Center for Advanced Language Proficiency Education and Research (CALPER)
Corpus Community Report #3
August 2009

Teddy Bofman **t-bofman@neiu.edu**
Paul Prez **p-prez@neiu.edu**

Northeastern Illinois University
Department of TESL/TEFL and English Language Program

Thai Pop Music: A Specialized Corpus for the Language Classroom

For those who assume that corpus linguistics is the exclusive domain of experts in the field of information technology, read on and you will see that armed with a personal computer and easy-to-use analytical software, language students, educators and researchers with a desire to explore naturally occurring language can reap the benefits of corpus-based methodologies. We also demonstrate that a non-Western script need not be a deterrent**,** as our research was conducted on Thai.

Large**,** general corpora are available for many languages, including Thai. A general corpus includes a wide variety of texts and language types and serves as a broad-spectrum reference. A specialized corpus is narrower in focus. It aims to provide information about a specific domain, genre or text type.

This Coprus Community Report describes how we designed, compiled and analyzed a small, specialized corpus of Thai language texts. We carried out our study with Thai second language pedagogy as our main focus. Our intent was to analyze the language of Thai pop music in order to provide information on the lexical items and patterns that appear in the lyrics and to guide teachers and students alike in their own investigations of such a corpus. Corpus-informed pedagogy offers exciting ways to explore and learn.

Our study looked at 400 popular songs that appeared on the Internet at **www.ethaimusic.com** in 2004. We compiled and analyzed our corpus of lyrics using Microsoft Word and Oxford University Press's WordSmith Tools 4.0, a suite of

programs that produces word lists, concordances and key word analyses. WordSmith Tools is capable of processing texts written in Thai as well as in a number of other non-Western scripts. Simply be sure to select Thai as the working language and to convert your data files into Unicode text files. You can easily convert files when using word processing software such as Microsoft Word by saving texts as plain text files (.txt) and selecting Unicode encoding.

Our corpus study was designed to aid teachers of Thai as a second language interested in utilizing pop music in their classes. We created word frequency lists to provide a profile of the lexical content of the corpus. Such a profile allows instructors to determine vocabulary level appropriateness and ultimate learning burden for students. We also modeled the process of identifying and selecting lexical items, lexical patterns and textual features for focused instruction. In addition, we demonstrated how the corpus can be used as a source for materials development.

Once we had collected the texts for our study we began to prepare them for use with WordSmith. One of the characteristics of Thai orthography must be dealt with when preparing texts for computer analysis. Thai is written without spaces between words**,** and this poses a challenge: a computer requires explicit word boundaries. Thus, our first task was to segment the long strings of Thai text into computer-recognizable units. A number of researchers are engaged in computer segmentation of Thai and computer-assisted segmentation is an option. However, we chose to segment the text manually, relying on our intuitions, dictionaries and the help of native speakers. Even native speakers of Thai do not agree on word boundaries; there is room for flexibility. We were satisfied that our segmentation was based on strategies that would allow the analytical software to produce optimal results for our purposes.

One strength of corpus analysis is that procedures can be customized to address the specific needs of a given project. In general, when segmenting**,** we favored splitting

over chunking so that certain morphemes would become more salient. For example, the abstract noun marker ความ /khwaam/ is a highly productive morpheme in Thai that corresponds roughly to '-ness' in English. It can occur with verbs and adjectives, forming lexical items such as ความรู้ /khwaam rúu/ (nominalizer + know), 'knowledge.' We chose to analyze /khwaam/ as one unit and an accompanying verb or adjective as a separate unit. Our goal was to determine how prominent the morpheme ความ /khwaam/ was in the lyrics we were analyzing. Indeed, ความ /khwaam/ ranked thirty-second in a frequency list of over 3,500 lexical items. Obviously, it is a morpheme that needs to be addressed in the classroom.

The process of manual segmentation, though time-consuming, allowed us to become quite familiar with the texts. As we segmented the texts, we noted various features that we wanted to investigate further. Tagging makes it possible to analyze and retrieve instances of specific linguistic phenomena. During the segmentation process, we identified and tagged a variety of items, including reduplications, Sanskrit terminology, and English words written in Thai script.

What did we achieve through our corpus analysis? We identified a core of high frequency lexical items which occurred and reoccurred throughout the lyrics. A mere 39 words represented 50% of the more than 3500 running words in the corpus**,** indicating that these songs constitute suitable material for relative beginners. Lexical analysis software is an invaluable tool for materials development. WordSmith Concord tool makes it possible to retrieve representative and meaningful examples of lexical items, collocations and chunks as well as tagged items. Using Concord we developed a variety of model exercises for explicit vocabulary instruction based on the most frequent lexical features. (See "Thai Pop Music: Corpus Analysis and Second Language Learning" for detailed information.)

Corpus linguistic methods allowed us to analyze the song lyrics in far greater detail and with far more accuracy than would otherwise be possible. Corpus research allows the language teacher the luxury of knowing exactly which lexis and lexical patterns warrant special focused attention. Authentic examples can be retrieved and presented to the learners. With minimal instruction learners can engage in data-driven learning and make discoveries on their own. Corpus linguistics provides the potential for dramatic changes in the way foreign languages are taught.