UNIVERSITY OF CALIFORNIA, LOS ANGELES

UCLA

BERKELEY · DAVIS · IRVINE · LOS ANGELES · RIVERSIDE · SAN DIEGO · SAN FRANCISCO

SANTA BARBARA · SANTA CRUZ

DEPARTMENT OF EAST ASIAN LANGUAGES & CULTURES
290 ROYCE HALL
Box 951540
LOS ANGELES, CA 90095-1540

November 20, 2003

Professor Yuphaphann Hoonchamlong
Department of Hawaiian and Indo-Pacific Languages and Literatures
University of Hawai'i at Manoa

Dear Dr. Yuphaphan:

This is to confirm that your paper "Thai Language Audio Resource Center Project: Thai Speech Database and Application in Web-Based Language Teaching" will appear in the Proceedings of the 13th Annual Meeting of the Southeast Asian Linguistic Society (May 2-4, 2003). This paper will serve an important resource for teachers and students of the Thai language. The editors of the volume are Dr. Andrew Simpson of University of London and Shoichi Iwasaki of University of California, Los Angeles. The publisher is the Program for Southeast Asian Studies, Arizona State University, Tempe, AZ. We are in the middle of editing the volume, and hope to have it published soon.

Sincerely,

Shoichi Iwasaki
Chief Organizer, the 13th Annual Meeting of the Southeast Asian Linguistic Society

Thai Language Audio Resource Center: Thai Speech Database and Application in Web-based Language Teaching

Yuphaphann Hoonchamlong
University of Hawaii- Manoa <*yuphapha@hawaii.edu*>
Sathaporn Koraksawet
Thammasat University <*stp@tu.ac.th*>
Rugchanok Janevarakul
Thammasat University <*rugchanokj@hotmail.com*>

Introduction

Recently, the importance of corpora or databases of language in research in linguistics, lexicography and natural language processing has gained increased recognition. The growth of the high speed internet enables the dissemination of and access to these text and multimedia language resources with ease and speed, making resource sharing among researchers from every corner of the world a possibility.

The development of audio resources of Thai language (especially speech) for research and development with dissemination via the internet is still in an early stage of evolution. This paper reports on **Thai Language Audio Resource Center,** one such Thai language resource now available via the web.(Hoonchamlong et al. 2002a, 2002b)

Thai Language Audio Resource Center (ThaiARC) project provides audio information (in the form of digitized audio files) on Thai language (especially speech) for academic research, disseminated via the Web as part of ThaiSARN (Thai Social/Scientific Academic Research Network) at **http://thaiarc.tu.ac.th**. ThaiARC also serves as an archive for these collections as reference and as a shared resource on Thai language for researchers.

The project pioneers the production and collection of various types of audio information on Thailand and the Thai language, such as royal speeches, academic lectures, oral literature, etc, for dissemination on the web. It also explores the application of speech data in web-based language teaching.

The audio files at ThaiARC are available in Wave (.wav) RealAudio (.ra) and mp3 formats, formats that are readily accessible by most computers. Besides being a repository for the collections of audio information on Thai Language, ThaiARC also provides electronic verbatim text/transcript and annotation accompanying each audio program, also distributed electronically through the web.

ThaiARC was funded by NECTEC (National Electronic and Computer Technology Center), NSTDA (National Science and Technology Development Agency), Ministry of Science, Technology and Environment of Thailand during the year 1997-2002. The project was carried out in two phases at Thammasat University which also hosts the project website.

Phase 1 of the project was conducted during 1997-1999 as a pilot phase to establish the audio resource center by studying appropriate technology to be used in each stage of database development, namely, data formatting, archiving, data accessing and retrieval. Samples of three groups of Thai audio information, namely, speech styles, regional folktales and poetry are made available. In addition, a Thai Language Page was also established especially to provide English information on the Thai language to foreigners.

Phase 2 of the project was conducted during 2000-2002. The objectives of Thai Language Audio Resource Center (THAIARC) project Phase 2 are

1) to establish a resource for Thai language audio information for research, by systematically collecting Thai audio data for linguistic research, starting from the tones of the four Thai dialects, which are the distinct characteristics of Thai language. The tonal word sets are collected according to linguistics research design for tones.

2) to investigate the forms and techniques most appropriate for web-based language instruction, and to apply the collected Thai audio information in developing a model web-based lesson for teaching Thai listening comprehension skills for foreigners.

At present, the following four groups of Thai speech data are available as an online voice sample library on the following topics, along with verbatim transcripts and annotations:

1) Thai Regional Dialects: samples of word sets demonstrating tonal variations among the 4 major Thai dialects (Northern, Northeastern, Southern, Central) and Standard Thai, with audio file search tool for easy
access, located at: http://thaiarc.tu.ac.th/host/thaiarc/dialect

2) Thai Regional Folktales: samples of Thai Folktales from the 4 major regions (Northern, Northeastern, Southern, Central), each with transcript and standard Thai translation.

3) Thai Poetry: samples of readings of various poetic styles and versifications. Explanations are available in both Thai and English.

4) Speech Styles: samples of the King's speeches for various occasions, for example, his Golden Jubilee Speech; also here are samples of various types of news broadcasts.

ThaiARC also features the **"Thai Language Page"** which aims to provide overview information about Thai language in various aspects for foreigners. This information includes the history of the Thai language, Thai alphabets and basic Thai phrases with audio files for tourists or those interested in Thai.

In addition, a sample application of **ThaiARC** audio and transcript data as "Web-based Instruction of Thai Listening Skills" is also demonstrated.

Figure 1. ThaiARC Home Page (May 2003)

Data Collection, Archiving and Dissemination

1) Thai Regional Dialects Word Sets

Thai is a tonal language belonging to the Tai language family, which includes languages

spoken in Assam, northern Burma, all of Thailand including the peninsula, Laos, Northern Vietnam and the Chinese provinces of Yunnan, Guizhou (Kweichow) and Guangxi(Kwangsi).

In Thai and Tai dialectology, tone systems and variations in tone systems are important identifying features of the various Thai/Tai dialects. Based on a method developed in the discipline of comparative and historical Thai linguistics, William Gedney (1972)'s "Proto Tai Tone Matrix" or commonly known as "Gedney's Tone Box" has been widely used as a tool for collecting word lists for determining the tone system of a Thai/Tai dialect (see Figure 2.). The tone box is based on the development of various Tai tone systems by tone splits and tone mergers from the reconstructed three tones (Tone A, B and C) of Proto-Tai. The tone splits and mergers were influenced by certain features of initial consonants of Proto-Tai such as voicing and aspiration, and also the types of syllables and vowel length. The word lists collected are words that are known to be Tai in origin, i.e. words describing daily life and environment in villages shared by related Tai dialects.

Proto Tai Tone Matrix (adapted from Gedney 1972:434)

Initial Consonants at time of split	Live (Smooth) Syllables			Dead (Checked) Syllables	
	A	B	C	Long Vowels (DL)	Short Vowels (DS)
Class 1. Voiceless Friction *s *hm *ph					
Class 2.Voiceless unaspirated stops *p *t					
Class 3. Glottal *ʔ *ʔb					
Class 4. Voiced *b *m					

Figure 2. Proto-Tai tone matrix

The following diagram (Figure 3), adapted from Tingsabadh 2001, show how Standard Thai consonant classes and tone markers correspond with Proto-Tai initial consonant classes and tones.

Proto Tai Initial Consonants at time of split	Standard Thai Consonants
Class 1. Voiceless Friction *s *hm *ph	"HIGH" ข ฉ ถ ฐ ผ ฝ ส ษ ศ ห (หม หน หง หย หร หล หว) kh ch th ph f s h m n ŋ y r l w
Class 2.Voiceless unaspirated stops *p *t	"MID" ก จ ต ฏ ป k c t p
Class 3. Glottal *ʔ *ʔb	บ ด ฎ อ b d ʔ
Class 4. Voiced *b *m	"LOW" ค ช ท ธ ฑ ฒ พ ฟ ช ฮ ม น ณ ง ย ญ ร ล ฬ ว kh ch th ph f s h m n ŋ y r l w

Proto Tai Tones	Standard Thai Tone Marks
A	No marks ex. ปา paa 'throw'
B	◌่ máy ʔèek ex. ป่า pàa 'forest'
C	◌๋ may thoo ex. ป๋า pâa 'aunt'

Figure 3. Proto Tai and Standard Thai consonant classes and tones

The following (Figure 4.) show the word lists, based on the Tone Box, that we use for collecting dialectal word samples.

Cons Class	Live Syllables A	Live Syllables B	Live Syllables C	Dead Syllables Long V. (DL)	Dead Syllables Short V. (DS)
1.	ขา khaa 'leg' หู huu 'ear' หัว hua 'head' หมา maa 'dog'	เข่า khaw 'knee' ไข่ khay 'egg' ผ่า phaa 'to cleave' ใหม่ may 'new'	หน้า naa 'face' เสื้อ sɯa 'blouse/shirt' ข้าว khaaw 'rice' ห้า haa 'five'	ศอก sɔɔk 'elbow' สาก saak 'pestle' หาบ haap 'carry with pole'	ผัก phak 'vegetable' สุก suk 'ripe, cooked' ขุด khut 'dig' หมัด mat 'flea'
2.	ตา taa 'eye' กิน kin 'eat' กา kaa 'crow' ปี pii 'year' ปลา plaa 'fish'	ไก่ kay 'chicken' เต่า taw 'turtle'	ก้าง kaaŋ 'fish bone' เก้า kaaw 'nine' ป้า paa 'older aunt' ใต้ tay 'underneath' ต้ม tom 'boil'	ปาก paak 'mouth' ตาก taak 'to dry' ปีก piik 'wing' แปด pɛɛt 'eight'	เป็ด pet 'duck' เจ็ด cet 'seven' กบ kop 'frog' เจ็บ cep 'to be hurt' ตัก tak 'to scoop'
Cons Class	A	B	C	DL	DS
3.	บิน bin 'fly' ดำ dam 'black' เอว ʔeew 'waist' ดาว daaw 'star' แดง dɛɛŋ 'red'	บ่า baa 'shoulder' อิ่ม ʔim 'to be full' บ่าว baaw 'young man/groom'	อ้า ʔaa 'to open' บ้า baa 'to be mad'	แดด dɛɛt 'sunlight' ดอก dɔɔk 'flower' บอด bɔɔt 'to be blind' อาบ ʔaap 'to bathe'	เบ็ด bet 'fishing rod' อก ʔok 'chest' ดิบ dip 'to be raw'
4.	มือ mɯɯ 'hand' นา naa 'face' งู ŋuu 'snake' ลุง luŋ 'uncle' ควาย khwaay 'buffalo'	พ่อ phɔɔ 'father' แม่ mɛɛ 'mother'	น้า naa 'younger aunt' ไม้ may 'wood' น้อง nɔɔŋ 'younger sibling' ม้า maa 'horse'	มืด mɯɯt 'to be dark' ลูก luuk 'offspring' เลือด lɯat 'blood' มีด miit 'knife'	มด mot 'ant' วัด wat 'temple' นก nok 'bird' ซัก sak 'to wash' มัด mat 'to tie'

Figure 4. Word list used in collecting dialect samples

The word sets were collected in May-August 2001 from the four major regional dialects in Thailand in addition to the Bangkok Thai or Standard Thai: A dialect that is a prominent dialect in each region was collected as a representative dialect.

1) Northern Thai (Kam Muang): represented by a dialect in Chiangmai province..
2) Northeastern Thai (Isan): represented by a dialect in Khonkaen and Mahasarakham
3) Southern Thai (Pak Tai): represented by a dialect in Nakhon Sithammarat.
4) Central Thai: represented by a dialect in Suphanburi

In addition to the word sets, Thai numbers in the above five dialects are also collected.

Our tone box word list comprises 78 words. For each dialect, we collected from six male and six female informants. Therefore, the total number of words collected are 4,680 words. As for the numbers, we collected 27 number words (0-10, 11, 20, 21, 30, 31, 40, 50, 60, 70, 80, 90, 100, 1000, 10000, 100000, 1000000) from the five dialects, from one male and one female informant.

The word sets were digitized into Wave (.wav) format. An interactive search program to access and listen to an audio file of a word in the word sets is provided. Users can set the following parameters to select an audio file (see Figure 5): a word from the tone box(1), region (2), gender of speakers (3) before clicking button 4 to submit the choices and clicking button 5 to play the audio of the desired word.

This search program was written in Javascript, therefore it works with a media player in any standard web browser such as Internet Explorer and Netscape.

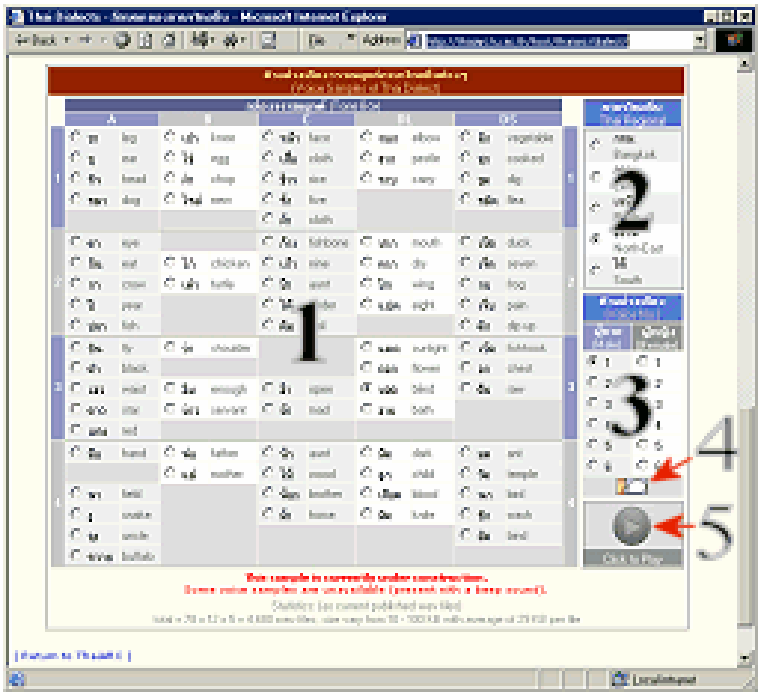

Figure 5: Interactive program for tone word sets

In addition, accompanying texts and excerpts from articles in Thai give an overview of the phonological characteristics of each regional dialect, the tone box, the process of tone word set collection are also provided.

2) **Thai Regional Folktales**

The samples of Northern and Southern folktales are selected and digitized from the audiotaped folktales collection of Social Research Institute of Chiangmai University and Thaksin Khadi Institute of Prince of Songkhla University respectively. The samples of Northeastern folktales are selected from the collection of Dr. Wajuppa Tossa of Mahasarakham University and collected

from the informants. The samples of central Thai (Suphanburi) Folktales are collected from informants.

Each folktale is presented as verbatim transcript in Thai orthography with a translation into Standard Thai.(see Figure 6). The verbatim transcript of each tale is presented in the left column and the Standard Thai translation in the right column. Users can select to listen to the tales in paragraph units or as whole tales. The audio file formats available are: Wave (.wav), Real Audio (.ra) and MPEG3 (.mp3).

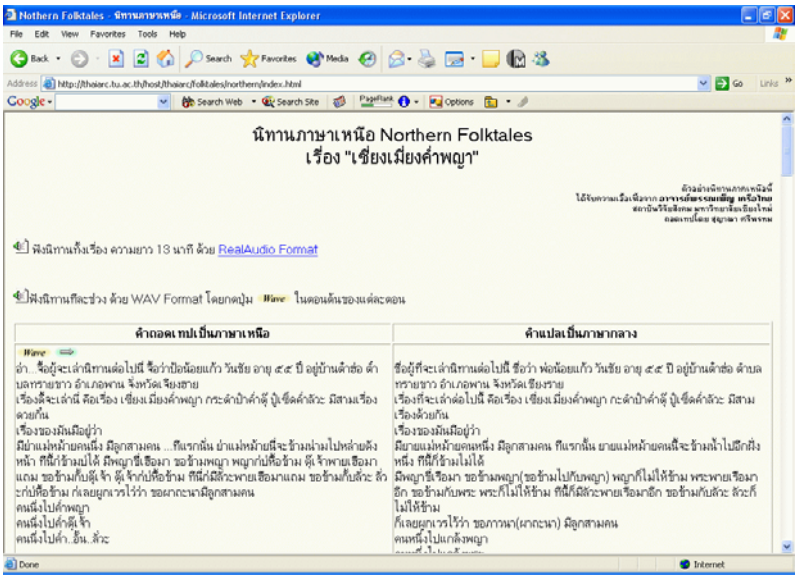

Figure 6. Northern Folktales page

3) Thai Poetry

Sample pieces of various types of versification of Thai poetry were selected from a Thai reference grammar by Thonglo (1972) and spoken by Thavorn Sikkhakosol, a lecturer at Thammasat University. Accompanying texts include diagrams of versification pattern of each type and excerpts from articles in Thai and English on Thai poetry. The audio files are in both wave and Real Audio format. Figure 7 shows a sample Thai poetry page.

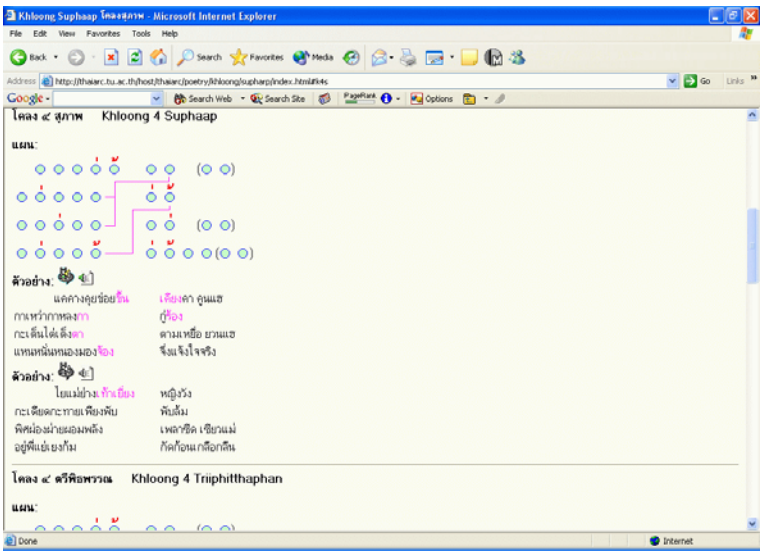

Figure 7 A page showing Khlong versification pattern and examples.

4) Speech Styles:

The audio samples of the speech styles which are available are:

1. the King's speeches for various occasions, for example, his Golden Jubilee Speech.
2. three types of news broadcasts: weather forecast, local news and foreign news.

A verbatim transcript are provided for each speech sample.

Sample Application of the Thai Speech Database: Web-based Instruction of Thai Listening Comprehension

Nowadays, the computers' role in education is well recognized as a new medium for presenting and delivering lessons. A commonly used term to describe such a role is "Computer Aided Instruction" or CAI. In the context of language learning/instruction by using computers, "Computer Assisted Language Learning" or CALL is the widely used term. This encompasses various technologies available for use with computers such as using computers with educational software on CD-ROMs, using computers to play various multimedia files (audio, video clips) and using networked computers to access language resources and data via the World Wide Web.

ThaiARC as a Thai Speech database that is accessible via the web can be a useful resource in teaching Thai to foreigners in addition to being a Thai language resource for researchers or the general public. The authentic speech samples of various genres and dialects of spoken Thai are available and their accompanying texts can be applied as teaching materials for various language skills, especially in listening comprehension skills.

We have explored the web technologies and their optimal uses in language instruction and have designed a sample web-based lesson in listening comprehension skill, using the available data from ThaiARC. In designing the web-based lesson, we have taken into account the features of web technologies and the language teaching methodology that can make full use of the simple web technologies available. To ensure accessibility for all users, the sample web-based lesson makes use of commonly available freeware and shareware.

Potentials of Web Technologies as Instructional Media

Web-based instruction is normally thought of as one type of Computer Aided Instruction. However, Horton (2000) pointed out the following differences between Web-based instruction and other types of CAI:

Typical CAIs are disk-based instruction: they use a computer in running instructional software on CD-ROMs which are normally interactive multimedia lessons, i.e., there are texts, graphics, animations, audio clips and also video-clips in a lesson; learners can get automatic feedback from lessons and exercises. However, computers are used as "standalone" computers this way, i.e., there is no communication and interaction among users via the computers in use - whether among learners or between instructors and learners. In addition, the content and exercises on the CD-ROMs are static and cannot be added to or changed or deleted.

On the other hand, to elaborate on Horton (2000: 19), web-based instruction has the following advantages over typical CAIs:

1) Centralized storage and maintenance: The web is used as storage for course lessons, exercises and data, and also the site for dissemination and delivery of instruction. The content of the lessons and exercises can easily be updated and changed anytime from anywhere by the authorized web maintainer.

2) Access to web-based resources: We can make use of the hypermedia link feature of the web page to access resources on other websites such as libraries, museums, articles from journals, newspapers, magazines, radio and television broadcast programs on the web, and so forth.
3) Collaboration mechanism: The web provides various means for learners to communicate with fellow learners, instructors and other web users. The communication could be "synchronous" (real-time or live) communication such as chat by text or by voice in chatrooms, or "asynchronous" communication such as e-mail.

The majority of language lessons and instructions available via the web at present do not fully utilize the full capability and potentials of the web; they only use the web as the media to disseminate interactive multimedia lessons and exercises in the same way that CD-ROMs are used as media.

Limitations of Web Technologies as Instructional Media

In maximizing the potential of the web in language learning, instructors should assess the web features that suit both the learners' technical and language skills and expertise. It would be to the learners' disadvantage and a waste of their effort for them to try to use the media without a grasp of the web technologies employed in the lessons and exercises. The following skills and expertise of learners, based on Horton's (2000: 337) observations, should be considered in designing web-based instruction:

1) Language fluency: Some types of communication requires high language proficiency, especially in the real-time or synchronous communication inherent in voice chat, audio- conferencing or video-conferencing. On the other hand, asynchronous communication such as e-mail or web-board posting allows users time to compose and revise a message, so they require a lesser degree of language proficiency. This point is especially important for web-based language lessons, which should choose among these modes of communication for those appropriate to the language level of learners.
2) Sound quality and accents: This is a point to consider in web-based lessons/activities that use sound or voice communication. The audio quality tends to degrade when transmitted over the internet. This, coupled with various accents in pronunciation that might not be familiar to learners can cause difficulty in comprehension and communication on the learners' part.
3) Typing skill: This is a point to consider in web-based lessons/activities that use the keyboard in communication such as e-mail, web-board posting, text chat, especially with the "real-time" communication by keyboard called for in text chatting. Learners need to have adequate typing skills in order to communicate effectively with keyboards. This should be of special consideration in web-based instruction of foreign language with non-typical scripts or keyboard layouts.
4) Web/Internet technical expertise: Each web technology may require different levels of internet technical skills from users, for example, web browsing and e-mailing require less technological skill than chat or voice e-mailing. Lessons and activities in web-based instruction should suit the learners' level of skill and comfort in these technologies.

A sample web-based Thai listening comprehension lesson.

The target audience of the sample listening comprehension lesson are Thai learners at intermediate high or advanced low level who can read and write simple Thai paragraphs. The listening skill lesson will acquaint the learners with one genre of authentic everyday Thai speech that can be accessed from the sources on the web. We have selected weather forecast news for use in this

sample lesson, which is a type of news speech styles that are available as sample Thai speech disseminated via ThaiARC.

Weather forecast news has many desirable characteristics for use as an authentic language material in a listening comprehension lesson. It has a number of formulaic expressions and a well-defined vocabulary range, with a fixed sequence of presentation. Therefore most of the information is predictable. Additionally, the information can be useful in daily life such as in trip planning. In addition, weather forecast news is an authentic piece of information that we have access to everyday.

The weather forecast news used in the lesson is from a broadcast from Radio Thailand as part of Radio Thailand daily morning news. Learners are introduced to vocabulary and expressions commonly used in talking about weather and various regions of Thailand.

The sample web-based lesson is located at: **http;//thaiarc.tu.ac.th/host/thaiarc/wbll/intro.htm**. There are four parts in this sample lesson.

Part 1: Home Page: Gives introduction to the lesson components, the objectives of the lesson and the sequence of activities in the lesson. (Figure 8)

Part 2: Vocabulary and Expression Explanation and Practice. (Figure 9) Learners can listen to each word and expression introduced. A map of Thailand is also presented as a visual aid. (Figure 10)

Part 3: Exercises: A set of 28 questions (multiple choice) in English to test the comprehension of the weather forecast news presented in the lesson. Learners will get immediate feed back from the system. (Figure 11)

Part 4: Assignment. Learners are directed to Radio Thailand web page to listen to the actual weather forecast of the day. (Figure 12)

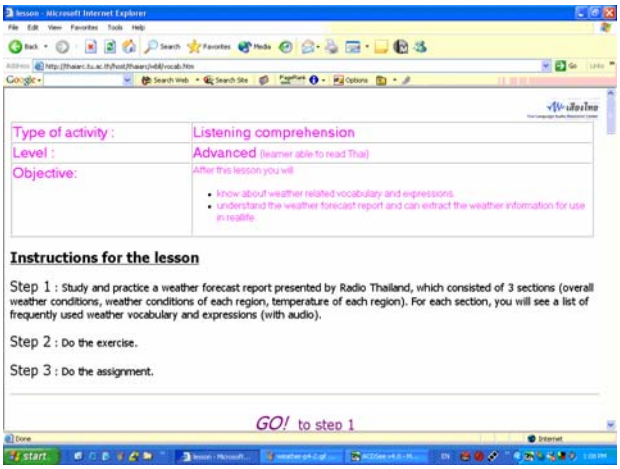

Figure 8. Lesson Introduction.

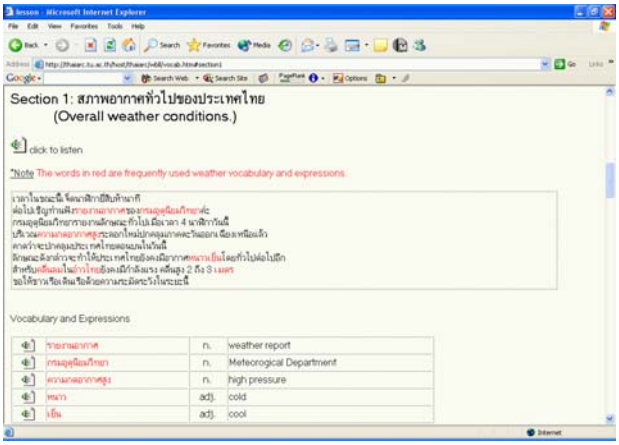

Figure 9**.** Vocabulary and expressions

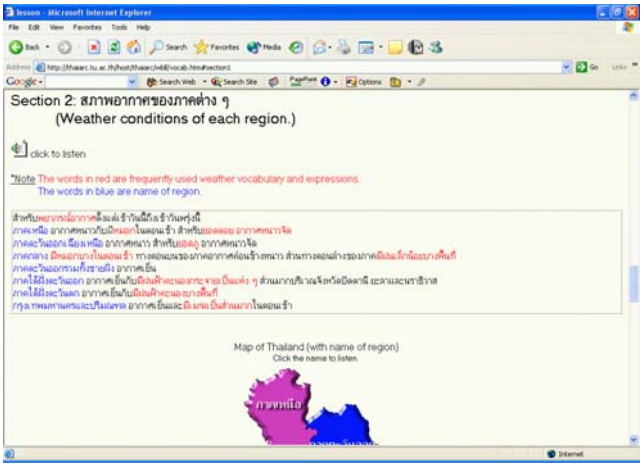

Figure 10. Weather news transcript with map of Thailand.

Figure 11. Comprehension test

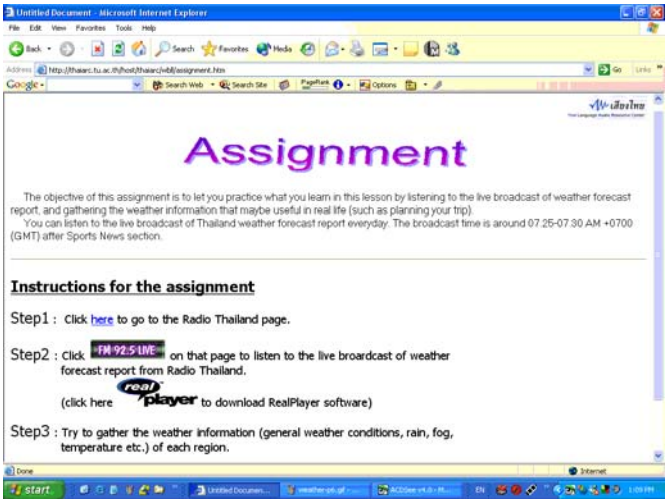

Figure 12. Assignment

Closing Remarks

We have outlined above the development of the Thai Language Audio Resource Center (ThaiARC) and the Thai speech data available at the ThaiARC website at: http://thaiarc.tu.ac.th. We invite researchers of Thai and those interested in Thai to visit the site and make use of the Thai speech data that we have provided for you.

Notes

ThaiARC research project was supported by Computer Network Research and Development Grant, National Electronics and Computer Technology Center, National Science and Technology Development Agency of Thailand. We would like to thank the Information Processing Institute for Education and Development of Thammasat University for hosting ThaiARC's website and providing technical support for the site. We also would like to acknowledge the grant from University Research Council of the University of Hawaii which enabled the first author to present this paper at SEALS 13th Annual Meeting.

References

Gedney, William. (1972) A checklist for determining tones in Tai dialects. In M. Estelle Smith (ed.) *Studies in linguistics in honor of George L. Trager*.(pp. 423-437) The Hague: Mouton.

Hoonchamlong, Yuphaphann et al. (2002a). *Thai Language Audio Resource Center (ThaiARC) Project Phase II.* (in Thai)**.** Unpublished Technical Report. National Electronic and Computer Technology Center (NECTEC) of Thailand.

Hoonchamlong, Yuphaphann et al. (2002b) Thai Language Audio Resource Center Project: Thai Speech Database and Application in Web-Based Language Teaching.(in Thai) *Proceedings of Annual Conference of National Electronic and Computer Center (NECTEC) August*

19th -21st 2002. Bangkok: NECTEC.
Horton, William. (2000) *Designing Web Based Training* . New York: John Wiley& Sons.
Thonglo, Kamchai (1972). *Principles of Thai Language*. (in Thai) Bangkok: Ruamsarn.