

Dr. San San Hnin Tun
Senior Lecturer (Burmese & French)
Department of Asian Studies & Romance Studies
Cornell University
e-mail: sht3@cornell.edu

Working with a spoken Burmese corpus

I recently completed a thesis on *Discourse marking systems in Burmese and in English, using a corpus-based approach*. It was a challenging project (given the nature of Burmese, a tonal language with its own orthography, which is not compatible with existing concordancing software suites), but a gratifying endeavour for me, for the findings not only reveal valuable insights into our understanding of discourse features in Burmese, but also pave newfound ways to analyze recurring patterns of natural spoken discourse in Burmese.

1. Corpus design

As the focus of my study involves discourse particles in spoken Burmese, the primary objective of the corpus design was to include a reasonably large variety of texts that are representative of natural spoken language. From the outset this seems to be a fairly straightforward task. However, given the cultural-political situation in Myanmar (for example, recording strangers may evoke suspicion) and for practical reasons (I can go to the country only infrequently, and usually for a short period), I initially opted to use pre-scripted speech such as radio plays, broadcast news and audio recordings collected by various individuals for their research purposes. During the course of my research project, I also managed to collect my own recordings with friends and family. In the end, the selection of Burmese data includes a variety of texts that are representative of natural spoken language – spontaneous as well as pre-scripted speech delivered in both spoken and written media – yielding approximately 250,000 tokens, a decent corpus size for the purpose of my study.

Since my study concerns a comparative analysis, the next step was a selection of a comparable English corpus from the existing corpora - at which point I was faced with a new (and unanticipated) challenge, which also turned out to expose an important revelation about some shortcomings of the current theories of corpus linguistics, mostly based on English, which use 'word' as a unit of analysis. Burmese on the other hand is a syllabic language in which syllables may represent free-standing as well as bound lexical items, which are highly polysemous and context-dependent. As a result, the definition of 'word' is problematic for Burmese, and a syllable count system is used for measuring Burmese texts [see 3. for further details].

In sum, in comparing Burmese and English corpora, which use two different systems of measuring corpus size, it was necessary to devise a normalization process. On average, there are approximately 2,500 syllables produced in ten minutes of audio recordings in Burmese, yielding 15,000 syllables per hour. Based on the existing corpus size and their transcripts, it can be calculated that in English, speech is produced at approximately 10,000 words per hour (figure verified by McCarthy, personal communication, in relation to the

CANCODE¹ corpus; see McCarthy 1998). Therefore it was assumed that syllable to word ratio is approximately [1.5:1]. In other words, every 15 syllables in Burmese are considered equal to 10 words in English for comparative analyses of my study, and the selected English corpus is composed of approximately 180,000 words.

2. Transcription

All data were first transcribed in Burmese². In order to ensure a maximum consistency, the texts are transcribed based on their standardized written form because most items in Burmese undergo sound change (mutations) according to their phonetic environment. For example, according to their neighbouring phonemes, the polite particle ပါ 'paa' can be pronounced /pa/ or /ba/, sentence final particle တယ် 'teeh' may be pronounced /teh/ or /deh/, and so forth. On the other hand, Burmese orthography is more stable, and therefore used as a preferred tool for transcription.

The Burmese transcription is then converted manually into Roman characters. To my knowledge there is no concordancing software program available for texts in Burmese script. In addition, Burmese being a tonal language, it was a challenge to find an appropriate transcription system given that the roman script is not exactly equipped to represent the phonemic nature of Burmese tones. Preliminary explorations with diacritics did not render satisfactory results, particularly because accent marks are not properly recognized in my chosen concordancing programs. I opted therefore to create a transliteration system that allows me to transcribe Burmese texts in roman characters that are compatible for use with my chosen software suite, *WordSmith Tools* (Scott, 1998). In the current transcription system, I use one to three vowels to represent the three tones in Burmese³, which can be described as follows:

- one vowel (e.g. /a/ /e/ or /ou/) represents a *short tone*,
- two vowels (e.g. /aa/ /ee/ or /ouu/) represent a *middle tone*, and
- three vowels (e.g. /aaa/ /eee/ or /ouuu/) represent a *long tone*

Table 1 illustrates examples of transcription for Burmese linguistic items representing the three tones.

	Short tone	Mid tone	High tone
Burmese orthography	တဲ	တယ်	တဲး
Transcription	teh	teeh	teeeh
Meaning	<i>It is said that</i>	Verb Sentence Marker	<i>a hut</i>

Table 1. Illustration of transcription system for three tones in Burmese

¹ CANCODE: Cambridge and Nottingham Corpus of Discourse in English.

² It is arguable that this step may be skipped, but it is my preference to keep the original transcript in Burmese as it serves as a (stable) reference throughout the study, for transcription as well as for the analyses.

³ The number of tones and their labels are disputable as they have not yet received a consensus. The three-tone system that I have chosen here represents the three tones that are typically reflected in the standardized Burmese orthography. Since prosodic features are not included in the analyses for this study, labels are also simplified as *short*, *medium*, and *long tones*.

3. Analyzing Burmese corpora with concordancing software Wordsmith

In doing corpus analysis of Burmese, it is important to note that there is no concordancing software available for Burmese, nor any word-processing program that does "word counts" for Burmese. In fact, the notion of "word", as perceived for languages that use an alphabetic script cannot be applied to Burmese, which uses a syllabic script. I am therefore using "syllables" as a measuring system for the Burmese corpus. Consequently, 'word count' in the *Wordsmith Tools* program (Scott 1998) has to be interpreted as syllable count for Burmese.

After frequency lists are mechanically generated with *WordSmith Tools*, particles to be examined have to be isolated manually because many particles are either homonyms or polysemous. For example 'paa' can be a polite particle or a main verb meaning *to be included*; and 'teeh' can be a sentence final particle or an adverbial intensifier, and so forth, depending on the context. Moreover, some syllables may be monosyllabic particles or a part of a 'word'. For instance, 'ka' may be the subject marker or a part of the noun 'kA lee' meaning *child* or the verb 'kA saaa' meaning *to play*.

From the early stages of analysing the Burmese corpus data it has come to my attention that while the computer programs and concordancing tools have facilitated corpus-analysis for English and other western languages alike, possibilities of machine generated analysis for Burmese are highly limited. As it has not been a straightforward matter to define the notion of 'word' in Burmese, Wordsmith tools can take us only as far as generating frequency lists (mostly for mono-syllabic lexical items⁴), and concordance files, from which extracting relevant information requires several steps of going through concordance lines one by one, and manually isolating the relevant instances. For instance, steps necessary to identify discourse functions of particle 'taw' among the 3,540 instances include:

- eliminating 'taw' tokens which do not fit into the category under investigation such as 'taw' in poly-syllabic words (e.g. 'kaan taw' *to pay homage to Buddha, elders*), 'taw' as an object pronoun *you* or a possessive adjective *your*,
- eliminating 'taw' with mainly grammatical functions (e.g. 'taw' expressing an English equivalent of *when*)
- distinguishing the meanings of 'ssoo taw' among 'NP/utterance + ssoo taw' meaning *since+Np is/utterance*, 'baa ppyiq lo leeh ssoo taw' meaning *because*, and an utterance-initial 'ssoo taw' which functions as a discourse connector such as *so*, etc.

4. Concluding remarks

In sum, analysing corpus in Burmese requires improvisation and modification of existing tools and methodologies, which are mostly based on English. It involves a labour intensive task of manual selection and extraction of relevant data. Nonetheless, final outcomes of my study offer many valuable insights and an interim solution for a corpus analysis involving a less-commonly

⁴ It is possible to generate a frequency list in clusters, but my preliminary analyses suggest that this method also involves a considerable amount of manual selection in order to extract relevant data, the clusters yielded on the frequency list are mostly incomprehensible and therefore do not seem useful for the present study.

studied language using non-roman script, until corpus linguistics grows into a more language sensitive, globally-applicable science.