Transforming Semi-Structured Indigenous Dictionary into Machine-Readable Dictionary

Bali Ranaivo-Malançon[1], Suhaila Saee[1,2], Rosita Mohamed Othman[1] and Jennifer Fiona Wilfred Busu[1]
[1]*Faculty of Computer Sciences and Information Technology,*
[2]*Institute of Social Informatics and Technological Innovations,*
Universiti Malaysia Sarawak.
ssuhaila@unimas.my

***Abstract*—Creating a machine-readable dictionary for an indigenous language is not an easy process and thus, transforming an existing indigenous dictionary into a machine-readable dictionary is one approach to speed up the process. This paper presents the sequential transformation of two bilingual Sarawak indigenous dictionaries, Melanau-Mukah-Malay and Iban-Malay dictionaries, from their initial semi-structured form into their structured representation. The transformation makes use of an OCR to convert the original PDF format of the dictionaries into HTML files, which is then analysed by the Python HTMLParser to extract only the content of the dictionaries. The extracted content is saved in plain text file. To understand the original structure of each dictionary, the textual units in the plain text file are converted into generic symbols. The observation of the collocations of the generic symbols yields to the writing of regular expressions that can delimit each dictionary element. The result is a machine-readable dictionary stored in comma-separated values format. The inspection of each column in the comma-separated values file indicates that the written regular expressions offer a good coverage of the different dictionary elements present in the studied dictionaries. Therefore, the proposed sequential transformation is efficient in accomplishing the conversion of a semi-structured indigenous dictionary into a structured machine-readable dictionary.**

***Index Terms*—Machine-Readable Dictionary; Semi-Structured Data; Indigenous Dictionary; Regular Expressions; Python.**

I. INTRODUCTION

Creating a machine-readable dictionary (MRD) for an indigenous language is a long-term process requiring a large amount of lexical knowledge, human resource, and a sufficient financial support. Thus, transforming an existing indigenous dictionary into an MRD is one approach to speed up the process. This paper presents the different steps needed for the transformation.

Numerous bilingual dictionaries of indigenous languages of Sarawak exist as either in printed or electronic form, which is usually the electronic version of the printed one. Bilingual MRDs are useful for many applications such as word-processing assistance, machine translation, generating parallel corpora [1], etc. However, transforming the electronic version of a printed dictionary into an MRD is not straightforward and can be challenging. The transformation process requires a good strategy to be re-usable for any similar type of dictionaries, with a minimum cost of processing and human intervention. To illustrate the application of the proposed transformation method, two bilingual Sarawak indigenous dictionaries, Melanau-Mukah-Malay and Iban-Malay dictionaries, are used as the case-study. "A bilingual dictionary consists of an alphabetical list of words or expressions in one language (the 'source language') for which, ideally, exact equivalents are given in another language (the 'target language')." [2]. Hence, the source languages are Melanau-Mukah and Iban and the target language is Malay Standard. The input dictionaries are in PDF (Portable Document Format) and their content is semi-structured that makes their transformation difficult. The target dictionaries (MRDs) are structured and stored in CSV (Comma-Separated Values) format. The proposed transformation process consists of converting the PDF file into a plain text file for programming language purpose, and then at identifying automatically the different textual units corresponding to dictionary elements for final storage. Since each indigenous dictionary has its own structure and information, the identification of the dictionary elements is the only step that is dictionary-dependent. The other steps (transformation of PDF into plain text and transformation of annotated plain text into CSV) are totally generic. In general, a dictionary has four types of structures. The megastructure concerns the entire structure of the dictionary. The macrostructure relates to the organisation of the dictionary (or lexical) entries. The microstructure concerns the consistent organisation of lexical information within lexical entries. The mesostructure refers to the set of relations that exist between lexical entries. To be able to identify each of these structures as well as the sub-structures, a rigorous and systematic method needs to be determined. The dictionary elements that need to be identified are the elements of the microstructure. Usually, they correspond to the headword, the pronunciation, the part of speech (POS) tag, and the various senses. However, other information can be available such as the etymology, examples, hyphenation, translations in a target language, etc.

II. RELATED WORK

The creation of a dictionary for indigenous, under-resourced, and endangered languages is part of their documentation. Today, with the spread of computer and smart devices, a printed dictionary is no more attractive for end-users, who are more interested in accessing information in a fast way. Thus, a dictionary needs to be in a digital form and the content needs to be accessible through "intelligent" search such as searching for all lexical entries sharing the same root (e.g. read, reader, reading, etc.) or belonging to a specific POS like verb. An MRD can offer the answers to these queries. However, creating an MRD from scratch is a