Journal of Southeast Asian Language Teaching

A refereed journal of the Council of Teachers of Southeast Asian Languages

Journal of Southeast Asian Language Teaching
Volume 12, Spring 2006

Computer Tools for Translating Southeast Asian Language: Illustrations from Thai

John Hartmann
Northern Illinois University

Abstract

(Please click on the links below)

The often messy and frustrating task of learning the art of translation can be managed and monitored by using several computer tools: 1) a concordance; 2) a word frequency count; 3) an interlinear translation program; 4) an on-line dictionary look-up, and 5) a words-not-found program. For students who are new to the task---and for the teacher who wants to monitor their work---the use of these tools can be employed to meet the needs of both. This article illustrates how the computer tools work with authentic texts as input. In Part A, the text to be translated by the student is a short news article on foreign affairs using several separate tools developed and tested in 2005 at Northern Illinois University. In Part B, a new tool, which we have dubbed "Paragraph Translator 2006," merges three tools into one: 1) interlinear translation; 2) a modified key-word-in-context concordance, and 3) word frequency counts in both numerical high-to-low count and alphabetic order. The single tool that has proven to be most useful to the learner and teacher is the word frequency count. It is used as a pre- and post-test to determine what vocabulary the student translator controls or does not control. Research on translation shows that if the students do not control 95% of the vocabulary in the target text, the task of translation will overwhelm them, and adjustments will be necessitated. Because the authentic Thai texts do not have words segmented, the task of identifying word boundaries and separating words with a text editor must be carried out in advance of employing any of the computer tools. The segmentation process itself forces the students to develop notions of what a word is in Thai and improves their reading and comprehension skills at the same time. The word frequency lists can be copied into a spreadsheet by the teacher or student and used by the latter

to construct a personal dictionary that can be expanded as more and more texts with newer vocabulary are added over time. The acquisition of new vocabulary, can thus be managed, monitored, and maintained in a computer file and shared with others as well. The most frequent error that novice translators make is in defining units (words, compounds, idioms) at the "word" level. That is why the interlinear, word-by-word "Paragraph Translator" is a particularly useful tool for the instructor, who can more readily point to the source of error that ends up in a free, finished, but flawed translation.

Introduction

For the past two years (2005-2006), I have been developing a new course on the use of translation tools for advanced students of Thai at Northern Illinois University. This article is an illustration of how the tools---by-and-large computer programs---work. Hence, the look and feel of this article will be less of a traditional "narrative discourse" than one usually expects in a traditional print journal format. Moreover, because it is a "how to" discussion, amplified with pictorial illustrations of the various computer programs, it should be read in the mode of a "procedural discourse" amplified by graphics. In some ways, it has the look of a Power Point Presentation, or an interactive New York Times on-line article, with clear borders between "slides" rather than print paragraph divisions. The border were added to clarify the steps in the "how to" procedures involved. Furthermore, because this is a on-line publication in an e-journal, the reader will have the luxury of actually going out into cyberspace by clicking on links that serve the "how to" function of the article. It should be noted that these tools can be used for languages other than Thai, as long as the fonts are UNICODE.

The "tools" were tested in a class of four on-campus students and one off-campus student who communicated by e-mail during both the fall and spring semesters of 2005-2006. One of the outcomes learned from using the tools was that they require close guidance in their implementation by the instructor, who, in turn, needed to possess some computer savvy. At the beginning of this project, all students worked on the same text, prepared in advance by the instructor. By the end of the project (second half of second semester), each student was working on her own text, in her own area of graduate level specialization.

Part A 2005

How to Use Thai-English Translation Tools

Illustrations Using News Articles for Advanced Thai Reading

The task of learning to translate can be managed by using several computer tools listed below.

Translation Tools

1) Concordance: a list of all the words, single word, partial word, or regular expression in a document as they appear in context.

2) Word Frequency: a list from high frequency to low; or in alphabetic order. The list can then be copied into an Excel Spreadsheet for adding definitions and used as a personal dictionary or for language proficiency testing and/or vocabulary building.

3) Interlinear Translation: a program for moving between word-by-word translations and a smooth translation.

4) On-line Thai Dictionary: a cut-and-paste dictionary lookup tool.

5) A Words-Not-Found program that lists the words in the source text that have no definitions in the on-line dictionary. The WNF list can then be used to update the on-line dictionary.

Experimental SourceText (ST) Data

Five short news articles on Foreign Affairs from Prawet Jantharat's web page of advanced Thai reading exercises were used as source text (ST) for utilizing the computer tools that will be illustrated in the following slides.

Please open the web page at:
http://siamwestdc.com/thairead/index.htm

Home Page of News Readings

Module 01 was used as "Source Text" for application of computer tools.


Click to see larger view

Preparation of a Concordance and Word Frequency List

Because the five readings of Module 1 selected as source text to be copied into the concordance and

word frequency programs all involve foreign affairs, they were combined into a single text, with no separation between words.

Thai, like other Indic-derived writing systems, does not employ spaces between words.

A sample of the original source text appears below. Note the absence of separation between words. Separations do appear between clauses, as seen in line 3.

Sample of Source Text – Without Separations Between Words

จีนยิ้มโสมแดงยุติโครงการนุก
มังกรจีนออกโรงแสดงความยินดีรัฐบาลโสมแดงเกาหลีเหนือยอมยุติโครงการพัฒนา
อาวุธนิวเคลียร์ชี้จะทำให้ช่องว่างของความขัดแย้งเล็กแคบลง และจะช่วยให้การ
เจรจายุติวิกฤตินิวเคลียร์บนคาบสมุทรเกาหลีครั้งต่อไปมีความคืบหน้ามากกว่าเดิม

Step 1. Separation of ‘‘Words” and Larger Units of Meaning

Because the concordance and word frequency programs require text input as individual “words,” spaces were inserted by hand (space bar) between each Thai word. The decision as to what constitutes a “word” in Thai is not always clear. Our overall standard was that a word is whatever is an entry in a Thai dictionary; but that didn’t always apply because a “word” can also be a semantic doublet or a multi-word expression that constitutes a unit of meaning, e.g., an idiomatic expression.

Separation of high frequency items such as, e.g., ความ, การ, can be sped up by using the “Find” tool in word processor (MS Word, WordPad, Notepad, etc.)

A portion of the source text with word divisions now appears as shown in the following.
Caution: Please save text as plain Unicode text. How to save, visit:
http://www.seasite.niu.edu/trans/thai/howto/wordprocessor.htm

จีน ยิ้ม โสมแดง ยุติ โครงการนุก มังกร จีน ออก โรง แสดง ความยินดี รัฐบาล โสม แดง เกาหลี เหนือ ยอม ยุติ โครงการ พัฒนา อาวุธ นิวเคลียร์ ชี้ จะ ทำให้ ช่อง ว่าง ของ ความ ขัดแย้ง เล็ก แคบ ลง และ จะ ช่วย ให้ การ เจรจา ยุติ วิกฤติ นิวเคลียร์ บน คาบ สมุทร เกาหลี ครั้ง ต่อไป มี ความ คืบหน้า มาก กว่า เดิม

Step 2: Cut and Paste Segmented Text Into Concordance Program Window. N.B. Any non-Thai characters, e.g. numerals, punctuation, etc, must be "found" and deleted from the input text first.

The segmented source text now becomes the input for the concordance program.

Click to see larger view

* The first step: With your cursor, highlight and copy the segmented source text from the word processor and paste it into the textbox window of the concordance program. Click on “Paste text to use”.

* The second step: We chose the default “Display all Words”.

* The third step: We chose “Whole word match” to display whole word in context of the pasted text.

* The fourth step: We chose “Context Size” to be 40. And click the “Submit” button. If everything goes well, the output concordance will appear with list of words in context of the pasted text. Visit the following link to view example:
www.seasite.niu.edu/trans/thai/howto/concordanceview.htm

Note: If we want to look for a single word in context of the pasted text, we have to choose “Enter a single word for display” in the second step and enter the word we want to find in the textbox immediately below it. The third and fourth step should be the same as above. View example:
www.seasite.niu.edu/trans/thai/howto/concordanceviewsingleword.htm

---

The Concordance

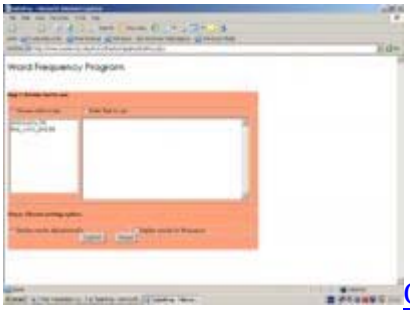
The output of the program, the concordance itself, appears at the following web address:
www.seasite.niu.edu/trans/thai/howto/concordanceoutput.htm.

The % number indicates where approximately in the text the word appears: e.g., 3% would be towards the beginning; 86% towards the end. (A more powerful (and complex) concordance could, with line numbering in original format, would give the exact line number reference in the original text.)

Word Frequency Program

Step 3: Cut and Paste Segmented Text into the Word Frequency Program Window

Cut and Paste Same Segmented Text as Input; Choose to sort alphabetically or numerically (Step 2).


Click to see large view

The output of the program, the word frequency itself, appears at the following web address: www.seasite.niu.edu/trans/thai/howto/wordfrequencyresult.htm.

Word Frequency Result in Excel Spreadsheet

Step 4: Cut and Paste the Word Frequency List into an Excel Spreadsheet.

Word Frequency List copied into an Excel Spreadsheet to be used as student Dictionary. The example appears at the following web address: www.seasite.niu.edu/trans/thai/howto/wordfrequencyinexcel.htm.

Utility of Word Frequency Lists

Perhaps of most utility and efficiency are word frequency lists.

The numerical listing of high-to-low frequency can be exploited in a variety of ways. High frequency words often have multiple meanings that can be discovered in exercises in the concordance, where the occurrences are brought together, or in the text itself. A typical exercise would be to copy and paste a word from the word frequency list into the “find” pop-up window of MS-Word and first locate the occurrences of the word of interest in the concordance. Examples of this kind of exercise are illustrated in the Appendix to this paper: Teaching/Learning using Collocations.

At the opposite end of the scale, the less frequent words, single meanings as opposed to multiple, are the rule. However, for the advanced student, these could well be the new words that need to be acquired or studied once they have been identified and counted. We have shown two words, one with one occurrence and one with only two occurrences. The concordance shows that both appear in the opening of the text, which is of interest in and of itself, part of the total meaningful context.

Words-Not-Found

Step 5: Cut and Paste the Segmented Text (free of non-Thai characters) into the Words-Not-Found Program.

www.seasite.niu.edu/thai/tdreader/wordchecker.htm

The output of this program will be a list of words that cannot be found in the online dictionary. It tells the instructor which words need to be added to the dictionary; it lets the student know that they will need to go to another source to look up the words not found in that particular text.

Interlinear Translation (IT)

(*N.B. A newer IT program dubbed "Paragraph Translator 2006" can be found in Part B of this article.)*

The translation process can proceed as follows:

The text of words, separated by spaces, becomes the input for NIU-IT (Interlinear Translation) program. You can use Notepad to separate words as describe at: http://www.seasite.niu.edu/trans/thai/howto/wordprocessor.htm.

Images of the homepage for IT and the page with the input document now appearing as a working translation document

Click to see large view

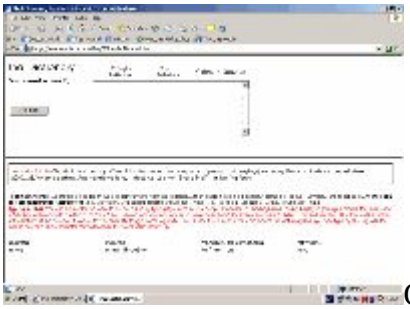
IT text auto-formatted for word-by-word and smooth translation

 Click to see large view

---

NIU Online Dictionary

There is one more tool that students can take advantage of, and that is an on-line dictionary. We are currently updating an earlier version for that purpose. The homepage for the online dictionary appears below.

 Click to see large view

---

Discussion

The translation tools we have discussed can be used by both the instructor and learner in a number of ways. In testing these tools with our students, we have found that, once they have been introduced to them and start using them, the task of translating actually becomes enjoyable and gives a feeling of control over what can be an onerous task.

The outputs (Concordance and Word Frequency List) can be modified and edited, and the increase in language proficiency can be monitored to a higher degree by the instructor and learner as well. Each student can build a personal dictionary to measure and monitor his/her own vocabulary acquisition.

Where students are involved in the process of “word divisions” (physically separating words in a continuous text for input into the word frequency and concordance programs), intensive word study takes place and raises questions that improve understanding and translation proficiency. What is a word? What is a compound? What is a semantic doublet and what are clues to their identification?

What is an elaborate expression? What is formal or unique to a certain kind of discourse? What is idiomatic? What is the “prior text” of the text being translated? These are just some of the questions that the translator must face.

Limitations: The IT (Interlinear Translation) program is limited to very basic translation work and is more suited for beginning and intermediate students. However it does give students the feel of moving from a word-by-word translation to a free translation and the teacher a means of monitoring the students word choice decisions, which are critical and the point at which many smooth translations end in mistranslations.

APPENDIX

Teaching/Learning Using Collocations

The illustrations that follow have been selected by a “copy + paste” into “find” from the Word Frequency List to search the Concordance.

High frequencies usually reveal multiple meanings and point out important collocations.

Low frequency words are items that usually have a single meaning or unique collocations and are candidates for quick look-up in the dictionary. The low frequency instances cited here turn out to be “puzzles.” In one case, a definition was not found in dictionaries, which suggests a very technical meaning or a new one of recent or rare appearance in the language. Later inquiry revealed the the word was part of a compound “SomDaeng” (literally “Red Ginseng”) an unusual reference to North Korea.

Sample collocations

context	keyword Gen./Mr.	context
ง ปักกิ่ง เมื่อ ม.ค. ว่า	นาย Mr.	กงฉวน Kongchuan โฆษก ประจำ กระทรวง
รม. ลา ออก ท้วง เลือกตั้ง	นาย	อาลี อับทาฮี Ali Abtahi รอง ประธานาธ
ทน ราษฎร สหรัฐฯ คน นำ โดย	นาย	เคิร์ต เวลดอน Kurt Weldon ส.ส. พรรค ร
และ เปรู อย่าง รุนแรง ของ	นาย	ฟิเดล คาสโตร Fidel Castro ประธานาธิบดี
พฤษภาคม ที่ ผ่านมา โดย	นาย	หลุยส์ เออร์เนสโต Louis Ernesto เดอร์เป
ขอ ปฏิเสธ ข้อ กล่าวหา ของ	นาย	คาสโตร Castro และ ขอ ลด ระดับ ทา
เดียวกัน ก็ ขอ ชื่นชม ที่	นายพล General	โคลิน เพาเวลล์ Colin Powell รมว. ต่างป

nuclear

ยุติ โครงการ พัฒนา อาวุธ 'weapon'	นิวเคลียร์	ชี้ จะ ทำ ให้ ช่อง ว่าง ข
ให้ การ เจรจา ยุติ วิกฤติ 'crisis'	นิวเคลียร์	บน คาบสมุทร เกาหลี ครั้ง
ยุติ โครงการ พัฒนา อาวุธ 'weapon'	นิวเคลียร์	ขณะเดียวกัน ก็ ขอ ชื่นชม
ให้ การ เจรจา ยุติ วิกฤติ 'crisis'	นิวเคลียร์	บน คาบสมุทร เกาหลี มี ควา

terminate/end

จีน ยิ้ม โสม แดง	ยุติ (vs. หยุด)	โครงการ 'program' นุก มังกร จีน ออก
โสม แดง เกาหลี เหนือ ยอม	ยุติ	โครงการ 'program' พัฒนา อาวุธ นิวเค
และ จะ ช่วย ให้ การ เจรจา	ยุติ	วิกฤติ 'crisis' นิวเคลียร์ บน คาบส
บาล เกาหลี เหนือ ตัดสินใจ	ยุติ	โครงการ 'program' พัฒนา อาวุธ นิวเค
อ ซึ่ง จะ ทำให้ การ เจรจา	ยุติ	วิกฤติ 'crisis' นิวเคลียร์ บน คาบส

and

ความ ขัดแย้ง เล็ก แคบ ลง VERB	และ	จะ ช่วย VERB ให้ การ เจรจา ยุต
ต สหรัฐ ประจำ กรุง ริยาด NOUN	และ	สถาน NOUN กงสุล สหรัฐ ประจำ เม
สหรัฐ ประจำ เมือง เจดดาห์ NOUN	และ	เมือง ดาร์เรน NOUN ประเทศ ซาอุ
ชาว อเมริกัน คน อังกฤษ NOUN	และ	คน ออสเตรเลีย NOUN คน ที่ ทำงาน ใ
วิจารณ์ รัฐบาล เม็กซิโก NOUN	และ	เปรู NOUN อย่าง รุนแรง ของ นาย
นใจ เรื่อง นโยบาย ทั้ง ใน NOUN	และ	NOUN ต่างประเทศ ของ เม็กซิโก พ
อ กล่าวหา ของ นาย คาสโตร NOUN PHRASE	และ	ขอ ลด ระดับ VERB PHRASE ทาง การ ทูต ร

Solving Semantic Puzzles: Two examples of compounding

????

จีน ยิ้ม	โสม	แดง ยุติ โครงการ นุก มังก
รง แสดง ความ ยินดี รัฐบาล	โสม	แดง เกาหลี เหนือ ยอม ยุติ
โครงการ นุก มังกร จีน ออก	โรง	แสดง ความ ยินดี รัฐบาล โส

Citation from the original text

จีนยิ้มโสมแดงยุติโครงการนุก

มังกรจีนออกโรงแสดงความยินดีรัฐบาลโสมแดงเกาหลีเหนือยอมยุติโครงการพัฒนา อาวุธนิวเคลียร์ชี้จะทำให้ช่องว่างของความขัดแย้งเล็กแคบลง

โสมแดง is a compound of "ginseng + red" and refers to North Korea.
ออกโรง is a compound of "enter onto + the stage."

มังกรจีน is a compound of “dragon + china”, i.e., The Chinese Dragon.

The collocations/expressions that refer to North Korea and China, respectively are, at the same time, ethnic epithets used by the Thai. The Chinese are like the dragon in a Chinese Opera and the North Koreans are pictured as ginseng root, which has the shape of a pair of human legs.

Final translation:
(Headline) China Smiles: “Red Ginseng” (North Korea) Terminates Its Nuclear Program

(Lead Sentence) The “Chinese Dragon” enters onto the stage to show its pleasure that the government of “Red Ginseng” North Korea has agree to cease the development of nuclear weapons.

---

www.Thai2English.com dictionary search produced the following results

The Thai words contained in your search "โสมแดง" are shown below. Click on any of the matches for a more complete definition.

Search Results ผลการค้นหา

โสม	Sorry, we could not find the word โสม in the dictionary.
แดง daeng	red;

Unresolved translation: The news article implies that “som daeng” refers to North Korea. However, in questioning two native Thai speakers, who are also avid soccer fans, they both claim that “som daeng” refers to South Korea. North Korea, they claim is called “som khao” or “White Ginseng.”

Preliminary Conclusions

Word Frequency Lists and Concordances are powerful tools that can be effectively and efficiently used to:

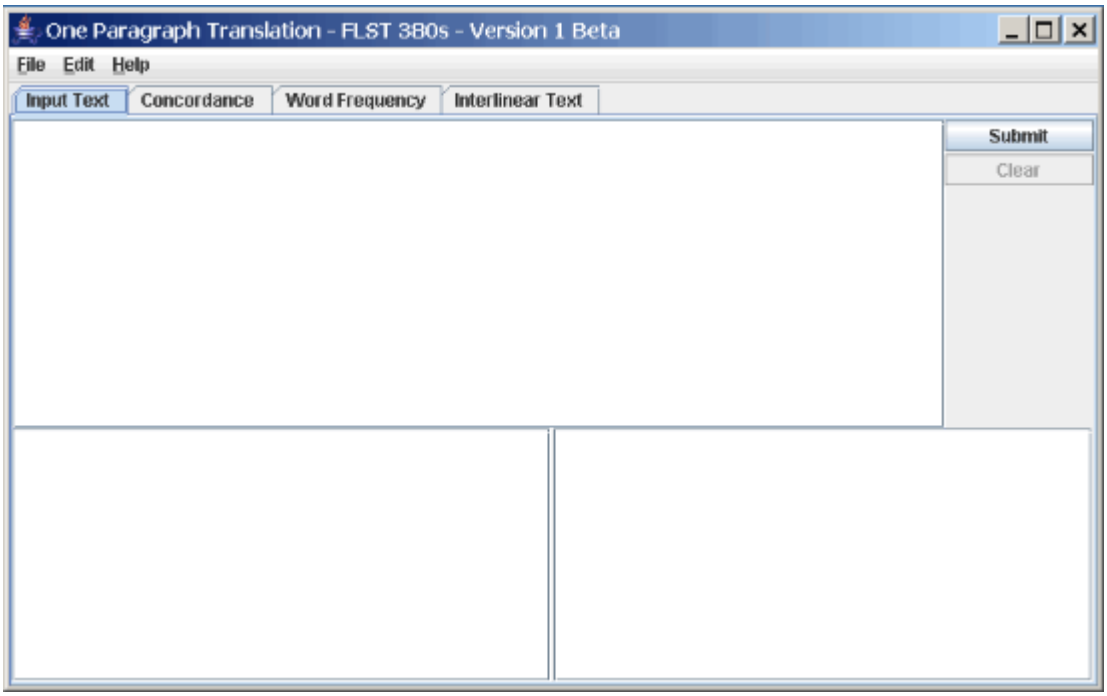
manage large text corpuses
monitor the learner’s acquisition of vocabulary
go beyond the information in dictionaries to discover collocations
discover solve semantic puzzles in the translation process

Part B 2006

Translation Application - for Paragraphs and Short Texts

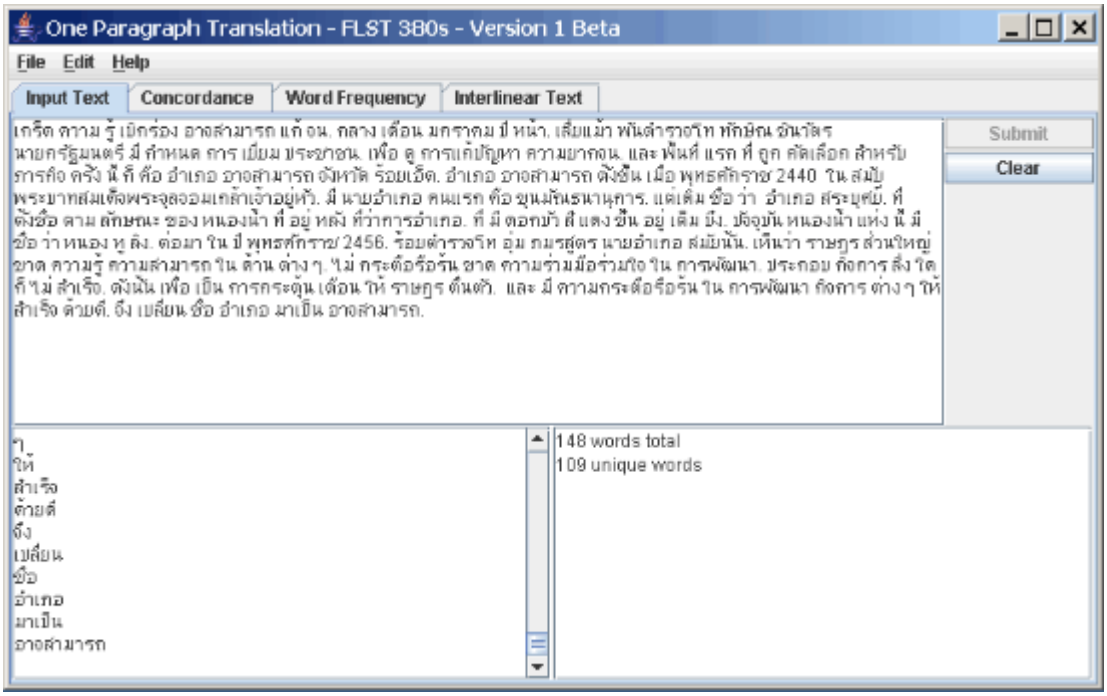
1. The application found in the link (Paragraph Translator) below requires Java Software.
2. If you have not downloaded and installed Java Software in your computer, please do so by going to http://java.sun.com.
3. If you have installed Java Software, then click One Paragraph Translation to launch translation application.

How to use the translation application:



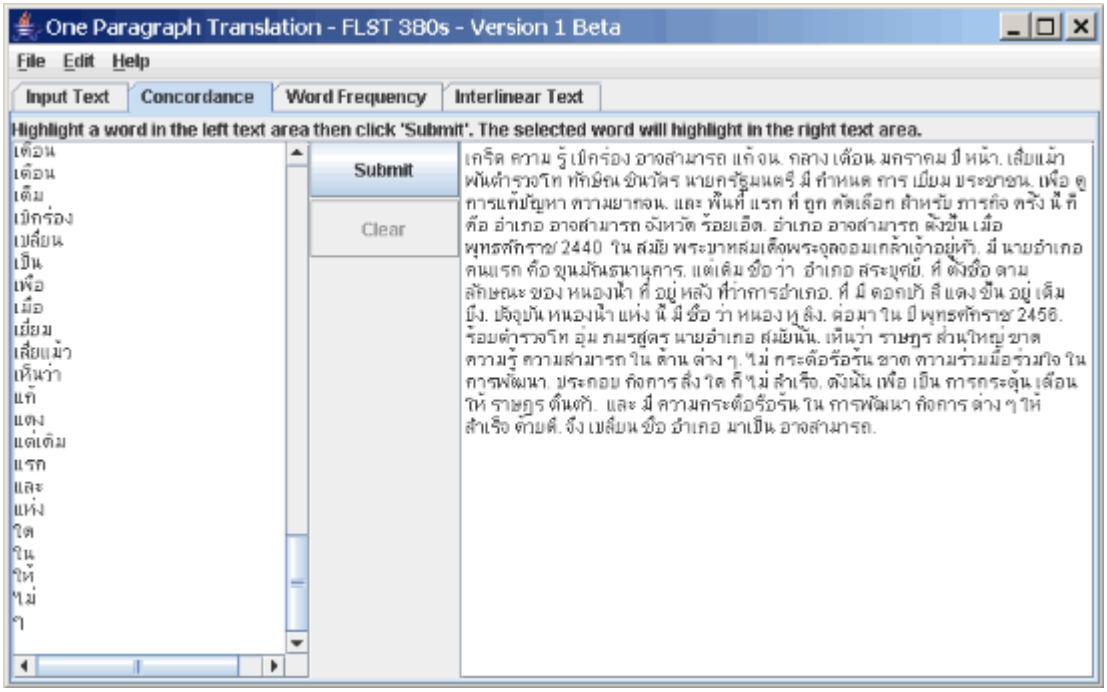
Input Text:

1. Segment a paragraph or other text into clauses or sentences and end each with a period ("."), but **do not use any carriage returns**. (If you use carriage returns, the program will not work and will display and error message to that effect once you "submit" your paragraph/text for processing.) Then, copy and paste segmented paragraph into the first box. The paragraph must end with a typed period (".")
2. Click submit.
3. Notice that the two lower boxes will contain (left box) all the words in the text in a single column and (right box) a count of total words and unique words. This is to tell you how many words are in the text and also to see if they have been segmented correctly. If not, return to the text and make corrections; then rerun the program.



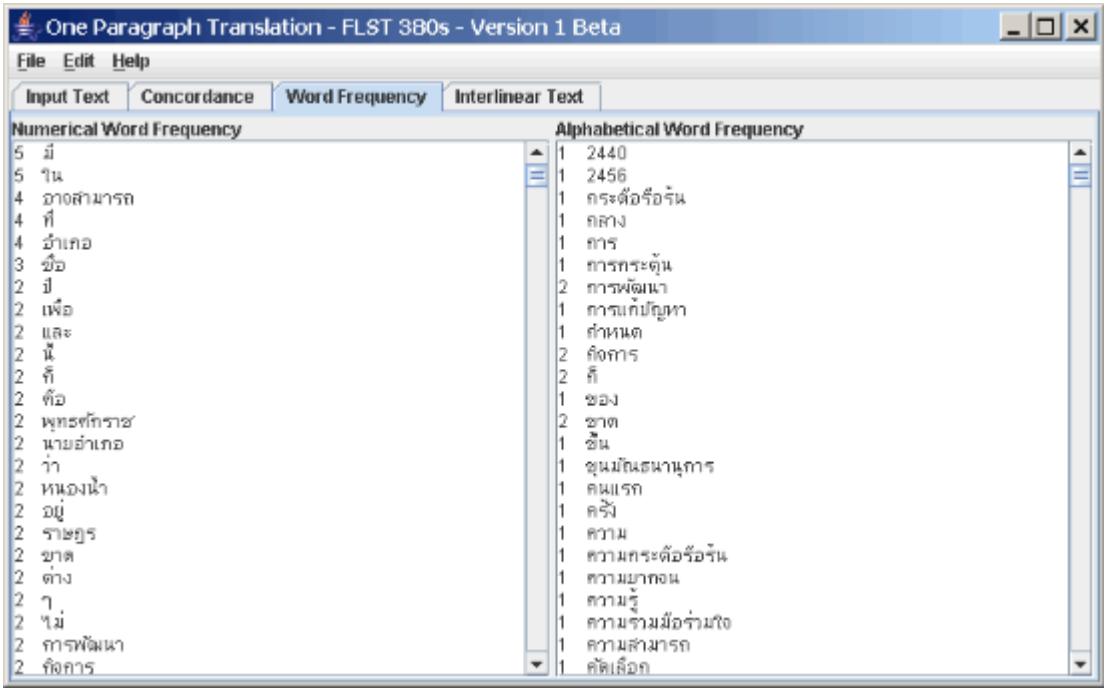
Concordance:

1. Click on the Concordance button. Notice that the input text appears in a box on the right; in the box on the left are the unique words in (Unicode) alphabetical order.
2. To see how any word is used in context, highlight a word on the left box.
3. Click Submit, and the same word will be highlighted in the text on the right.
4. Click Clear to continue the process of highlighting and submitting new words.



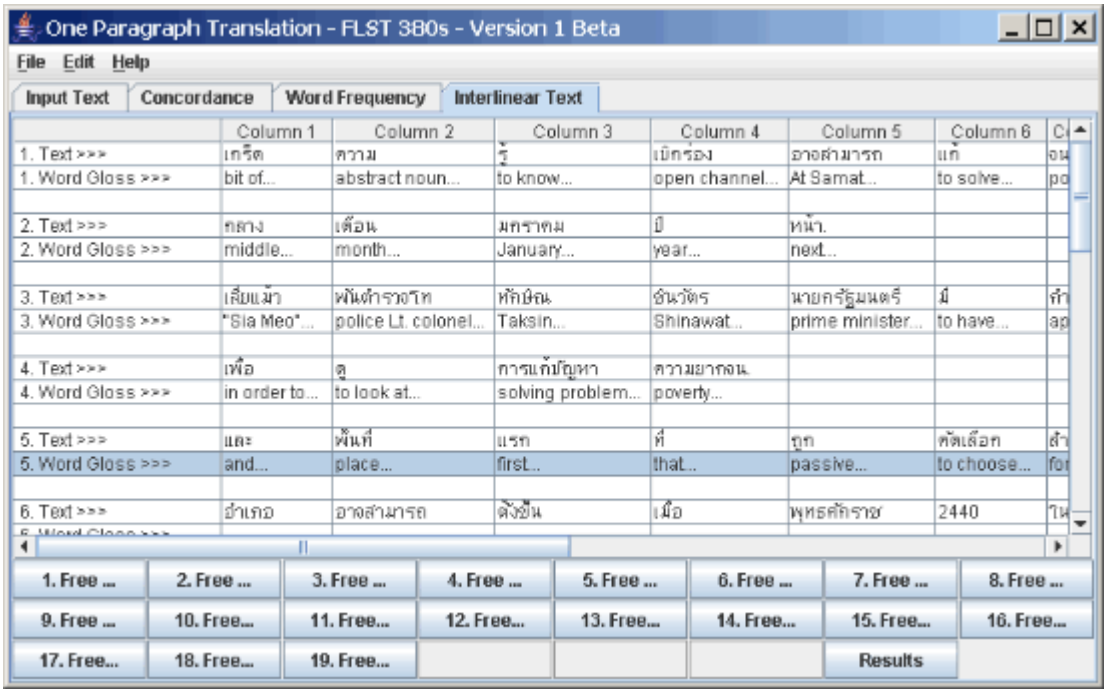
Word Frequency:

1. Click on the Word Frequency button. Two boxes will open up. The one on the left lists words in numerical order; the one on the right lists words in (Unicode) alphabetical order. You can select (ctrl-a), copy (ctrl-c), and paste (ctrl-v) either list into a Word document or Excel spreadsheet, save and print or use on-screen to fill in the definitions and save for further study.



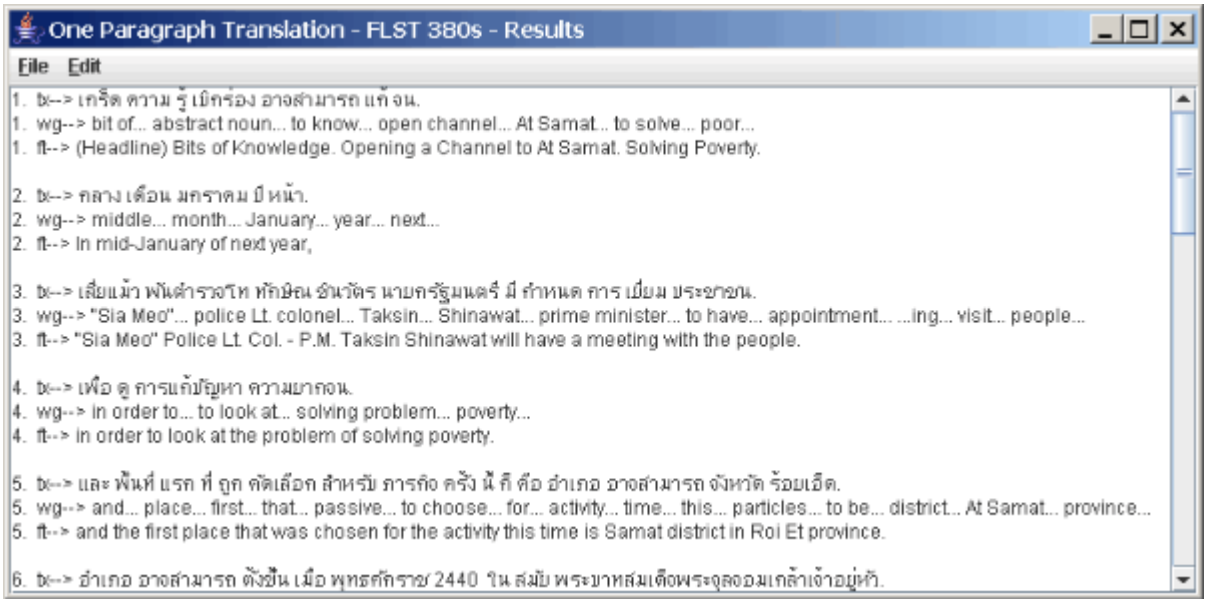
Interlinear Translation:

1. Click on the Interlinear Text button. The text has been formatted for translation at two levels: word-by-word and free translation.
2. For word-by-word translation, place the cursor under the word you want to define and type in the meaning. Do this for all the words.

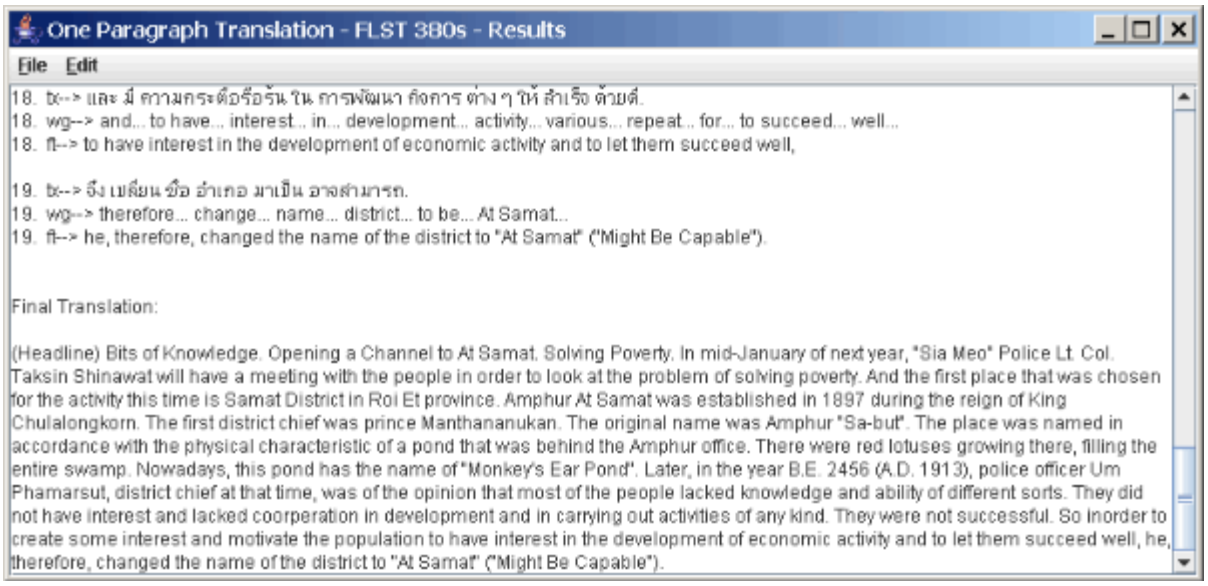


3. For a free translation, click on the free translation button at the bottom of the screen corresponding to each sentence number.
4. Type your free translation.
5. Click the Results button to view your finished translation.

Final Results:



Free Translation as Continuous Text (This can be edited. See steps 1 and 2 below.)



1. Select all the final results output, copy, and paste into a Word document and save it with a file name you choose.
2. From there, you can edit it, use "Track" to make an exchange corrections (teacher and student), and print it out.