recognition (OCR)—we now have an opportunity to digitize and preserve Amchigale Konkani. By creating a digital corpus of Konkani words, equivalent meanings in English, Marathi and Hindi, and sentences, we can digitize the language aiding in its revival and preserving the rich cultural heritage of those who speak and write in Konkani.

The importance of language preservation extends beyond Konkani. Many indigenous languages around the world face similar challenges. If we successfully develop a framework to revive Konkani using AI, the same approach can be applied to other endangered languages.

Problem Statement

The goal of this initiative is to create (a) Amchigale Konkani dictionary by having others contribute Konkani words written in English and Devanagari along with it meaning in English. (b) Create a **Language Model (LLM)** for Amchigale Konkani (hereafter referred to as Konkani). (c) Create a Speech to Text and Text to Speech version of Konkani. We have the following constraints:

1. **Limited Digital Resources**:
    - Some Konkani is available in printed text, written in Devanagari script. We were able to use a combination of OCR and LLM to digitize about 100 pages of written Konkani. Since these were scanned pages that were digitized, and since LLMs do not understand Konkani inherently, some of these transcriptions may not be completely accurate.

- About 2000 **Konkani words** written either in Devnagiri or English has been captured along with their meanings in English or Marathi. This has been done via crowdsourcing from expert humans who speak Konkani.

2. **Human Expertise Available**:

    - A group of fluent Konkani speakers is willing to contribute by writing Konkani words into English or and Devangiri and then writing the English meaning. They can also write entire phrases. We would like to use this for the dictionary as well as to feed words to the language model.

3. **Potential Audio Resources**:

    - There are **100 hours of spoken Konkani content on YouTube**, which could be converted into text using speech-to-text AI.

    - However, since no existing Konkani language model exists, it is unclear how accurately AI can transcribe spoken Konkani words into Devanagari script. The limited attempts we made with existing multi-modal LLMs found that the LLM was transcribing the spoken Konkani word into the nearest Marathi or Hindi word.

4. **Linguistic Structure Considerations**:

    - Konkani shares a significant vocabulary overlap with **Sanskrit**.

    - Its **grammar and syntax** are structurally similar to **Marathi, Hindi, and Kannada**, which could aid in developing an AI model.
    - There are a few books written that teach Konkani to beginners. We could potentially use these books to train LLM on how to use Konkani.