NTU corpus of Formosan languages: 
A state-of-the-art report

LILY I-WEN SU, LI-MAY SUNG, SHUPING HUANG, 
FUHUI HSIEH and ZHEMIN LIN

The NTU Corpus of Formosan Languages (http://corpus.linguistics.ntu.edu .tw) started as one of the sub-projects of the Multimedia Laboratory established by the Center for Information and Electronics Technologies. The corpus integrates the professional and academic works of various departments and colleges at National Taiwan University, with a view to establish a standard for the creation of linguistic corpora databases through the application of information technology on linguistics research. The development of natural language processing techniques and dynamic web technology has generated wide interest in the construction of an integrated platform which enables people to submit, browse and search among collected texts against established corpora.

For seriously endangered languages, corpus documentation plays a crucial role in preserving the precious linguistic and cultural data. The NTU Corpus of Formosan Languages is constructed for the purpose of preserving Austronesian Languages spoken in Taiwan, many of which are on the verge of extinction. It is the result of our effort in providing a multilingual on-line corpus with multimedia capability to meet the needs of both linguists and those interested. It features: (i) discourse-oriented documentation that goes beyond sentential level, (ii) retrieval with cross-referencing capability, (iii) ethnolinguistic notes preserving precious cultural information, and (iv) multi-media data with interoperability with different systems for the greatest interest of the users.

To record faithfully discourse data collected, we incorporate in our corpus interactional linguistic features such as pauses, hesitations, fillers, repetitions, false starts etc. Approaching language from a functional approach, we believe linguistic features as such are non-trivial in that they may reveal cognitively and psychologically important details of language in use. Investigating spoken data in natural context makes it possible for us to explore issues beyond the scope offered by the traditional syntactic studies. For example, linguistic
issues used to be the central concern of the formalists, such as the development of ergativity (cf. DuBois 1987; DeLancey 1981), transitivity (Hopper and Thompson 1980), and the issues related to the identification of lexical category (Hopper and Thompson 1984), now assume a new meaning when the discourse motives are ascertained.

In our corpus, all utterances are segmented and coded according to Chafe’s (1987) idea of the so-called Intonation Unit (IU). Also referred to as contour-defined tone unit, or tone group, IU is “a stretch of speech uttered under a single coherent intonation contour.” As an information-structure unit, IU is not merely a physiological unit of spoken discourse; it actually serves the function of packing the conveyed messages into prosodic chunks of information by the speakers. At present, a total of 4186 IUs are available from our corpus online.

As a special feature, our corpus allows web users to query any specific target word or morpheme in English, Chinese or any single one of the Formosan languages uploaded. With 32 elicited texts (totaling 2 hours and 30 minutes) from Kavalan, SaiSIyat, Amis, Tsou, and Sakizaya, our Search Engine can search for a specific lexeme in any of the five Formosan languages. By typing a chosen linguistic form, the system may generate instances of the lexeme in all texts of the language specified. The Search function can further perform cross-linguistic searches. By typing the desired lexeme and selecting the “all languages” option, all occurrences of this lexeme in the Formosan languages contained can be retrieved. The multilingual search function as such is useful for comparative and cross-linguistic studies.

Unlike many existing corpora, NTU Corpus of Formosan Languages includes field notes jogged down during the course of fieldwork. The notes contain information that may hint at and give clues to how each grammatical form may be related to its social action, providing linguistic “evidence” for the “socially-shared cognition” (Schegloff 1991) unique to the speech community. Data from the notes are of socio-cultural values significant for ethno-linguistic or anthropological studies, since preserving these soon-to-be-extinct Formosan languages in Taiwan has been from the very beginning one major mission we wish to accomplish through our corpus.

An additional special feature of our corpus is the presentation of data in multimedia format, integrating the metadata, the texts, as well as the multimedia contents via the web interface. The media contents include visual as well as audio outputs, enabling the users to listen to the actual spoken sound and to watch, even at the IU level, the recorded film of each text unit.

Considering the enormous amount of time, money, and manpower invested in corpus construction, it only makes sense if the data can be shared by the linguistic community. The interoperability between different systems is made possible by the Extensible Mark-up Language (XML), a simple and flexible language which is a de facto standard on the web. Our system
can count the number of lexical units when new texts are uploaded, while compiling them automatically into an online dictionary. The dictionary can be printed out, retrieved, and manipulated for further research.

With the advances in storage, character-encoding and web technology, the development of an online platform for corpora has great impact on the accessibility, quality and quantity of linguistic field data. The spirit behind the design of our NTU Corpus of Formosan languages is to house virtually unlimited quantities of linguistic data for easy accessibility, with the extension potential to document all Formosan languages, once the proper feature vectors are set. More Formosan texts are now being annotated and are scheduled to be available online soon.

Notes
1. The participating NTU departments and colleges are: Graduate Institute of Linguistics (GLE), Department of Information Management, Department of Library and Information Science, Department of Computer Science and Information Engineering, Department of Electrical Engineering, Department of Journalism, and Department of Drama and Theater, College of Electrical Engineering and Computer Science, College of Liberal Arts, College of Social Sciences, and College of Management.
2. Not all 98 texts (with a total length of 8 hours and 24 minutes) elicited from the six Formosan languages (Socondi plus the five already uploaded online – Kavalan, Saisiyat, Ama, Tai and Salinan) are currently available from our website. Our text-collecting is made possible primarily through funding granted by the National Science Council of Taiwan, and also through work completed over the past few years in the GLE-required course, Fieldwork, taught by Li-Mei Sung, the second author of the present report. Some of the research results of the NSC projects are reported in Huang, Su and Sung (2000, 2002, 2003, 2004), Huang and Sung (2005), and Huang, Snakos and Sung (1999).
3. Extensible Markup Language (XML) is a simple, very flexible text format designed to handle the exchange of a wide variety of data on the Web. See http://www.w3.org/XML/ for a detailed explanation.

References
Chafe, Wallace L.
DeLancey, Scott
Du Bois, John W.
Hopper, Paul J. and Sandra A. Thompson
1984 The discourse basis for lexical categories in Universal Grammar. Language 60, 702–752.
Huang, Shuan-fen, Lily I-wen Su and Li-May Sung
Huang, Shuan-fen and Li-May Sung
Huang, Shuan-fen, Jozsef Szakos, and Li-May Sung
Schegloff, Emanuel A.