

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; Chamniyom, 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 “ບ' ອນຊາຍບ” /bòn-khāay-pii/. Most Thai people without a background in Lao can't understand “ບ' ອນຊາຍບ,” but they know the meaning of some morphemes, such as “ບ' ອນ” /khāay/, which is “a casino” and “ຊາຍ” is “to sell.” Thus, they can infer the meaning of “ບ' ອນຊາຍບ”.

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, “ຊາຍ” is a transitive verb, so the tagging uses the abbreviation Vt. with the parentheses. To wit: [ຊາຍ]/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 & Soonthornphisaj, 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: “ປ່ອງຍີ່ມ” /pòŋ-yîam/; “ແວ່ນແຍງ” /wên-yēŋ/; “ອ້ານແປງໂມງ” /hān-pēŋ-moŋ/; “ຜູ້ຈ້າງຫນ້າ” /phū-tāŋ-nā/; “ຫ້ອງກາຮັບອົງກົນຕວາມສງບ” /hōŋ-kān-pōŋ-kan-khwaam-sa-ŋòb/; “ສບຸຜູ້ນ” /sa-bū-fūn/; “ວັນສົງຊາຮຸ້ນ” /wan-sāŋ-khān-khuen/; “ໃບຍື່ງຍື່ນ” /bay-yāŋ-yuen/; “ບ່ອນຊາຍປີ້” /bòŋ-khāy-pīi/; and “ຂົກກັດໂລ” /chāk-kh f-lāy/. 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 1
BASIC INFORMATION REGARDING THE SAMPLE GROUP

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

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.

TABLE 2
INFERRED WORDS AND SCENARIOS

Inferred words	Scenarios
1. ບໍ່ ອັງເຍ ັ ຍັມ /pòŋ-y àm/	1.) pregnancy 2.) fat 3.) wealthy 4.) expand 5.) hole 6.) scorpion 7.) appointment 8.) journey 9.) hospital 10.) admiration
2. ຄວ ັ ນ ຄ ຍ ງ /wên-yēŋ/	1.) mirror 2.) glasses 3.) light 4.) argument 5.) diagonal 6.) evening 7.) vegetable 8.) slice
3. ອັ ັ ນ ຄ ບ ງ ໂມງ /h áan-pēŋ-moon/	1.) time 2.) clock repair shop 3.) grocery 4.) watermelon 5.) dressing table 6.) countryside 7.) front row 8.) stage 9.) late 10.) mosquito net
4. ຜ ັ ຈ ັ ງ ຫ ັ ນ ັ ັ /ph ú-tàan-n ài/	1.) stranger 2.) guest 3.) representative 4.) foreigner 5.) male 6.) make-up artist 7.) window 8.) prime minister 9.) village headman
5. ຫ ັ ອ ງ ກ ັ ສ ບ ັ ອ ງ ກ ັ ນ ຕ ວ າ ມ ສ ງ ບ /hòŋ-kæn-pòŋ-kan-khwaam-sa-ŋòb/	1.) police 2.) police booth 3.) police station 4.) room 4.1) bedroom 4.2) darkroom 4.3) control room 4.4) bathroom 4.5) karaoke room 4.6) first aid room 4.7) security room 4.8) operating room 5.) banquet hall 6.) office 7.) sound 8.) sanctum 9.) Ministry of Justice 10.) mortuary 11.) insurance 12.) pub
6. ສ ບ ັ ຜ ັ ນ /sa-b ú-fùn/	1.) powder 2.) detergent 3.) liquid soap 4.) talc 5.) tanaka 6.) sponge 7.) dust 8.) body scrub 9.) cosmetic 10.) face scrub 11.) dishwashing liquid soap 12.) blackboard eraser
7. ວ ັ ນ ສ ັ ງ ຊ ັ ຄ ຊ ັ ນ /wan-sǎŋ-khāan-khuen/	1.) day 1.1) birthday 1.2) marriage day 1.3) Thai new year holiday (Songkran) 1.4) the last day of the Thai new year holiday 1.5) the third day of Thai new year day 1.6) new year day 1.7) The death day 1.8) Visakha Bucha day 1.9) good-weather day 1.10) holiday 1.11) Buddhist holy day 1.12) important day 1.13) new year day in Laos 1.14) cremation day 1.15) Loy Krathong day 1.16) family day 2.) get well 3.) older 4.) waxing moon
8. ໃ ບ ຍ ັ ງ ຍ ັ ນ /bay-yân-yuen/	1.) tree 2.) basil 3.) leaf 3.1) tea leave 3.2) banana leave 3.3) green leaves of spotted sicklefish 4.) certificate 4.1) birth certificate 4.2) death certificate 4.3) warranty certificate 4.4) marriage certificate 5.) government documents 5.1) driver's license 5.2) identification card 5.3) house registration 5.4) title deed 5.5) traffic ticket 5.6) member card 6.) security 7) banknote 8.) money (noun) 9.) long life
9. ບ ັ ອ ນ ຊ ັ ຍ ັ ປ ັ /bòn-khāay-pii/	1.) casino 2.) pawnshop 3.) brothel 4.) market 5.) shell market 6.) grocery 7.) drug store 8.) ticket office 9.) bank 10.) coin shop
10. ຂ ັ ກ ຕ ັ ດ ັ ດ ັ ດ /ch ák-kh í-lây/	1.) thinking 2.) calculator 3.) abacus 4.) clear 5.) machine 6.) sewing machine 7.) computer 8.) scarecrow 9.) exile 10.) drive out 11.) boomerang 12.) revolution 13.) bicycle 14.) miss 15.) hate

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. ບໍ່ ອັງເຍ ັ ຍັມ /pòŋ-y àm/

Ten scenarios were created from the word “ບໍ່ ອັງເຍ ັ ຍັມ” /pòŋ yám/. In Lao language, “ບໍ່ ອັງເຍ ັ ຍັມ” /pòŋ yám/ means “a window.” In Thai language, “ปล่อง” /pòŋ/ means “a bulge,” such as in one’s stomach, and “เยี่ยม” /y àm/ means “to visit.” The examples from the scenario are “pregnancy, fat, and hospital”.

2. ຄວ ັ ນ ຄ ຍ ງ /wên-yēŋ/

Seven scenarios were created from the word “ຄວ ັ ນ ຄ ຍ ງ” /wên-yēŋ/. In Lao language, “ຄວ ັ ນ ຄ ຍ ງ” /wên-yēŋ/ means “a mirror.” In Thai language, “แว่น” /wên/ means “the glasses” and “แหยง” /yēŋ/ means “to push into something.” The examples from the scenario are “mirror, glasses and light.”

3. ອັ ັ ນ ຄ ບ ງ ໂມງ /h áan-pēŋ-moon/

Nine scenarios were created from the word “ອັ ັ ນ ຄ ບ ງ ໂມງ” /h áan-pēŋ-moon/. In Lao language, “ອັ ັ ນ ຄ ບ ງ ໂມງ” /h áan-pēŋ-moon/ means “a clock repair shop.” In Thai language, “ร้าน” /hāan/ means “a shop (dialect).” “แพง” /pēŋ/ has no meaning in standard Thai (but the meaning is “expensive” in dialect). “แปลง” /plēŋ/ means “an area and changing.” “โมง” /moon/ is “a unit of time.” The examples from the scenario are “eight o’clock, clock repair shop, and stage.”

4. ຜ ັ ຈ ັ ງ ຫ ັ ນ ັ ັ /ph ú-tàan-n ài/

Nine scenarios were created from the word “ຜ ັ ຈ ັ ງ ຫ ັ ນ ັ ັ” /ph ú-tàan-n ài/. In Lao language, “ຜ ັ ຈ ັ ງ ຫ ັ ນ ັ ັ” /ph ú-tàan-n ài/ means “a representative.” In Thai language, “ผู้” /phûu/ means “a person and male” in colloquial language. “จ ำ ງ ຫ ັ ນ ັ ັ” /ph ú-tàan-n ài/ has no meaning in Thai. The examples from the scenario are “representative, guest and male”.

5. ຫ ັ ອ ງ ກ ັ ສ ບ ັ ອ ງ ກ ັ ນ ຕ ວ າ ມ ສ ງ ບ /hòŋ-kæn-pòŋ-kan-khwaam-sa-ŋòb/

6. ಸಬ್ ' ಫನ್ ' ಬ /sa-bùu-fùn/

7. វង្សានុវង្ស /wan-sǎŋ-khǎan-khuen/

8. 𐄀𐄁𐄂 𐄃𐄄𐄅 /bay-yân-yuen/

9. ບ ' ອນຊາຍບ ^{໒໒} /bòn-khăay-pii/

10.ຂໍ້ກຮຕໍ່ໄລ່ /ch ək-kh f-l ây/

VI. DISCUSSION AND IMPLEMENTATION

Considering dialectology, there are some effects from dialect such as “ฮ่าน” /háan/ and “ແປງ” /pɛɲ/. 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 “ບ' ອງເຍີ້ມ” (ticket office). This word may catch the eyes of Thai people because “ปี่” /pii/ is a slang word. This issue is connected to studies by Srinarawat (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”/thǒwng-pòŋ/) and he or she may use the sentence “ท้องป่องผิดปกติทำอย่างไร” /thǒwng-pòŋ-ph ɪp̚k̚k̚at ɪtham-yàaŋray/ (What should I do when I have a big stomach?) in the search engine as shown in Figure 1.

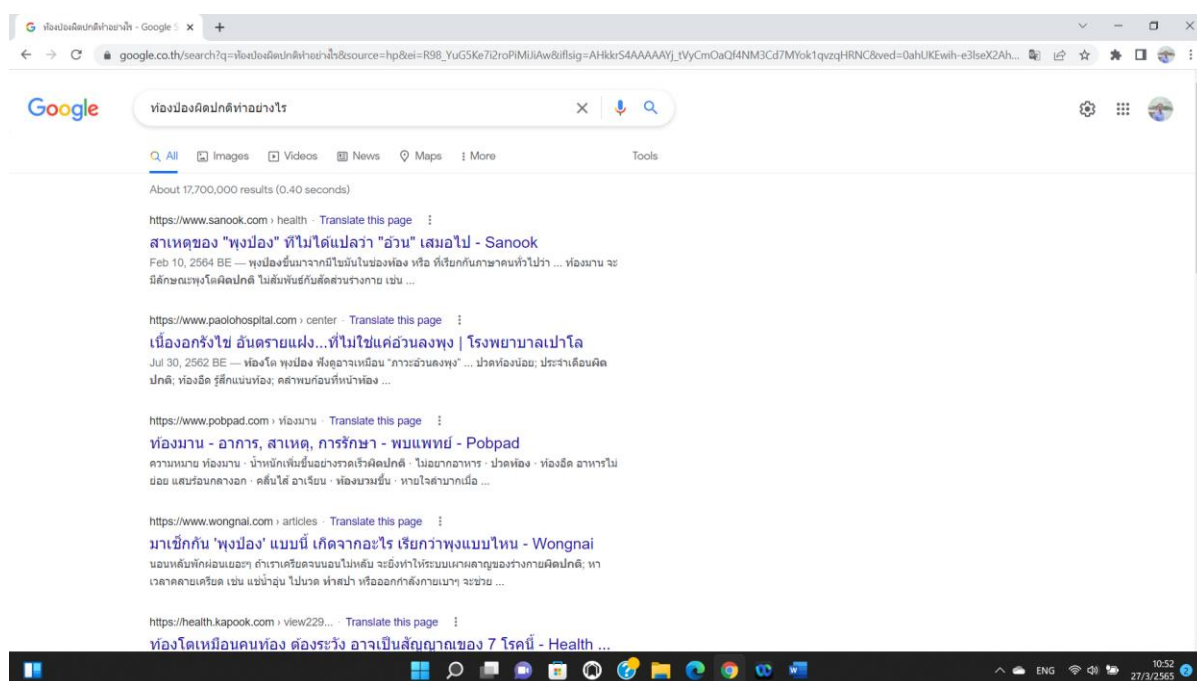


Figure 1 Presents a Sentence That Was Searched by Using a Search Engine.

According to Figure 1, we see the morpheme “ป่อง” /pòŋ/ in the sentence “ท้องป่องผิดปกติทำอย่างไร” /thǒwng-pòŋ-ph ɪp̚k̚k̚at ɪtham-yàaŋray/. So, the inferred word “ບ' ອງເຍີ້ມ” can be divided into two morphemes “ບ' ອງ + ເຍີ້ມ,” and the scenario of “ບ' ອງ” consists of 1.) pregnancy 2.) fat 3.) wealthy 4.) expand 5.) hole and 6.) scorpion. This is shown in the table below.

TABLE 3
SCENARIOS OF “ບ' ອງ”

Inferred words	Scenarios
1. ບ' ອງເຍີ້ມ /pòŋ-y ɪm/	1.) pregnancy 2.) fat 3.) wealthy 4.) expand 5.) hole 6.) scorpion

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 “ບ' ອງ.” 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” /pay-y ɪm-yàaŋray/ was used as a search sentence and “เยี่ยม” /y ɪm/ was a part of it, the data is displayed as Figure 2.

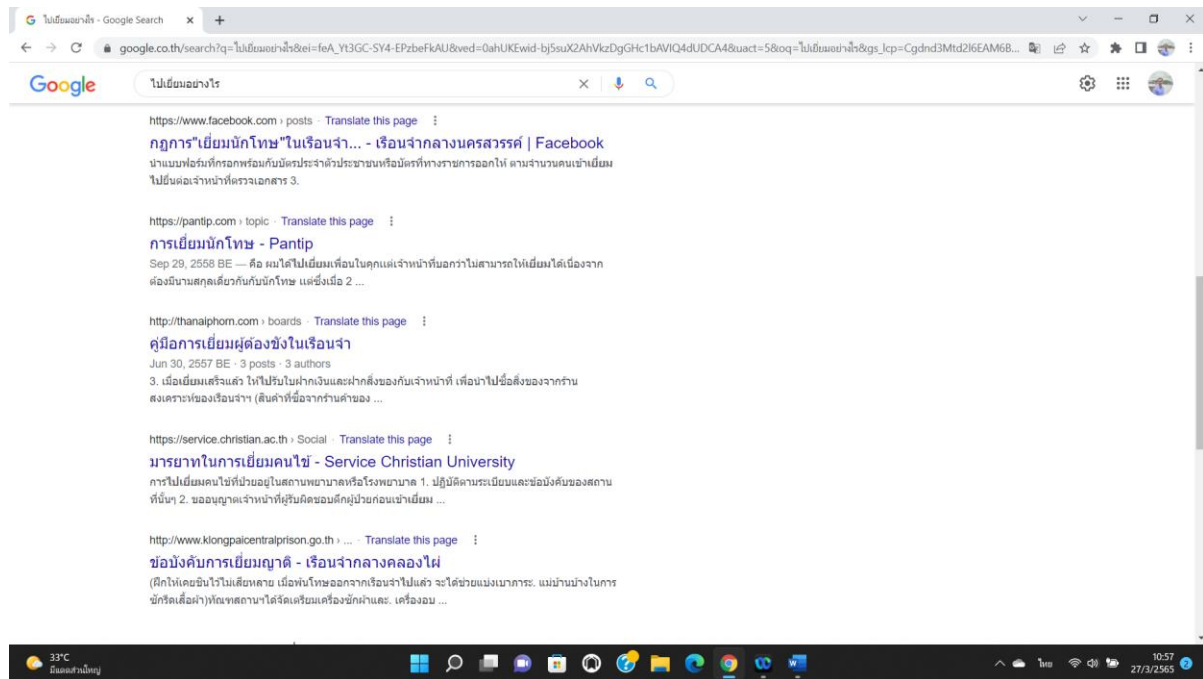


Figure 2 Presents a Sentence That was Searched by Using a Search Engine.

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

TABLE 4
SCENARIOS OF “ เยี่ยม ”

Inferred words	Scenarios
1. ไปเยี่ยม /pòŋ-y àm/	7.) appointment 8.) journey 9.) hospital and 10.) admiration

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” /khúk/, 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 “เยี่ยม” /y àm/ 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

- [1] Arun, N. (2011). Artificial Intelligence and the Application. *Executive Journal*, 30(4), 167-171.
- [2] Baker, M. (1992). *In other words*. London: Routledge.
- [3] Benson, M. (1985). Collocations and Idioms. In Ilson, R. (Ed.), *Dictionaries, lexicography and language learning*. Oxford: Pergamon Press.
- [4] Bloomfield. (1933). *Language*. University of Chicago Press.
- [5] Cambridge, (2021). *Online Dictionary*. Retrieved 27 March, 2022 from <https://dictionary.cambridge.org/>
- [6] Chamniyom, R. (2003). *Kan sueksa kham bubphabod thi klaimachak khamkariya nai phasa Thai*. [A Study of Preposition converted from Verbs in Thai]. M.A. Thesis, Department of Thai Language, Silpakorn University, Thailand.

- [7] Charoensuk, J. (2006). *Kanbaeng Khobkhed arnupark pariched nai phasa Thai doy chai kham rabunai lae khorsonthed choeng wakayasumphan*. [Thai elementary discourse unit segmentation by using discourse segmentation cues and syntactic information]. M.A. Thesis, Department of computer engineering, Kasetsart University.
- [8] Choomthong, D. (2017). Characteristics of Slang Terms in Online Discussion among Thai Stock Investors. *NIDA Journal of Language and Communication*, 22(30), 81-105.
- [9] Fillmore, C. (1968). *The Case for Case*. In *Universals in Linguistic Theory*.
- [10] Fillmore, C. (1971). *Some Problems for Case Grammar*. In O'Brien, RJ, ed.
- [11] Fillmore, C. and B. T. Atkins. (1992). *Frames, Fields and Contrasts: New Essays in Semantics and Lexical Organization*. Lawrence Erlbaum Associates.
- [12] Fillmore and Baker. (2009). *A Frames Approach to Semantic Analysis*. Oxford Handbook of Linguistic Analysis. Oxford University Press. 313-340.
- [13] Gruber, T. R. (1993). A translation approach to portable ontologies. *Knowledge Acquisition*, 5(2):199-220.
- [14] Gruber, T. R. (1995). Toward principles for the design of ontologies used for knowledge sharing. *International Journal of Human-Computer Studies*, 43(4-5), 907-928.
- [15] Intasaw, N. (2013). *Kanyak arnupark phasa Thai duay kanchai baebchamlong supportwektor machine*. [Thai Clause Segmentation Using a Support Vector Machine Model]. M.A. Thesis, Department of Linguistics, Chulalongkorn University, Thailand.
- [16] Intratat, C. (1996). *Krabuankan thi kumkariya kraipen kham bubphabot*. [Grammaticalization of verbs into prepositions in Thai]. Doctoral Dissertation, Department of Linguistics, Chulalongkorn University, Thailand.
- [17] Jeon, N. et.al. (2019). Kan priabthiab withikan niyam khamsub nai phodchananukrom chabub ratchabundittayasathan porsor 2554 kab watchananukrom phasa Laos sathabun withayasarangchad porsor 2555. [The Comparison Between the Lexical Definition Used in the Thai Dictionary of the Royal Institute 2011 and the Lao Dictionary of the National Science Council 2012]. *Sripatum Review of Humanities and Social Sciences*, 19(2). 7-20.
- [18] Johnson, M. (1987). *The body in the mind: The bodily basis of meaning, imagination, and reason*. University of Chicago Press.
- [19] Kanjanawasee, S. (1991). *Kan sarup choeng sahed thi maichai kan todlong*. [Causal Inferences in Nonexperimental Research]. Educational Research News. Educational Research Department, Office of the National Education Commission, 9-13.
- [20] Kawtrakul, A., et al. (2002). "A State of the Art of Thai Language Resources and Thai Language Behavior Analysis and Modeling". In *Proceedings of the International Conference on Natural Language Processing Post COLING 2002 Workshop*, 1-8.
- [21] Khruahong, S. et al. (2015). Ontology Design for Thailand Travel Industry. *International Journal of Knowledge Engineering*, 1(3), 191-196.
- [22] Luo, Y. (2020). Morphology in Kra-Dai Languages. *Linguistics Oxford Research Encyclopedias*. Oxford Research Encyclopaedia of Linguistics, Oxford University Press. "https://doi.org/10.1093/acrefore/9780199384655.013.529"
- [23] Leenoi, T. (2008). *Kansang kruea krai kham Thai khong manotad penthan ruam khong entity lamdap thi nueng duay withikan pae songthang lae kanchai phodchananukrom thi sangdauwithikan taektangkan*. [The Construction of Thai Wordnet of 1st Order Entity Common Base Concepts Using a Bi-Directional Translation Method and with Dictionaries of Different Compilational Approaches]. M.A. Thesis, Department of Linguistics, Chulalongkorn University, Thailand.
- [24] León, M.A.L. (2015). Corpus design and compilation process for the preparation of a bilingual glossary (English-Spanish) in the logistics and maritime transport field: LogisTRANS. *Proceedings of 32nd International Conference of the Spanish Association of Applied Linguistics (AESLA)*, 293 – 299.
- [25] Miller, G. (1998). An introduction to Wordnet. University of Princeton. Odeunmi, A and F. Unuabonah. 2013. "Generic Structure of Nigerian and South Africa Quasi-Judicial Public Hearings. Language, Discourse and Society". *Language & Society, RC 25 of the International Sociological Association*, 2(2), 82-98.
- [26] Nakon, C. (2000). *Rabobkham lae wakayasumphan nai phasa Lao krang mooban Wanglao tambon Non-Krot amphoe Muang changwat Nakhon Sawan*. [Morphology and syntax of Lao Khrang at Wanglao village, Tambon Non-Krot, Muang district, Nakhon Sawan province]. M.A. Thesis, Department of Oriental Epigraphy, Silpakorn University, Thailand.
- [27] National Social Sciences Council. (2012). *Watchananukrom phasa Lao*. [Lao Dictionary]. Vientian: National Publication.
- [28] Nooted, J. and Potibal, P. (2019). Phumisart Phasathin khong kham thi mee payanchanaton kuabkram /mr/ lae /pl/ nai phasa Thai thin tai: lumnam thalaysab Songkhla. [Dialect Geography of lexemes with Initial clusters /mr/ and /ml/ in Southern Thai Dialect: Songkhla Lake Basin]. *Journal of Humanities and Social Sciences Surathani Rajabhat University*, 11(1), 191-215.
- [29] Panmeta, N. (2011). *Waiyakorn Thai*. [Thai Grammar]. Faculty of Arts, Chulalongkorn University, Thailand.
- [30] Phetsiri, C. (2010). *Kan sakad kwamroo kiawkab sabphakun tangya khong phued samunpai chak eakkasarn phasa Thai puea sanab sanun kan tob khamtham attanomat*. [Knowledge extraction of medicinal properties of Thai herbs from Thai texts for supporting automatic question - answering system]. Research report. Dhurakij Pundit University.
- [31] Phosai, K. (2009). *Kanwikhro ha kuam mai fang lae kan rianroo kreangchak sumrab rabob tam tob attanomat*. [Latent semantic analysis and machine learning for Thai question answering system]. M.A. Thesis, Department of Computer Sciences, Thammasat University, Thailand.
- [32] Prapin, W. (1996). *Kan suksa priabthiab kham sub phasa Lao song in changwat Nakhon Pathom, Ratchaburi lae Phetchaburi*. [Comparative lexical study of Lao-Song in Nakhonpathom, Ratchaburi and Phetchaburi province]. M.A. Thesis, Department of Thai Epigraphy, Silpakorn University, Thailand.
- [33] Rattanaprasert, W. (1985). *Kham and laksana kham nai phasa Lao wiang nai changwat Chachoengsao*. [Word classes and word types of Lao-Wiang language in Chachoeng Sao province]. M.A. Thesis, Department of Oriental Epigraphy, Silpakorn University, Thailand.
- [34] Rungruang, J. (2012). A Comparative Study of Thai and Chinese Internet Language. *Journal of International Studies*, 2(2), 33-46.
- [35] Saengsupawat, P. et al. (2014). Ontology-based knowledge acquisition for Thai ingredient substitution. *ARNP Journal of Engineering and Applied Sciences*, 9(9), 1461-1468.

- [36] Sila, S. (1975). *Priabthiab kham taektang rawang phasa Krungthep kap phasa Kubua*. [A comparative study of Krungthep and Kubua language]. B.A. Thesis, Faculty of Archaeology, Silpakorn University, Thailand.
- [37] Sinclair, J. (1991). *Corpus concordance collocation*. Hong Kong: Oxford University Press.
- [38] Srinarawat, D. (2007). Thai political slang: formation and attitudes towards usage. *International Journal of the Sociology of Language*, 186, 95- 107. "https://doi.org/10.1515/IJSL.2007.044"
- [39] Suktharachan, M. (2017). *Kansrang krobmanothad phasa Thai chak klangkhormun kankaset phuea pen than kwam roo sumrub kanpramuangpon duay computer*. The development of Thai concept frame in agricultural domain for knowledge based computational processing. Doctoral Dissertation, Department of Linguistics, Kasetsart University, Thailand.
- [40] Tantisripreecha, T. and Soonthornphisaj, N. (2010). A Study of Thai Succession Law Ontology on Supreme Court Sentences Retrieval. In *Proceedings of the International Multi Conference on Engineers and Computer Scientists*, 146-151.
- [41] Thipsena, R. et.al. (2014). Kanchamnak glum khamtham attanomat bon kradan sonthana dau chai technique mueang khor kwam. [Automatic Question Classification on Webboard Using Text Mining Techniques]. *Science and Technology Journal Mahasarakham University*, 33(5), 493-502.
- [42] Tungkwampian, W. et al. (2015). Development of Thai Herbal Medicine Knowledge Base Using Ontology Technique. *The Thai Journal of Pharmaceutical Sciences*, 39(3), 102-109.
- [43] Wijasika, A. and Srivihok, A. (2014). *Thai Food Safety Document Searching System by Ontology*. 2nd International Conference on Research in Science, Engineering and Technology (ICRSET'2014), 11-15.
- [44] Winch and Gingell. (1994). Dialect interference and difficulties with writing: An investigation in St. Lucian primary schools. *Language and Education*, 8(3), 157-182. "https://doi.org/10.1080/09500789409541388"
- [45] Wutthikorn, Y. (2010). *Kan khayai than kwamru baeb attanomat sumrab rabob tham-tob attanomat*. [Automatic knowledge base expansion for dialog system]. M.A. Thesis, Department of Computer Sciences, Thammasat University, Thailand.
- [46] Yensamut, P. (1981). *Kham lae Kwammai nai phasa Laosong*. [Words and meaning in Lao Song]. M.A. Thesis, Department of Thai Epigraphy, Silpakorn University, Thailand.
- [47] Yuyen. V. (1997). Kansueksa Priabthiab kwam taektang rawang phasa Thai kub Phasa Lao lae khorbokphrong nai kanchai phasa Thai khong naksueksa Lao thi Mahawithayalai Kasetsart. [A Contrastive Study of Thai and Lao Languages as well as The Errors in Thai Usage of Lao Students at Kasetsart University]. *Manutsayasad Wichakan Journal*, 5(1), 35-54.

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