A Cross-Scenario Data Set Applying to Thai and Lao Language Processing

Gornkrit Meemongkol
International College for Sustainability Studies, Srinakharinwirot University, Bangkok, Thailand

Abstract—The purpose of this study was to create a cross-scenario data set for applying to Thai and Lao language processing. Our method involved 113 people, and a questionnaire was given to each person to collect the data. The people were asked to infer the meaning of ten Lao words in the Thai language. The data were analyzed by employing the framework of Bloomfield (1933), Benson (1985), Johnson (1987), Sinclair (1991), Baker (1992) and Miller (1998). They stated that the schema is the concept of a word’s meaning in a person’s mind. This derives from the individual personal experience. The results showed that a cross-scenario data set can be created from these Lao-to-Thai inferences. Each scenario consists of profuse lexical features that are consistent with words in Thai and Lao language. This study will be beneficial for language-processing developers as well as linguists in the future.

Index Terms—cross-scenario, data set, Thai and Lao language, language processing

I. INTRODUCTION

Recently, there has been a growing interest in language-processing development, which aims to develop language-processing machines to enhance communication. One example is using machine translations and search engines. Indeed, language usage is a significant element of language processing’s effectiveness (Arun, 2011, p. 170).

The development of language processing has rapidly grown and taken on a key role in this era of globalization. However, it remains a classic problem in the field of computational linguistics and still needs to advance its potential. Previous studies have confirmed that it is a major problem facing Thai developers (Charoensuk, 2006; Leenoi, 2008; Phosai, 2009; Intasaw, 2013; Thipsena, 2014). For example, it has been limited to processing the language. Moreover, especially with the Thai language, some linguistic mistakes still occur involving language processing when using a search engine, as well as with automatic machine translations. Previous studies have produced some significant findings regarding obstacles to language processing, such as the study by Kawtrakul (2002). She showed that the main problems include unknown words, compound words, complex words, and the great variety of word meanings.

To solve such problems, a deep study of semantics is very useful. Corpus-based semantics is a method of data collection for most language-processing systems. It is language storage for imitative human language. For example, automatic questions-and-answers machines require language data to correctly answer the questions of the language-processing tool’s users (Wutthikorn, 2010). So, a study of the words and languages in corpus should be categorized by meanings and classified in each set by lexical fields. Suktharachan (2017) called concept frame of words, which is a set of words in a related domain or a set of words in the pattern of the WordNet database. All words are linked by the meaning for extending the language boundary. (Miller, 1998) To illustrate, the word “sadness” is in the set of those relating to melancholia. Therefore, “sadness” can extend to the words in the other sets which are related to melancholia. In this study, it is called “a cross-scenario data set”.

A cross-scenario data set is the core data for language-processing development. It can support the processes of natural language-processing systems, but how to create a cross-scenario data set is a complex matter. It should be created from human cognition because thinking is the original source of language-producing. Language-processing systems must reflect an understanding of languages. Lao is similar to some Thai dialects, and they are in the same language family of Tai-Kadai (Sila, 1975; Yensamut; 1981; Rattanaprasert; 1985; Prapin, 1996). This study will use Laos words as a tool of experiment and create a cross-scenario data set because Thai people can understand many Lao morphemes. They can infer unknown whole words by using the meaning of some morphemes as a clue. It is a challenge to pursue the development of a cross-scenario data set by using a non-native language as an effective sample group. The results will be beneficial to both Thai and Lao language-processing development.

II. OBJECTIVE

The purpose of this study is to create a cross-scenario data set applying to Thai and Lao language processing.

III. LITERATURE REVIEW

Phonetics and phonology are studies of sound systems in languages. In Thai, when the sound of words is changed, the meaning of some words will be changed, as well. For instance, the initial cluster /mr/, /ml/ in the southern Thai
dialect can be changed into various patterns (Nooteed & Potibal, 2019). This factor can shift the meaning to other meaning concepts which can’t be guessed from the original meaning.

Morphology (the study of word structures) and syntax (the study of sentence structures) focused on word formations and types of sentences demonstrate that Thai and Lao are distinct languages. There is no grammatical agreement among words in their sentences. For example, there are no gender, number, tense, and voice agreements between the subject and verb. Also, Thai words have various parts of speech and their function can be adapted. In addition, most morphemes have a meaning in themselves (Intratrat, 1996; Channiyom, 2003; Panmeta, 2011). So, most Thai people can infer the meaning of Lao words by using some morphemes as a clue. For example, the word “ວ່າ ວາງ/ี่້າ,” which is “a casino” and “фер” is “to sell.” Thus, they can infer the meaning of “ວ່າ ວາງ/ี่້າ.”

Semantics (the study of words’ meanings) is a significant conceptual framework among several linguistic theories. In other words, the scenarios in this research should be created based on semantic theory. Miller (1998) explained that the mechanism of word meanings is divided into reversal domains and they will be linked like a net or wordnet. Moreover, in semantic theory, homonymy refers to words which have the same alphabetical writing but different meanings. Homophones are words which have the same alphabetical writing but different pronunciations. Polysemy refers to one word with several meanings. Synonymy refers to a group of words which have similar meanings. Antonymy refers to a group of words which have the opposite meaning. Hyponymy is a group of words which have a hierarchy of meaning. Meronymy is a group of words which have the components of meaning; collocation is a group of words which based on the rules of the language (Bloomfield, 1933; Benson, 1985; Sinclair 1991; Baker 1992). Johnson (1987) stated that the image schema is the concept of a word’s meaning in a person’s mind. This derives from the personal experience of each individual. For example, when we hear or think about the utterance “glasses in a case,” we can extend the meaning to: A) glasses will be protected and B) glasses can’t move. The words in the scenarios regarding protection may be related to the meanings in one’s brain, to our cognition. For this study, cognitive semantics is a valuable way to explore and create the cross-scenarios. Because of this, the theories stated above will be focused on and employed in this study as a core theoretical, conceptual framework for data analysis. Inference also plays a key role, as it involves a guess that you make or an opinion that you form based on the information that you have (Cambridge, 2021). It is consistent with Kanchanawasee (1991) who emphasized that causation produces the effects, and inferences can be connected with other things in the world when we apply our background knowledge.

Corpus preparation is a process for creating a corpus. The processes rely on the theory and framework selected by developers, depending on the purposes of each corpus. For instance, the corpus preparation of agricultural domains uses word segmentations and tagging of parts of speech, along with phrases and case grammar as a main step (León, 2015). Likewise, bags of words are sets of words in the same domain for use in natural-language processing. The work of Phetsiri (2010) indicated that tagging is a process of corpus preparation. However, a scenario can consist of many words. For example, a scenario regarding the “marriage ceremony” may consist of the terms: bride, groom, guest, master of ceremonies, bouquet, banquet, music band, etc.

Tagging is a technique for identifying the details of the language in the corpus. For example, “แก้ว” is a transitive verb, so the tagging uses the abbreviation Vt. with the parentheses. To wit: [แก้ว]Vt. (transitive verb). For phrase tagging, the main purpose is to identify the boundaries of the phrase. Case grammar is a theory proposed by Fillmore (1968, 1971) regarding the relation among the words in each sentence. In general, AGENT case refers to the subject or a person who did the action, and OBJECT case refers to the object or a person who receives the action. In the corpus preparation, tagging uses case abbreviations. Specifically, AGT is the abbreviation of “AGENT case,” and OBJ is the abbreviation of “OBJECT case” (Kawtrakul, 2002). As a rule, there are nine cases, as laid out by Fillmore and Baker (2009). But our research will justify AGT and OBJ cases only if they are the main cases of the words in the cross-scenarios. It is the knowledge of the language-processing system and an agreement about how to process the language within a corpus (Gruber, 1993, 1995). Additionally, our aim is to develop the cross-scenarios into a kind of ontology – a term from computer science focused on the study of words’ meanings. Several previous studies have confirmed that the cross-scenarios employed by semantic theory are effective regarding language-processing development (Tantisripreecha & Soonthornphaisaj, 2010; Saengsupawat, 2014; Wijasika & Srivihok, 2014; Khruahong, 2015; Tungkwampian, 2015).

Regarding dialectology as well as comparative Thai and Lao, unfortunately, the research related to this study has been quite limited. Generally, most such studies have focused on Laotian words borrowing by Thai and errors of Thai language used by Lao students (Yuyen, 1997; Nakon 2002). Another kind of study involving meaning identification in Thai and Lao dictionaries of Thai and Lao vocabulary was also conducted (Jeon, 2020). Conversely, we have found no study on meaning inferences by using Lao words. Consequently, the study of Thai words’ meanings by using Lao words is a very challenging task.

IV. METHODS
A. Samples

According to my pilot study, it confirmed that only one word can be distributed to many scenarios. So, ten words are sufficient as a qualified sample group. It was found that, if this research use too many words for the study, it’s quite hard to categorize the scenarios consistently. Because of enormous words, one may not categorize all the words exactly. Thus, ten Lao words were selected by purposive sampling techniques from a Lao-Thai dictionary (the National Social Sciences Council, 2012). The criterion was that native Thai people must know the meaning of some morphemes in the words, but they may not know the meaning of the whole words. So, I chose these ten words: “ປ່ອງເຍີ່ຍມ” /pŋ-yîam/; “ແວ່ນແຍງ” /wîn-yëng/; “ອ້ານແປງໄມງ” /há an-pîng-mîoŋ/; “ຫ້ອງກາຮປ້ອງກິນຕວາມສງບ” /hîng-kæ n-pîng-kan-kîawm-sa-nâb/; “ຊິດຊ່ວຍຮັບ” /sa-bûu-fûn/; “ວິນສິງຊາຮຊຄ້ນ” /wan-sîng-sîaw-khuen/; “ຜຼ້ຈ່າງຫນ້າ” /phû u-tàaŋ-nâ a/; “ຫ້ອງກາຮປ້ອງກິນຕວາມສງບ” /hîng-kæ n-pîng-kan-kîawm-sa-nâb/; “ຂິກຮຕຶດໄລ່” /chà k-khí t-lâ y/. For quality confirmation, all of them were sent to three experts for verification before they were used in this study.

B. Data Collection

Ten Lao words were customized into the online questionnaires, and the sample group was asked to guess the meaning of the Laotian words in Thai. Questionnaires were sent to 120 people as an experimental sampling group via an online-chat application and email. Unfortunately, due to the Covid-19 pandemic, I needed to use online questionnaires, and I couldn’t predict the number of responders. For this reason, I decided to send the questionnaire to 120 people. All of them could complete the questionnaire on a computer and by using a smartphone. Moreover, the selection criteria of the sample group included: (1) The person is a native Thai. (2) He or she has no background knowledge of Lao, and (3) The individuals live in Thailand and don’t have experience working or studying in Laos.

C. Data Analysis

To categorize and analyze the data based on the conceptual frameworks, the cross-scenarios were created. Next, all of the results were reported in the academic language, and phonetic transcription used the IPA system via a free, automatic phonetic-transcription tool at https://thai-notes.com/tools/thai2ipa.html

V. RESULTS

<table>
<thead>
<tr>
<th>Factors</th>
<th>Details (people)</th>
<th>Total (number of people)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. gender</td>
<td>Male = 80, Female = 33</td>
<td>113</td>
</tr>
<tr>
<td>2. range of ages</td>
<td>15-25 = 7; 26-35 = 35; 36-45 = 48; 46-55 = 21; 56-65 = 2</td>
<td>113</td>
</tr>
<tr>
<td>3. background in the Laotian language</td>
<td>No = 113</td>
<td>113</td>
</tr>
<tr>
<td>4. region of hometown in Thailand</td>
<td>Central = 63; Northern = 19; Northeastern = 12; Southern = 9; Eastern = 8; Western = 2</td>
<td>113</td>
</tr>
</tbody>
</table>

As you can see in Table 1, the total was 113 people. That means that seven people did not response to the questionnaire. The group included 80 men and 33 women. Most participants ranged from 36-45 years old. Most of them didn’t have background in the Lao language and most came from the north of Thailand. The factors of the sample groups were not discussed in this study because this research focused on cross-scenario data-set construction to be applied only to language processing.
As you can see in Table 2, a brief explanation of the results and the meanings of the inferred words will be described. The details are as follows:

1. ป่าค่ำ / pà-yàm/ 
   Ten scenarios were created from the word “ป่าค่ำ / pà-yàm/’. In Lao language, “ป่าค่ำ / pà-yàm/ means “a window.” In Thai language, “ที่่า” / pà-yàm/ means “a bulge,” such as in one’s stomach, and “ไป / yàm/ means “to visit.” The examples from the scenario are “pregnancy, fat, and hospital”.

2. ค่ำต่ำ / wên-yêep/ 
   Seven scenarios were created from the word “ค่ำต่ำ / wên-yêep/’. In Lao language, “ค่ำต่ำ / wên-yêep/ means “a mirror.” In Thai language, “แว่น” / wên/ means “the glasses” and “แว่น” / wên/ means “to push into something.”. The examples from the scenario are “mirror, glasses and light.”

3. ห้ามปนปี / háam-pèn-moon/ 
   Nine scenarios were created from the word “ห้ามปนปี / háam-pèn-moon/’. In Lao language, “ห้ามปนปี / háam-pèn-moon/ means “a clock repair shop.” In Thai language, “ห้าม / háam/ means “a shop (dialect).” “ปนปี / pèn/ has no meaning in standard Thai (but the meaning is “expensive” in dialect).” “ปนปี / pèn/ means “an area and changing.” “ปนปี / pèn/ is “a unit of time.” The examples from the scenario are “eight o’clock, clock repair shop, and stage.”

4. ผู้พาต้า / phûu-tàn-nàa/ 
   Nine scenarios were created from the word “ผู้พาต้า / phûu-tàn-nàa/’. In Lao language, “ผู้พาต้า / phûu-tàn-nàa/ means “a representative.” In Thai language, “ผู้ / phûu/ means “a person and male” in colloquial language. “ผู้ / phûu/ /phûu-tàn-nàa/ has no meaning in Thai. The examples from the scenario are “representative, guest and male”.

5. ฮ่องกง - ฮ่องกง / hông-keng-pêng-kang-khwaam-sa-gôb/
Twelve scenarios were created from the word “ฟืน” /sá-phùu-fún/. In Lao language, “ฟืน” /sá-phùu-fún/ means “detergent.” In Thai language, “ฟืน” /sá-phùu-fún/ means “soap” and “ฟืน” /fùn/ means “dust. The examples from the scenario are “detergent, talc, and dust”.

VI. DISCUSSION AND IMPLEMENTATION

From the results, according to an overview of linguistic perspectives, there are various linguistic-competency aspects that interfere with human cognition. Regarding the sound system, the results of this study are that the clusters /ml/ and /mr/ significantly affect cognition, especially when the morphemes have the same meaning. The conclusion of my study is consistent with that of Nooteed and Potibal (2019).

In terms of morphology, clearly, word formation is a standard characteristic of the Thai and Lao languages. It is no surprise that some morphemes can be separated and still maintain the same meaning. In other words, the root and/or stem of words can shift the meaning into another scenario, and the results reflect the fact that prefix, suffix, and infix are the main factors of word formation in Thai and Lao. This conclusion is consistent with the study by Lao (2020). Moreover, my study confirmed that some Thai word-formation rules can be applied to Lao language processing.

In semantics, my plausible explanation is direct meaning is as important as indirect meaning. Most people infer the overall meaning based on both the direct and indirect meaning of words. I have concluded that all scenarios should be categorized in the pattern of ontology (Tantisripreecha & Soonthornphisaj, 2010; Saengsupawat, 2014; Wijasika & Srivihok, 2014; Khruahong, 2015; Tungkwampian, 2015). An especially interesting phenomenon is taboo words. They play a key role in the inferring of meanings. Though taboo words are not accepted as a formal aspect of language, people use them so much. For this reason, they should be included in the corpus for language processing.

Considering dialectology, there are some effects from dialect such as “ด่าน” /hān/ and “เผียว” /peey/. This fact is consistent with the study by Winch and Gingell (1994), who stressed that dialect interferes with language usage. In my view, because they focused on children in secondary school, it’s obvious that dialect has a major effect on them. However, all of the sample group in my study were adults, so there is little such interference to be made. Furthermore,
several people in my study said that they can infer the meaning of Laotian words because they have visited Laos and they saw the words on travel websites, blogs, and vlogs, especially the word “ບ່ອນຊາຍປີ້” (ticket office). This word may catch the eyes of Thai people because “ปี้” /pi/ is a slang word. This issue is connected to studies by Srinawat (2007), Rungruang (2012), and Choomthong (2017). They emphasized that many slang words are used in social media and newspapers, and that such terms draw the attention of readers.

As a result, the significance of my study involves how to implement the results of developers to natural language-processing systems. An example of this is the automatic question-and-answer machine, machine translations, speech recognition, search engines, etc. Two examples of the implementation will be illustrated when the results of this study are applied by using a search engine. Although search engines don’t have a problem now, their data sources may be extended, thereby enhancing their proficiency. Moreover, because the Thai and Lao languages are similar, the results may apply to use with both languages. Some examples are as follows:

When a person wants to find further information about “big stomach” (in Thai; “ท้องป่อง”/thōŋ-pǒŋ/) and he or she may use the sentence “ท้องป่องผิดปกติทำอย่างไร” /thōŋ-pǒŋ-phì tpò kkatì-tham-yàaŋray/ (What should I do when I have a big stomach?) in the search engine as shown in Figure 1.

![Figure 1 Presents a Sentence That Was Searched by Using a Search Engine.](image)

According to Figure 1, we see the morpheme “ป่อง” /pǒŋ/ in the sentence “ท้องป่องผิดปกติทำอย่างไร” /thōŋ-pǒŋ-phì tpò kkatì-tham-yàaŋray/. So, the inferred word “ป้อง-ยี่ม” /pǒŋ-yīm/ can be divided into two morphemes “ป้อง + ยี่ม,” and the scenario of “ป้อง” consists of 1.) pregnancy 2.) fat 3.) wealthy 4.) expand 5.) hole and 6.) scorpion. This is shown in the table below.

<table>
<thead>
<tr>
<th>Table 3: Scenarios of “ป้อง”</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inferred words</td>
</tr>
<tr>
<td>ป้อง-ยี่ม /pǒŋ-yīm/</td>
</tr>
</tbody>
</table>

As you can see in Table 3, the information on the screen should show the information related to 1.) pregnancy 2.) fat 3.) wealthy 4.) expand 5.) hole and 6.) scorpion. The users can get several options and kinds of information. Moreover, the computer can scope the scenarios which are related to the search words “ป้อง” Additionally, it can extend the information to other words which users may want to use to get further information.

In one case, when the sentence “ไปเยี่ยมอย่างไร” /pài-yīm-yàaŋray/ was used as a search sentence and “เยี่ยม” /yīm/ was a part of it, the data is displayed as Figure 2.
Figure 2 Presents a Sentence That was Searched by Using a Search Engine.

According to Figure 2, we see the morpheme “ເຍີ່ຍມ/yî am/ is in the word “ປ່ອງເຍີ່ຍມ.” So, the inferred word “ປ່ອງເຍີ່ຍມ” can be divided into two morphemes “ປ່ອງ + ເຍີ່ຍມ,” and the scenario of “ເຍີ່ຍມ” consists of 7.) appointment 8.) journey 9.) hospital and 10.) admiration. These are shown in the table below.

<table>
<thead>
<tr>
<th>Inferred words</th>
<th>Scenarios</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. ທ່ອງ/phi-yî am/</td>
<td>7.) appointment 8.) journey 9.) hospital and 10.) admiration</td>
</tr>
</tbody>
</table>

As you can see in Table 4, the scenarios consist of 7.) appointment 8.) journey 9.) hospital and 10.) admiration. So, the computer should show the data related to appointment, journey, hospital, and admiration on the screen for users. Significantly, there was no word meaning related to “prison” in the answer from the sample group. On the other hand, as you can see in Picture 2, a search engine presents many lists of web-page data related to “the prison” and the word “เรือนจำ/ruan-cham/ and “คุก/khûk/ which means “a prison” in Thai language. This ties into a major challenge of this study: Why don’t people in this case think about a prison? There may be a meaning transference because “ເຍີ່ຍມ/yî am/ can be used in the situation like: The cousins go to the prison to visit someone. The bags of words (word lists) in each scenario are so important, and future studies should focus on bags of words in deep detail.

To sum up, the results illustrated that a cross-scenario data set can be created from these Lao-to-Thai inferences. It composes of related words that are useful for Thai and Lao language processing. This study will be advantageous for language processing developing.

ACKNOWLEDGEMENTS

I greatly appreciate the International College for Sustainability Studies at Srinakharinwirot University in Bangkok for giving me a research grant.

REFERENCES


Gornkrit Meemongkol is a lecturer at International College for Sustainability Studies, Srinakharinwirot University, Bangkok, Thailand. He has earned a Ph.D. in applied linguistics. His research interests are on applied linguistics, computational linguistics, semantics, pragmatics, sociolinguistics and linguistic landscape. To contact him at email: gornkrit@g.swu.ac.th