Towards Electronic Version of the Royin Thai Dictionary from Information-Heavily Semi-structured Data Source

Taneth Ruangrajitpakor[†a)], Adisak Kingkaewkanthong[†], and Thepchai Supnithi[†]

National Electronics and Computer Technology Center, National Science and Technology Development Agency, Thailand

Abstract— **As to provide knowledge of Thai words, the Royin dictionary has been decided to become digitised. In this work, processes of extracting information from printing version of the dictionary are described. Since the information source is in semi-structured format, an automatic method of type detection is used to extract respective details into database. Patterns and format of the source are fully used in consequence as a hint for extraction. Moreover, ambiguities and their solution in extracting process are discussed. As a result, lexical entries are systematically stored with distinguishable details, and entries are connected with other by interoperable relations. From evaluation, the automatic extraction processes can handle more than 80% of entries in overall, and the remaining ambiguous entries were sent to experts for decision-making.**

Index Terms— Dictionary Development, Electronic Dictionary; Information Extraction, Semi-structure Source

I. Introduction

Lexical information from a monolingual dictionary is one of basic data in natural language processing. A monolingual dictionary aims to describe lexical information in several aspects including syntactic usage, semantic meaning and examples. In Thai-land, Royin dictionary [1] is the most referable Thai monolingual dictionary since it has been published by Office of the Royal Society in which gathers top-notch scholars in many fields to inform lexical details in their respective genre. The first version dictionary was first published in 1982 in a printed version, and the processes of development were carefully and manually conducted. Since then, missing words and new words have been added to the content. Revisions of content have been made several times passing down by respectable technical committees to improve a quality of the dictionary. In the current version (2011 edition), the dictionary apparently be-comes the main lexical reference for Thai and provides not only syntactic and semantic usage, but also trivia of words such as word etymology (origin of word), semantic relation (synonym, hypernym and hyponym), register (a variety of a language used in a particular social setting), and word form (short form, full form, abbreviation form).

From the beginning, all contents have been manually crafted and solely designed for printed version. Hence, symbols (period, parenthesis, semicolon and comma) and formats (bold font, italic font, whitespace and line-break) are used to notify different types of lexical information. The notations along with contents make the data to become semi-structured. However, with the continuous revisions and more information types, the notations and patterns used in dictionary content become limited and complex in a printing version. Moreover, the development of computer networks has made the urge and request for a dictionary to become electronic data. For convenience in distribution of Thai lexical knowledge, the Office of the Royal Society decided to develop their dictionaries in an electronic version alongside with a printing version.

With semi-structure of the original content, we aim to automatically extract various types of information using existing pattern and format as a clue. In this work, knowledge of words such as semantic relation is carefully maintained within the ex-traction process and is managed in machine-readable form for further retrieval. To assure high quality of the generated results, incomplete and ambiguity within content are plan to be semi-automatically processed. We expect that the developed dictionary will be one of reliable Thai lexical resources to enhance researches of Thai natural language processing.

The rest of this paper is organised as follows. Section II provides background related to Royin Dictionary and electronic dictionary. Section III gives details on extraction methods designed to capture information from semi-structure. Section IV shows results of the method and discussion. Last, Section V provides conclusion and a plan for future work.

[†] The authors are with National Electronics and Computer Technology Center, National Science and Technology Development Agency, 112 Phahon-Yothin Rd., Klong-Luang, Pathumthani, 12120, Thailand.
a) taneth.rua@nectec.or.th

TABLE I
Various types of information provided in Royin Dictionary

Type	**Description**	**Notation and Format**	**Example**
Head Word*	-	bold, bigger font size, first word on the line	-
Sub Word	an expanding word(s) of the head word	bold	-
Part of Speech*	a syntactic category in accordance with its functions	abbreviated notation from POS list in parentheses, located after head word	น. (noun) ก. (verb) ว. (adjective or adverb)
Register	a variety of a language used for a particular purpose or in a particular social setting	short word notation from a list in brackets after head word	โบ. (old word) น. (verb)
Domain	a specific field where a word is used	short word notation from a list in brackets after head word (may appear in the same brackets with register)	คอม. (computer domain) น. (verb)
Definition*	meaning	normal text in explanation, semicolon is used to separate several different meanings while comma is used to split close meanings.	-
Etymology	the origin of words informing borrowing word or meaning from another languages	an abbreviated notation from a language list, may include the original word in the referring languages	
Usage Example	-	a normal text after the last explanation and often initial with word “เช่น”	
Reference	-	a book name in parentheses where the definition is referred to	-
Citation	-	a book name in parentheses for referring original of the example quote	-
Word Relation	a relation to other words	a normal text after the last explanation along with a keyword such as “ก็ว่า” (synonym) or “เทียบ” (see also)	-
Extra Information	information related to the word such as image, cultural history related to the word, word forms (short form or abbreviation), etc.	a normal text with or without specified keyword within explanation (inconsistent format)	-

II. Background

a. Royin Dictionary

Royin dictionary [1] is a Thai monolingual dictionary published by the Office of the Royal Society. It has been recognised for reliable reference for Thai lexical information and utilised throughout Thai government sections in official documents. The first version was created in 1982 in a printing version, and it has been revised and improved by adding more emerging words since then. The last version was 2011 edition in a printing version. The content in the later versions was developed in Adobe InDesign format recommended from the publisher for printing purpose. The content in the printing version is denoted with symbols and formats to inform several types of lexical information as shown in Fig 1.

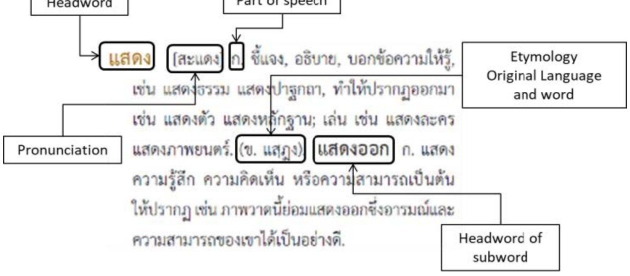

Headword
Part of speech
แสดง
[สะแดง]
ก.
ชี้แจง, อธิบาย, บอกข้อความให้รู้,
เช่น แสดงธรรม แสดงปาฐกถา, ทำให้ปรากฏออกมา
เช่น แสดงตัว แสดงหลักฐาน; เล่น เช่น แสดงละคร
แสดงภาพยนตร์. (ข. แสฎง) แสดงออก ก. แสดง
ความรู้สึก ความคิดเห็น หรือความสามารถเป็นต้น
ให้ปรากฏ เช่น ภาพวาดนี้ย่อมแสดงออกซึ่งอารมณ์และ
ความสามารถของเขาได้เป็นอย่างดี.
Etymology
Original Language
and word
Pronunciation
Headword of
subword

Fig. 1. An example of Royin Dictionary in printing version

Despite being semi-structured data, notations and formats can be used in combination and cause confusion in extracting information. Due to space limitation, we list some of frequent-used information types and their example in Table I. Please note that the rolls in Table I denoted with star (*) refer to mandatory information while the rest is optional. Please also be noted that in Thai, most of content-modifying words (adjectives and adverbs) in the same form can be used to modify or express attributes to noun and verb without prefix or suffix. Hence, both adjective and adverb in Thai are acceptably grouped together in the same syntactic part-of-speech.

With provided information based on Table I, we can see that same notations and formats are assigned for many information types. Despite these patterns are highly accurate (above 90%) in annotation, content is still difficult to be extracted to structured data without fully understanding in linguistic information. Moreover, some details such as etymology or word relations are magnificently rich with content. For examples of content-richness, there are details of how the word was formed from original words of two different languages, or how the current spell-out form changes from the original language.

b. A Development of Electronic Dictionary

To develop an electronic dictionary, there are three main approaches: manual, automatic and semi-automatic. For manual approach, developers can ask native linguistic

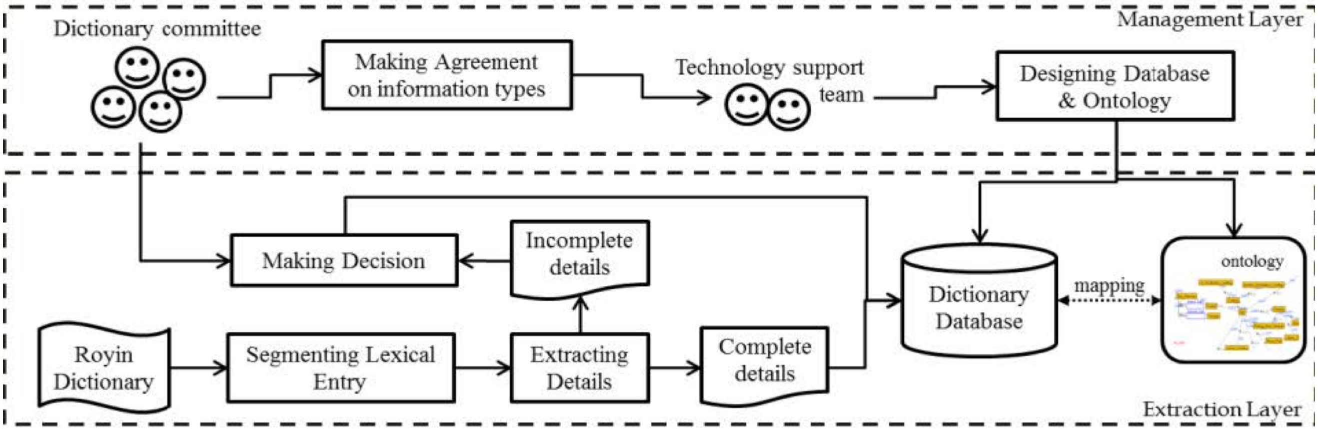


Fig. 2. An overview of a development of electronic Royin dictionary

experts or use words from corpora to gather words for initial data and fill details manually. This approach has an advantage of data with high quality but trading off to the high cost and time consuming. The remarkable difficulty in this approach is to expand their lexical coverage. To solve such issues, some developers choose to manually add entries or apply a strategy to ask users in their community to reduce experts' burden in adding new words. The renowned example of electronic dictionary developed with this approach is Lexitron [2][3]. Automatic approach, in the other hand, attempts to use existing information from open resources such Wikipedia or thesaurus as a base and to categories information into designated fields. This approach is fast and effective in developing, but the obtained details are required to be approved. Moreover, the risk of incorrectness in details is apparently high, and it rarely provides information in expert-level knowledge (such as word etymology and variation of words) since the information will require several types of linguistic expertise in analysis. Last, the automatic approach requires a very high amount of reliable resources and often cannot be done for resourceless languages. Semi-automatic approach in developing electronic dictionary is the most anticipated approach since it gains ad-vantages of both manual and automatic approach. By combining helps from linguistic experts and finding clues from resources, a developed electronic dictionary contains with reliability and high-level knowledge of lexicons with lower cost and burden. However, this approach requires a base resource with well format and pattern as initial resource in which is rare and often under right-protection. An example of electronic dictionary developed using this approach is [4].

III. Methodology

This work aims to extract data from Royin Dictionary to a database schema. Though, our ideal is fully automatic processes of extraction, truthfully some parts may still require expert decision to disambiguate confusing results. In this work, the main and only input is the Adobe InDesign file of the Royin Dictionary 2011 edition [1]. Information in the content will be recognised and extracted separately into types and store into a database schema designed in accordance to an agreement with the responsible committee of the Office of the Royal Society. Moreover, an ontology [5] is strictly designed in order to standardise and generalise the information. An over-view of the processes is sketched in Fig. 2.

a. *A Design of Information Types*

Since the Royin Dictionary contains various types of information, we need to make an agreement on selected types and details. From several meetings, a linguistic expert committee and technical supporting team together made a consensus on selected types based on a standard of database schema following ISO1951:2007 [6]. To make it more generalised and comfortably interoperable, an ontology schema was designed on top of the database as a representation of lexical information and relations

With the designed ontology, we can see that a lexical entry is the main concept of the whole. However, not all of the information types are directly linked to an entry concept but is related to details. For example, pronunciation and etymology concept belongs to a headword concept while example and citation belongs to a description concept since description is differentiated based on word-sense, not a headword. With an ontology schema, cloudy implicit relations within dictionary are made more visible and understandable, and this eventually helped in discussion and agreement processes.

b. *A Segmentation of Lexical Entry*

At first, the Royin dictionary is in a printing format based on *Adobe InDesign CS 6*. For extraction, the format is exported into a universal tag format, XML. Since con-tents in a dictionary are about word-senses in a lexical entry, we aim to separate entries in XML notation.

In fact, the Royin dictionary has been designed to gather words and their respective expanded words in the same entry since a nature of Thai language is to expand more meaning by combining other stems after the core meaning word. In this work, we aim to divide the core meaning and its expanded meaning into different entries as main sense and expanded sense for query purpose, respectively. To keep the knowledge of lexicon relationship, the main sense and expanded sense are designed to link with each other in a form of hypernym and hyponym.

To detect main sense and its expanded sense(s), we simply used a format pattern and a symbol notation. In the Royin case, the solid format of using full-stop (.) symbol following by a word in bold is the hint for separating these entries. As a result, an amount of entries is dramatically increased from the original while still keeps hierarchical relation. Eventually, the entries become easier to process further for assigning semantically relation and detail extraction.

c. *Information Detection and Extraction*

Format For each lexical entry, details are possible to be within entry or inherited from its main entries. This situation happened because the printing version attempts to reduce the repetition of same information appearing in the same entry. However, this will decrease a quality of data if some known details are missing. Hence, we attempt to assign these details considering hierarchy relation from the source.

In this process, all details are extracted in consequential manner since some separation will solve ambiguity in overlapping symbol and format used. Within entry, symbol such as brackets, parentheses, dashes, commas and semicolons are heavily used while only italic can be the only hint in contextual format. For details, the following types of information are a major extracting goal.

- Part of speech
- Local usage
- Domain
- Pronunciation
- Definition
- Etymology
    - Original language
    - Original word(s)
- Example
- Reference
- Citation
- Form
    - Short form
    - Full form
    - Abbreviation form
    - Royal word form
- Word relation
    - Synonym
    - Hypernym-Hyponym
    - See also

These are major contributed details of the entries. Though there are several more details, but those are minority and will be stored in additional note field for further analysis and processing. Despite being semi-structured, some details can be used in the exactly same format or located together, such as local usage and domain in which given in brackets after a headword. These ambiguous issues generally occur for most types. Thus, the major tasks in this work are to disambiguate the confusing. From observation, we can classify ambiguity in extracting into types as follow.

1) An Ambiguity by Bracket-type

Bracket-type punctuation refers to bracket and round-bracket punctuation that is always used in pair forming close section in a lexical entry. The found ambiguity with this type is 1) missing opener or closer bracket, 2) different types of information notified in the same type bracket and the same format, and 3) joining different types of information together in the same bracket set. Furthermore, the types of information in this issue are split into two kinds: close set and open set. The closed set is for specific type such as part-of-speech. The open set refers to freely unspecific text.

For the first one, it happens in a notable number for a close set, but very few for open set. It can be easily detected and solved by checking for a pair for a closed set, and fortunately this issue of open set is always located to the end of entries for effort-less solving. The second and last issue are common in the dictionary for only a closed set, such as domain and local usage, since the printing version should not contain redundancy brackets nearing each other. For the case, a list of possible instances of the information types was asked from the committee as reference for differentiating information types.

2) Inequivalent Relations among Entries

Another issue in extracting Royin data is to maintain relations of lexical entries to other. Relations assigned in the dictionary are split into two kinds: one-way relation and round-trip relation. For one-way relation, we detect the markers for inferring a relation type, such as "ดู" and "ดูประกอบ" both signify the relation see also, and "เรียกเต็มว่า" refers to the relation of the target entry of its full form. By knowing the relations from markers, most of the one-way related entries can be extracted with less trouble. Furthermore, the round-trip relations (such as synonym) require examination of same markers from all referring entries to assure the correctness. For this process, we encounter two issues. The former is a missing of a marker from the designated entry, and the latter is different types of a marker used to relate one another. In these cases, we should not automatically solve the issues since these will directly affect accuracy of the dictionary output. Thus, we collected these troublesome entries and consulted with responsible committees. For the latter type, about 60% of the round-trip marked entries were detected as errors.

3) Usage Example Details

Among the meaning content, usage examples of a word in sense specific are given. The examples are hinted with the keyword „เช่น□ (for example) as an initial marker for this kind of detail. However, all contexts after the keyword are not always usage ex-ample, but they can be a part of description such as providing a subtype as an example. This leads to ambiguity in detection this kind of detail.

TABLE II
Extraction result comparing to original printing version source

Information Type	Printing Ver.	Electronic Ver.	Process and Change Made
Lexical Entry	20,944 entries	52,099 entries	Main sense and its expanded senses were separated into individual instances for searching and linking purpose.
Definition	In consequent text split with semicolon along with example	Numbering definitions and attaching with example to respective sense	Splitting definitions and numbering them; usage examples and concept examples[1] are disambiguated by checking if the headword appears after the marker "เช่น" (for example) or not
Local Usage (Dialect)	13 dialects	16 dialects	Dialect names referring to same dialect were unified, and some new dialects were found in the process from within content and got approved.
Book Reference (Reference and Citation)	229 books in a list in preface	68 books currently in used; 6 new books not from the list but existing in content	Detection and splitting books into two types: Reference and Citation. Reference is for dictionaries for referring lexical meaning and detail. Citation includes poetry copies and historical inscriptions for citing original quotes.
Semantic Relation	Some entries skip all details but link for synonymy	Approved semantic relation among entries and adding omitted details of entries from synonym	Automatic detection of used marker (see section 3.2) was performed and applying details from its synonymous entries for omitting details to complete information of entries.
Etymology (Loan Word and Its Change)	Provision with the same pattern as dialect and additional note	Found different internal patterns (see section 3.3.3)	Automatic Detection for etymology field; 78% of details were extracted correctly while the rest required experts for decision-making.

To handle the issue, a rule is designed to match pattern and text for distinguishing usage example and description containing the keyword. Several rules are designed based on many suggestions from experts from Royin. The designed rules for usage example detection are as follows.

- Detecting 'เช่น' and keep the following texts until finding period (.) or comma (,) as a candidate detail
  1. A candidate contains more than 25 strings.
  2. A candidate must not locate among commas.
  3. A candidate must contain a headword of the entry.

The first rule is mandatory while other rules are optional. The rules (rule#1-3) are from different experts. In this work, we try to select among or combine these rules and see which rule(s) is best in which circumstances.

4) Unique Details

The most complex issue in this process is details in etymology. Please note that Thai words may be originated from various languages from across cultures in the past. Some words may be directly loaned while some words were originated by combining original words from same or different languages. In Royin dictionary, these details were carefully studied and provided for users. The denotation for the etymology information is located at the last of an entry in round brackets referring to only the headword of the current entry, and the possible patterns of the information are listed below:

1) language
2) language original_word
3) language, language
4) language; language
5) language original_word + original_word
6) language original_word, original_word
7) language original_word + language original_word

Each pattern has its own different specific meaning. The first one is chosen to use for loan words that carry on the same spelling or pronunciation while the second one informs an original word that is slightly different in spelling or pronunciation. For the third and fourth pattern, the languages in the closed family such as Pali and Sansakrit are separated with comma as in the third pattern while the fourth pattern signifies the languages in different family. The fifth pattern informs the word created by combining two original words from the same language, and the sixth is about two uncertain original words from the same language. The seventh pattern shows a combination of different words from different original language. Unfortunately, these patterns can be integrated in a case of several possibilities of word originality.

All details are given in the set of round brackets. The language is provided in abbreviated form in a closed set and can be retrieved from the list. The original words are, however, a freely open type. In the current situation, the language in etymology can be semi-automatically handled by referring to the list despite some same abbreviations are used for different languages. The original language once extracted is schematised as attribute of a headword of a lexical entry. For multi-languages, different fields are required to store data separately. The original word(s) in the second and fifth pattern were extracted and linked to the original language for signifying the relation from word to language to headword, consequentially. However, details of the seventh pattern are difficult to keep with given knowledge with separation of languages and words; hence, we keep the details together for now in one data field.

IV. Results and Evaluations

In this section, we show the results of detail extraction and the change from printing version. We also conducted an experiment to evaluate our automatic processes.

a. Extraction Result

Based on the statistics of the printing version of Royin Dictionary, we found the difference in numbers from disambiguating patterns and error formats. The numbers and changes of notable information types are given in Table II.

b. Evaluation Result

In this part, we split a testing into 2 groups. The first one is an evaluation of details that we have a gold standard from experts, and the second is about details that we manually counted a correct result. The first group contains a detail of usage example while the rest details are in the second group.

1) Experiment on Usage Example Detection

Based on section 3.3.3, there are several available rules to detect and extract the details. We attempt to find which rule or combined rules perform best in usage example detection. Since we obtained the manual result from expert, it will be used to calculate for precision and recall for detection. The results are given in Table III.

TABLE III
Experiment results of Experiment on Usage Example Detection

Rules	Precision	Recall
1	0.93	0.77
2	0.87	0.56
3	1.0	0.98
1+2	0.83	0.49
2+3	0.85	0.53
1+3	0.93	0.75
1+2+3	0.77	0.41

From the results, the rule that gave the best result was the use of only rule#3. Moreover, all combinations of rules returned lower precision and recall than the single rule used. This can be implied that several conditions from combining rules lead to less matching and were worse than a single effective rule. From analysis, we found that rule#3 can effectively detect a correct detail from all candidates because a usage example (as name implies) should contain the headword among the words in context. However, the missing examples to reduce the recall were those entries with a special Thai word case. The case is that Thai word developed from Pali and Sanskrit language can be combined and some of vowels can be reduced or transformed in the combination. For example, Thai word "กุศล" and "อุบาย" are respectively combined to "กุศโลบาย" to form a word of a more complex meaning. The usage examples that the rule#3 missed were all in this case and the headword cannot be found in an example. The issue however does not occur frequently as there were 12 cases in total so manually handling was acceptable.

TABLE IV
Overall Accuracy results

Types	Accuracy
Main word - Sub word detection	• Detection of main word: 100% • Detection of sub word: 99.36%
Word Relation	• One-way relation: 99.87% • Round-trip relation: 76.11%
Reference and Citation	• Known books: 100% • Unknown books: 33.33%
Etymology	• Original language: 98.24% • Original word: 68.92%
Common Details (domain, local usage, POS, and definition)	• 98.83%

2) Overall Evaluation

By detecting format and symbols used in content, automatic processes were evaluated to find accuracy of information extraction. The accuracy was calculated for information types and provided in Table IV.

From the accuracy results, most of information types were effectively handled. However, the mistakes in most cases were from inconsistent format and patterns or typos. The lowest accuracy was the extraction of entries with unknown books for reference. This case was accurate for 2 correct out of 6 cases since the detection can only get hints from format and pattern but did not match a reference from the list. Another interesting case was the detection of original words for etymology. The cases, which the automatic extraction was unable to solve, are from either multiple languages or multiple original words. These cases however are difficult even for humans who are not a linguistic expert.

V. Conclusion

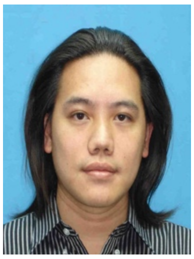
In this paper, a development of electronic version of famous Royin dictionary is presented. By extracting several types of information from the last printing version, various details of Thai lexicon are stored in the database. The database is designed to support lexical information in many aspects, and it is made with the thought of general and interoperable usage by following ontological engineering. The printing version, which is a source of our work, is well formatted from thoughtful consideration regarding users' comprehensibility. Formats and pattern from this semi-structured data are used as a clue to detect types of information in the content. The aim of the detection is to accurately extract several types of information (such as semantic relation, word history and usage). Several difficulties in detection were found and solved by either automatic process or expert decision. The results of extraction were the expansion of lexical entries with keeping semantic relation given in the source and the machine-readable classified information types of the details. By evaluating the methods, we found that the applied processes were reliable and yielded high accuracy, and the errors in extraction were from inconsistency of pattern and format from the source. To make use of the electronic version of the Royin Dictionary, we with the contracted to the Office of the Royal Society plan to include the detailed information into Royin Dictionary mobile application and web-based dictionary service. Incomplete data such as partial pronunciation of the source will be fixed or added for data completion. Current semi-automatic process will be learned for finding knowledge in decision-making to create decision-support framework to reduce experts' burden in solving similar cases.

Acknowledgement

This research has been done under funding and contract from Office of the Royal Society. We would like to thank the Royin Dictionary Committee for providing comments and suggestion in developing the extraction process.

References

1. Office of the Royal Society, Royin Thai Dictionary, 2011. (in Thai)
2. Palingoon P., Chantanapraiwan P., Theerawattanasuk S., Charoenporn T. and Sornlertlumvanich V., Qualitative and Quantitative Approaches in Bilingual Corpus-Based Dictionary, In Proc. Of The 5th Symposium on Natural Language Processing 2002 & Ori-ental COCOSDA Workshop 2002, 2002.
3. LEXiTRON Thai-English Electronic Dictionary, Available online at http://lexitron.nectec.or.th
4. Ruangrajitpakorn T. and Supnithi T., Pali-Thai Dictionary: A semi-automatic approach on form-based to content-based structure, In Proc. Of International Workshop on e-Learning Tools, Techniques and Applications for Cultural Heritage in The 17th International Con-ference on Computers in Education 2009 (ICCE 2009), 2009.
5. Kozaki K., Hayashi Y., Sasajima M., Tarumi S., and Mizoguchi R., Understanding Seman-tic Web applications,In Proc. of the 3rd Asian Semantic Web Conference (ASWC2008), 2008, pp. 524–539.
6. ISO 1951:2007 - Presentation/representation of entries in dictionaries -- Requirements, recommendations and information, Available online at https://www.iso.org/standard/36609.html



Mr. Taneth Ruangrajitpakorn was borned in Bangkok, Thailand. He received his B.A. degree in Religious Study from College of Religious Study, Mahidol University, Thailand. He received M.A. degree in Computational Linguistics from Department of Linguistics, Faculty of Arts, Chulalongkorn University, Thailand. He has been working in Language and Semantic Technology Lab at NECTEC in Thailand since 2003. Currently, he is a Ph.D. student in Computer Science at Faculty of Science and Technology, Thammasat University, Thailand.

Adisak Kingkaewkanthong received his Bachelor of Engineering degree in Computer Engineering from Suranaree University of Technology in 2010. He has been working with Language and Semantic Technology Laboratory, NECTEC, Thailand since 2011.

Thepchai Supnithi received the B.S. degree in Mathematics from Chulalongkorn University in 1992. He received the M.S. and Ph.D. degrees in Engineering from the Osaka University in 1997 and 2001, respectively. He is currently the head of the Language and Semantic Technology Lab at NECTEC in Thailand.