Saint Louis University
School of Accountancy, Management, Computing and Information Studies
Computer Science and Computer Applications Department

CSE30 Prelim Project: Building a Tuwali Lexicon as a Foundation for an Interactive Language Learning Tool for Elementary Students

Submitted by:

BAUTISTA, MARY ANTOINNETTE	2234109@slu.edu.ph
BUSTRIA, JAZPER LOUIE YSMAEL	2235111@slu.edu.ph
DEL ROSARIO, EDWARD NATHAN	2220081@slu.edu.ph
LAGUERTA, HAROLD	2235904@slu.edu.ph
PALACAY, ABIGAIL	2227408@slu.edu.ph
RIBOROSO, RHOTRE MATTHIEU	2235192@slu.edu.ph
URBIZTONDO, KARL JASPER	2233067@slu.edu.ph

Submitted to:
Sir Dalos D. Miguel

September 2025

INTRODUCTION

The Cordillera Administrative Region (CAR) in the Philippines is home to many indigenous languages that reflect its people's culture, traditions, and identity. One of these is Tuwali Ifugao, often called Tuwali, which is spoken mainly in the province of Ifugao, particularly in the southern part of the province, with Kiangan serving as its primary center of use (Taleon, 2020; Komisyon sa Wikang Filipino, 2018).

Although it might initially appear that Tuwali lacks structured linguistic resources, the language has in fact been the subject of substantial scholarly documentation. Existing works include a Tuwali ifugao dictionary and grammar sketch, detailed analysis of its morphology (Ballard, 2007) and phonological sketch (Roe, 2020). Its typological characteristics have been described using the concept of cross-referencing, a term to describe the language's morphological and syntactic patterns (Ballard, 2007).

At present, some online materials contain information about Tuwali. For example, the Tuwali Ifugao Dictionary on Webonary provides vernacular to English word entries, while Omniglot documents the number system of the language. Although helpful, these sources are limited and do not yet provide the resources needed for learning and wider language processing. This shows the need to build more organized and accessible resources to preserve and promote the language in modern contexts.
Digital tools such as electronic lexicons play an important role in this effort. A lexicon organizes words, meanings, and other details that can be used for both human learning and computational processing. In recent years, lexicons have become increasingly important for building Natural Language Processing (NLP) applications such as text analysis, translation, and educational software (Huang, Liu, & Zou, 2020). When designed for young learners, lexicons can be adapted into interactive, multimedia platforms that make language acquisition more engaging and help strengthen interest in indigenous languages (Living Dictionaries; Digitizing the Higaonon Language, 2025).

This paper aims to develop a Tuwali electronic lexicon as the foundation for an interactive language learning tool for elementary students. The project begins with two main data sources: the Tuwali Ifugao Dictionary from Webonary and the Numbers in Tuwali Ifugao from Omniglot. These serve as the initial foundation of the lexicon, which may later be improved or expanded. By creating this lexicon, the study contributes to both the preservation of Tuwali and the creation of resources that can be used for education and possible future NLP applications.

This project aims to create an initial digital lexicon for Tuwali Ifugao by combining traditional and modern language resources with systematic linguistic processing. The objective is not only to document vocabulary but also to organize it in a way that can serve as the foundation for an interactive language learning tool designed for elementary students.

To achieve this, the project will:

1. **Gather Resources**
   Collect language materials from reliable sources such as dictionaries, and other written texts. These will serve as the primary dataset for building the lexicon.

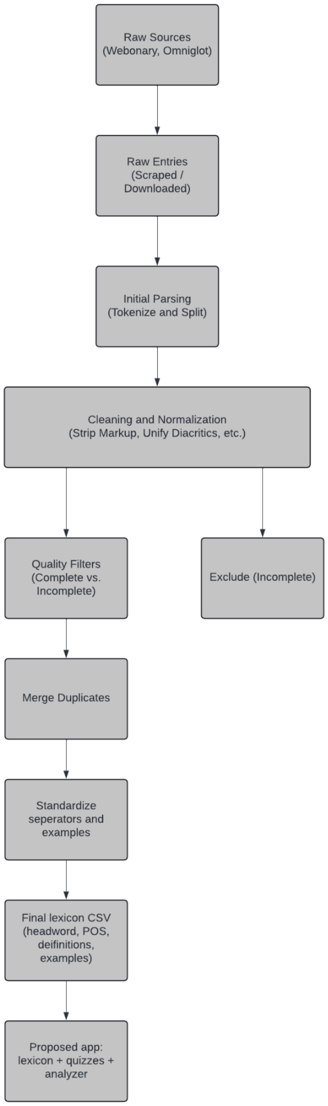
2. **Process and Organize Data**
   Clean and standardize the gathered resources by addressing spelling variations, formatting inconsistencies, and duplication to prepare the data for systematic analysis.

3. **Construct a Working Lexicon**
   Develop a structured lexicon that records each word along with its definitions, part of speech (POS) tag, and example sentences to show its usage in context.

4. **Support Future Application Development**
   Lay the groundwork for an interactive language learning tool tailored to elementary students by ensuring the lexicon is structured in a way that can later be integrated into digital applications

METHODOLOGY

The data gathering focused on extracting lexical entries from the Tuwali Dictionary from Webonary. Entries were collected for each letter of the alphabet to ensure comprehensible coverage. Number translations were also gathered in Omniglot, an online encyclopedia and writing system - which includes languages like Tuwali. This includes the cardinal and ordinal numbers which can be found in the standard English language. For each entry the following fields were extracted where present: Tuwali *headword*, *part of speech*, *English definitions*, *example sentences*, and *Tagalog equivalents/tag_definition*. These records were compiled into a working dataset for preprocessing analysis.



Raw Sources (Webonary, Omniglot)
Raw Entries (Scraped / Downloaded)
Initial Parsing (Tokenize and Split)
Cleaning and Normalization (Strip Markup, Unify Diacritics, etc.)
Quality Filters (Complete vs. Incomplete)
Exclude (Incomplete)
Merge Duplicates
Standardize seperators and examples
Final lexicon CSV (headword, POS, deifinitions, examples)
Proposed app: lexicon + quizzes + analyzer

Figure 1: Data collection and preprocessing pipeline

1. **Data Preprocessing**

   Following the data gathering phase, the raw lexical data underwent a preprocessing procedure to ensure its quality, consistency, and usability. The preprocessing objectives were to: (1) standardize formatting for deterministic parsing, (2) remove or flag incomplete and non-Tuwali items, and (3) reduce redundancy so the lexicon is reliable for downstream use (search, quizzes, morphological checks).

2. **Data Processing and Cleaning Procedure**

   The dataset from the Tuwali lexicon CSV was loaded into Python using *pandas* for inspection and cleaning. The following structured cleaning steps were applied.

   - Normalize definitions, separators and capitalization.
     Replace mixed punctuation with a single separator(;), collapse repeated separators, trim whitespaces, and capitalize the first letter of each definition.
   - Clean *examples*.
     Split example sentences on the "|" delimiter, trim whitespace, remove duplicated sentences within the cell, and apply light spacing fixes. Rejoin examples using the "|" as the canonical separator to preserve multiple illustrative sentences per headword while ensuring each example is readable.
   - Normalize *tag_definition*
     Apply the same separator normalization and deduplication used for English definitions, trim whitespace, and standardize capitalization. Tagalog equivalents can be used for bilingual presentation.
   - Convert blank strings to missing values and remove low-information rows.
     Replace empty/whitespace strings with proper missing values and drop rows with an excessively large number of missing fields (entries missing headword, definitions, and examples).
   - Remove duplicates
     Drop exact duplicate rows. For near-duplicates, merge fields so that multiple definitions and examples are combined into the record.

Example Raw

word	part_of_speech	definitions	examples	tag_definition
balang	comm	an insect.; insect, small; tiny insect	Aaa!Kanana ot mun-oga.| Aaa! Kanana ot muna-oga.	mga panalo; mga panalo

Example Cleaned Entry

word	part_of_speech	definitions	examples	tag_definition
balang	comm	An insect; Insect, small; Tiny insect	Aaa! Kanana ot mun-oga.	Mga panalo

RESULTS AND DISCUSSION

After the data cleaning process, the researchers are able to create a Tuwali language lexicon. The final output (Table 1) has a number of 5343 rows of words and 5 columns consisting of the word, part of speech, definitions, examples and tag definition.

The entry is organized into the following items:

- Word - the Tuwali term that is to be translated to its English/Filipino definition
- Part of speech - a category classification for words based on the grammar (e.g. adj (adjective), comm (common noun), trans (transitive verb), intrans (intransitive verb), sta (stative verb), proc (process verb), adjunct, det (determiner), prop (proper noun), interj (interjection), neg (negation), adv (adverb))
- Definitions - the definition of the Tuwali word into English
- Examples - shows the application of the Tuwali word into a simple Tuwali sentence
- Tag definition - the definition of the Tuwali term into its tagalog definition

word	part_of_speech	definitions	examples	tag_definition
aaa	adjunct	Expression of fear; Ahh!	Aaa! Kanana ot mun-oga. →	Pagpapahayag ng takot; Ahh!

			Ahh, he said, then fell.	
…	…	…	…	…

Table 1. Final Lexicon sample data

The Tuwali Lexicon contributes a great deal to language conservation, education, and linguistics. Its bilingual definition style (Tuwali–>English–Filipino) can be easily accessed by native speakers and other individuals beyond Ifugao province. The lexicon connects local knowledge with broader audiences by providing equivalent meanings in national and global languages.

Linguistically, the lexicon brings out distinctive aspects of Tuwali Ifugao like reduplication, verb richness in morphology, and interplay between indigenous and borrowed vocabulary. Placing contextual examples in the lexicon enhances pedagogical strength so learners can see definitions and witness words used authentically. Practical use in this form benefits language revitalization projects, as contextual learning is crucial to proficiency.

In building the lexicon, the researchers encountered some challenges. We are not able to incorporate pronunciation, synonyms/antonyms, and the words' actual translation in Filipino or English due to the scarcity of web based dictionaries for the Tuwali language. This lexicon captures the majority of commonly used words in Tuwali but in comparison to the dictionaries of Filipino, Cebuano, Ilocano, this lexicon falls behind since most translated words in these languages are approximately 25,000 to more than 100,000 words translated from the local language to the English language. Spelling standardization was challenging since Tuwali words can potentially have alternative orthographies within communities. There were also some words that were difficult to translate into Filipino or English without sacrificing cultural specificity, requiring careful paraphrasing. Bearing these challenges in mind, the lexicon demonstrates that it is possible to document an indigenous language both linguistically correct and culturally sensitive.

In general, this lexicon emphasizes the importance of indigenous vocabularies. It preserves traditional knowledge and provides the foundation for incorporating local languages into modern education and digital media. Therefore, the Tuwali lexicon is a linguistic record and a cultural bridge that reinforces identity while opening pathways for intercultural understanding.

PROPOSED NLP TOOL

The proposed tool for this project is a web-based application specifically designed for elementary students to learn the Tuwali language. In the Philippine educational curriculum, particularly at the elementary level, students are introduced to subjects like Mother Tongue, which emphasize the importance of learning and appreciating regional dialects and indigenous languages. This tool directly supports that initiative by providing a platform where students can explore, learn, and deepen their understanding of Tuwali through a digital lexicon and interactive learning activities.

The goal of this tool is to help students develop vocabulary and language awareness by providing access to a structured online lexicon or dictionary, supplemented with quizzes and possibly speech features. It is designed not only to preserve the language but also to enhance and polish the Tuwali proficiency of young learners in a more creative and engaging manner.

Furthermore, the development of this tool contributes to the larger body of knowledge on indigenous language integration in modern education. By modeling how traditional linguistic resources (e.g., dictionaries, Bible translations, community documentation) can be transformed into interactive digital learning platforms, this study proposes an innovative approach for promoting Philippine local languages keeping both history and cultural knowledge intact while embracing technological advancements.

The application would include three key features namely:

1. Online Lexicon

   The main component of the application is a lexicon that includes:

   a. Tuwali Words
   b. Meaning of the words
   c. Part of speech
   d. Sample use of the word in a sentence

This lexicon serves as the foundation for all language learning features in the web application. It is designed to be easy to navigate for elementary students, with a clean user interface, searchable word entries, and categorization by topic

2. Morphological Analyzer

   To enhance the linguistic depth of the platform, a morphological analyzer is proposed as a future extension. This component will help break down words into roots and affixes, aiding in the understanding of word formation and grammar. Moreover, it helps learners grasp basic morphological rules of Tuwali, and prepares the lexicon for future Natural Language Processing (NLP) applications, such as automated translation or language modeling.

3. Interactive Quiz Feature

   To promote engagement and retention, the tool will include interactive quizzes that allow students to test their knowledge through:

   a. Multiple-choice questions
   b. Matching words with meanings
   c. Sentence completion exercises

   These quizzes will be aligned with the Mother Tongue curriculum to ensure relevance to classroom learning. They will also incorporate gamified elements (e.g., points, levels, badges) to motivate students and make language learning more enjoyable. Furthermore this will ultimately help polish the students' proficiency in learning and talking the language.

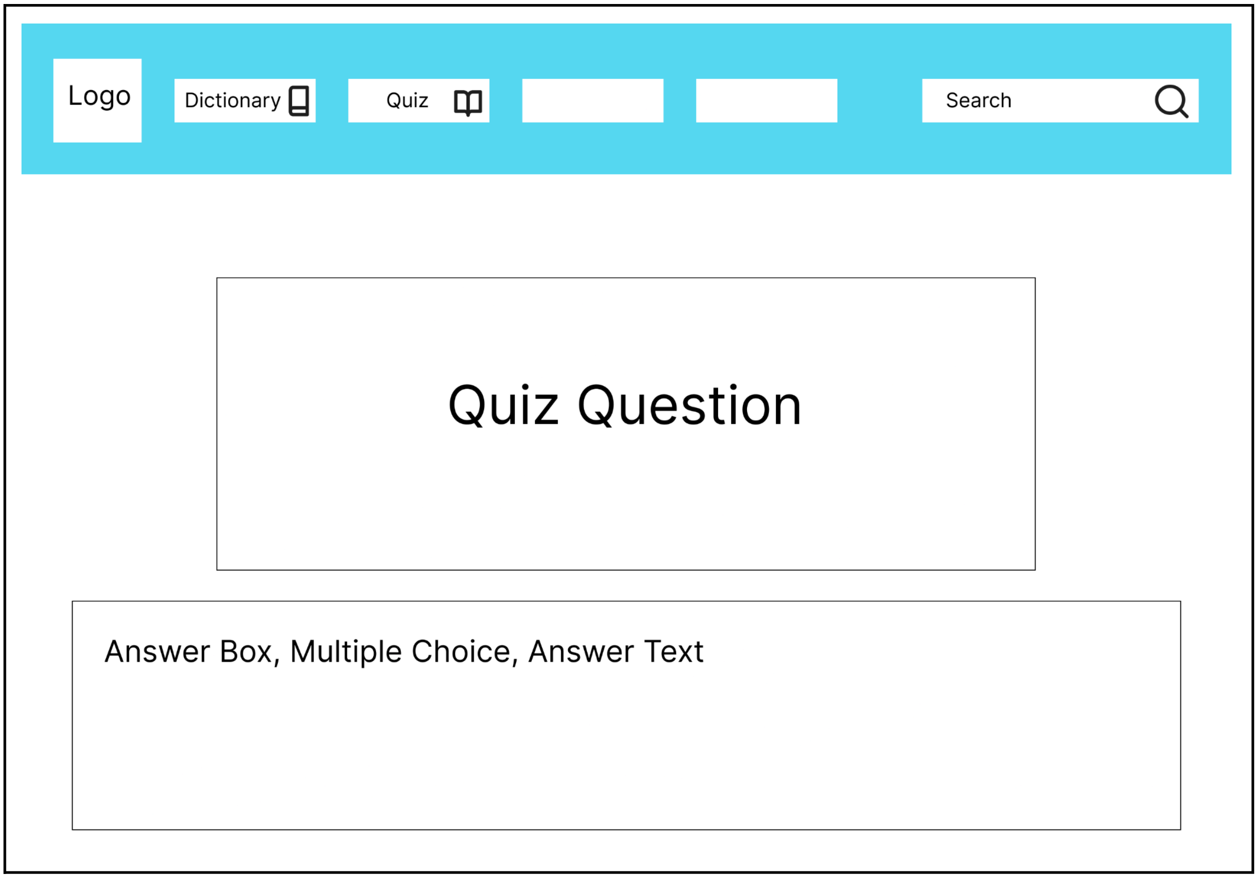


Figure 2: Interactive Quiz Feature

This prototype is the User Interface for the quiz feature of the application. It would contain the quiz questions in the question box and on the bottom part of the question would be the potential ways to answer the questions may it be either multiple choice, identification or fill in the blanks.

Beyond education, the tool has potential use cases in e-governance and community institutions, particularly in Ifugao and other Tuwali-speaking areas. The structured lexicon and linguistic tools can assist in:

a. Creating a blueprint for applying the research to different context and languages here in the Philippines
b. Promoting civic engagement through local-language materials
c. Supporting language revitalization programs by LGUs and NGOs

This tool is designed with scalability in mind. While its primary users are elementary school students, the architecture allows for future expansion, such as:

a. Integrating with DepEd's digital learning platforms
b. Supporting other Philippine indigenous languages
c. Adding more complex NLP tools like syntax analyzers or translation engines

It may also serve as a model for digital language preservation efforts across other indigenous communities in the Philippines, encouraging local governments and academic institutions to invest in similar projects.

In summary, the proposed NLP tool is both a learning aid and a language preservation platform, tailored for the elementary level but extensible for broader societal impact. It empowers students to learn Tuwali in a modern, interactive way while also laying the groundwork for future applications in education, governance, and technology.

CONCLUSION

In conclusion, the creation of the Tuwali lexicon application will preserve the knowledge of the tuwali-ifugao language and support the preservation of the linguistic heritage as well as creating an avenue for research and advancement in developing educational applications and tools for the modern day learning system. By systematically gathering lexical entries from reliable sources and applying different preprocessing techniques and structuring the data in a furnished dataset, the study produced a lexicon of more than five thousand words with English and Filipino translations. This not only strengthens the accessibility of Tuwali for learners but also ensures that the language is documented in a way that is linguistically accurate and culturally sensitive.

Future Researchers can use this study as a means to create other technological developments and application for the preservation of the linguistic heritage of other dialects here in the Philippines as well as a way for children and even adults to learn about a certain language and local dialect with features such as a morphological analyzer and an interactive online quiz to be able to further polish their knowledge base. With features such as an online lexicon, morphological analyzer, and interactive quizzes, the proposed tool highlights the potential of integrating traditional knowledge into digital platforms for education, governance, and community engagement. Ultimately, this project not only contributes to language preservation but also empowers future generations to embrace their heritage while fostering intercultural understanding through technology.

REFERENCES

Daigneault, A. L., & Anderson, G. D. S. (2023). Living Dictionaries: A platform for Indigenous and under-resourced languages. *Dictionaries: Journal of the Dictionary Society of North America, 44*(2), 57–74. https://muse.jhu.edu/article/915065

Huang, Z., Liu, F., & Zou, Y. (2020). Federated learning for spoken language understanding. Proceedings of the 28th International Conference on Computational Linguistics (pp. 3467–3478). International Committee on Computational Linguistics. https://aclanthology.org/2020.coling-main.313

Komisyon sa Wikang Filipino. (2018, August 17). KWF unveils 1st language marker in Cordillera. Loren Legarda Official Website. https://lorenlegarda.com.ph/kwf-unveils-1st-language-marker-in-cordillera/

Morphology / Typology (Philippine Journal of Linguistics, 2007) Ballard, E. (2007). Morphophonemics and the typology of the focus system in Tuwali Ifugao. Philippine Journal of Linguistics, 38(1–2), 1–30. Philippine Social Science Council. https://pssc.org.ph/wp-content/pssc-archives/Philippine%20Journal%20of%20Linguistics/2007/JUNE-DEC%202007,%20VOL%2038%20No%201-2.pdf

Roe, R. (2020). A phonological sketch of Tuwali Ifugao. Philippine Journal of Linguistics, 51(2), 45–68. https://www.researchgate.net/publication/344170392_A_Phonological_Sketch_of_Tuwali_Ifugao

Taleon, K. A. (2020). A phonological sketch of Tuwali Ifugao. University of the Philippines Diliman. https://www.researchgate.net/publication/344170392_A_Phonological_Sketch_of_Tuwali_Ifugao