Revitalizing Kalamang: Language Expansion through Large Language Models

Aaditya Bhargav *aaditya23006@iiitd.ac.in* **Md. Abuzar Khan** *abuzark@iiitd.ac.in*
Mohit *mohit21542@iiitd.ac.in* **Shah Jayshil Ketankumar** *jayshil23138@iiitd.ac.in*
Vani Mittal *vani23102@iiitd.ac.in* **Vishal Singh** *vishal21575@iiitd.ac.in*

Abstract

Machine translation (MT) has made significant strides in recent years, with large language models (LLMs) leading to improved performance across languages. However, translating low-resource languages remains an ongoing challenge, particularly for languages with little to no digital presence. In this survey, we analyze the latest research advancements in low-resource MT, with a focus on "A Benchmark for Learning to Translate a New Language from One Grammar Book" by introducing methods that leverage minimal resources, including single grammar books, bilingual word lists, and small parallel corpora [1]. We explore how these findings integrate with novel approaches in multilingual translation, vocabulary optimization, and multimodal methods. The paper highlights the limitations of current MT models, particularly with languages such as Kalamang and Zhuang, and proposes future directions that can address these limitations through more effective use of external knowledge, non-parametric memory, and better contextual understanding.

1. Introduction

Machine translation (MT) has undergone a paradigm shift with the development of large, pre-trained models capable of translating numerous languages with remarkable accuracy. Yet, low-resource languages remain largely underserved, primarily due to the scarcity of parallel corpora, grammars, and linguistic resources. With languages such as Kalamang, which have fewer than 200 speakers, translation models must rely on minimal resources, including grammar books and small parallel corpora [1].

This paper surveys the current landscape of low-resource MT with a central focus on the recent benchmark proposed by the paper *"A Benchmark for Learning to Translate a New Language from One Grammar Book"* (MTOB) [1]. The survey covers key methodologies such as in-context learning, instruction inference, vocabulary optimization, and multimodal models, presenting insights into how these approaches can contribute to the translation of extremely low-resource languages.

2. Problem Definition and Scope

Low-resource language translation presents a unique challenge. Most successful MT models depend on vast amounts of parallel data, which is unattainable for most of the world's languages. Many low-resource languages are spoken by small populations and lack sufficient written resources. In cases such as Kalamang, with fewer than 200 speakers, translation models must rely on minimal resources, including grammar books and bilingual word lists [1].

The scope of this survey centers around methodologies designed to address this problem, particularly:

- Learning a language's structure and translation using minimal resources like grammar books and bilingual word lists [1].
- Approaches to improving MT quality in extremely low-resource settings using external knowledge bases and multimodal inputs [4].

3. Literature Review

A Benchmark for Learning to Translate a New Language from One Grammar Book

The paper "A Benchmark for Learning to Translate a New Language from One Grammar Book" introduces MTOB (Machine Translation from One Book), a novel benchmark for translating between English and Kalamang, a low-resource language with fewer than 200 speakers. The unique aspect of this benchmark is that it mimics how humans learn languages by relying solely on a single grammar book, word lists, and a small corpus rather than large datasets. The study, which evaluates models such as LLaMA-2 and GPT-4, tests in-context learning and lightweight finetuning for translation tasks, comparing the performance to a human baseline. The work highlights challenges such as data scarcity, hallucination, and difficulty retrieving useful context, particularly with model finetuning. The paper also outlines future goals, including improving translation accuracy for low-resource languages, expanding the approach to other typologically diverse languages, and exploring multimodal models to support endangered language preservation and revitalization. [1].

An Incomplete Loop: Instruction Inference, Instruction Following, and In-Context Learning in Language Models

The paper "An Incomplete Loop: Instruction Inference, Instruction Following, and In-Context Learning in Language Models" explores the reasoning capabilities of language models (LMs) across deductive, inductive, and abductive reasoning, with a focus on tasks such as hypothesis proposal, in-context learning, and instruction following. The authors experiment with models like GPT-3.5-turbo, GPT-4, and LLaMA-2 for tasks like linear function learning and artificial language translation, specifically translating Kalamang. A novel aspect of the work is the integration of instruction inference within reasoning tasks, allowing the model to generate and refine instructions during problem-solving, contrasting prior approaches that relied on instruction back-translation. However, the study highlights challenges, particularly in abductive reasoning, which proves to be a weak point in current LMs, and suggests future directions such as advanced hypothesis verification and autonomous learning mechanisms to enhance reasoning consistency and accuracy. [2].

Low-Resource NMT with Smaller Vocabulary Sizes

The abstract you provided discusses a study on optimizing subword vocabulary sizes in neural machine translation (NMT) for low-resource languages. The authors highlight that current state-of-the-art models like Transformers, when applied to low-resource languages, show significantly lower performance than

on high-resource languages. They attribute this to the model's sensitivity to hyperparameters, particularly the subword vocabulary size. Their experiments demonstrate that using smaller vocabularies, as low as 1k tokens, leads to faster training, smaller model sizes, and better translation quality. In their experiments with languages like English-Akkadian, Lower Sorbian-German, and English-Manipuri, they found that smaller vocabularies not only improve the ChrF scores by up to 322% but also reduce model size by 66% and training time by up to 17%. This suggests that smaller vocabularies may be more effective in low-resource conditions than the default vocabulary size of 32k, commonly used in machine translation. The study concludes by advocating for careful selection of vocabulary sizes in NMT, especially when dealing with under-resourced languages, to maximize model efficiency and performance. [3].

Multimodal Machine Translation for Manipuri

This paper explores recent advancements in low-resource machine translation, particularly the inclusion of additional modalities such as visual and auditory inputs, which have been shown to boost accuracy. Studies highlight the effectiveness of multimodal approaches, with notable improvements in BLEU scores. Research has explored the integration of images and audio with text to enhance machine understanding and context for translation, addressing the challenge of limited written resources in underrepresented languages. Additionally, multimodal machine translation has been applied in various languages, indicating potential for improved performance across multiple language pairs. [4].

Chinese-Centric Neural Machine Translation for Low-Resource Languages

F This study, which focuses on Chinese as a hub language for low-resource translation, introduces bilingual curriculum learning, contrastive learning, and noise-robust methods to improve MT for low-resource languages like Zhuang [5]. Monolingual data and novel loss functions, such as the In-Trust loss, demonstrate the effectiveness of handling noisy, low-resource data. It highlights how NMT models like Transformer and Recurrent Neural Networks have been successful in resource-rich environments, but challenges persist in low-resource languages due to limited parallel settings. Recent Recents in data augmentation, transfer learning, and the use of monolingual data have improved The paper also discusses how contrastive learning and auxiliary language data can be used to address these challenges, providing insights into improving NMT systems. [5].

KARD: Knowledge-Augmented Reasoning Distillation

This paper addresses the key developments in large language models (LLMs) and their application to knowledge-intensive reasoning tasks. The authors discuss how LLMs, such as GPT-3.5 and others, have demonstrated remarkable performance in various domains, especially those requiring deep reasoning and domain-specific knowledge. However, they also emphasize the challenges of deploying LLMs in real-world scenarios, such as the high computational costs and privacy concerns. The paper highlights previous efforts to distill reasoning abilities from LLMs into smaller language models, noting that these approaches often fall short due to the limited capacity of smaller models to memorize knowledge. This motivates the need for techniques like Knowledge-Augmented Reasoning Distillation (KARD), which leverages external knowledge bases to supplement small models. The review also touches upon reasoning distillation, retrieval-augmented language models, and the limitations of existing methods in addressing

knowledge-intensive tasks. The use of neural rankers to improve document retrieval for reasoning is discussed as an innovative contribution to the field. [6].

Teaching Large Language Models an Unseen Language on the Fly

This paper focuses on the challenges of adapting large LLMs (LLMs) to low-resource languages, which have limited available data. Traditional methods, such as continual pre-training on monolingual texts (Yong et al., 2023) and the use of adapters like MAD-X (Pfeiffer et al., 2020), have been employed to improve model performance in these settings. Other approaches, such as supervised fine-tuning (SFT) using cross-lingual instructions (Cahyawijaya et al., 2023), have shown some success, but these require a substantial amount of training data. For machine translation, researchers have explored prompting-based methods, such as DIPMT (Ghazvininejad et al., 2023), which use dictionaries to aid translation in low-resource languages. However, the methods often assume some baseline knowledge of the target language, which is lacking for extremely low-resource languages like Zhuang. This paper addresses these gaps by introducing DIPMT++, a framework that enables LLMs to learn a new language solely through prompting, expanding on previous work with strategies like bilingual lexicon induction and synonym expansion to enhance performance in unseen languages. [7].

4. Overview

Our project aims to create a Cross-Linguistic Data Formats (CLDF) dataset for the Kalamang language, which involves processing racldfbencCLDFBencheve-structured CLDF tables. The project's code leverages *pydictionaria* and *cldfbencCLDFBench*eve this transformation, ensuring compliance with the CLDF Dictionary module's standards.

The project involves three key steps:

1. **Downloading or accessing raw data files**.
2. **Preprocessing and transforming the raw data**.
3. **Generating the CLDF dataset**.

This report outlines the key components, goals, and expected baseline outcomes of the project, covering preprocessing, dataset generation, logging, and final outputs.

1. Setup and Environment

We have the following dependencies:

- *pydictionaria*: To process the raw.sfm files and assist in converting them to the CLDF format.
- *A tool for managing and creating CLDF datasets is called CLDFBench.*
- *sfm2cldf*: Helps in processing. sfm data to CLDF schema.

To ensure the correct environment setup:

- We install the required libraries (*cldfbench*, *pydictionaria*, *sfm2cldf*) via pip.

- We upgrade pip to avoid compatibility issues with older versions.

For testing purposes, we install pytest-cldf, which ensures that we can validate the output CLDF files.

➢ **Pydictionaria**: provides tools for processing. SFMm (standard format marker) files are commonly used in lexicographic and linguistic datasets. This library offers utilities for preprocessing, cleaning, and transforming raw data into a format compatible with CLDF.
➢ **CLDFBench**: A toolkit designed for managing and creating CLDF datasets. It offers support for creating different types of linguistic datasets in the CLDF format, such as dictionaries, wordlists, or typological databases.
➢ **sfm2cldf**: Specifically used to convert.sfm data to CLDF schema. This can be included in Pydictionaria or installed as a standalone package.

2. Raw Data Handling and Preprocessing

Data Sources:

- Raw.sfm files, such as db.sfm and examples.sfm, are the primary sources of linguistic data. These files contain entries with markers representing various linguistic features (e.g., headwords, meanings, parts of speech).

➢ **db.sfm**: Contains lexical entries with their associated properties (e.g., headwords, parts of speech, glosses, etc.).
➢ **examples.sfm**: Includes example sentences or usage data related to the lexical entries.

- **Marker Mapping**: Markers in the.sfm files represent linguistic properties, such as part of speech (lx, va, sn, etc.). These markers are mapped using the marker_map, either provided in the md.json file or defaulting to sfm2cldf. DEFAULT_MARKER_MAP.

Preprocessing Tasks:

- **Dropping Irrelevant Entries:** The DropTracker class is used to remove unwanted data such as variant entries (e.g., marked as 'MLY' or 'pending').
- **Marker Merging and Modifications:** Entries with similar markers (e.g., 'pc_Eng' and 'pc_Kar') are merged, and various linguistic markers (e.g.,'mn','vet') are transformed.
- **Semantic Domain Parsing:** Semantic domains (e.g., 'aquatic life', 'birds') are parsed and standardized, ensuring that the dataset uses recognized categories.
- **Variant Handling**: Entries with specific markers (e.g.,'sp. var. of') are filtered out using the filter_sp_var function.

The preprocessing pipeline ensures that only relevant, cleaned, and normalized entries are passed to the next stage for dataset generation.

3. CLDF dataset generation

Structure: Using the CLDF Dictionary Module, the project converts the cleaned data into a CLDF-compliant dataset. The dataset generation involves the following components:

Tables Generated:

- **LanguageTable:** Contains metadata about the Kalamang language (name, ISO code, glottocode).
- **EntryTable:** Contains lexical entries with their associated headwords, parts of speech, meanings, and other linguistic features.
- **SenseTable:** holds semantic senses linked to each entry.
- **ExampleTable:** Provides usage examples for certain lexical items.
- **media.csv:** Stores any associated media files (if applicable, though optional for the baseline).

Schema Definition: The sfm2cldf.make_cldf_schema() function is used to create a CLDF schema that defines the structure and properties of the dataset, ensuring that the output files are valid and conform to CLDF standards.

Ensuring Data Integrity:

- **Column Titles**: Proper column titles are attached to the tables to ensure that the output is well structured.
- **Removing Senseless Entries**: Entries that do not have associated senses are filtered out to maintain data consistency.

Logging:

Logging is a critical part of the dataset generation process, ensuring that any errors or warnings are captured and can be debugged. The system generates several log files:

- **cldf.log**: Tracks the overall dataset processing and logs issues encountered.
- **examples.log**: Logs issues related to example sentences.
- **glosses.log**: Logs issues related to glosses in the dataset.

The logging mechanism is crucial for maintaining transparency during data transformation and debugging errors when they occur.

4. Validation and Testing

CLDF Validation: We perform testing with pytest-cldf to ensure the generated CLDF files are valid. The command! *Python test.py --cldf-metadata=cldf/cldf-metadata.json* validates the metadata and the overall structure of the output files.

Expected validation outcomes include:

- **No missing required columns** in the CLDF tables.
- **Proper column titles are** attached to the CLDF files, ensuring that each table has the correct headers.
- **No senseless entries:** Entries that don't have associated senses are removed to maintain data integrity.

Key Technical Considerations

1. **Data Integrity: Preprocessing functions ensure the data's integrity by filtering out invalid entries, merging markers correctly, and ensuring that the output dataset contains only useful, clean data.**
2. **Modular Design**: The dataset generation process is modular, allowing simple adjustments to preprocessing steps, schema generation, and validation. This is important for scalability when processing larger or more complex datasets.
3. **Automation**: The pipeline is designed to handle data ingestion, preprocessing, and output generation in an automated manner. This reduces the likelihood of human error and ensures that the dataset can be reproduced consistently.

5. Results

Entries.csv -

ID	Language_ID	Headword	Part_Of_Speech	Contains	Entry_IDs	Etymology	Main_Entry	Pronunciation	Source_Language	Variant_Form
LX000001	kgv	Ø	Ditransitive verb							
a_1	kgv	a	Interjection					a ; ah ; ā		
a_3	kgv	a	Interjection							
LX000002	kgv	=a	Grammatical marker					a		
LX000003	kgv	a'a	Interjection					a?a		
adat	kgv	adat	Noun							
ade	kgv	ade	Interjection							
adi	kgv	adi	Interjection							
adu	kgv	adu	Interjection					aˈdu		
afukat	kgv	afukat	Noun					afuˈkat͡		apukat ; alfukat ; afokat ; alpukat
ahat	kgv	ahat	Noun			al'ahad		aˈhat͡	Arabic	
-ahutak	kgv	-ahutak	Grammatical marker		-tain			aˈhutak͡		-autak ; -sutak
ajar	kgv	ajar	Verb						Malay or other AN	
ak	kgv	ak	Noun					ak͡		
akal	kgv	akal	Noun						Malay or other AN	
aknar	kgv	aknar	Noun					akˈnar		
aknar_kangun	kgv	aknar kangun	Noun					akˈnar ˌkaŋun		
akpis	kgv	akpis	Noun	-pis ; ak				akˈpis		
LX000004	kgv	*al	Noun					ˈal		
LX000005	kgv	*al	Classifier					ˈal		
alangan	kgv	alangan	Verb						Malay or other AN	
alanganrep	kgv	alanganrep	Verb	alangan ; rep						

Kalamang Dataset: entries.csv File Breakdown

The entries.csv file contains the Kalamang language's core linguistic entries. Each row represents a distinct lexical entry (word or phrase), accompanied by various details that provide context and metadata. The following is a detailed description of the columns in the file:

1. **ID**:
    - A unique identifier for each lexical entry (e.g., LX000001, LX000002).
    - The "LX" prefix indicates these are lexical entries, which can be cross-referenced with other tables (e.g., senses or examples).
2. **Language_ID**:
    - It indicates the language's ISO 639-3 code.
    - For Kalamang, it is kgv, ensuring each entry is correctly associated with the language.
3. **Headword**:
    - Shows the main form of the lexical item in Kalamang (e.g., gadadat as a noun, **afukafukat).** another noun).
    - In dictionaries, the headword is typically a word's citation form.
4. **Part_Of_Speech** Each entry's grammatical category is indicated. ry.
    - Examples include:
        - **A ditransitive verb is one that takes two objects (e.g., "give" in English).**
        - **Noun**: A word referring to a person, place, thing, or idea.

        - **Interjection**: words used to express emotions or exclamations.
    - This classification aids in understanding the function of the language's words.
5. **Related_Entries**:
    - Lists other entries or related subentries contained within this entry.
    - For example, it can reference related words or variants (e.g., **alanga rep** indicates references to multiple items).
6. **Entry_IDs**:
    - May cross-reference specific entries, though this column appears empty in the visible portion.
    - When populated, it could contain references to other related entries.
7. **Etymology**:
    - Provides the word's origin, explaining where it comes from.
    - For example, the letters f and ahad in the word afukat may indicate Arabic influence.
    - This provides historical and linguistic context for each entry.
8. **Main_Entry**:
    - Specifies whether the word is the main entry for a sequence of related words or variants.
    - Some entries, such as afukat, do not specify a main entry in the visible rows.
9. **Pronunciation**:
    - The phonetic representation of the headword is included.
    - For example, the entry **a** has the pronunciation listed as **a; ah; à**, indicating multiple pronunciations.
    - **Afukat** has the pronunciation **afu ka?**.
10. **Source_Language**:
    - Denotes the source language for loanwords or borrowed terms.
    - For example, **afukat** is noted as having Arabic as its source language, suggesting a borrowed term.
11. **Variant_Form**:
    - Lists variations of the headword.
    - For instance, **afukat** has variants such as **apukat** and **aflukat**.
    - Variants indicate alternate spellings, dialectal differences, or historical forms of the word.

Senses.csv -

ID	Description	Entry_ID	Antonym	Media_IDs	Scientific_Name	Semantic_Domain	Synonym	alt_translation1
SN000001	to give	LX000001						kasih
SN000002	[filler]	a_1						[kata pengisi]
SN000003	[interjection]	a_3						[kata seru]
SN000004	[focus]	LX000002						[fokus]
SN000005	[agreement interjection]; yes	LX000003						[kata seru persetujuan]; iya
SN000006	tradition	adat				culture and communication		adat
SN000007	[interjection expressing contempt or dissatisfaction]	ade				values and emotions		[aduh, adeh]
SN000008	[interjection expressing pain or discomfort]	adi				values and emotions		adih
SN000009	[interjection for sudden pain or surprise]	adu				values and emotions		[kata seru]; aduh
SN000010	avocado	afukat		c009d8495a5cc7d05422c7593e65895e ; 49dc39d38efb1424bd8cff60bdae7c60 ; e31d9bf4d827ece0cf2fae2e26e7a003		food, cooking, fire ; plants		alpukat
SN000011	Sunday	ahat				location, direction, time		hari minggu
SN000012	[quantifying pronoun suffix]; alone	-ahutak				bodily states, colours, dimensions, quantity		sendiri
SN000013	to teach	ajar				work		ajar
SN000014	to continue	ajar						ajar

Kalamang Dataset: `senses.csv` File Breakdown

The `senses.csv` file provides information about the meanings (senses) associated with the lexical entries from `entries.csv`. Each row represents a specific sense (meaning) of a lexical entry and includes various details such as synonyms, antonyms, semantic domains, and other metadata. The following is a detailed description of the columns in the file:

1. **ID**:
    - Contains unique identifiers for each sense (e.g., SN000001, SN000002).
    - The prefix "SN" indicates that these senses are associated with specific lexical entries. Each sense has its own unique ID, which can be used to reference other related data (such as examples).
2. **Description**:
    - Provides a short description or definition of the sense in the language.
    - Examples include:
        - **to give** for the entry LX000001.
        - **avocado** for the entry **afukat**.
        - **tradition** for the entry **adat**.
    - These descriptions convey the meaning of the word in the specific context of this sense.
3. **Entry_ID**:
    - The sense is linked to a specific lexical entry in entries.csv.
    - For example, SN000001 refers to entry LX000001, whereas SN000009 refers to entry adu.

    - This cross-referencing ensures that each sense is associated with a headword from the `entries. csv`.
4. **Antonym**:
    - Lists antonyms (words with opposite meanings) when applicable.
    - In the visible portion of the file, this column is mostly empty. When populated, it would provide semantic relationships between entries, highlighting contrasting meanings.
5. **Media_IDs**:
    - References media files associated with the sense.
    - For example, the meaning **SN000010** for the word **afukat** (avocado) contains several media IDs, such as **c0d98d495af6cc7405427c759a3f658e**, which may point to images, audio files, or videos illustrating the sense.
    - Multiple media IDs can be separated by semicolons if more than one media item is related to the sense.
6. **Scientific_Name**:
    - Relevant when the sense refers to a scientific or botanical entity.
    - For instance, in **SN000010** for **afukat**, the scientific name of the avocado is associated with the sense.
    - This column is particularly useful for plants, animals, or other natural entities where a scientific name can clarify the entry.
7. **Semantic_Domain**:
    - Categorizes the sense into a specific semantic domain.
    - Examples include:
        - **culture and communication** for the word **adat** (tradition).
        - **food, cooking, fire; plants** for the word **afukat** (avocado).
    - Semantic domains help group words into broader categories of meaning, such as values, emotions, physical objects, or actions.
8. **Synonym**:
    - Lists synonyms (words with similar meanings).
    - For example, the sense for **SN000001** (to give) has the synonym **kasih**.
    - The sense for **SN000002** (filter) lists the synonym **kata pongisi**.
    - Synonyms help illustrate relationships between words with similar meanings within the language.
9. **Alt_translation1**:
    - May list alternate translations or explanations for the sense.
    - For example:
        - The sense **SN000002** has an alternate translation, **kata seru**.
        - **SN000003** also has a related form **fokus**.
    - These alternate translations can provide additional meanings or nuances of the word.

Examples.csv -

ID	Language_ID	Primary_Text	Analyzed_Word	Gloss	Translated_Text	Meta_Language_ID	Comment	Sense_IDs	Sources	alt_translation1
XV000001	kgv	Ma sandalbon ladanbona ditamanunggi			Dia kasih sandal dengan baju sama teman.			SN000001	stim4 16.1	He gives his friend shoes and sandals-
XV000002	kgv	Kara gonggungnin.	ka\t=at\t=a\tgonggung\t=nin\t.	2SG\t=OBJ\t=FOC\tcall; call.out\t=NEG\t	[They] didn't call you.			SN000004		[Dong] tidak panggil kau.
XV000003	kgv	Mu owatko afokarat parua.			They over there pluck avocados.			SN000010	Fajaria Yarkuran	Mereka di sana petik alpukat.
XV000004	kgv	Hari ahat Unyil esun mu he yecie.			Sunday Unyil's father and family will return.			SN000011	Fajaria Yarkuran	Hari minggu Unyil pu bapak dong su pulang.
XV000005	kgv	Anahutak owatko melalu reba.	an\t-ahutak\towatko\tmelalu\treba\t.	1SG\talone\tF.DIST.LOC\tsit\tPROG\t	I alone sit over there.			SN000012	Fajaria Yarkuran	Saya sendiri duduk di sana.
XV000006	kgv	Ak, ak, ak, suagi wilao laur.			Go seawards, seawards, seawards, there's tuna playing in the sea.			SN000015	Fajaria Yarkuran	Pi di laut, ada komo di laut.
XV000007	kgv	Aknaran anggon ning.			My chest hurts.			SN000017	Fajaria Yarkuran	Saya punya dada sakit.
XV000008	kgv	An kiemalunara rep.			I'm getting strings for a basket.			SN000020	Fajaria Yarkuran	Saya ambil tali keranjang.
XV000009	kgv	Mu he Almaherangga yecieni bot.			They are returning to Halmahera.			SN000026	Fajaria Yarkuran	Mereka su pulang di Halmahera.
XV000010	kgv	Tumun me emun amunat nani mindi bon min.			The child drinks its mother's milk until it sleeps.			SN000027	Fajaria Yarkuran	Anak itu minum susu ibunya sampai tidur.
XV000011	kgv	Ma emun ambelun neing.			Her mother's nipple is sore.			SN000028	Fajaria Yarkuran	Dia pu mama pu ujung susu sakit.
XV000012	kgv	Mu amdirat Kamburkoa paruo.			They are making a garden at Kambur.			SN000030	Fajaria Yarkuran	Mereka bikin kebun di Kambur.
XV000013	kgv	Mustafa esun Kamberarkoa amdirat komaruk.			Mustafa's father is burning a garden patch at Kamberar.			SN000031	Fajaria Yarkuran	Mustafa pu bapak bakar kebun di Kamberar.

Kalamang Dataset: `examples.csv` File Breakdown

The `examples.csv` file contains example sentences or phrases that provide context for the lexical entries and senses from `entries.csv` and `senses.csv`. These examples are essential for understanding how words are used in real contexts. The following is a detailed description of the columns in the file:

1. **ID**:
    - Contains unique identifiers for each example (e.g., XV000001, XV000002).
    - The "XV" prefix indicates that these are examples. Each example is assigned a unique ID, which can be referenced elsewhere, such as in the senses table.
2. **Language_ID**:
    - The ISO 639-3 code for the language, kgv for Kalamang, is contained.
    - This associates the example with the correct language.
3. **Primary_Text**:
    - This contains a Kalamang example sentence or phrase.
    - These are real-world usages of the word in context. Examples include:
        - **Ina kasih sandal ladanan bataannungai**: A sentence involving the verb "to give."
        - **Kara pongisungin**: A phrase that indicates interaction between speakers.
        - **Hari adat Unyil esun minhe tu vesie**: A sentence involving cultural references.
    - These examples are crucial for demonstrating how words are used in everyday speech.
4. **Analyzed_Word**:
    - Provides a morphological analysis of the words in the example sentence.

    - For instance, in **XV000002**, the analyzed word is **ka=t** followed by grammatical markers like **OBJ:1FOC=call** and **NEG** to show the structure of the phrase.
    - These breakdowns explain the individual components of the phrase in linguistic terms, showing how the word is structured grammatically.
5. **Gloss**:
    - Provides a word-by-word or morpheme-by-morpheme gloss of the example.
    - For example, **2SG1.OBJ=FOC=call.out=NEG** breaks down the example phrase into its grammatical parts.
    - These glosses are useful for linguists or learners who want to understand the language's grammar in greater detail.
6. **Translated_Text**:
    - Provides the translation of the example sentence into another language (likely English or Indonesian).
    - For example, **Ina kasih sandal dengan baju sama temnan** translates to "He gives his friend shoes and sandals."
    - **Sunday Unyil's father and family will return** translates a sentence involving cultural or social events.
    - The translations help readers who do not speak the language understand the meaning of the examples.
7. **Meta_Language_ID**:
    - Might represent a meta-language reference, though it is empty in the visible portion of the file.
    - If populated, it could provide additional metadata about the language or the example's source.
8. **Comment**:
    - Contains comments related to the example.
    - For example, **He gives his friend shoes and sandals** is a comment on the meaning or context of the sentence.
    - Comments help clarify cultural or linguistic nuances that may not be immediately obvious from the text.
9. **Sense_IDs**:
    - Links the example to specific senses in the `senses.csv` file.
    - For instance, **SN000001** links the example to the sense for the verb "to give" (found in the `senses.csv` file).
    - Multiple sense IDs (e.g., **SN000010**, **SN000002**) can be associated with the same example if it illustrates multiple senses.
    - This cross-referencing is essential for tying the example to the specific meaning or usage of a word.
10. **Sources**:
    - contains references to the example's source.
    - For example, many of the examples are sourced from **Fajriani Yakiran**, likely a field researcher or informant.
    - Sources indicate where the example was collected or who provided the example, which is important for linguistic research and verification.

11. **Alt_translation1**:
    - Provides alternative translations or comments on the example.
    - For instance, **He gives his friend shoes and sandals** might have a slightly different or additional translation, providing multiple ways to interpret the example.
    - These alternate translations help give additional context or nuance to the meaning.

Media.csv -

ID	Name	Description	Media_Type	Download_URL	Language_ID	size
f92bc00b1ff3429b55926a6e1634e303	'arar_natperahu.JPG		image/jpeg	https://cdstar.eva.mpg.de/bitstreams/EAEA0-68A5-27E3-D959-0/_arar_natperahu.JPG	kgv	3490274
c96129f895505e8e593b1235b0411b7c	'gorip_tefar laut.jpg		image/jpeg	https://cdstar.eva.mpg.de/bitstreams/EAEA0-6166-6337-62D2-0/_gorip_tefar_laut.jpg	kgv	206404
7900b2fb2b67974bd6fc81d9dc855f15	'keibar_penahan seman_two long sticks parallel to boat.jpg		image/jpeg	https://cdstar.eva.mpg.de/bitstreams/EAEA0-8323-6644-787E-0/_keibar_penahan_seman_two_long_sticks_parallel_to_boat.jpg	kgv	1854457
74f3096370893716 24d8d3b07ef07d91	'musing_mengiwang pasir.jpg		image/jpeg	https://cdstar.eva.mpg.de/bitstreams/EAEA0-3640-1D5A-5379-0/_musing_mengiwang_pasir.jpg	kgv	1978382
9c2399433e585eb4adb8a04c127fe3ca	'mutam_rarout kutu.jpg		image/jpeg	https://cdstar.eva.mpg.de/bitstreams/EAEA0-8215-0806-8EC3-0/_mutam_rarout_kutu.jpg	kgv	760257
8fbafa3b4a29934dbfb4939421881a6b	'siram_ikan layar.jpg		image/jpeg	https://cdstar.eva.mpg.de/bitstreams/EAEA0-B41A-8C5B-57E3-0/_siram_ikan_layar.jpg	kgv	726300
8c1e5aa9c4d8aef80821aa9aadb88ff2	'torak_momar.jpg		image/jpeg	https://cdstar.eva.mpg.de/bitstreams/EAEA0-D5EA-8B30-E2D1-0/_torak_momar.jpg	kgv	242858
0fc05587920f25f027e2790be8f736c5	'weswes_sj siput.jpg		image/jpeg	https://cdstar.eva.mpg.de/bitstreams/EAEA0-3EFE-8C88-7231-0/_weswes_sj_siput.jpg	kgv	610452
ac9fcd4270dddab374b7eb255628b61f	IMG_1444.JPG		image/jpeg	https://cdstar.eva.mpg.de/bitstreams/EAEA0-867E-D86E-A8CB-0/IMG_1444.JPG	kgv	9049580
be6890c1b930ff557f383f7f19f9dc0e	IMG_2157.JPG		image/jpeg	https://cdstar.eva.mpg.de/bitstreams/EAEA0-4A6D-5FCB-119F-0/IMG_2157.JPG	kgv	3180699
1f84305dd3915fbe3c7ffdeb75898b6f	IMG_2734.JPG	Cape Lapangan – Lapangan Karimun	image/jpeg	https://cdstar.eva.mpg.de/bitstreams/EAEA0-08AA-53C6-63BF-0/IMG_2734.JPG	kgv	2708202
7b29329045454273ca9d1e097566dfaf	_bungakupukupu_bastardvervain (2).JPG		image/jpeg	https://cdstar.eva.mpg.de/bitstreams/EAEA0-36CC-E897-F7C5-0/_bungakupukupu_bastardvervain__2_.JPG	kgv	6979511
49dc39d38efb1424bd8cff60bdae7c60	afokat.JPG	a sprouted avocado pit – afukat narun kosten	image/jpeg	https://cdstar.eva.mpg.de/bitstreams/EAEA0-8523-EED4-BB8C-0/afokat.JPG	kgv	2484155
c009d8495a5cc7d05422c7593e65895e	afukat.jpg	avocado fruit – afukat naun	image/jpeg	https://cdstar.eva.mpg.de/bitstreams/EAEA0-0964-428C-D523-0/afukat.jpg	kgv	661498
8e92a421a0b78869f5bbc5167f1b5096	alar.jpg		image/jpeg	https://cdstar.eva.mpg.de/bitstreams/EAEA0-2AB9-B803-693A-0/alar.jpg	kgv	267908
e31d9bf4d827ece0cf2fae2e26e7a003	alfokat arun.jpg	avocado tree – afukat arun	image/jpeg	https://cdstar.eva.mpg.de/bitstreams/EAEA0-8DC5-05BE-4439-0/alfokat_arun.jpg	kgv	1050149
21482a224712f697a03fcf5a2e0f76ac	alun_tali.jpg		image/jpeg	https://cdstar.eva.mpg.de/bitstreams/EAEA0-E554-3248-5D1C-0/alun_tali.jpg	kgv	1126147

Kalamang Dataset: `media.csv` File Breakdown

The `media.csv` file contains metadata about media files associated with various entries and senses from the `entries.csv` and `senses.csv` files. Each row represents a different media file (image, audio, video, etc.) that illustrates or provides additional context for a word or example in the linguistic dataset. The following is a detailed description of the columns in the file:

1. **ID**:
    - Contains a unique identifier for each media item (e.g., **f82b00fb114943e45b25f91e35d76512**, **c961289195505e9e5b5931235014017b**).
    - These IDs are used to reference and associate the media files with specific entries or senses. Each media file has a unique identifier, ensuring that it can be linked unambiguously to other data.
2. **Na**:
    - Contains the name of the media file (e.g., **suara_natpatrahu.jpg**, **borp_tefar_latui.jpg**, **kelar_penahan_seman_wow_long_sticks_parallel_to_boat.jpg**).

    - The name gives a hint about the media file's content and typically corresponds to the file name or a descriptive label.
3. **Description**:
    - Provides a short description of the media file's content.
    - For example, **suara_natpatrahu.jpg** might refer to an image related to the sound or depiction of a specific cultural item, while **borp_tefar_latui.jpg** may refer to a related image of a cultural or physical object.
    - Descriptions like **Cape Lapanga—pangan Karimun** offer additional context, such as geographic locations or specific items being depicted. These descriptions help users understand what the media file represents.
4. **Media_Type**:
    - Specifies the type of media file (e.g., **image/jpeg** indicates the media file is a JPEG image).
    - If other media types like audio files, videos, or PDFs were included, they would also appear in this column. This information helps users or applications handle the media files correctly (e.g., loading an image, playing audio, etc.).
5. **Download_URL**:
    - Contains a URL where the media file can be downloaded or accessed.
    - Each file is hosted online and can be accessed via its specific URL. These URLs are essential for users who want to view or download the media associated with the linguistic entries or senses.
6. **Language_ID**:
    - The ISO 639-3 code for the language, kgv for Kalamang, is contained.
    - This associates the media files with the Kalamang language, ensuring that they are linked to the correct language in a multilingual dataset.
7. **Size**:
    - Indicates the media file's size in bytes.
    - For example, **344027** bytes (approximately 344 KB) for **suara_natpatrahu.jpg** and **849464** bytes (approximately 849 KB) for **borp_tefar_latui.jpg**.
    - The file's size can give users an idea of the bandwidth or storage requirements when downloading or handling media files.

6. Conclusion

The **Kalamang CLDF Dataset Project** has successfully transformed raw linguistic data into a structured, CLDF-compliant format, ensuring comprehensive documentation of the Kalamang language. From the initial setup—where we installed key dependencies like pydictionaria and cldfbench—to the data processing pipeline, each step was designed to clean, normalize, and map the raw.sfmraw.sfmn to meaningful linguistic data.

The project efficiently organized lexical entries, their meanings (senses), and examples into well-structured CSV files. Through preprocessing, we handled variant forms, etymology, semantic domains, and media associations, ensuring each entry was enriched with relevant linguistic metadata. Additionally, logging mechanisms provided detailed insight into the data transformation process, allowing us to track the quality and accuracy of the dataset.

The baseline result is a fully validated, structured dataset ready for linguistic analysis, following CLDF standards. This dataset lays a strong foundation for further research and exploration of the Kalamang language, providing insights into its vocabulary, meanings, and usage in a clear and consistent format.

7. References

1. A Benchmark for Learning to Translate a New Language from One Grammar Book. https://arxiv.org/pdf/2309.16575
2. An Incomplete Loop: Instruction Inference, Instruction Following, and In-Context Learning in Language Models. https://arxiv.org/pdf/2404.03028
3. Low-Resource NMT with Smaller Vocabulary Sizes. https://link.springer.com/chapter/10.1007/978-3-031-70563-2_15
4. Multimodal Machine Translation for Manipuri. https://link.springer.com/article/10.1007/s11042-023-15721-2
5. Chinese-Centric Neural Machine Translation for Low-Resource Languages. https://www.sciencedirect.com/science/article/abs/pii/S0885230823000852
6. KARD: Knowledge-Augmented Reasoning Distillation. https://github.com/Nardien/KARD. https://arxiv.org/pdf/2305.18395
7. Teaching Large Language Models an Unseen Language on the Fly https://arxiv.org/pdf/2402.19167