Nation Building Through Lends a Hand of ICT Innovation: Preliminary Approach to the Multilingual Dictionary for the Prosperity of the Shan State

[1] Yuzana, [1] Darli Myint Aung, [1]Hsu Mon Kyi, [1]Swe Swe Aung, [1]Amy Aung, [1]Chan Myae Aye, [1]Yin Nyein Aye
[1] University of Computer Studies, Taunggyi, Myanmar
yuzana.yzn@gmail.com

Abstract—**Educational prospects for the regions which are the outside of the main capital in Myanmar are reluctant to reach to the goal. The educational, socioeconomic developments of the local people in the Shan state has been encountered as a lot of challenges in everyday life and the vast areas are beyond the opportunities. According to 2014 official census statistics, the percentage of illiterate in the urban area is only 15% of the population of the Shan state whereas 42.1% are illiterate in the rural area. It is only about 19.75% of primary school students in Shan State reach high school. And also 3.43% of the total population of Shan state who completed the highest level of education such as College and University. This is our main contribution of the research for the ethnic students weak in Myanmar language and English language consequently they have rare chances to connect to the science and technology. As lends a hand of ICT, this paper describes the preliminary approach for the creation of the machine readable dictionary, can provide the meaning of Myanmar, English and Pa-O language words vice versa. In this paper we proposed dictionary building design process model and the framework for the development of the multilingual dictionary that used the traditional building method, such as theory driven process. As Pa-O language is such a kind of an under-resourced language, the data resources are rarely hard to find. Hitherto, there is no officially printed dictionary for Pa-O language to Myanmar language and also English language as well. As a preliminary approach we performed the data collection, the data identification stairs of the proposed design process model in this paper. We collected 26174 Pa-O words with related meaning of Myanmar words and analyzed these words, we discovered the four inference observations presented in this paper. We noticed that Pa-O language have 33 consonants same as Myanmar language whereas the phonetics, some vowels, medial are dissimilar. The Pa-O language words started with the consonant of [A, အ] are observed as the most widely used word about 4782 words and the second most are started with the consonant of [Ta, တ] about 2612 words. Our purpose is being the University of Computer Studies (Taunggyi) in the Shan state, we craft with the support of Information technology innovation to this region. It enforces not only widen the ICT sectors but also for arising homogeneous developing and the enormous prosperity of the economic, education and social sectors in all areas of mountain and land in Myanmar.**

Keywords: natural language processing, machine readable dictionary, education. ICT

I. INTRODUCTION

Most of the Shan State is a hilly plateau, which together with the higher mountains in the north and south forms the Shan Hills. Shan State borders China to the north, Laos to the east, and Thailand to the south. Shan State covers 155,800 km², almost a quarter of the total area of Myanmar. The state gets its name from the Shan people, one of several ethnic groups that inhabit the area. Shan State is largely rural, with only three cities of significant size: Lashio, Kengtung, and the capital, Taunggyi. Taunggyi is 150.7 km north east of the nation's capital Naypyitaw [1].

The distribution of the Pa-O people around Myanmar is in the Southwestern Shan State in and surrounding townships of Taunggyi and Pa'O Self-Administered Zone, in the Kayah State in Loikaw, such as the Hopong township, the Hsi Hseng township and the Pinlaung township [3]. And in the Kayin State, areas near Shan State border, in the Mon State near Thaton and the Bago Division near Taungoo. The native speaker around Myanmar is about 1.8 million according to the 2014 census [2].

The valleys and tableland are inhabited by the Pa-O people. The large majority inhabits the south-western part of Shan State in Myanmar. The geographical centre of the Pa-O could be considered the mountains around the towns of Taunggyi and Kalaw. The smaller number of the Pa-O people live in Thaton,Mon State, Karen State and as far south as the Tanintharyi (formerly Tenasserim) Division on the Gulf of Martahan in the southern tip of Myanmar. They used the same Pa-O language from the Shan State. Additionally, the 900 Pa-O people live inside Thailand, inhabiting four villages within the Muang District of Mae Hong Son Province. The Pa-O is also called “Taungthu” which means 'hill people'. The British colonialists called them Black Karen because most Pa-O women wear black or dark blue dresses. Many Pa-O do not even know of this connection to the Karen, however, and consider themselves a unique people group. One key area in which the Pa-O differ from other Karen groups is in their religious beliefs. While most of the Karen people are either animists or Christians, the Pa-O people have been a strong Buddhist group for many centuries [7].

Our University is situated between Taunggyi Township and Hopong Township (Pa-O Self Administered Zone) both Township 6 miles far apart from our university. Majority of the students are Shan and Pa-O students, the University is within the Pa-O villages surroundings. As we all know that the agriculture is the staple business in their region, the farm, the mountain and the scenery is gifted from nature and they can attract the tourists to upgrade their economic status and farming business. The students of Shan State can be divided into nine primary ethnic groups: Shan, Pa-O, Intha, Lahu, Lisu, Taungyo, Danu, Ta'ang, Ahka and Jinghpaw (Kachin). The population of the Pa-O people in southern Shan State is about 858000, who speak Pa-O language [8]. Currently, the Pa-O students from our university are about 100 among the 600 students from the first year to the final. The problem statements in this area are the educational, socioeconomic developments depend on of the local people's background education, intellectual skill and knowledge and our motivation is we wish for them to widen their intellectual capacity reach to some international extend. The contribution of our research is to building up the multilingual that has never been before in Myanmar and this can prop up not only for the education and social but also for the economic development of the local people and also to encourage the tourism in the Shan state. This facilitates in easy communication not only for the local people but also for the outsiders.

Nowadays, the tourist arrivals in Myanmar are rising dramatically. Especially in Shan State because Shan State has famous places such as In Le Lake, Pindaya Pagoda. And the famous Pagoda for the Kakku Pagoda, Htam-sam Cave, and the Mae Nae Mountains and valleys which is about 8000 above sea level which is situated at Hopong Township (Pa-O Self Administered Zone). Currently it is the most popular places for trekking and hiking not only for the foreigner but also for the local youths among social media. And also have the famous festivals such as the Taunggyi Fireballoon Festival, the Phaung-Daw-Oo Pagoda Festival have been attracted most of the tourist from the all over the world. And also the Pa-O famous ceremony is the novice ordination ceremony, called poi sang long. They consider ordination as a novice even greater merit than ordination as a monk. During the first day of the this ceremony, the sounds of drums, gongs and cymbals echo between the mountain ridges, when boys with shaved heads are taken from their homes to the temple by using horses or cars[7].

Therefore travel to this area, the communication is the vital role in this situation. The basic requirement in this area is the dictionary and machine translation system from their native language to Myanmar and also English language. This is the very first construction of the e-dictionary that has never been before in both Shan state and Myanmar. This system is the break new ground the natural language processing project for the UCS Taunggyi, Shan state. This paper structure as follows: Section II presents related works. Section III describes the theory background. Section IV discusses the proposed system and Section V provides the results of the experiment.

II. Related Work

The Myanmar researchers proposed the RSS dictionary which is also the same as our ethnic multilingual dictionary, they intend for the ethnic of the Rakhine language to Myanmar and English, the Myanmar to Rakhine and English and the last one is the English to Rakhine and Myanmar. The input words accepted as Zawgyi and Unicode. This can be used from the website and from the phone. We tested this system, it is very useful expect for this point for example if the user enters the query in Myanmar (ကပတုာရုု) called 'baby', the item displayed from the text box (ကပတုာရုု), logically the meaning is the same, literally it is different for contain of (၁) but the result showed no item found in paper, that means system perform well only for the exact match [4]. The ECTACO multilingual online dictionary can translate English into many languages and include part of speech tags. This dictionaries are very useful for user while travelling, studying languages, communicating with foreigners and in many other situations. They have devices such as the talking, non talking dictionaries, PDA electronic dictionaries for more than 200 languages based on advanced Text-to Speech and speech recognition technologies and have the functions for business organizer. The resources are bidirectional vocabularies of up to (1, 000,000) words and more [5]. The authors from [6] presented designing and implementing bi-lingual lexicon in both directions Arabic and English which is to be used in machine translation and language processing. This paper proposed based on WordNet lexical database with a semantic and commonsense knowledge and using the cloud computing for mobile dictionary. The first phase is used to collect and download the data from online English. Therefore, the authors classified the dictionary by creating a list of meaning expressions and classifying these meaning in order of their concepts by defining the relations between the words in each concept. And also in order to achieve scalability and interoperability of mobile users they used SQL Azure. The system dictionary is developed and tested in Android mobile platform. The performance measures of the system are the access and response time and real time test for the display result. The authors from [12] described the dictionary is called MUHIT which is an interactive multilingual dictionary web application it is easily accessible to all users. They discussed a trial to build a multilingual harmonized dictionary that contains more than 40 languages, with special reference to Arabic which represents about 20% of the whole size of the dictionary. It provided users with full linguistic description to each lexical item that is useful to many natural language processing tasks such as multilingual translation and cross-language synonym search. This dictionary was built depending on two language resources the International Corpus of Arabic (ICA) and the English WordNet 3.0 used in [12]. The manual development of lexicon was not available; the natural language processing researchers have turned to the machine readable conventional dictionaries (MRDs) as a highly structured and substantial source of lexical information. In that paper the authors pointed out the disadvantages of the MRDs present information in a manner which relied on the linguistic skills and background knowledge of the user. The advanced learner dictionary organized into database which has undergone very systematic error checking. As a continue successful work was built of lexical knowledge base by using this MRD sources has been based on more codified information, such as headword orthography, part of speech codes, grammatical codes and pronunciation fields proposed in [13].

III. Theory Background

Three methods to dictionary building have been identified such as the manual, the semiautomatic and the automatic. There are the three core activities in the dictionary building process: developing categories, identifying entries and categorizing entries [11]. Our method is manual method also called theory driven process. We did not focus on the activities of the developing categories because the words we arranged in the consonants lexicographic order.

A. Theory Driven Process /Manual Method

Rooted in the traditional content analysis method, manual dictionary building is usually a theory driven process, which is similar to the process of developing a coding schema. Since the core activities are conducted manually, this approach requires the highest domain knowledge and the lowest programming knowledge. In addition, it does not rely on a large corpus, and typically results in dictionaries with small sizes. Because it is a theory-driven process, the manual dictionary building approach usually results in dictionaries with high abstractions and low variations. The dictionaries developed using a manual approach usually have a theory-based and systematic category structure and are less probable to have unexpected categories or entries [11].

B. Data Driven/Automatic Method

Automatic dictionary building is rooted in the field of computational linguistics which focuses on modeling language. In general, it involves extracting key words and/or phrases automatically based on learning algorithms and subsequently evaluating the resulting dictionary through experiments or comparing it with existing dictionaries. Compared to the manual approach, automatic dictionary building requires the lowest domain knowledge, but the highest programming knowledge. It can handle very large corpora and produce 'big' dictionaries. However, because it is a data-driven process, the resulting dictionaries may not correspond to theory and can result in unexpected categories and/or entries [11].

C. Hybrid/ Semi-automatic Method

In the semi-automatic approach used corpus based approach which collection text from related articles, news with their corresponding project. Additionally, do the preprocessing steps (a, an, the, stop word removal) and also perform cut off criterion (words duplicate removed, words cleaning). The final results produce unique words. Three features of the corpus could be considered to decide whether the corpus is "adequate". First, the corpus should be relevant. It should include the contents which are consistent with the theme of the dictionary being built. Second, the corpus should be appropriate. Third, the corpus should be complete. For example, to develop a category structure, researchers could initially propose or adopt some categories based on theory and then modify them based on the result of automatic topic extraction from a corpus. To identify the entries, one can first narrow down the scope of the corpus by setting up a frequency criterion with the help of text analysis software. Although the semi-automatic approach is not as computationally efficient as the automatic approach, it is self-justified by its accessibility: one does not need a programming background to adopt it [11].

Each of the three dictionary building approaches has its advantages and disadvantages, and the choice of the appropriate one should be made based on the objectives of the research project under consideration. Therefore, in this paper our proposed framework is to develop a design process model for theory driven process for the building of the multilingual dictionary.

IV. Proposed System And Model

Design science is a research paradigm that focuses on problem-solving. It aims to create artifacts (i.e., construct, model, method, or instantiation) to solve identified problems and serve human purposes. According to the core activities of design science research are 'build' (construct an artifact for a specific purpose) and 'evaluate' (determine how well the artifact performs) [11]. Thus, we solve the challenges of the problem statement by proposed the design process model for building of the dictionary for the local people.

A. Proposed Dictionary Building Design Process Model

Our model catches idea from the inspiration from the [11] but redesign for our proposed system. Our design process model contain five phases as shown in Fig.1 The first phase is the problem identification and motivation that is as we aforementioned in the abstract, the ethnic people has been far leg behind the education, economic and social because they run into a lot of challenges in daily life that's hesitant to achieve to the prosperity. The second stage is the objective clarification, it is yes, indeed our objective is very clear for we intend to the heterogeneous developing in all sectors for these region of the ethnic people. The third phase we design and develop the system that is explained in Fig.2. And then the developing step, the third phase will finish, we will demonstrate the dictionary and last we will validate the reliability and accuracy of the extraction of words.

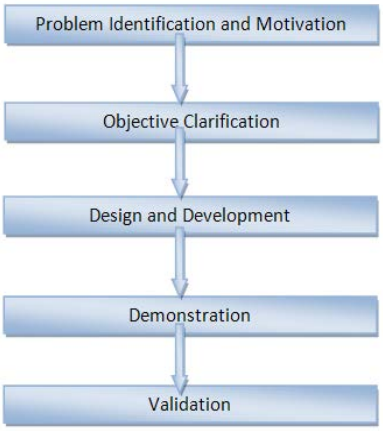

Figure 1. Proposed dictionary building design process model

B. The Third Phase: Design and Development

In order to perform the third phase of the design process model, we proposed the overview of the machine readable multilingual dictionary creation footsteps is illustrated in Fig.2. The first stair of the system is the data collection step. Thus we all known that the Pa-O language is such a kind of an under-resourced language, this footstep accomplished with the help of the language experts and the local people [10].

The second stair is the data identification and the third stair is the data will be inserted into the system with machine accessible code. And the last stair is the development of the dictionary with the functionalities and utilities. The language training is essential role in the construction stair, the developers, who are not local ethnic, will not understand the ethnic languages, even though the local people, teachers and students who are not familiar with their native languages. According to the overview framework we reached at the data collection stair and identification stair which gathered more than 20000 words [10], this paper focused on the finding the observations from this steps.

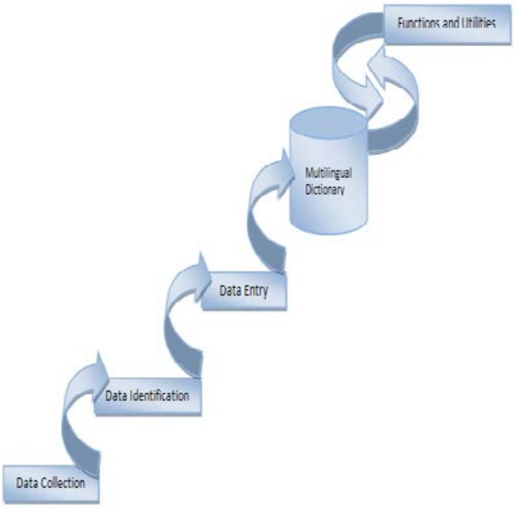

Figure 2. Overview of the proposed system

C. Flow Chart of the Proposed Framework

So we proposed the developing framework for the system that performs well for searching of the words illustrated in Fig. 3. Firstly, the system checks the password for authentication whether it is the admin or the user. If it is admin, he/she can conduct the database operation of the insert, update and delete for the words in the three languages. If it is user, the user enters with input query words. The user can query with three kinds of search.

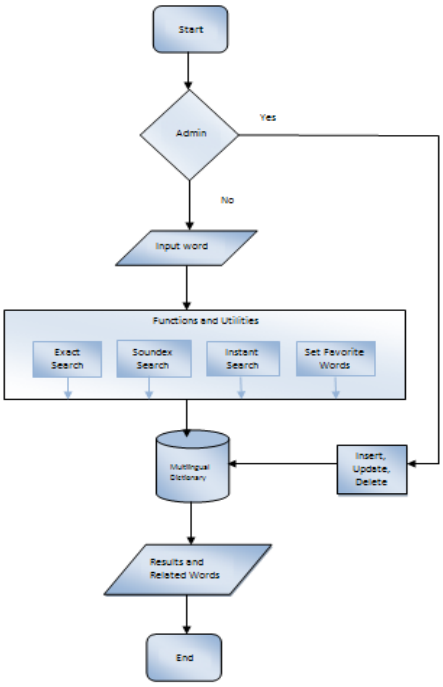

Figure 3. The system flow chart of the proposed framework

D. The System Input and Output

The system input can be accepted words as three languages as the source language, and the output can be fabricated words in two languages as the target language. Additionally, the user can be listened the speech of the output word of the two target languages. The dictionary input and output stage consider as three folds. Firstly, the source language is Pa-O and the target language is Myanmar and English. Second, the source language is Myanmar and target language is Pa-O and English. At last, the source language is English and target language is Pa-O and Myanmar depicted in Fig.4. The output shows the desire result and additionally the result shows with the accompaniment words of the corresponding query words.

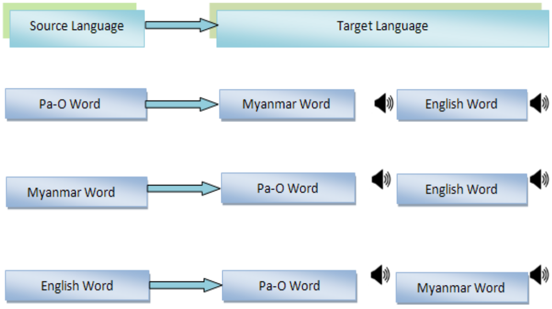

Figure 4. Input and output of the system

E. System Functionalities and Utilities

The system supplies with a mixture of searches and some utilities for the users. As we all know that this dictionary is a kind of multi-lingual support, the main feature of the system which provides ability to search words against three languages. The search processes contain the exact search, soundex (sounds alike) search and the instant search. The system provides exact search for the case of the correct spelling input from the users, but incase for the wrong spelling the system does not return the result. For this case, we consider soundex (sounds alike) search to give the possible approximate results. The instant search when the system starts at the main screen to reduce the users' time consumption. As for the utilities for users the system provides random words and meanings to users for the efficient sharing of the knowledge of the languages. The system provides favorites function which enables user to set favorite words in the system. The system displays related words based on the search words on the details screen.

V. Experimental Result

Many researchers use various performance measure methods for evaluation of their system. For our system, the developing phase is under construction status. If the system will finish, we will use precision and recall for using performance measure. We reach at the middle of the design and development phase, to explicate the data collection and data identification steps of our system, an ‘analysis of concept’ is provided in this section for the evaluation. It has been widely used in research areas, such as engineering, business development, software development, as well as design science research [9].

A. *Proof of Concept from the Data Collection*

For the data collection phase, we collected with the 26174 Pa-O words [10] with Myanmar language meaning. This Fig.5 shows that the number of collected the Pa-O language words which are categorized in the consonants and Independent vowels which used in the system.

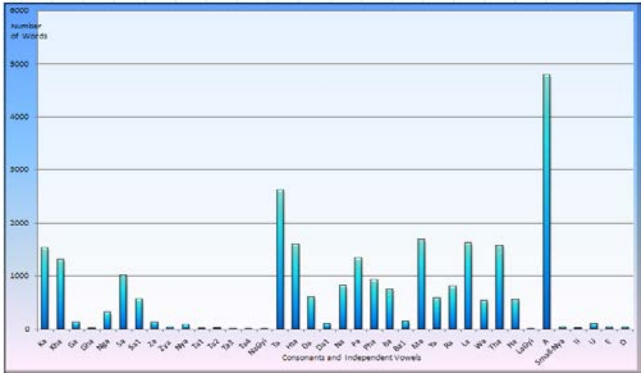

Figure 5. Categorization of Pa-O language collected words

B. *Analysis of Concept from the Data Identification*

Thus we discovered that there are four observed things inference by analyzing the collected data in the identification phase.

Inference 1.

We realized the facts in the Pa-O language is that the most widely used words are started with the consonant of [A, အ] for the 4782 words, the second most use words are the consonant of [Ta, တ] which are about 2612 words. The third, fourth, fifth, sixth of their widely used corresponding order and their respective collected words are as illustrated in Table. I. Perhaps, there may be some words left in the unexpected situation. And the most rarely used words starts with consonants are in the seventh order that are about between 1 to 9 words in our system are also shown in Table. I. That can also be seen as the flat area of the bar chart of the Fig.5.

TABLE I. USAGE ORDER OF WORDS START WITH CONSONANT

The Most Widely Usage Order	Number of words	Consonant in Pa-O language	Description
1	More than 4800	အ	A
2	More than 2600	တ	Ta
3	Between 1000-1700	က, ခ, စ, ထ, ပ, မ, လ, သ	Ka, Kha, Sa, Hta, Pa, Ma, La, Tha
4	Between 500-999	ဆ, ဒ, န, ဗ, ဖ, ယ, ရ, ဟ,ဝ	Sa1, Da, Na, Ba, Pha,Ya, Ra, Ha,Wa
5	Between 100-499	ဂ, င, ဇ, ဘ	Ga, Nga, Za, Ba1
6	Between 10-99	ဈ, ည, ဌ, ဓ	Zya, Nya, Ta2, Da1
7	Between 1-9	ဃ, ဋ, ဍ, ဎ, ဏ,ဠ	Gha, Ta1, Ta3, Ta4, NaGyi, LaGyi

Inference 2

The alphabets of the Pa-O language are same as the Myanmar Language alphabets which are 33 consonants whereas the phonetics is different. Additionally, the vowels, medial are also different. The independent vowels are also the same as Myanmar language which are [II, ဣ], [U, ဥ], [E, ဧ], [O, ဩ].

Inference 3

Some of the Pa-O language words which start with independent vowels which are approximately 164 words collected in our system as shown in Fig. 5.

Inference 4

The consonant [Small Nya, ဉ] are used in Pa-O language. The usage is about 30 words as shown in Fig.5.

VI. CONCLUSION

To our knowledge, it is the first Pa-O-Myanmar-English dictionary. This developing process achieved like an initial phase and still need further modifications but we believe that this innovation conduct the leverage of the prosperity of the Shan State. As far as we can see that the creation of the multilingual dictionary, as a preliminary approach we performed data collection and data identification steps, collection of more than 26174 Pa-O words with relevant Myanmar meaning and conduct the analysis in this paper. The future work will be the finding of the associated English words and construct the machine accessible dictionary. This dictionary can also be used by further research of developing and applying in text mining, text analysis and natural language processing research. We trust that our research assist in the steps forward for the individual skills link up with machines by using their mother language.

REFERENCES

[1] https://en.wikipedia.org/ Shan state.
[2] https://en.wikipedia.org/ Pa-O language.
[3] https://en.wikipedia.org/wiki/PaO_Self-Administered_Zone.
[4] https://www.rrsdic.com/, Rakhine- Myanmar- English online dictionary.
[5] https://www.ectaco.com/, ECTACO, multilingual online dictionary.
[6] Hassanin M. Al-Barhamtoshy, "Designing and implementing bi-lingual mobile dictionary to be used in machine translation, Research-gate publication ID-262396657
[7] Joshua Project, "Pa-O in Myanmar, Burma", A Ministry of Frontier Ventures, 2019 Joshua Project.
[8] Ministry of Immigration and Population ,"2014 Myanmar population and housing census, Shan state report", volume 3-M, May 2015
[9] Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A design science research methodology for information systems research. Journal of Management Information Systems, 24(3), 45-77.
[10] Pa-O literature and cultural development association, Taunggyi, Shan State.
[11] Qi Deng, Michael J. Hine ," Inside the black box of dictionary building for text analytics: a design science approach "Journal of International Technology and Information Management ,Volume 27, Issue 3 Article 7 ,1-1-2019.
[12] Sameh Alansary, "MUHIT: A multilingual harmonized dictionary", Department of Phonetics and Linguistics, Faculty of Arts, Alexandria University, Egypt.
[13] Ted Briscoe, "Lexical issues in natural language processing", University of Cambridge Computer Laboratory, ACQUILEX working paper No 041.