CREATING THE THAI NATIONAL CORPUS

Wirote Aroonmanakun[1]

Abstract

This paper reports on the progress of Thai National Corpus development. The TNC is designed as a general corpus of standard Thai. Only written texts are collected in the first phase. It aims to include at least eighty million words. Various text types produced by various authors are included in the TNC so that it would closely represent written language in general. Texts are word segmented and tagged following the Text Encoding Initiative (TEI) guidelines on text encoding. The TNC was designed as a resource for general applications, such as lexicography, language teaching, and linguistic research. In addition, the TNC is designed to be comparable to the British National Corpus so that a comparative study between the two languages is also possible.

1. Introduction

Based on the definitions of corpora made by many scholars (Sinclair 2005:23, McEnery and Wilson 2001:32, Kennedy 1998:1), we can see corpora as the language data collected according to certain criteria to represent a language under examination. Corpora have been widely used in various fields of study. Lexicographers use corpora to differentiate senses of polysemous words, and to spot coinages. Beside its use for compiling dictionaries, corpora can be used for compiling other language resources like grammar books and text books. Some language teachers even think that the use of corpora would cause a major change in language learning. Corpora are resources for students to explore and discover the authentic uses of languages. The new learning process, known as data-driven learning (Johns 1991), is more student-centered. The similar impact is also hold in the views of scholars in translation studies (Baker 1993). Corpora of translations (parallel corpora) play an important role in revealing the nature of translation. Translated texts are no longer regarded as merely a duplication of the source texts, but they are a kind of language phenomenon worth studying in its own right. In linguistics, corpora are no doubt used as the basis of research, especially for empirical research.

In addition to the use of corpora for theoretical interests, some people use corpora for practical purposes. Translation professions use corpora of translation known as translation memory to facilitate translation tasks. Computational linguists and language engineers use corpora as the training data for their natural language processing (NLP) systems. In general, NLP systems using statistical information of language extracted from a corpus are more robust than systems using manually crafted rules when processing naturally occurring texts. Therefore, it suffices to say that corpora have been receiving much attention in various fields of study.

---

[1] Assistant Professor, Department of Linguistics, Faculty of Arts, Chulalongkorn University, Bangkok, Thailand.

1.1 The need for the Thai National Corpus

Though corpora are useful for many fields of study, there are few public Thai corpora. Most of Thai corpora are in-house products. Many of them are small in size and designed for a specific purpose. Only a few Thai corpora are released to the public. The most recognized one is the Orchid corpus from NECTEC[2]. It is a part-of-speech tagged corpus. Texts are collected from papers published in proceedings. Therefore, it is a specialized corpus for Thai language processing applications rather than for general applications. A more general corpus is found in the Thai Concordance Online service page at the department of Linguistics, Chulalongkorn University[3]. Though the corpora provided are large in size, they do not represent the Thai language in general. Most of the texts are newspaper articles. They are opportunistic rather than balanced corpora. In addition, these corpora can only be searched online, but cannot be released to the public due to the copyright constraints.

On the other hand, in other countries, many large and general corpora have been created. The British National Corpus (BNC) is the first of its kind. It was released in 1994. With the size of 100 million words, in which 10 million words are spoken data and 90 million words are written texts, it is sufficient for studying British English in general. Many research papers, such as Aston (1996), Bernardini (1997), Kilgarriff (1997), Rayson et al. (1997), etc., have used the BNC in their studies. In addition, many commercial English dictionaries have been using the BNC for defining word senses and illustrating word usages. Besides corpora of English, corpora of other languages have also been developed in a similar fashion, such as American National Corpus, Czech National Corpus, Hellenic National Corpus, National Corpus of Irish, Hungarian National Corpus, Slovak National Corpus, and Croatian National Corpus.

In Thailand, although Thai is the national language and everybody has learnt Thai since elementary school, many people sometimes are not certain about using or spelling some words. Resources like Thai dictionaries are not currently as informative as English dictionaries. Creating a large corpus of the Thai language such as the Thai National Corpus will be a significant progress in Thai language resource development. Not only would lexicographers have authentic examples of the language use, but ordinary people would also be able to explore and discover the complexity of the Thai language. It is widely known that King Bhumibol has urged Thai people to realize the importance of the Thai language and to pay more attention to problems of Thai language usages. The day he gave the speech on this matter at Chulalongkorn University (29 July 1962) was later, in 1999, declared as the Thai national language day. The creation of the TNC can serve this objective. The TNC, which represents standard Thai in general, would be an important resource and would be useful in many applications, such as corpus-based lexicography, linguistic research, Thai language teaching, and Thai

---

[2] Orchid Corpus
http://www.links.nectec.or.th/orchid/
[Accessed 2006-10-24]

[3] Thai Concordance Online
http://www.arts.chula.ac.th/~ling/ThaiConc/
[Accessed 2006-10-24]

language processing. Thus, the creation of the TNC is one of many activities that Thai people could do to celebrate the King's 80th birthday in 2007. The project of the TNC is under the patronage of H.R.H Princess Maha Chakri Sirindhorn. The project is led by the department of linguistics, Chulalongkorn University, with cooperation from publishing companies and the IBM Thailand Company.

1.2 Goal of the TNC

As a general corpus, the TNC should include various text types in different domains. The corpus should also include both spoken and written data. However, collecting spoken data requires much work. Thus, the project is aimed at collecting only written texts in its first phase. Eighty million words are the minimal size of the corpus (the number is chosen to celebrate the King's 80th birthday). The TNC is designed to be balance and representative of the standard Thai language. That means the components inside the corpus should correspond to the actual proportions of naturally occurring data, and should include all text types. However, it is unlikely or impossible to know the correct proportion of each text type. What we could do is to estimate and inform users of our decisions. For this project, we decided to make the TNC comparable to the BNC in terms of its domain and medium proportions (see the next section). By doing this, it would be possible to do a comparative study between Thai and British English.

Since Thai texts are written without word boundary markers, the corpus should be at least word segmented. This is necessary to enable precise searching. For example, when searching for the word *มา* 'come', users would not want to retrieve words like *มาก* 'many', *จุดหมาย* 'aim', *สามารถ* 'able', *สมาคม* 'association', etc. Contextual information of the texts, such as author's information, medium, genre, domain, published time etc., should also be stored with each text so that users can search only texts that meet their requirements.

Next, the corpus should be marked with the standard markup language proposed by the TEI (Text Encoding Initiatives). By doing this, it will ensure that the TNC is compatible and conform to the international standard. It will then be easier for foreign scholars who want to use the TNC in their studies. Finally, in order to maximize the use of TNC, software for searching the TNC should also be developed and released together with the corpus. Though the TNC conforms to XML format, it does not mean that any XML searching software can be used. General software may not be able to process Thai language properly. Thai concordance program should be developed especially for sorting and searching Thai words.

2. Design of the TNC

Since the TNC is designed to be comparable with the BNC, the criteria used to select written texts here will be similar to those of the BNC. In addition to the three main criteria used in the BNC, domain, medium, and time, an additional criterion, that of genre, will be used in this project for reasons that will be explained below.

On the dimension of domain, the TNC plans to have 75% of its samples from informative texts and 25% of its samples from imaginative texts. A higher number of informative texts is based on the belief

that general people read or write informative texts, e.g. newspapers, journals, text books, reports, etc., more often than imaginative texts, e.g. novels, short stories, or poems. On the medium dimension, it is planned to have 60% of samples from books, 25% from journal and newspaper, 5-10% from other published works, e.g. brochures, leaflets, 5-10% from unpublished works, e.g. letters, notes, memos, and about 5% from written texts published on the internet. The "internet" medium is added in place of "written to be spoken" medium used in the BNC because many texts are now published on the web. On the dimension of time, since the corpus is a synchronic corpus, texts are selected mainly from those produced in the last ten years, roughly the period 1998-2007. A small portion (less than 10%) is reserved for texts produced more than 10 years ago but less than 20 years ago (1988-1997). An exception can be made for some imaginative texts written before 1988. They can be included in the corpus if it was still reprinted after 1988. The following is the list of domain, medium and time classifications used in the TNC.

In addition to the three main criteria, written texts are also selected and categorized on the basis of their genres. After the BNC was released, Lee (2001) argued that the domain criterion used in he found that some genres are under-represented or missing. This is because texts were not chosen on the basis of genres in the first place. Therefore, in this project, genre classification will be added as one criterion of text selection. Texts will be selected to cover all genres.

2.1 Genre classification

Lee (2001) discussed the overlapping meanings and the confused usages among terms like "genres", "domains", "registers", and "styles". Genre is a complex concept. It is not solely about situated linguistic patterns (register), co-occurrences of linguistic features (text types), subject fields (domain), or text-structures, but includes aspects of all of these things. He suggested that genres should be used as the criterion for text categorization when compiling a corpus.

Table 1 : Weights of Domain, Medium and Time in TNC

Domain		**Medium**	
Imaginative	25%	Book	60%
Informative	75%	Periodical	20%
Applied science		Published miscellanea	5-10%
Arts		Unpublished miscellanea	5-10%
Belief and thought		Internet	5%
Commerce and finance			
Leisure		**Time**	
Natural and pure science		1998-2007 (2541-2550 B.E.)	90-100%
Social science		1988-1997 (2531-2540 B.E.)	0-10%
World affairs		* before 1988 (-2531 B.E.)	0-5%

I prefer to use the term genre to describe groups of texts collected and compiled for corpora or corpus-based studies. Such groups are all more or less conventionally recognisable as text categories, and are associated with typical configurations of power, ideology, and social purposes, which are dynamic/negotiated aspects of situated language use. (Lee 2001:47)

Lee proposed a set of genres to be used with the BNC. His set of genres is the result after studying various genres used in other corpora, e.g. London-Lund Corpus, Lancaster-Oslo/Bergen Corpus, International Corpus of English. In this project, we will follow Lee's idea of categorizing texts into different genres based on external factors like the purpose of communication, participants, and the settings of communication. We also believe that genres are language-specific because the communicative settings could be different in different cultures. We expect that texts in the same genre shares the same characteristics of language usages, e.g. discourse structure, sentence patterns, etc. In this project, the genres adapted from Lee (2001) appear in Table 2.

2.2 Text selection

After corpus structure is designed, we need to collect texts that fit into the designed structure. There are many issues to be considered when selecting texts. Obviously it is easier and cheaper to select texts that are already in electronic forms, like those found in the internet or publisher houses. But there are other factors to be considered as well. Since we cannot use all texts from the same genre, texts should be ranked based on their significance. Texts read by a lot of people, texts produced by famous writers, and texts recognized as valuable work should be more important than the others. Thus, when selecting texts, these factors will be used to decide what should be included in the TNC first.

After texts are selected, copyright owners of each text will be contacted and asked to sign a permission agreement form. To make this process easier, the same form will be used for all owners. Although this is a time consuming process, it is necessary if the TNC will be released to the public later.

During the process of compiling corpus, a report stating the number of words and texts in different genres, domains, and mediums will be generated regularly. This will help people who do the text selection task to clearly see what is still missing or underrepresented in the corpus.

2.3 Sample size

Ideally, the whole text should be kept in the corpus to ensure that information related to discourse structure is not lost when sampling the data. But doing that would force us to include fewer texts, since the time and budget are constant. This leads to sample only some parts of the text. Sampling size can vary, but the maximum size will not exceed 40,000 words or about 80 pages of A4 paper. Texts are randomly selected either from the beginning, the middle, the end, or selected from many sections. If the entire text is less than 40,000 words, only 90% of the text will be used. Using only a part of written text will make it easier to ask for the permission from copyright owners. This method of sampling is similar to the method used in the BNC (see BNC Handbook). But for texts that are less than 40,000 words and copyright-free, such as government documents, the whole text can be stored in the corpus.

Table 2 : Genres in TNC (adapted from Lee, 2001)

Genres	Sub-genres
Academic	*Humanities, e.g. Philosophy, History, Literature, Art, Music*
	Medicine
	Natural Sciences, e.g. Physics, Chemistry, Biology
	Political Science - Law – Education
	Social Sciences, e.g. Psychology, Sociology, Linguistics
	Technology & Engineering, e.g. Computing, Engineering
Non-Academic	*Humanities*
	Medicine
	Natural Sciences
	Political Science - Law – Education
	Social Sciences
	Technology & Engineering
Administration	
Advertisement	
Biography	
Commerce - Finance – Economics	
Religion	*(not philosophy)*
Institutional Documents	
Instructional – DIY	
Law & Regulation	
Essay	*School*
	University
Letter	*Personal*
	Professional
Blog	
Magazine	
Newspaper	*Editorial – views*
	Agriculture news
	Crime news
	Economic news
	Education news
	Entertainment news
	Foreign news
	Local news
	Politics news
	Sciences & Technology news
	Society news

	Sports news
	Royal family news
	Miscellaneous
Fiction	*Drama*
	Poetry
	Prose
	Short Stories
Miscellanea	

3. Encoding the TNC

Encoding is a normal practice when creating a corpus. With encoding, we can add contextual information as well as the analysis of the texts into the corpus. As a result, users can do more than plain text searches. For example, users can search texts in a certain domain or in a particular genre. And with linguistic annotation like part-of-speech tagged, users can search with some syntactic constraints. The standard for encoding electronic texts has been proposed by the TEI (Text Encoding Initiative), in which XML is used to define a markup language. Thus, each element in the text will always be marked with a start tag and an end tag as in these examples:

<name type="person">อานันท์ ปัญญารชุน</name>

<w tran="khlaN0khOO2muun0">คลังข้อมูล</w>

Tags are in the < > symbols. The end tag is differentiated from the start tag by adding / in front of the tag name. Attributes of the element are defined in the start tag in the format of attribute_name="attribute_value". In addition to these notations, the set of tag names and attributes have to be defined so that anyone can use and share the same set of tag names and attributes. This is what we call a markup language. In this TNC project, we use the TEI guideline, "TEI P4", as the markup language. Three types of information are marked in the document: documentation of encoded data, primary data, and linguistic annotation. Documentation of encoded data is the markup used for contextual information about the text. Primary data refers to the basic elements in the text, such as paragraphs, sections, sentences, etc. Linguistic annotation is the markup used for linguistic analysis, such as parts of speech, sentence structures, etc. The first two types are the minimum requirements for marking up texts. The structure of each document is represented in the following tags:

<tncDoc xml:id="DocName">
<tncHeader> ...markup for contextual information </tncHeader>
<text> ...body text, markup for primary data e.g. <p> and linguistic analysis e.g. <w>, <name> .... </text>
</tncDoc>

For linguistic annotation, we mark word boundaries and transcriptions for every word. Information of parts-of-speech will not be marked at present. The following is an example of markup in a document.

<w tran="kha1na1thii2">ขณะที่</w><w tran="khaa2">ค่า</w><w tran="rak3saa4">รักษา</w><w tran="suuan1">ส่วน</w><w tran="k@@n0">เกิน</w><w tran="thii2">ที่</w><w tran="rooN0pha3jaa0baan0">โรงพยาบาล</w><w tran="ton2saN4kat1">ต้นสังกัด</w><w tran="tOON2">ต้อง</w><w tran="caaj1">จ่าย</w>

We recognize that marking tags manually is a difficult and time-consuming task, so for this project, three programs are used for tagging language data and contextual information. TNC Tagger is used for segmenting words and marking basic tags <*w*> and <*p*> in the text. TNC Header is used for inputting contextual information and generating header tag for each text. Output from TNC Tagger will be combined with the header tag as an XML document. TNC Editor is used for editing the XML document. Deleting, inserting and modifying tags can be done easily with this editor. These software packages will control how tags are marked in the document. The following page is an example of the header tag.

4. Inputting Data

Since the TNC will be used for language study, only texts will be kept in a document file. Other information such as tables, graphs, pictures will be omitted. An empty tag <*gap desc="...."/*> will be inserted to describe what is omitted from the original text. Texts are input mainly from typing. Texts which are already in electronic forms like web pages, MS Word files will be converted into plain text files. Texts are marked with basic information like paragraphs <*p*>, names <*name*>, foreign words <*foreign*>, etc. In addition to language data, contextual information of each text file, such as genre, domain, medium, period, author, audience, publication, etc., will be input and saved in the header tag <*header*> through the TNCHeader program.

Beside information in the header and primary data, we also have to decide what linguistic analysis we will impose on the text. In this first phase, only word boundary and pronunciation will be marked in the texts, e.g. <*w tran="thot3lOON0"*>*ทดลอง*</*w*>. Problems on word segmentation and error corrections will be discussed as follows:

<tncHeader type="text" status="new" date.updated="10-9-2549"><fileDesc><titleStmt><title> ใต้ดาวมฤตยู
Sample containing about 27458 words (domain: Imaginative) </title><respStmt><resp>data entry by </resp><name> Dept. of Linguistics
</name></respStmt></titleStmt><editionStmt><para>TNC First Edition</para></editionStmt><extent>Approximately 713 Kbytes running text, containing about 27458 Thai orthographically-defined words; for encoding details see <tagUsage> element. </extent><publicationStmt><distributor>Thai National Corpus Project</distributor><address>
<addrLine>Phayathai Rd., Bangkok 10330</addrLine>
<addrLine>Telephone: +662 2184918</addrLine>
<addrLine>Facsimile: +662 2184918</addrLine>
<addrLine>Internet mail: tnc.ac.th</addrLine></address>
<idno type="tnc">PRNV01</idno><availability status="restricted"><para> THIS TEXT IS AVAILABLE only as part of Thai National Corpus. It is your responsibility, as a user, to ensure that an End User Licence is in place. Terms of the End User Licence are set out in the corpus header, which is likely to have a file name similar to "corphdr". Distribution of any part of the corpus under the terms of the Licence must include a copy of the corpus header. Distribution of this corpus text under the terms of the Licence must include this header embodying this notice. </para></availability><date value="10-9-2549">10-9-2549</date></publicationStmt><sourceDesc><biblStruct><monogr><title> ใต้ดาวมฤตยู </title><author> เสนีย์ เสาวพงศ์</author><imprint><publisher> สำนักพิมพ์มติชน </publisher><pubPlace> กรุงเทพฯ</pubPlace><date vvalue="2545">2545 </date>
</imprint><biblScope>pp. 228-327</biblScope></monogr></biblStruct></sourceDesc></fileDesc><encodingDesc>
<projectDesc><para> See the project description in the corpus header for information about the Thai National Corpus project. </para></projectDesc><tagsDecl>
<tagUsage gi="w" occurs="27507"></tagUsage><tagUsage gi="p" occurs="1031"></tagUsage><tagUsage gi="c" occurs="2328"></tagUsage></tagsDecl></encodingDesc><profileDesc>
<textClass><catRef target="alltim3 wrisam0 wrimed1 wridom1 wriaag0 wriase1 wriad0 wriaty3 wriaud4 writas3"/><classCode scheme="DLeeMod">W_fict_prose </classCode>
<keywords><term>(none)</term></keywords></textClass></profileDesc>
<revisionDesc><change><date>10-9-2549</date><respStmt><resp>ed</resp>
<name>Ling CU</name></respStmt><para>First entering of data</para></change>
</revisionDesc></tncHeader>

The header tag is marked by <tncHeader>. It consists of four tags, <fileDesc>,<encodingDesc>,<profileDesc>, and <revisionDesc>. They are used to encode information of creation and the source text, information of encoding, contextual information, and revision respectively.

4.1 Word segmentation

Word segmentation is the basic problem of processing the Thai language. Since the writing in Thai does not indicate word boundary, the data has to be word segmented either by hand or by a program. Word segmented data will enable users to search for words correctly. For example,

when searching for the word *รก* 'messy', if word boundary is not marked, users might get undesired results like *แทรก* 'insert', *เกษตรกร* 'farmer', *นรก* 'hell', etc. In this project, we use a Thai word segmentation program (Aroonmanakun, 2002) as a tool for marking word boundary. However, some segmentation errors may occur, mostly due to segmenting proper names and transliterated words. For example, a personal name *กมลพรรณ* is segmented as two words <w tran="ka1mon0">*กมล*</w><w tran="phan0">*พรรณ*</w>. A transliterated word *แอดมิชชั่น 'admission'* is incorrectly segmented as two words <w tran="?xxt1">*แอด*</w><w tran="mit3chan2">*มิชชั่น*</w>. These errors will need to be manually corrected. To ease the task of editing tagged data, a special program, TNC Editor, was developed by a staff from the IBM Thailand Co.ltd.

4.2 Error correction

When entering data into the corpus, we seek maximally correct data input. If the data are incorrect, the result of analysis could be affected by the included errors with a certain degree. But in real life, errors could occur while the data are processed. Errors could arise from mistyping, especially when the texts are retyped from a book. Some errors are already in original publications, if the publishers or the writers did not carefully edit the texts. There are also questions concerning what could be counted as an error in the data, and how those errors could be detected. The following is the method used for handling different error types in this project.

1. Typo error type I. This is an unintentional typo that produces an ill-formed string. These errors are easier to detect and most people would agree that they should be corrected. For example, *รถเสีเมื่อเช้า* is ill-formed because a consonant character is missing after *เสี*. This string cannot be parsed and read. It should be edited as *รถเสียเมื่อเช้า* 'car, broken, this morning'.

2. Typo error type II. This is an intentional typo that produces an ill-formed string. Even if the string produced from this type is ill-formed with respect to orthography rules, they are written intentionally to intensify meaning. For example, *ยากกกกก* -'difficult' is a word in which the last consonant is repeated to intensify the degree of difficulty.

3. Hidden error. This is also an unintentional typo error because the actual text should be something else. But the error does not produce an ill-formed string. The string can be parsed and readable. But its meaning could be strange because the actual word is mistaken as another word. For example, the phrase *เครียดลอด 3 ปี* is well-formed because it can be read as four words *เครียด ลอด 3 ปี*, 'stress, go under, 3, year'. But its meaning is anomalous. Thus, it should be changed to *เครียด ตลอด 3 ปี*, 'stress, throughout, 3, year'. This type of error is called "hidden error" because it could not be detected by simply applying orthography rules. To correct this type of error, manual editing might be required.

4. Variation error. This type is not exactly an error. It is a variation of written form produced by different authors. From a prescriptive view, it could be viewed as an error and should be corrected. Some variations are a result of the lack of knowledge in spelling. For example, some

people write the word *โลกาภิวัตน์* 'globalization' incorrectly as *โลกาภิวัฒน์*. Some write the word that does not conform to orthographic rules, e.g. *แซ่ด*, which should be written as *แซด* 'buzzing'. It is possible that they do not know how to spell these words, which makes it an unintentional error (type I). Preserving these errors would provide us authentic information, which will be very useful for studying spelling problems. Nevertheless, since the TNC is expected to be a reference of Thai language usages, keeping these variations could confuse users who want to know the correct or standard form of writing. Therefore, these variations should be corrected. If anyone wants to study spelling problems, they could do it easily by using other resources like the internet to find various forms of spelling.

However, we do not think that all variations of writing are errors. Variations caused by different transliteration methods should be kept as they are. When transliterating foreign words, it is likely that they are written differently despite the fact that a guideline for transliteration to Thai has been proposed by the Royal Institute. For example, the word *internet* is found written as *"อินเตอร์เน็ต"*, *"อินเตอร์เนต"*, *"อินเตอร์เนท"*, *"อินเตอร์เน็ท"*, *"อินเทอร์เน็ต"*, *"อินเทอร์เนต"*, or *"อินเทอร์เน็ท"*. All of these variations are not seen as errors and therefore are not modified.

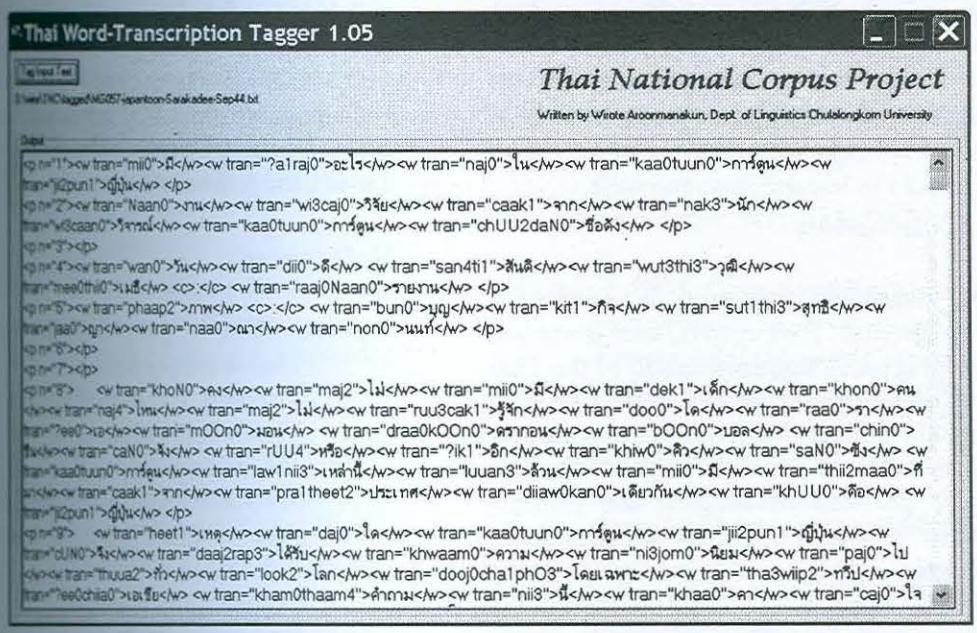
5. Segmentation errors. These are errors caused by the segmentation program. It is likely that the program would segment proper names incorrectly. For example, the name *น.พ.วินัย สวัสดิวร* is segmented as <w tran="nOO0phOO0">น.พ.</w><w tran="wi3naj0">วินัย</w> <w tran="sa1wat1">สวัส</w><w tran="di1">ดิ</w><w tran="wOOn0">วร</w>, instead of <w tran="nOO0phOO0">น.พ.</w><w tran="wi3naj0">วิ นั ย</w><w tran="sa1wat1di1wOOn0">สวัสดิวร</w>. These errors have to be manually corrected.

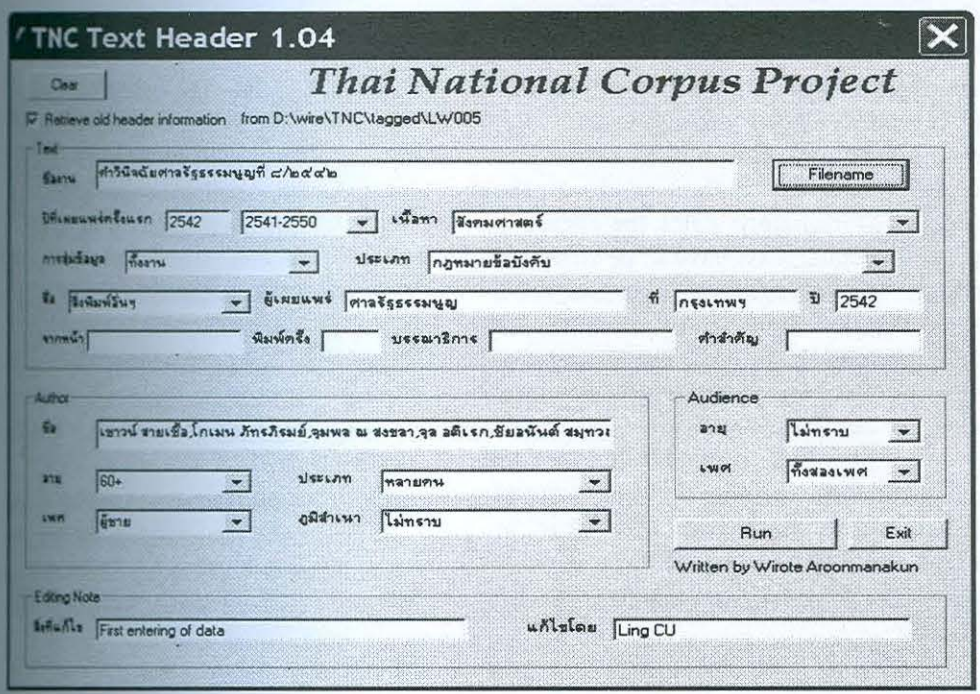
To correct errors caused by typos, we could compare the same text typed by two typists. But this method would double the expense of typing. Therefore, we seek to detect typos indirectly by using the TNC Tagger program. Basically, the program will segment words in the text. If a typo causes an ill-formed character sequence, the program will fail to segment that character sequence. Then, a pop-up screen asking for a correction of that string sequence will appear. If it is typo type I, the correct word will be typed in. If it is typo type II, the intentionally incorrect word will be tagged manually. After the program finishes segmenting words, the program will show a list of unknown words (words that are not found in the dictionary) and words that occur only once in the file. This word list will be used for spotting errors that are not typos. At this stage, TNC Editor will be used for manually editing the document, especially the hidden, variation, and segmentation errors.

4.3 Software for creating TNC

A number of software programs have been written to create the TNC. As described earlier, TNC Tagger is the program used for segmenting words and transcribing pronunciation of each word. Texts will be marked with <w> and <p> tags by the TNC Tagger. Then, TNC Header program

will be used to input contextual information. This information will be stored in the header tag *<header>* as declared by the TEI. Header and text then are combined into an XML document. This XML document will be checked for any hidden errors and edited by using the TNC Editor, which is an XML editor designed specifically for the project. Screen shots of these programs are shown below.





Besides the programs used for inputting and editing the document, other programs have been developed for reporting the current status of the project. They will show the proportion of texts in different dimensions so that we could know which genre, domain, medium, author sex, author age, etc., are underrepresented at the moment. Then, we can decide what texts should be acquired next.

5. Conclusion

This paper has discussed the need of a large general Thai corpus, and described the design and implementation of the Thai National Corpus, which is a new and ongoing project. The TNC is expected to be partially released in July 2007. The major problem of creating the national corpus is neither technical nor linguistic, but practical. Since the concept of using corpora is not well-known to the public, obtaining copyright permission from authors and publishers requires explaining what the corpora are and what we want from them. This process takes time and patience. Nevertheless, many writers, after they understood our need, have been willing to give all of their works to the project. Unfortunately, since the corpus is a general corpus, texts have to be collected from as many writers as possible. This means that we could only include 3-5 works from each author. The process of collecting texts still has a long way to go. Once that has been done, all tagged texts have to be manually checked. It is undeniable that creating a large general corpus requires much work. However, in the end, the TNC will prove to be a major resource for Thai language studies.

Acknowledgments—The TNC project is under the patronage of H.R.H Princess Maha Chakri Sirindhorn. It is led by Assoc. Prof. Kingkarn Thepkanjana, the head of the linguistics department, with collaboration from many researchers, such as Assist. Prof. Prima Mallikamas, Assist. Prof.Manop Wongsaisuwan, Mr. Wichai Patipaporn, Mr. Vichai Watanathawornwong (IBM Thailand Co.ltd.), and Mr. Kachain Tansiri, the project manager.

References

Aroonmanakun, W. 2002. Collocation and Thai Word Segmentation. In Thanaruk Theeramunkong and Virach Sornlertlamvanich (eds.) *Proceedings of the Fifth Symposium on Natural Language Processing & The Fifth Oriental COCOSDA Workshop.* Pathumthani: Sirindhorn International Institute of Technology. 68-75.

Aston, G. 1996. The British National Corpus as a language learner resource. Paper presented at *TALC 96*. Available online from http://www.natcorp.ox.ac.uk/archive/papers/aston96a.htm [Accessed 2006-10-24]

Aston, G. and L. Burnard. 1998. *The BNC Handbook: Exploring the British National Corpus with SARA.*

Baker, M. 1993. Corpus Linguistics and translation Studies: Implications and Applications. In Mona Baker, Gill Francis and Elena Tognini

Bonelli (eds.) *Text and Technology in honor of John Sinclair*. Amsterdam/Philadelphia: John Benjamins, 223-250.

Bernardini, Silvia 1997. A 'trainee' translator's perspective on corpora. Paper presented at the first international conference on *Corpus Use and Learning to translate*, Bertinoro, 14-15 November 1997.

McEnery, T. and A. Wilson. 2001. *Corpus Linguistics*. 2nd Edition. Edinburgh : Edinburgh University Press.

Johns, T. 1991. Should you be persuaded - two samples of data-driven learning materials?. In T. Johns and P. King (eds.). *ELR Journal: Classroom Concordancing* 4: 1-16. Centre for English Language Studies, The University of Birmingham.

Kennedy, G. 1998. *An Introduction to Corpus Linguistics*. London: Longman.

Kilgarriff, A. 1997. Putting frequencies into the dictionary. *International Journal of Lexicography*, 1997, 10: 135-155.

Lee, D. 2001. Genres, registers, text types, domains and styles: clarifying the concepts and navigating a path through the BNC jungle. *Language Learning & Technology*, 5(3): 37-72.

Orchid Corpus http://www.links.nectec.or.th/orchid/[Accessed [2006-10-24]

Rayson, P., Leech, G., and Hodges, M. 1997. Social differentiation in the use of English vocabulary: some analyses of the conversational component of the British National Corpus. *International Journal of Corpus Linguistics*. 2(1), 133 - 152.

Sinclair, J. 2005. Corpus and Text - Basic Principles. In M. Wynne. (ed.) *Developing Linguistic Corpora: a Guide to Good Practice*. Oxford : Oxbow Books, 1-16. Available online from http://ahds.ac.uk/linguistic-corpora/ [Accessed 2006-09-14].

Thai Concordance Online. http://www.arts.chula.ac.th/~ling/ThaiConc/ [Accessed 2006-10-24]

The TEI guidelines. http://www.tei-c.org/Guidelines2/ [Accessed 2006-09-14].