METHODOLOGY

SO1 - Creation and Evaluation of a Bilingual English–Itaukei (Fijian) Parallel Corpus

Overview

This research involves the creation and evaluation of a bilingual English–iTaukei (Fijian) parallel corpus to support machine translation research in a low-resource language context. The work was carried out in four main stages: data collection, corpus creation & harmonization, corpus combination & structuring, corpus analysis, and corpus quality evaluation.

1. Data Collection

Parallel data was collected from real-world sources across multiple domains to ensure linguistic diversity and real-world applicability:

- **Legal** (Fiji Constitution)
- **Medical** (patient instructions, clinical texts, public health posters and awareness documents)
- **Dictionary/Definitions** (lexical mappings/entries)
- **Idioms** (figurative expressions-(literal and semantic meanings))
- **Conversational** (educational dialogue materials)
- **Religious** (Book of Genesis EN↔FJ verses)

2. Corpus Construction & Harmonization

All documents were converted into a canonical schema containing:

- Domain and subdomain
- Source and target language
- Translation direction
- Source and target text

A Python pipeline automatically:

- Extracted text from XLSX/CSV/PDF/DOCX sources
- Normalised text (encoding fix, trimming, whitespace)

- Detected and separated English vs Fijian content
- Added metadata and removed empty/noisy rows

3. Corpus Combination and Structuring

All domain-specific datasets were merged into a single combined corpus using a consistent data structure.

Two additional annotations were automatically added:

- Sentence length type (short, medium, long)
- Sentence function (e.g., instruction, warning, narrative, definition), tailored to each domain

This allows analysis not only by domain, but also by communicative function.

4. Corpus Analysis

The combined corpus was analysed to understand its characteristics, including:

- Size and domain distribution
- Balance between translation directions
- Sentence length patterns across domains
- Distribution of sentence functions

This analysis provides insight into the linguistic diversity and complexity of the corpus.

5. Corpus Quality Evaluation

Corpus quality was evaluated using **three complementary approaches**:

1. **Automatic checks**
   Basic rule-based checks were used to detect encoding errors, duplicated entries, unusually short sentences, and potential misalignments.
2. **Semantic similarity scoring**
   Sentence-level semantic similarity was computed to identify pairs that are likely to be poorly aligned.

3. **Manual evaluation**
   A representative sample of sentence pairs was manually evaluated for adequacy, fluency, and meaning preservation. Agreement between annotators was measured to ensure reliability.

Based on these evaluations, a cleaned version of the corpus was produced, while problematic entries were retained separately for transparency.

Conclusion:
The resulting corpus is clean, diverse, and well-aligned, providing a credible foundation for MT evaluation, error analysis, and linguistic research on the iTaukei language.