Thai National Corpus: A Progress Report

Wirote Aroonmanakun
Department of Linguistics
Chulalongkorn University
awirote@chula.ac.th

Kachen Tansiri
Thai National Corpus Project
Chulalongkorn University
kc.tansiri@gmail.com

Pairit Nittayanuparp
Thai National Corpus Project
Chulalongkorn University
cherngx@gmail.com

Abstract

This paper presents problems and solutions in developing Thai National Corpus (TNC). TNC is designed to be a comparable corpus of British National Corpus. The project aims to collect eighty million words. Since 2006, the project can now collect only fourteen million words. The data is accessible from the TNC Web. Delay in creating the TNC is mainly caused from obtaining authorization of copyright texts. Methods used for collecting data and the results are discussed. Errors during the process of encoding data and how to handle these errors will be described.

1 Thai National Corpus

Thai National Corpus (TNC) is a general corpus of the standard Thai language (Aroonmanakun, 2007). It is designed to be comparable to the British National Corpus (Aston and Burnard, 1998) in terms of its domain and medium proportions. However, only written texts are collected in the TNC, and the corpus size is targeted at eighty million words. In addition to domain and medium criteria, texts are also selected and categorized on the basis of their genres. We adopted Lee's idea of categorizing texts into different genres based on external factors like the purpose of communication, participants, and the settings of communication (Lee 2001). Texts in the same genre share the same characteristics of language usages, e.g. discourse structure, sentence patterns, etc. Moreover, since TNC is a representative of the standard Thai language at present, 90% of the texts will be texts produced before 1998. The rest 10% can be texts produced before 1998 if they are published recently. Therefore, the structure of TNC is shaped on the dimensions of domain, medium, genres and time (see Table 1). Texts that fit into the designed portion of these criteria will be selected. After that, copyright holders of each text will be contacted and asked to sign a permission form. To make this process easier, the same form is used for all copyright holders. When authorization is granted, texts are randomly selected either from the beginning, the middle, the end, or selected from many sections. Sampling size can vary, but the maximum size will not exceed 40,000 words or about 80 pages of A4 paper.

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="kot1maaj4">กฎหมาย</w><w tran="thaN3">ทั้ง</w> <w>3</w> <w tran="cha1bap1">ฉบับ</w><w tran="mii0">มี

</w><w tran="lak3sa1na1">ลักษณะ</w><w tran="mUUan4">เหมือน</w><w tran="kan0">กัน</w><w tran="juu1">อยู่</w><w tran="jaaN1">อย่าง</w><w tran="nUN1">หนึ่ง</w>

We recognize that marking tags manually is a difficult and a time-consuming task, so for this project, two 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. Word segmentation and transcription program proposed in Aroonmanakun and Rivepiboon (2004) is used as a tagger. 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.

2 Data collection

This section explains methods of data collection and the outcomes. First, we thought that texts could be collected easily from publishers. So, we first wrote a letter to three major publishers asking for collaboration. We thought that they would be able to provide us lot of texts in electronic formats. So, we asked them to give us a list of their publications and mark for us items that they have in electronic forms. It turned out that they did not even understand much about the corpus and the purpose of collecting texts. Thus, we did not receive positive responses as expected. Only one publisher was able to give us the list of their publications. The rest asked us to be more specific about the texts we want. The fault is ours because we did not make clear what texts that we want and how their rights on the texts will be protected. Thus, corresponding with the publishers did not go smoothly and quickly as it should be. We also learned that the publishers are not the owners of all the texts. It depends on the agreement signed between the authors and the publishers. Normally, the author is the copyright holder. Publishers may hold the copyright for a certain period agreed by both parties.

Later, before we wrote to a publisher asking for their helps, we searched and listed the title and the number of pages that we want from each text. Project details and samples of concordance output were enclosed to give them a better understanding of the project. And we only asked the publishers to collaborate by providing us the contact address of the copyright holder of each text. This time we received a positive response from many publishers. From twenty two publishers we contacted, only one publisher officially refused to collaborate for their own reasons. Fourteen publishers did not response. Seven of them sent us the information we requested. After we received the contact addresses from the publishers, we then wrote a letter directly to the author. A permission form in which selected publications are listed was attached in the letter. We asked them to sign a permission form and return it in the enclosed envelope. To make them feel easier to support us, we informed them that they may remove their works from the TNC anytime by writing a letter informing us to do so. We did not even ask for a copy of the book or the article. We will look for those texts and typing them in ourselves. By doing this, we did not put a burden on the copyright owners. In addition, we contacted the P.E.N International-Thailand Centre, which is the association of publishers, editors, and novelists in Thailand, asking for contact addresses of novelists. For academic writers, we searched for their contact addresses from university websites. Of those 780 authors we had contacted, 250 of them granted us the permission to use their texts. We suspected that the address list we received from the P.E.N International-Thailand Centre may be out-of-date because we received only 41 replies from 278 requests to novelists.

For texts that are not copyrighted in Thai, e.g. news reports, documents from governments, laws and orders etc., they are collected preferably from those that are available in the internet.

After texts were saved in electronic format and catalogued in the database, they were parsed by the TNC Tagger program. Texts will be word segmented and marked basic tags as described in the previous section. The process is not fully automatic. The program will ask a user to make a correction if any chunk of texts could not be parsed. This usually happened because there was a spelling error within that text chunk. After the text is parsed, contextual information of the text will be inserted by using the TNC Header program. With these two programs, texts are converted into an XML format that conforms to the TEI P4 standard. Some problems occurred during this process will be discuss in section 4.

3 TNC web

It is now clear that collecting eighty million words is a long time process. At present, only fourteen million words are processed in the TNC. Nevertheless, it is a good idea to make the corpus

accessible to the public. So, we had been developing a web interface to search the TNC, or the TNC web[1].

TNC web is a web interface for concordance software that will show not only keyword-in-context but also collocations and distributions of the keyword. When users enter a keyword, the distribution of keyword in five major genres will be shown on the right window. Users can click on the frequency of occurrence in any genre on this window. A concordance window will then be displayed underneath. Users can filter the search by specifying a genre, a domain, published year, authors' age range, and authors' gender. By doing this, users can search for the occurrence of the keyword in any specific context. Figure 1 shows the screen of concordance search from TNC web.

Collocation is searched by clicking on the icon "COLLOCATE". Collocations within 1-3 words on the left and right contexts will be ranked by statistical measure. Frequency of occurrence in five major genres will also be shown. Users can click on these numbers to see the concordance context. Figures 2 and 3 shows the collocation of the keyword วิ่ง – 'run' using log-likelihood and mutual information .

To make the processing time acceptable, the XML data was converted into MySQL database and PHP scripting language was used for web development. Co-occurrences of words are also stored as precache data. By doing this, the size of the data storage gets larger. The XML data of 14 million words, which is about 365 megabytes, is expanded to 2,064 megabytes on the server.

Though at present, the TNC is not balance and does not have a proportion of texts as planned, making it searchable through the web is still a useful idea. Users can get authentic data in various genres. And it would be easier for us to explain to the public what the TNC is and how it can be used.

4 Problems

The difficulties of creating the TNC are grounded on management rather than technical problems. The most difficult part is to get copyright texts. Unexpected errors during the process of creating an annotation text are also another problem causing a delay in creating the TNC.

4.1 Getting more texts

Though the use of corpora is quite well known to academics, it is little known to the public at large. Without understanding from the people especially writers and publishers, it is not easy to get the support and collaboration from them. This is the main obstruction causing a delay in creating the TNC. Implementing TNC web is one method of getting TNC known to the public. Another strategy that we plan to do is to publicize the project and praise those who contributed their texts to the project. At this moment, a number of famous novelists had granted us the permission to include parts of their novels in the TNC. We could use these names to make other people feel that it is a privilege to have their texts as a part of TNC.

Another strategy of promoting TNC is to show its worth. We plan to publish a series of linguistic papers that use TNC as data of analysis, and demonstrate how basic information like word frequency and collocations in different genres can be used for teaching the Thai language.

4.2 Validating data

The delay in creating the TNC is also caused during the process of encoding data. As stated earlier in section 2, texts have to be parsed and encoded as XML data. During this process, different types of errors are found. These have to be handled to make the data correct and consistent.

System errors (unintentional): 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'.

System errors (intentional): 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.

Hidden errors: 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 รถตากลางถนน is well-

[1] http://www.arts.chula.ac.th/~ling/ tnc2/

formed because it can be read as four words รถ ตา กลาง ถนน, 'car, grandfather, middle, street'. But its meaning is anomalous. Thus, it should be changed to รถ ตาย กลาง ถนน, 'car, broken, middle, street' - 'the car was broken in the middle of the street. 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.

Variation of writing: 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. 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 and removed from the TNC. However, these variations will be saved in an error log file for further use of spelling problems.[2]

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.

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 นายวันชัย กู้ประเสริฐ is segmented as <w tran="naaj0">นาย</w><w tran="wan0">วัน</w><w tran="chaj0">ชัย</w><w tran="kuu2">กู้</w><w tran="pra1s@@t1">ประเสริฐ</w>, instead of <w tran="naaj0">นาย</w><w tran="wan0chaj0">วันชัย</w> <w tran="kuu2pra1s@@t1">กู้ประเสริฐ</w>. A Thai named entity recognition module is needed to handle this problem. But before the module is included in the TNC tagger, 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 an unintentional system error, the correct word will be typed in. If it is an intentional system error, the intentionally incorrect word will be tagged manually. After the program finishes segmenting words, the program will create 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 by the TNC Editor program for spotting errors that are not typos. TNC Editor will be used for manually editing the document, especially the hidden, variation, and segmentation errors.

4.3 Obtaining authorization

Acquiring permission from the copyright holders is a time consuming process. We once thought of a way to use copyright text under a condition of "fair use" stated in the copyright protection act in Thailand. According to the act, any writing is automatically protected by the law throughout the life of the creator plus fifty years after the author dies. However, some works are not copyrighted, such as news reports which are facts rather than opinions; constitution and laws; rules, regulation, reports or documents issued by government organizations, etc.

On section 32 of the copyright protection act, certain uses of copyright materials are not considered a violation of copyright law, such as making a copy of text for research purpose without making a profit, making a copy for private use, for criticism with an acknowledgement of the writer, for teaching or educational purpose without making a profit, etc. But all these activities must not affect the benefits that the copyright holders should have received from their works.

[2] Thanks to Dr. Virach Sornlertlamvanich for making this suggestion.

In addition, on section 33, it is stated that a reasonable and acceptable part of a copyright work can be copied or cited if the copyright owner is acknowledged. Therefore, we had consulted an eminent law firm whether our project can make use of these exceptions of the Thai copyright law. Is it possible to argue that the texts we collected are used for educational/research purpose and no profit is generated from the TNC? In addition, users can see the bibliographic reference of each concordance line. Thus, is it possible to conclude that our uses of copyright texts are under the condition of "fair use"? However, the lawyers thought that we cannot use those argumentations since the text size we collected could be up to 40,000 words. Although the reference to the source text is shown to the users, the text length is greater than acceptable level. The TNC project is the project for creating a new database. Texts collected in this project are not used for criticism or for the study of those texts per se. Our activity in collecting copyright texts could affect the benefits the copyright holder should have. Thus, the creation of a corpus is not under the conditions of sections 32 and 33. At the end, the law firm advised us to continue asking for authorization from the copyright holder as we have been doing.

5 Future plan

We plan to run three tasks concurrently: cleaning up data, expanding data, and utilizing the corpus. For cleaning up data, Thai named entity recognition module will be implemented to reduce errors of word segmentation. But at the end, TNC Editor is needed to clean up segmented data manually. The program is now under development by IBM Thailand Co.,Ltd. For expanding data, more publishers and writers are being contacted. Copyright texts are now constantly being added into the corpus. But to increase the growth rate of the corpus size, we would prefer to have people submitting their works themselves. We hope that by making the corpus searchable online and revealing famous writers who had contributed their works will make people feel that it is the prestige to have their works included in the corpus. It remains to be seen whether our plan to publicize the TNC project will be successful. And finally, to increase the worth of TNC, we will encourage linguists to use TNC as the basis of Thai language studies. Basic facts like word lists in different genres will be released. We also hope that new Thai language resources like dictionaries and grammar books could be produced based on the actual usages found in the TNC.

6 Conclusion

In this paper we described the current status of the TNC project and the problems causing the delay of collecting data. The future work will still be focused on collecting more texts, both copyright and non-copyright material. We hope to fill the TNC with texts according to the designed proportion in the dimensions of domain, medium, and genres. We hope that our publicizing plan, making the TNC known to the public and praising those who contributed their texts, would easy the process of text collection.

Given that there are a huge number of texts available on the internet, it would be easier to collect texts from the internet without going through the process of obtaining authorization from the copyright holders. In fact, many corpora have been collected directly from the web (Baroni and Ueyama, 2006; Fletcher, 2007), or the web itself has been used as a corpus (Killgarriff and Grefenstettey, 2003). It might be true that natural language processing research can use web as data source for their works effectively. Nevertheless, we think that by getting authorization from text owners, we could fully distribute the source data. And this is necessary for linguistic analysis. In addition, by manually selecting and categorizing data to be included in the corpus, users can look for similarity and difference between different text settings. Therefore, we believe that the creation of TNC will still be fruitful for research especially on Thai linguistic analysis.

Acknowledgments

The TNC project under the Royal Patronage of HRH Princess Maha Chakri Sirindhorn is a project of the Linguistic Department, Faculty of Arts, Chulalongkorn University with the collaboration of several researchers and publishers in Thailand. The project remains highly indebted to Mr. Domnern Garden for his invaluable contributions in support of this project which lasted until the final day of his life.

References

Aroonmanakun, W. 2007. Creating the Thai National Corpus. *Manusaya.* Special Issue No.13, 4-17.

Aroonmanakun, W., and W. Rivepiboon. 2004. A Unified Model of Thai Word Segmentation and Romanization. In *Proceedings of The 18th Pacific*

Asia Conference on Language, Information and Computation, Dec 8-10, 2004, Tokyo, Japan. 205-214.

Aston, G. and L. Burnard. 1998. *The BNC handbook: exploring the British National Corpus with SARA*. Edinburgh: Edinburgh University Press.

Baroni, M. and M. Ueyama. 2006. Building general- and special-purpose corpora byWeb crawling. In *Proceedings 13th NIJL International Symposium, Language Corpora: Their Compilation and Application*, Tokyo, Japan, 31-40.

Fletcher, William H. 2007. Implementing a BNC-Compare-able Web Corpus. In *Proceedings of the 3rd web as corpus workshop, incorporating cleaneval*, Louvain-la-Neuve, Belgium, 15-16 September 2007, 43-56.

Killgarriff, A, and G. Grefenstettey. 2003. Web as Corpus. In *Computational Linguistics* 9(3): 333-347.

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.

TEI guidelines. http://www.tei-c.org/Guidelines/ [Accessed 2009-04-24].

TNC web. http://www.arts.chula.ac.th/~ling/TNC/ [Accessed 2009-04-24].

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-present (2541-2550)*	90-100%
Social science		*1988-1997 (2531-2540)*	0-10%
World affairs		** before 1988 (-2531)*	0-5%

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
Advertisement	
Biography - Experiences	
Commerce - Finance – Economics	
Religion	
Institutional Documents	
Instructional – DIY	
Law & Regulation	
Essay	*School*
	University
Letter	*Personal*
	Professional
Blog	
Magazine	
News report	
Editorial - Opinion	
Interview – Question & Answer	
Prepared speech	
Fiction	*Drama*
	Poetry
	Prose
	Short Stories
Miscellanea	

Table 1: Design of Thai National Corpus

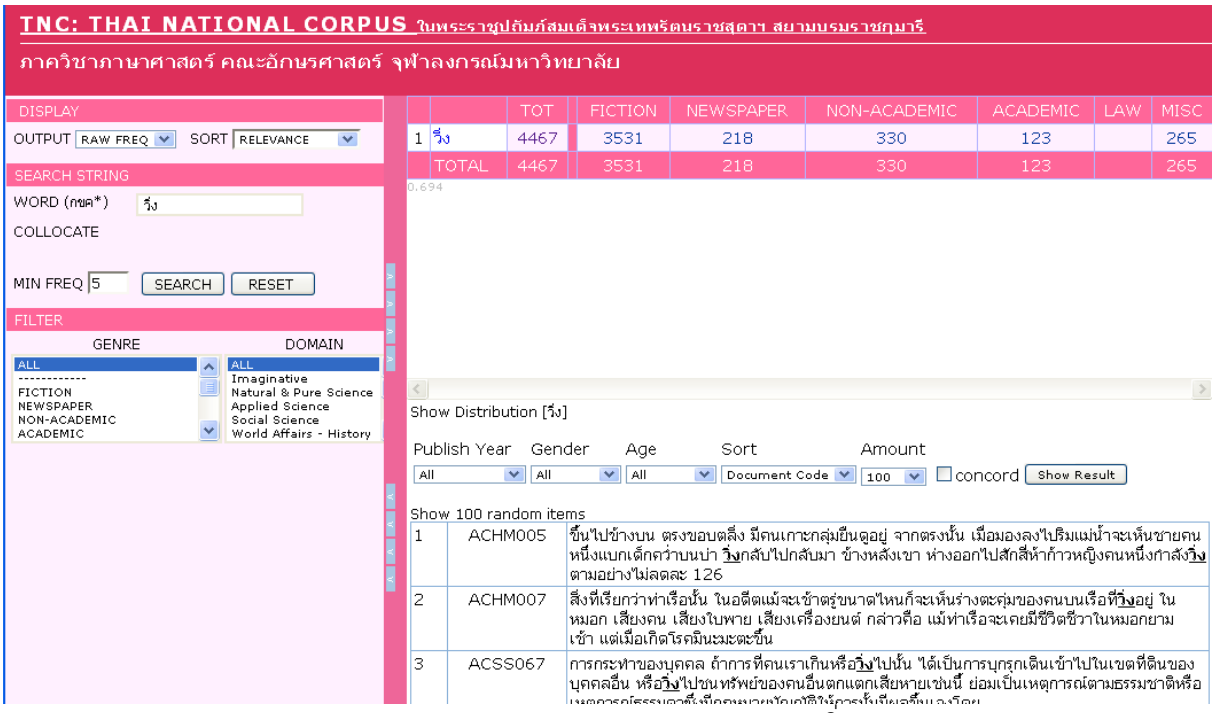

Figure 1: Concordance search result of the word วิ่ง ‘run’

		TOT	FICTION	NEWSPAPER	NON-ACADEMIC	ACADEMIC	LAW	MISC	ALL	%	DL
1	ไป	1183	1009	25	72	17		60	159851	0.74	5473.54
2	หนี	326	222	43	27	9		25	3557	9.17	3087.76
3	มา	712	588	26	53	11		34	142992	0.50	2672.56
4	เข้า	470	410	19	24	2		15	39011	1.20	2541.20
5	รีบ	267	234	9	20			4	6075	4.40	2121.40
6	ออก	390	342	16	16	4		12	40169	0.97	1936.89
7	ตาม	370	303	9	24	4		30	53701	0.69	1590.33
8	ก็	419	363	3	29	4		20	128433	0.33	1217.25

Figure 2: Collocation of the word วิ่ง ‘run’ using Dunning’s Log-likelihood

		TOT	FICTION	NEWSPAPER	NON-ACADEMIC	ACADEMIC	LAW	MISC	ALL	%	MI
1	จู๋จี๋	15	12		2			1	17	88.24	9.42
2	ตื๋อ	17	12		5				20	85.00	9.36
3	แจ๋น	27	23	1	2			1	47	57.45	8.80
4	ปรู๊ด	14	14						27	51.85	8.65
5	กระหืดกระหอบ	22	22						46	47.83	8.53
6	เหยาะ	30	28					2	64	46.88	8.50
7	เริด	5	5						15	33.33	8.01
8	กวด	10	6			1		3	38	26.32	7.67

Figure 3: Collocation of the word วิ่ง ‘run’ using Mutual Information