Computational Linguistics and Chinese Language Processing
Vol. 17, No. 1, March 2012, pp. 1-26

Development and Testing of Transcription Software for a Southern Min Spoken Corpus

Jia-Cing Ruan*, Chiung-Wen Hsu*, James Myers*, and Jane S. Tsay*

Abstract

The usual challenges of transcribing spoken language are compounded for Southern Min (Taiwanese) because it lacks a generally accepted orthography. This study reports the development and testing of software tools for assisting such transcription. Three tools are compared, each representing a different type of interface with our corpus-based Southern Min lexicon (Tsay, 2007): our original Chinese character-based tool (Segmentor), the first version of a romanization-based lexicon entry tool called Adult-Corpus Romanization Input Program (ACRIP 1.0), and a revised version of ACRIP that accepts both character and romanization inputs and integrates them with sound files (ACRIP 2.0). In two experiments, naive native speakers of Southern Min were asked to transcribe passages from our corpus of adult spoken Southern Min (Tsay and Myers, in progress), using one or more of these tools. Experiment 1 showed no disadvantage for romanization-based compared with character-based transcription even for untrained transcribers. Experiment 2 showed significant advantages of the new mixed-system tool (ACRIP 2.0) over both Segmentor and ACRIP 1.0, in both speed and accuracy of transcription. Experiment 2 also showed that only minimal additional training brought dramatic improvements in both speed and accuracy. These results suggest that the transcription of non-Mandarin Sinitic languages benefits from flexible, integrated software tools.

Keywords: Speech Transcription, Southern Min, Taiwanese, Romanization, Key-in Systems.

* Graduate Institute of Linguistics, National Chung Cheng University, Minshiung, Chiayi 62102, Taiwan
Telephone: (05) 272-0411 ext. 21510; Fax: (05) 272-1654
E-mail: Lngmyers@ccu.edu.tw
The author for correspondence is James Myers.

1. Introduction

1.1 Constructing a Southern Min Speech Corpus

As with any language, corpora of spoken Southern Min (Taiwanese) have many uses, both scientific and practical. Corpora of written Southern Min exist (e.g., Iunn, 2003a,b, 2005, based on novels, prose, dramas, and poems; the Southern Min Archives of Academia Sinica, 2002; Ministry of Education, 2010, with word frequency statistics), but Southern Min, unlike Mandarin, is virtually never written at all. For this reason, there has been increasing interest in corpora of spoken Southern Min, including the NCCU corpus of spoken Chinese (Chui, 2009), which includes everyday conversation in Southern Min, and ForSDat (Formosa Speech Database) of Lyu, Liang, & Chiang (2004), which is a multilingual speech corpus for Southern Min, Hakka and Mandarin.

One area where a spoken corpus is essential is in the study of first language acquisition. This consideration motivated the construction of the Taiwanese Child Language Corpus (TAICORP; Tsay, 2007), which contains about two million morphemes in half a million utterances, based on about 330 hours of recordings of spontaneous conversations between children and their caretakers. Speech corpora are also essential for understanding the use of language in adult conversation, motivating our corpus of adult spoken Southern Min (Tsay & Myers, in progress), based on spontaneous conversations from radio broadcasts in Chiayi county. Except for the coastal towns, the majority of the population (including the hosts and guests in the radio programs recorded) in this area speak a variety of Southern Min historically derived from that spoken in Zhangzhou in Southern Fujian, although due to language contact over the years this variety has been mixed with the other variety historically derived from Quanzhou Southern Min. As of December 2011, the completely double-checked and confirmed portion of this corpus has almost 800,000 word tokens (詞), based on about 3,800 minutes of recordings.

Both TAICORP and the Taiwanese Spoken Corpus are transcribed in cognate Chinese characters (本字) wherever applicable, and otherwise in the romanization system of the Ministry of Education (MOE), Taiwan (Ministry of Education, 2008). The most important features of the MOE transcription notation for the present discussion are the marking of coda glottal stop with “h” (e.g., 肉 <bah4> ‘meat’), the marking of vowel nasality with “nn” (e.g., 甜 <tinn1> ‘sweet’), and the marking of tone categories with digits (e.g., 詩 <si1> ‘poem’ vs. 時 <si5> ‘time’).

These two corpora have been used to generate a lexical bank, which as of December 2011, has approximately 20,000 entries. Each entry contains four elements (see Table 1): (1) the word written in Chinese characters (or romanization if no corresponding characters exist), with homographs distinguished with numerals; (2) the pronunciations in romanization

(including possible alternative pronunciations, typically due to borrowings from the Quanzhou variety of Southern Min); (3) near-synonyms or an explanatory definition in Mandarin; and (4) an example. Elements (3) and (4) are used to disambiguate homographic or homophonic entries.

Table 1. Sample entries in Southern Min lexicon.

Characters	Pronunciation	Explanation	Example
愛 1	ai3	喜歡、愛	你 有 愛 1 食 糖仔 oo02。
愛 2	ai3	需要(加單賓)	這 1 愛 2 兩 1 支 la0。

1.2 Challenges in Transcribing Southern Min

The usual challenges of transcribing a spoken language are compounded for Southern Min because it lacks a conventionalized orthography. With sufficient training in any adequate orthography, character-based or romanization-based, it should be possible for a native transcriber to write Southern Min as easily as Mandarin. Thus it is essential for Southern Min transcription to be assisted by some sort of automated orthography checker, to confirm that transcribers are consistent and to give hints when they get stuck.

The Southern Min lexicon we have been developing plays a key role in this orthography checking. Any entry can be accessed either via Chinese characters (if available) or via romanization, and once it is accessed, the explanation can confirm to the transcriber that the intended entry has been found. If an entry is not found, this either means that the transcriber has misspelled the word, or that the word has not previously appeared in the corpus.

For several years, transcribers for the Taiwanese Spoken Corpus have relied on a set of independent software tools developed for TAICORP (designed by James Myers and Jane Tsay, and written by Ming-Chung Chang and Charles Jie): a lexical access tool, a transcription tool, and a segmentation tool. For convenience we will call this package of tools Segmentor. As described in Tsay (2007), Segmentor requires the user first to transcribe speech into Chinese characters (wherever possible), and then run a program to segment the character strings into words defined by the lexicon, resulting in segmented text as shown in Appendix C, where each word is represented both in characters and in romanization within < > brackets. If any mistake is found at this point (i.e., if the program cannot find a word in the lexicon), the transcriber performs the above process again. Initial transcription is in Chinese characters, rather than Southern Min romanization, because we assumed that our student transcribers have many years of experience using Mandarin key-in systems and no experience with a systematic Southern Min key-in system.

However, transcribing Southern Min using Chinese characters has a number of shortcomings. First, transcribers must choose the correct Chinese characters (本字), which

may be low-frequency characters in Mandarin, even for high-frequency Southern Min morphemes (e.g., 囥 <khng3>, glossed as “放”, “to put/place/lay”). Second, most transcribers use phonetic key-in systems for Chinese characters, so they must mentally activate the Mandarin pronunciation, not the Southern Min pronunciation, to key in a character. Third, even if the characters are familiar from Mandarin, the Southern Min compound may not be, so they cannot rely on word auto-completion tools (e.g., 鐵齒 <thih4khi2>, glossed as “不聽勸/不信邪”, “stubborn” is a compound in Southern Min but not in Mandarin). Fourth, there are many common words in Southern Min that have no Chinese character form at all (e.g., chit4tho5 “to play”).

Segmentor also has limitations of its own. First, although the segmented text shows the romanization, this can only help transcribers uniquely identify words if they clearly recall which tone digit goes with which tone category, but we have found that native speakers have great trouble doing this. Second, because Segmentor only supports ANSI format text files, while the lexicon file is in UTF-8 format, it does not support Southern Min morphemes that must be written with Chinese characters outside of the traditional Mandarin set. Although this problem can be solved by incorporating Unicode BuWanJiHua (http://uao.cpatch.org/), the resulting transcription still cannot be properly handled by the segmentation tool, since its server settings support only Big5, not UTF-8. Finally, the source code of the segmentation program is no longer available for updating.

The purpose of this study, then, was to develop a new tool for transcribing Southern Min. Our intuition was that transcription might be more efficient if the student assistants could transcribe text word by word, rather than relying on a segmentation program, and directly in Southern Min romanization, rather than indirectly via Mandarin. Because new assistants have no prior experience writing a standardized Southern Min romanization system, a new software tool must provide considerable assistance. In particular, the tool cannot require users to enter tone digits, which are very hard to remember, and should use auto-completion so that users need only enter part of a compound word for it to be accessed from the lexicon.

In 2010, during the period of our study, the Ministry of Education released an input system for transforming Southern Min romanization into cognate Chinese characters (本字, or 漢字 in their terms); see Ministry of Education (2012) for the latest version of this system. The MOE is to be applauded for producing a very useful and flexible writing tool. However, it does not suffice for the transcribers of spoken corpora, who would benefit from being able to interact directly and simultaneously with sound files, the written corpus, and full lexical entries (including both character and romanized transcriptions, as well as other information for distinguishing among homonyms). In the remainder of this paper, we describe the development of just such a system (ACRIP), and demonstrate its effectiveness in experiments on naive participants learning to transcribe with it.

2. Adult-Corpus Romanization Input Program (ACRIP)

The key weakness of romanization input is that it requires student transcribers to be very familiar with the MOE Southern Min romanization system, and to be consciously aware of phonemic contrasts that do not exist in Mandarin, and hence are not associated with writing in their usual experience (despite their fluency with perceiving and producing Southern Min aurally and orally). The Adult-Corpus Romanization Input Program (ACRIP) helps transcribers in a number of ways when using the romanization system, by exploiting our large and growing corpus-based dictionary of Southern Min. The program was written by the first author in Microsoft Visual Basic 6.0, running in Microsoft Windows.

2.1 ACRIP Architecture

The architecture of ACRIP is presented in Figure 1.

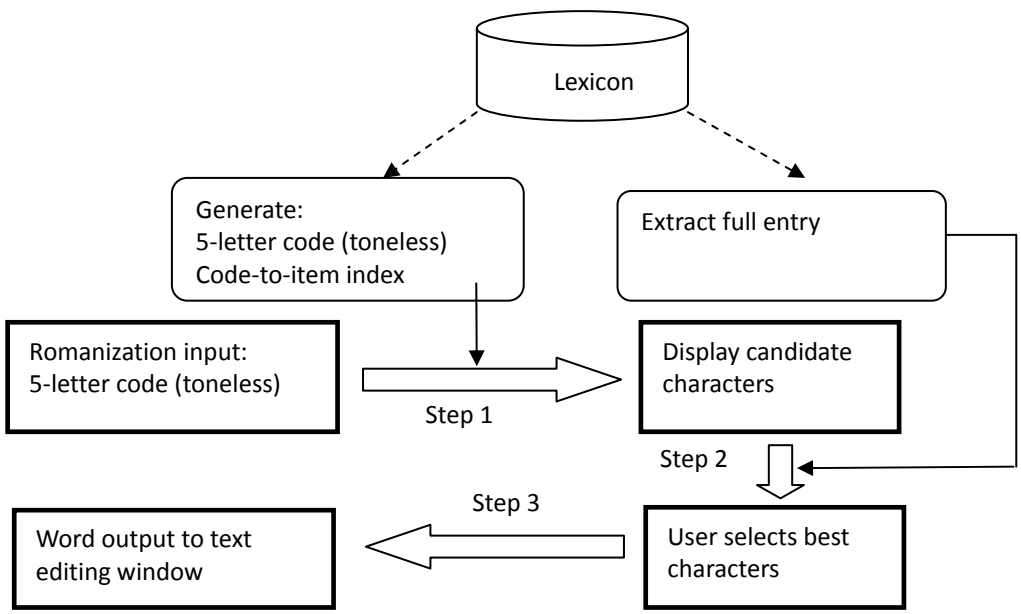


Figure 1. ACRIP architecture diagram.

The original corpus-based lexicon was edited to add a code of up to five letters for each entry, and a code-to-item index was established to link codes to candidate character-based entries, which were then linked to the other three elements of the entry (details are described in section 2.3). Each code is simply the first letters (up to five) of the romanization of a word, thus permitting a form of auto-completion: users only need to enter short strings of letters, without tone digits, to access full Southern Min words. More precisely, by entering a code, users get a list of candidate items, and then select the best item as the output according to the

other elements in the entry (including explanation and example). When new entries are added to the lexicon, the coding can be updated automatically using an Excel macro.

2.2 The Main Interface for ACRIP 1.0

ACRIP integrates many functions for the transcription of Southern Min. The first version of this program, ACRIP 1.0, has the main interface shown in Figure 2 (ACRIP 2.0 retains the same functions, but adds others).

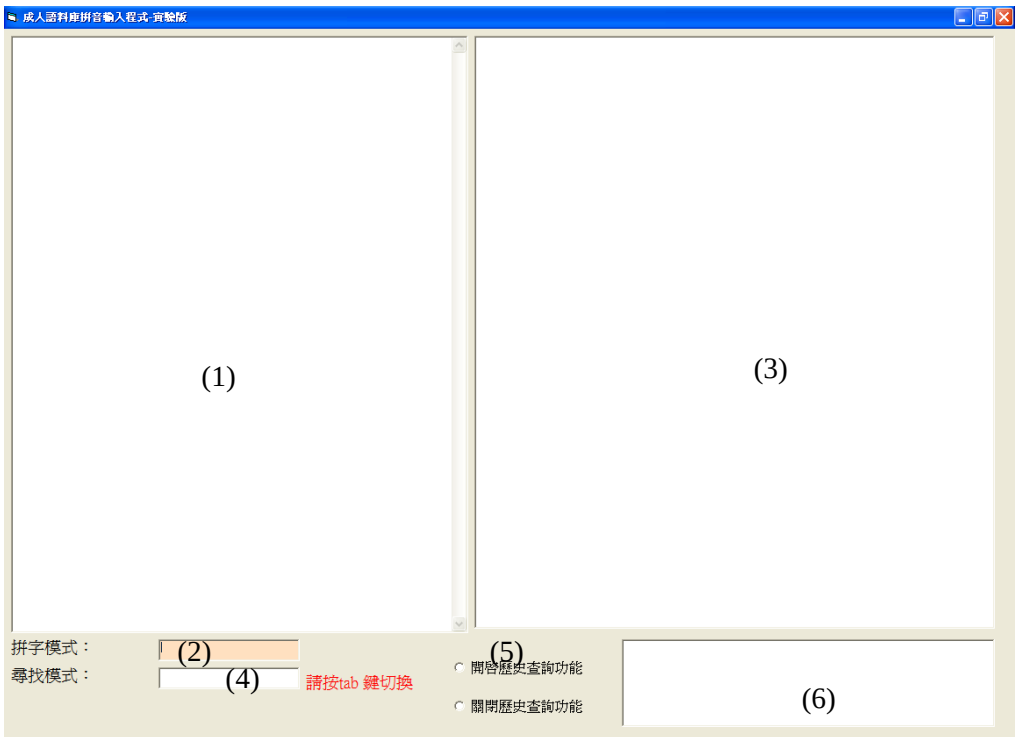


Figure 2. The main interface of ACRIP 1.0.

In contrast to the Segmentor tools, ACRIP integrates the three processes of accessing the lexicon, writing the transcription, and segmenting transcribed utterances into words, into a single interface. The corpus is transcribed by entering and checking one word (詞) at a time. The components of the ACRIP interface are as follows (identified by the numbers shown in Figure 2).

(1) Text editing window

This is the output window for segmented transcribed utterances (see Figure 3). The other components of ACRIP are designed to help the user fill this window with completed transcriptions. After transcriptions are complete, users can manually edit the contents of this window, or select the contents to copy or cut them to other editing programs.

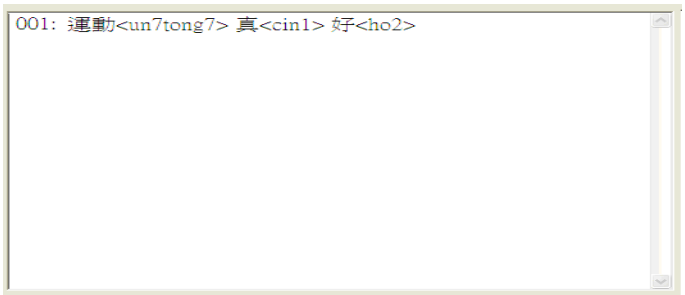


Figure 3. Window for text editing.

(2) Romanization search box

Transcribers enter up to five letters, without tone digits, to represent the word they hear in the spoken corpus. The words in the lexicon matching the first five letters will show up in the word candidate window. The example in Figure 4 shows the entry "unton", which is associated with the entry 運動<un7tong7>.

Figure 4. Text box for romanization input.

(3) Word candidate window

After entering a romanization code, all candidates in the lexicon with this code are shown in this window (see Figure 5). Users can then select the best candidate item to paste into the transcription being completed in the text editing window.

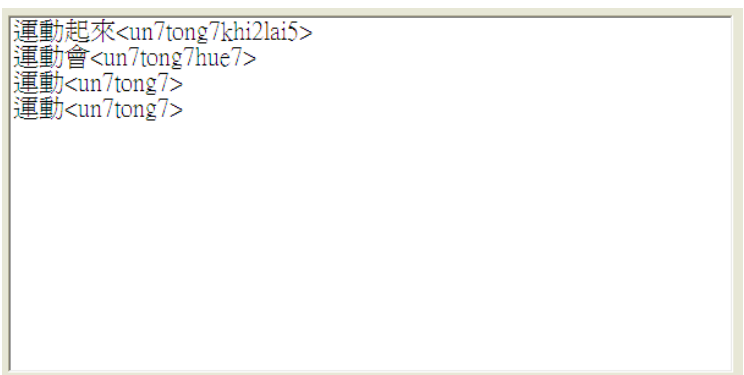


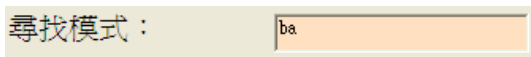
Figure 5. Window for candidate items

(4) Incremental romanization search box

This provides letter-by-letter search of romanization code for beginning users. This tool is helpful because pilot studies showed that the most difficult segments to perceive were the voiced onset obstruents (e.g., /b/ for 賣<be2> "sell", /g/ for 牛<gu5> "cow") and voiceless

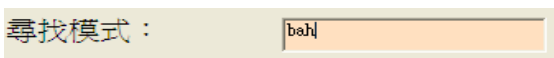
coda stops (e.g., /p/ for 汁<ciap4> “juice”, /t/ for 結<kat4> “knots”, /k/ for 角<kak4> “chunk”, glottal stop for 肉<bah4> “meat”). For example, transcribers often have trouble hearing glottal stop codas, as in the word 肉 (correctly transcribed in the MOE system as “bah4”). As shown in Figure 6, entering just the letters “ba” (a) only brings up the choices “ba5” (麻) and “ba7” (密) (b), immediately showing the transcriber that a coda is needed. Adding “h” (c) will then immediately change the list to the intended “bah4” (肉) (d).

(a) First two key presses:



(b) Resulting display:

(c) One more key press:



(d) Changed display:

Figure 6. Incremental romanization search.

(5) Toggle to save/erase work history

By turning on this function, users can avoid having to type the same code repeatedly for frequently occurring words. Instead, users can double-click strings in the work history to make them appear in the word candidate window. In the example shown in Figure 7, a user accessed the item 電腦<tian7nau2> by entering the code “tiann”. If the user needs to enter this item again, the user does not need to re-type the code, but can simply double click the string listed in the historical record. Users can also toggle this function off, erasing the work history.

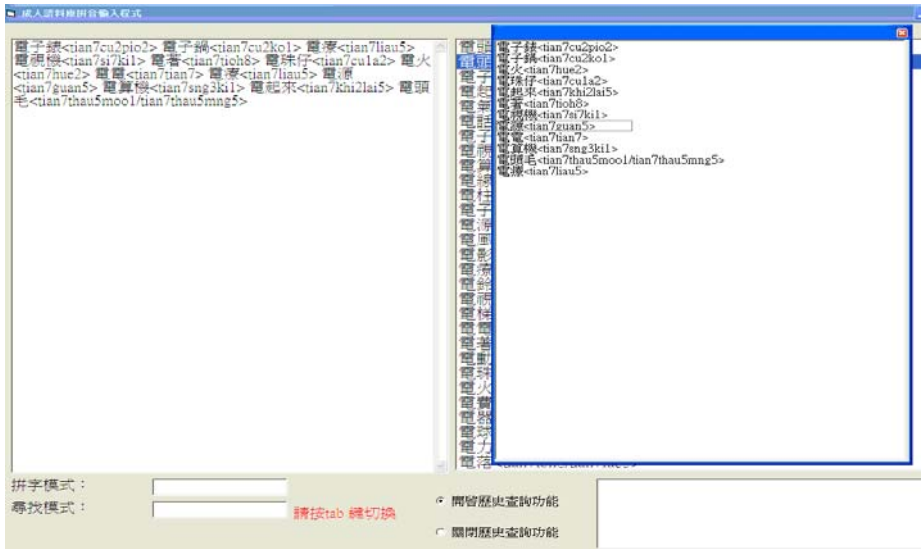


Figure 7. Using the history window.

(6) Pop-up lexical entry display window

After the list of candidate words has appeared in the candidate word window, there may be homonyms, as for example 愛 1 and 愛 2 shown earlier in Table 1. Prior to the development of ACRIP, transcribers would need to memorize the difference or to shift to a separate lexicon program to look them up. ACRIP's built-in lexical entry display window appears as a pop-up when users choose any item in the word candidate window and press the space bar. This tool helps disambiguate the intended word and saves time by not requiring users to change to a separate program or to retype items for lexical look-up (see Figures 8 and 9).

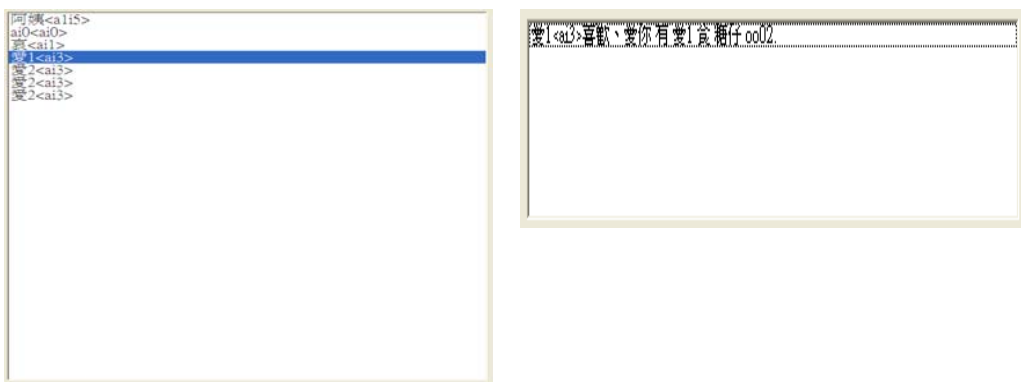


Figure 8. Looking up 愛 1

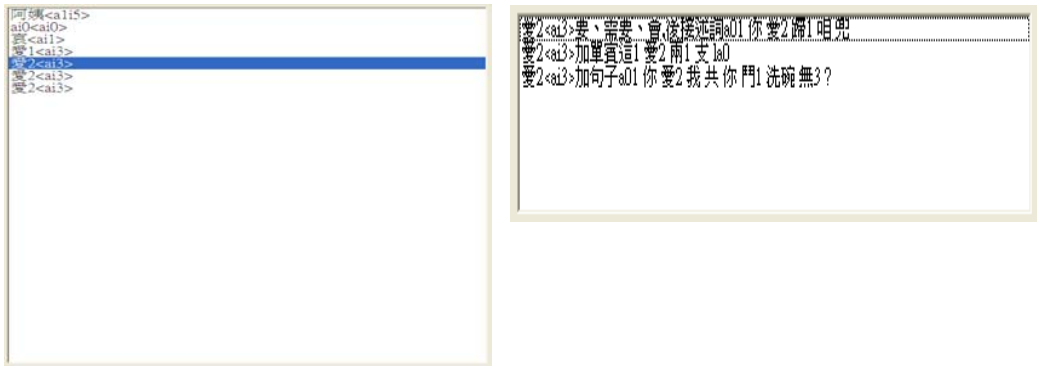


Figure 9. Looking up 愛2

2.3 Generation of the Romanization Input Codes and Code-to-item Index

In the development of ACRIP, the input romanization codes were generated from our original corpus-based lexicon by first deleting the tone digits and then extracting the first letters (up to five) as input code. This recoding was precompiled to speed up actual use of ACRIP (i.e., codes are stored in the lexicon rather than generated online).

One challenge faced when generating the input code was that the lexicon has many items that have alternative pronunciations, and therefore different romanizations, as shown in Table 2.

Table 2. Alternative pronunciations in a lexical entry.

Characters	Pronunciation	Explanation	Example
密密	ba7ba7/bat8bat8	滿滿	指緊密無縫

In this case, ‘baba’ and ‘batba’ are both codes for the entry ‘密密’. This problem was handled by editing the character and pronunciation elements of the lexical entries (using global replace in Microsoft Word and a macro in Microsoft Excel) to generate separate lexical entries for alternative pronunciations, so that each could be accessed separately.

After generating the romanization input code for each entry, we then incorporated them into the lexicon file using another macro in Microsoft Excel. The result was a file in which each lexical entry had a fifth element, representing the input code, as illustrated in Table 3.

Table 3. Revised lexical entries including romanization input code.

Input code	Characters	Pronunciation	Explanation	Example
baba	密密	ba7ba7	滿滿	指緊密無縫
batba	密密	bat8bat8	滿滿	指緊密無縫

3. Experiment 1: ACRIP 1.0 vs. Segmentor

In order to test whether ACRIP 1.0 improved the speed and accuracy of transcription of Southern Min using word-by-word romanization entry, we ran an experiment to compare it with the original Segmentor package for Chinese character transcription with post hoc segmentation. Naive native speakers of Southern Min transcribed short passages using both systems, and we examined the speed and accuracy of their transcriptions.

3.1 Methods

3.1.1 Participants

Twenty college students at National Chung Cheng University, who acquired Southern Min before kindergarten and without prior linguistic training, took part in the experiment. They were paid for their participation.

3.1.2 Design and Materials

The experiment had three phases: romanization training, romanization practice, and transcription testing. The romanization training phase used 30 nonlexical syllables that conformed to the phonotactic constraints of Southern Min (i.e., they were accidental gaps); see Appendix A. The romanization practice phase used 50 high-frequency Southern Min lexical items that together contain all of the segments and tone categories available in the phonological system of Southern Min (see Appendix B).

For transcription testing, two auditory passages were selected from the corpus of adult spoken Southern Min, Passages A and B; see Appendix C. Each passage was about 35 seconds long; based on piloting, we estimated that each would take less than an hour to transcribe. The two passages, which had already been transcribed and checked by our assistants, had roughly the same number of word tokens (Passage A: 129; Passage B: 122). The words were also matched in token frequency (based on our entire corpus), so we expected them to be approximately equal in transcription difficulty.

The transcription phase of the experiment used a Latin square design, balancing the presentation order of the two passages and the order of the two transcription systems across four groups of participants (five participants per group). Thus there was no confound among passage, order, or transcription method.

3.1.3 Procedure

In the romanization training phase, which lasted about an hour, the 30 nonlexical syllables were presented auditorily using Windows Media Player, and participant responses were made

by pen and paper. Feedback on correctness was immediately given by the experimenter (second author). The purpose of this phase was to familiarize participants with the contrasting onsets, vowels, codas, and tones of Southern Min, with special focus on codas (e.g., distinguishing glottal stop from /k/).

In the romanization practice phase, which also lasted about an hour, the 50 Southern Min words were presented in random order, both auditorily and visually, using E-Prime 2.0 (Schneider, Eschman & Zuccolotto, 2002). Participants were asked to transcribe the lexical items by typing romanization. Before they made their response, participants were allowed to play the word up to ten times. When they typed their response, subjects received feedback on the correctness of their transcription.

In the transcription testing phase, participants transcribed the two corpus passages, in their assigned order (see 3.1.2). Segmentor was used to transcribe using Chinese characters, with post-hoc segmentation, while ACRIP was used to transcribe word-by-word using romanization. All participants were given no more than one hour to transcribe each passage. Thus the entire experiment took approximately four hours for each participant.

3.2 Results

Separate by-participant analyses were conducted on transcription speed and accuracy. In both analyses, the independent variables were Passage (A vs. B) and Transcription System (Segmentor/characters vs. ACRIP/romanization). Our focus was on the effect of transcription system, with Passage included in the analysis merely to test for possible confounds.

The mean number of transcribed words (transcription speed) and percentage of mistranscribed words (error rate) are shown in Table 4.

Table 4. Mean number of transcribed words and percentage of mistranscribed words for the two transcription systems.

System	Transcribed words	Mistranscription rate (%)
Segmentor	92.15	36.94
ACRIP 1.0	83.85	38.11

Both measures formed normal distributions, so a parametric test was used. We chose linear mixed-effects regression modeling because it is more flexible than analysis of variance (Baayen, 2008). Passage and Transcription System (both within-participant) were coded as effect variables (i.e., their values were coded as -1 vs. 1), and their interaction was included in the analyses. As is standard with this test, we computed *p* values from Markov chain Monte Carlo samples (using the pvals.fnc function of the languageR package; Baayen, 2008) in R (R Development Core Team, 2011).

As shown in Table 4, the use of ACRIP 1.0 was associated with slightly fewer transcribed words than Segmentor and a slightly higher error rate, but neither difference was statistically significant (*p*s > .1). The only significant effect was a main effect for Passage on the number of transcribed words ($B = 12.4$, $p = .0001$), but this was merely because Passage A had more words (129) than Passage B (122). There were no other main effects and no interactions for either measure.

3.3 Discussion

The results showed no significant effects of transcription method on the number of transcribed words or transcription accuracy. Putting these null results in a positive light, we found no evidence that romanization-based transcription of Southern Min is inherently less efficient or error-prone than character-based transcription. Of course, these null results may also relate to a floor effect for both transcription methods: two hours of training, and one hour of transcription per passage, may not be enough for a naive transcriber to develop adequate competence, regardless of which system is used.

Each software tool has its own problems. As we mentioned earlier, Segmentor requires users to translate the heard Southern Min into Mandarin so that they can enter Chinese characters, and they also get feedback only as the segmentation tool is run, not word by word. Moreover, even after typing a word in Chinese characters, they may have to choose among a list of candidate Southern Min words distinguished partly by Southern Min romanization. Using Segmentor also requires users to enter the etymologically correct characters (本字), which are often unfamiliar to naive users (assuming any character form exists at all), so that it is not uncommon for them to type a semantically or phonologically related character instead of the correct one.

Nevertheless, ACRIP 1.0 has its limitations too. Although romanization entry solves the above problems in principle, naive transcribers are far more familiar with Chinese characters than with Southern Min romanization. Opinions on whether learning this romanization system is worthwhile seemed to be divided across the participants. After the experiment, a survey was emailed to participants to ask for their opinion about the two transcription tools. Of the five participants who replied, three acknowledged the efficiency of the romanization system and agreed that if they had had more practice with it, they would have been able to do the transcription more quickly with it than with Chinese character entry. However, the other two thought that using Chinese characters as input was more intuitive to them and saved time compared with correcting mistakes in their romanized entries.

4. ACRIP 2.0

Based on the results of Experiment 1, some novice transcribers still seem to need an option for Chinese character word entry. Therefore, we modified the input program to combine ACRIP 1.0 with the advantages of Segmentor, calling the new version ACRIP 2.0 (also written in Microsoft Visual Basic 6.0 by the first author). The main interface of ACRIP 2.0 is shown in Figure 10.

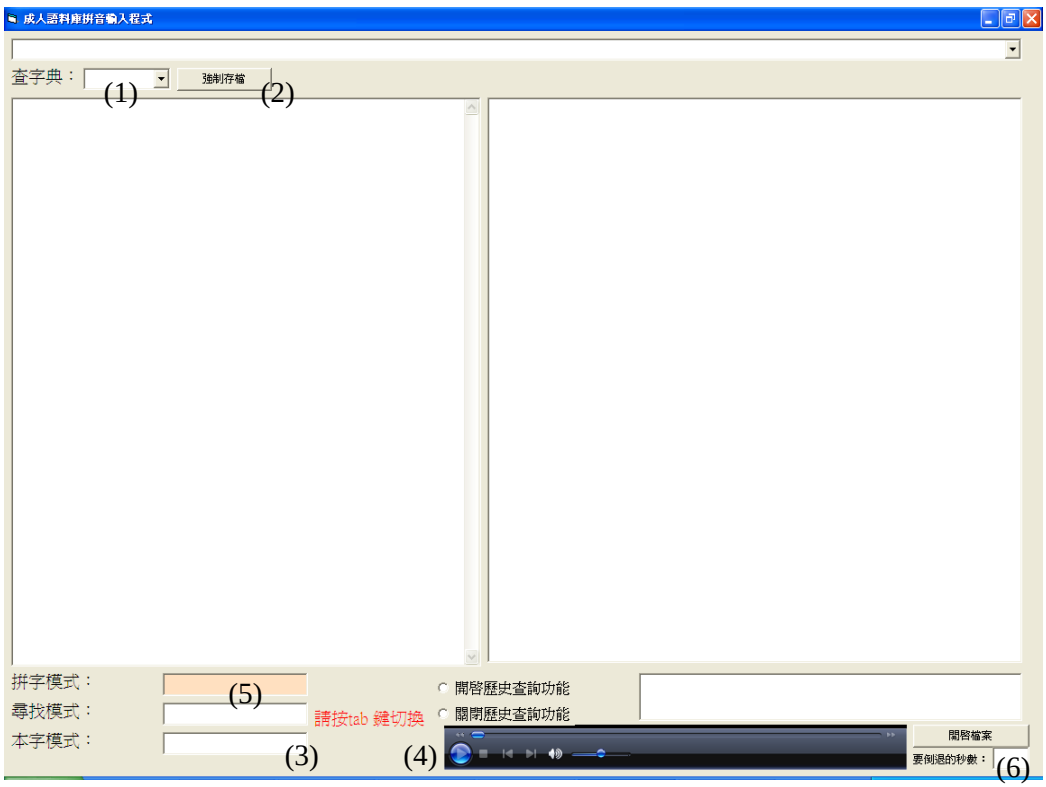


Figure 10. The main interface of ACRIP 2.0.

ACRIP 2.0 maintains all of the components of ACRIP 1.0, but adds the following new ones (see number labels in Figure 10).

(1) Integrated lexicon search box

Users can use this function to look up an item in the Southern Min lexicon by entering any of the four elements of an entry: Chinese characters, Southern Min romanization, Mandarin near-synonyms, or the explanatory example or definition (see Figure 11).

Figure 11. Looking up 愛2 in the integrated lexicon interface

(2) Auto-save into the editing area

For safety, this new function allows users to save data in the text editing window at any time. In addition, an automatic function operates invisibly to save data in the text editing window whenever any changes are made in this window.

(3) Incremental Chinese character search box

This provides a fuzzy search for lexical entries via the first character of the Chinese character element. For example, as shown in Figure 12, if a user enters “電” (a), the output list will be all items in the lexicon with Chinese character elements beginning with “電” (b).

(a) Character insertion:

(b) Resulting display in candidate item window:

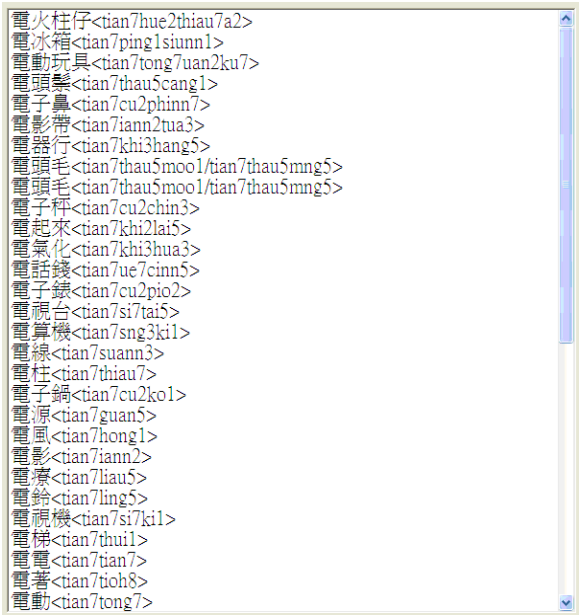


Figure 12. Incremental Chinese character search.

(4) Integrated Microsoft Windows Media Player

ACRIP 2.0 interfaces directly with Microsoft Windows Media Player so that users can play the portion of the audio file that they are currently transcribing.

(5) Play/stop the sound file

This function is attached to the romanization search box, and permits readily accessible keyboard control. When users press ESC, Microsoft Windows Media Player will play the sound file, and when they press ESC again, Microsoft Windows Player will stop playing.

(6) Automatic rewind timer

This function provides an automatic rewind operation which saves users the trouble of having to rewind sound files manually while replaying speech files during transcription. For example, if the timer is set to 3 seconds, when the sound file is off and users press ESC, Microsoft Windows Media Player will automatically rewind 3 seconds before replaying the speech file.

ACRIP 2.0 is intended to create a unified environment for the transcription of speech files. We observed that when using ACRIP 1.0, naive transcribers frequently needed to shift from this program to Microsoft Windows Media Player (in order to press the play/stop button and locate the time point they would like to replay in a speech file), and to the dictionary files (to look up items in Chinese characters when they did not know the Southern Min romanization). ACRIP 2.0 is designed to minimize the time needed to switch between these tasks: users first set up a default rewind time in the timer (6), and operate (4) and (5) via the ESC key (thus saving even more time by avoiding the need to use the mouse).

By permitting Chinese character search, including fuzzy search, and integrating Microsoft Windows Media Player for playing back speech files, users have more flexibility in entry options, have more powerful help tools, and can save time by not having to shift to other programs.

5. Experiment 2: ACRIP 2.0 vs. Segmentor and ACRIP 1.0

We hoped that the added features of ACRIP 2.0 would make it a much more efficient tool than either ACRIP 1.0 or Segmentor. To test this, we asked a new set of naive native speakers of Southern Min use ACRIP 2.0 to transcribe the same passages tested in Experiment 1. We also tested whether additional training brought any further improvements in speed and/or accuracy with using ACRIP 2.0.

5.1 Methods

5.1.1 Participants

Twenty college students at National Chung Cheng University, who acquired Southern Min before kindergarten and without prior linguistic training, took part in the experiment. None of the participants in Experiment 2 took part in Experiment 1. All participants were paid for first-session training and testing, and the half who received second-session training and testing were paid an additional fee.

5.1.2 Design and Materials

Experiment 2 had the same three phases as Experiment 1. The romanization training, romanization practice, and first-session transcription phases used the same materials as in Experiment 1. For the second-session transcription, two new passages, Passages C and D, were selected from the corpus of adult spoken Southern Min; see Appendix C. Both passages are about 39 seconds long, approximately the same length as Passages A and B, and had already been transcribed and checked. As with these earlier passages, we expected that the two new passages should take less than an hour to transcribe. The two passages have roughly an equal number of word tokens as the two passages in the Experiment 1 (Passage A: 129; Passage B: 122; Passage C: 123; Passage D: 121), and the words were matched in token frequency.

In the first session, half (10) of the participants transcribed Passage A before Passage B, while the other half transcribed the passages in the reverse order. To test the effect of additional training, half (10) of these participants were invited to join the second session, where half of these (5) transcribed Passage C before Passage D, while the other half transcribed the passages in the reverse order.

5.1.3 Procedure

The procedure for both sessions of Experiment 2 was identical to the procedure in Experiment 1, except that ACRIP 2.0 was the only transcription tool used. In both the first and second sessions, there was a romanization training phase, a romanization practice phase, and a transcription testing phase, each taking about an hour. Thus each experimental session lasted approximately three hours.

5.2 Results

We first compared the results for ACRIP 2.0 (the first phase of Experiment 2) with those for Segmentor and ACRIP 1.0 (Experiment 1), performing separate between-group by-participant analyses on transcription speed and accuracy. In all analyses, the independent variables were Passage (A vs. B) and Transcription System (ACRIP 2.0 vs. Segmentor, and ACRIP 2.0 vs. ACRIP 1.0). Our focus was on the effect of software tool: the mixed-system ACRIP 2.0 as compared with the Chinese character system Segmentor and with the romanization system ACRIP 1.0.

Table 5 shows the mean number of transcribed words (transcription speed) and percentage of mistranscribed words (error rate) for Experiment 1 (repeated from Table 4) and for the twenty participants in the first session of Experiment 2.

Table 5. Mean number of transcribed words and percentage of mistranscribed words for the three transcription systems.

System	Transcribed words	Mistranscription rate (%)
Segmentor	92.15	36.94
ACRIP 1.0	83.85	38.11
ACRIP 2.0	104.9	23.27

As can be seen in Table 5, ACRIP 2.0 yielded both a greater number of transcribed words and a lower mistranscription rate than either of the other two transcription tools. In two separate analyses, we compared ACRIP 2.0 with Segmentor and with ACRIP 1.0. Because the comparisons were being across different groups of participants, we used ordinary linear regression (equivalent to ANOVA, but chosen to facilitate comparison with the analyses used for Experiment 1). For each analysis, Passage and Transcription System were coded as effect variables, and their interaction was included in the analyses.

Both measures showed a statistically significant benefit of ACRIP 2.0 over Segmentor (number of transcribed words: $B = 6.375$, $p = .02$; mistranscription rate: $B = -6.83375$, $p = .002$). Similar positive results were found in the comparison of ACRIP 2.0 with ACRIP 1.0 (number of transcribed words: $B = 10.53$, $p = .0004$; mistranscription rate: $B = -7.42$, $p = .004$). significant main effect of Passage ($B = 14.175$, $p < .00001$). In addition, for the number of transcribed words, there were significant main effects of Passage (comparison with Segmentor: $B = 14.175$, $p < .00001$; comparison with ACRIP 1.0: $B = 12.23$, $p < .0001$), but again this was merely because Passage A had a few more words than Passage B. There were no other main effects and no interactions.

We then examined the effect of additional training with ACRIP 2.0 for the ten participants who received a second session of training and testing. The mean number of transcribed words (transcription speed) and percentage of mistranscribed words (error rate) for these ten participants are shown in Table 6.

Table 6. Mean number of transcribed words and percentage of mistranscribed words as a function of training on ACRIP 2.0.

Training	Transcribed words	Mistranscription rate (%)
First session	104.9	23.27
Second session	118.4	14.12

As shown in Table 6, additional training both increased the number of transcribed words and reduced the mistranscription rate. We analyzed both measures with Experience (-1 = first session, 1 = second session) as the only independent variable (Passage was confounded with session, since the first session used only Passages A and B and the second session used only

Passages C and D). Because Experience was a within-participant factor, we again used linear mixed-effects modeling with *p* values computed using Markov chain Monte Carlo samples. The results showed that the improvement in mistranscription rate was statistically significant (B = -5.38, p = .002) and the improvement in the number of transcribed words was marginally so (B = 6.46, p = .08).

5.3 Discussion

The results showed that transcription errors were significantly reduced when participants used the multi-functional, mixed-entry tool ACRIP 2.0, compared either with the character-based Segmentor or the romanization-based ACRIP 1.0. The number of transcribed words completed within the hour-long session also increased with the new tool.

Moreover, with additional training, transcriptions improved still further, with slightly more completed words and an even lower mistranscription rate. Projecting linearly, the drop in mistranscription rate from 23% to 14% from the first three-hour session to the second predicts that near-perfect accuracy could be attained with merely one further three-hour session. More realistically, of course, errors can never be expected to be eliminated entirely, so as is standard practice in the transcription of spoken corpora, the work of one transcriber must always be checked by another.

6. Conclusions

In this paper, we compared three software tools for assisting the transcription of the Taiwanese Spoken Corpus by interfacing with our Southern Min lexical bank. Segmentor requires users to transcribe passages as a string of Chinese characters, with segmentation performed later. The first version of Adult-Corpus Romanization Input Program (ACRIP 1.0) requires users to transcribe word by word, using romanization. The revised version, ACRIP 2.0, requires users to transcribe word by word, but permits them to input words either with Chinese characters or with romanization. In both versions of ACRIP, romanization input can be made without tone digits, and can use a form of auto-completion so that even longer words can be accessed with up to five letters. ACRIP 2.0 adds more flexibility to the input methods and also interfaces directly with Microsoft Windows Media Player so that audio files can be played and replayed from the same interface as word entry.

Our experiments found no significant disadvantage in using romanization entry compared with Chinese character entry, despite the native transcribers being much more familiar with the latter orthographic system. More importantly, ACRIP 2.0 was shown to permit significantly faster and more accurate transcriptions than either Segmentor or ACRIP 1.0. Efficiency and accuracy increased even more with only three additional hours of training. Since conducting this study, our trained graduate assistants use only ACRIP 2.0 as they

continue to transcribe sound files for the Taiwanese Spoken Corpus.

Of all of the innovations of ACRIP, the most surprising for compilers of Chinese speech corpora may be its use of word-based and romanization-based input. Chinese text is traditionally entered into a computer character by character, supplemented by auto-completion for multi-character words where relevant. Yet as our results suggest, this may not be the most efficient method for transcribing fluent speech in Southern Min, a language with a distinct lexicon and phonology from Mandarin.

Nevertheless, given the great increase in performance of ACRIP 2.0 over ACRIP 1.0, it seems that a major strength of the tool lies more in its transcription-specific interface rather than in the type of transcription notation. That is, accuracy and speed were improved in large part because ACRIP 2.0 makes it possible for transcribers to have direct and simultaneous access to sound files, written corpus fragments, and full lexical entries. It is conceivable that additional benefits may result by integrating ACRIP 2.0 more fully into the MOE Southern Min writing tool (Ministry of Education, 2012), but this has yet to be tested.

Given this success, it seems reasonable to ask whether an ACRIP-like corpus transcription tool would applicable to other languages like Hakka or Formosa languages. For the most part, the new functions in ACRIP 2.0 are designed for facilitating the mechanics of transcription regardless of language. The only feature that may be less universally applicable is the 'Incremental Chinese character search box' function, which is not relevant for languages without cognate characters.

We hope that our findings will encourage compilers of other non-Mandarin Sinitic spoken corpora to explore the greater efficiency of input systems beyond the traditional Chinese character-based systems.

References

Academia Sinica. (2002). Southern Min archives: A database of historical change and language distribution. *National Digital Archives Program*. (Retrieved 2010/10/25) http://southernmin.sinica.edu.tw/

Baayen, R. H. (2008). *Analyzing linguistic data: A practical introduction to statistics using R.* Cambridge, UK: Cambridge University Press.

Chui, K, & Lai, H. L. (2009). The NCCU corpus of spoken Chinese: Mandarin, Hakka, and Southern Min. *Taiwan Journal of Linguistics,* 6(2),119-144.

Iunn, U. G. (2003a). *Online Taiwanese syllable dictionary.* (Retrieved 2010/10/25) http://iug.csie.dahan.edu.tw/TG/jitian/.

Iunn, U. G. (2003b). *Online Taiwanese concordancer system.* (Retrieved 2010/10/25) http://iug.csie.dahan.edu.tw/TG/concordance/.

Iunn, U. G. (2005). *Taiwanese corpus collection and corpus based syllable / word frequency counts for written Taiwanese*. (Retrieved 2010/10/25) http://iug.csie.dahan.edu.tw/giankiu/keoe/KKH/guliau-supin/guliau-supin.asp.

Lyu, R. Y., Liang, M. S., & Chaing, Y. C. (2004). Toward constructing a multilingual speech corpus for Taiwanese (Min-nan), Hakka, and Mandarin. *International Journal of Computational Linguistics and Chinese Language Processing*, 9(2), 1-12.

Ministry of Education. (2008). 臺灣閩南語羅馬字拼音方案使用手冊. (Retrieved 2011/04/11) http://www.edu.tw/files/bulletin/M0001/tshiutsheh.pdf

Ministry of Education. (2010). 教育部臺灣閩南語字詞頻統計. (Retrieved 2010/10/25) http://203.64.42.97/bang-cham/thau-iah.php

Ministry of Education. (2012). *Taiwan Southern Min Hanzi Input*, version 2.1 (Retrieved 2012/2/9) http://www.edu.tw/mandr/download.aspx?download_sn=3015&pages=0&site_content_sn=3364

R Development Core Team (2011). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Schneider, W., Eschman, A., & Zuccolotto, A. (2002). *E-prime reference guide.* Pittsburgh: Psychology Software Tools Inc.

Tsay, J. (2007). Construction and automatization of a Minnan child speech corpus with some research findings. *International Journal of Computational Linguistics and Chinese Language Processing*, 12(4), 411-442.

Tsay, J., & Myers, J. (in progress) *Taiwanese Spoken Corpus*. National Chung Cheng University, Chia-Yi, Taiwan.

Appendix A: Fake syllables for romanization training.

bai3	counn7	giem5	hounn3	jeunn7
khoop8	luek8	neunn3	nuiunn5	phoong5
pou7	teinn5	thoinn2	suat8	tot8
bam2	liak4	cei3	ngut8	jen3
pim3	ken5	hoi3	ngang3	coong5
nuoong2	bom2	gooi2	kiai1	sion1

Appendix B: Real syllables for romanization practice.

剝	pak4	合	hap8	芳	phang1	慢	ban7
莫	mai3	針	ciam1	姆	m2	焦	ta1
踢	that4	讀	thak8	雷	lui5	軟	nng2
零	lan5	鹹	kiam5	國	kok4	囥	khng3
牙	ge5	夾	ngeh4	摸	bong1	黃	ng5
蝦	he5	割	kuah4	走	cau2	食	ciah8
手	chiu2	深	chim1	衫	sann1	仙	sian1
爪	jiau3	南	lam5	路	loo7	無	bo5
煎	cuann1	病	penn7	傷	siong1	歹	phainn2
原	guan5	廟	bio7	唱	chiunn3	橫	huainn5
市	chi7	鮮	chinn1	膽	tann2		
枕頭	cim2thau5	田嬰	chan5enn1	硬拗	nge7au2	泡茶	phau3te5
踅街	seh8ke1	庄跤	cng1kha1	避雨	phiah4hoo3		

Appendix C: Passages from the Southern Min Spoken Corpus. The passages here have been modified by hand to remove alternative pronunciations listed in the lexical bank but not used by the speakers in these passages.

Passage A (Duration: 36sec)

Participants: 001 (hostess 1), 002 (hostess 2)

Filename: RC002

002: 阿媽<a1ma2> e0<e0> 話<ue7>。

001: 分享著<hun1hiang2tioh8> 老祖先<lau7coo2sian1> 所<soo2> 流傳<liu5thuan5> e0<e0> 智慧<ti3hui7> e0<e0> 話<ue7>，m0<m0>，這 1<ce1> 咱<lan2> e0<e0> [m 開場白]。來<lai5>，啥人<siann2lang5> 先<sing1> 講<kong2>？

002: 啥人<siann2lang5> 先<sing1> 講<kong2> ne0<ne0>？

001: m0<m0>，我<gua2> 先<sing1> 來<lai5> 講<kong2> 好<ho2> a02<a0>。

002: 好<ho2> 好<ho2> 好<ho2>，你<li2> 先<sing1> 講<kong2>。

001: henn0<henn0> 我<gua2> 欲<beh4> 講<kong2> 這 2<cit4> 句<ku3> hoonn0<hoonn0>，伊<i1> 講<kong2>，食<ciah8> 人 1<lang5> 一 1<cit8> 斤<kin1>，嘛<ma7> 就<to7> 還<hing5> 人 1<lang5> 四<si3> 兩 2<niu2>。

002: oo0<oo0>，食<ciah8> 人 1<lang5> 一 1<cit8> 斤<kin1>，嘛<ma7> 就<to7> 還<hing5> 人 1<lang5> 四<si3> 兩 2<niu2>。

001: henn0<henn0> a02<a0> hoonn0<hoonn0>。

002: hm0<hm0> hm0<hm0>。

001: 這 2le0<cit4e0> 這 2<cit4> 句<ku3> 話<ue7> 所<soo2> 講<kong2> e0<e0>，就是<to7si7> 講<kong2> hoonn0<hoonn0>，咱<lan2> 做人<co3lang5>，這 1<ce1> 人 1<lang5> 佮<kah4> 人 1<lang5> 咧<teh4> 交際<kau1ce3> hoonn0<hoonn0>，咧<teh4> 交往<kau1ong2> e0<e0> 這 2<cit4> 个<e5> 過程<kue3ting5> le02<le0>，總是<cong2si7> hoonn0<hoonn0>，愛 2<ai3> [m 禮尚往來] la0<la0>。就是<to7si7> 講<kong2> hoonn0<hoonn0>，愛 2<ai3> 有來有去<u7lai5u7khi3> la0<la0>。譬論<phi3lun7> 講<kong2>，

002: 未當<be7tang3> 單仔 1<kan1na7> 食<ciah8> 人 1<lang5> e0<e0>，嘛<ma7> 愛 2<ai3> 分 2<pun1> 人 1<lang5> 食<ciah8> la0<la0> hoonn0<hoonn0>。

001: 著 1<tioh8> 著 1<tioh8> 著 1<tioh8> 著 1<tioh8> 著 1<tioh8> hoonn0<hoonn0>。

Passage B (Duration: 35sec)

Participants: 001 (hostess 1), 002 (hostess 2),

Filename: RC002

001: 這 1<ce1> 受<siu7> 人 1<lang5> e0<e0> 恩惠<un1hui7> ne0<ne0>，就<to7> 愛 2<ai3> 知影<cai1iann2> 回報<hue5po3> hoonn0<hoonn0>。

002: henn0<henn0> a02<a0>。

001: e01<e0> 當然<tong1jian5>，

002: 未當<be7tang3> 講<kong2> hoonn0<hoonn0>，受<siu7> 人 1<lang5> e0<e0> 恩惠 <un1hui7>，猶閣<a2koh4> 開始<khai1si2> 佇<ti7> 後壁<au7piah4> 共人<kang9> 創空<chong3khang1> 按呢<an3ne1>。

001: 敢 1<kam2> 會 1<e7> 創空<chong3khang1>？未<be7> la0<la0>，可能<kho2ling5> 是 <si7> hoonn0<hoonn0>， [若是<na7si7> 上界<siong7kai3> i] 比較<pi2kau3> 較 <khah4> 人 1<lang5> 無 1<bo5> 法度<huat4too7> 接受<ciap4siu7> 就是<to7si7> 講 <kong2> hoonn0<hoonn0>，a01<a0>，onn0<onn0>，算講<sng3kong2>，受<siu7> 人 1<lang5> e0<e0> 恩惠<un1hui7>，a01<a0> 伊<i1> 閣<koh4> 毋<m7> 知影 <cai1iann2> 講<kong2> hoonn0<hoonn0>，欲<beh4> 來<lai5> [m 知恩圖報] la0<la0> hoonn0<hoonn0>。

002: m0hm0<m0hm0>。

001: henn0<henn0>。e01<e0> 當然<tong1jian5> 今仔 2<cim2a2> 現代<hian7tai7> hoonn0<hoonn0>，就是<to7si7> 講<kong2>，社會<sia7hue7> 上<siong7>，真<cin1> 濟<ce7> 人 1<lang5> 就是<to7si7> 講<kong2>，咧<teh4> 幫助<pang1coo7> 別人 <pat8lang5> hoonn0<hoonn0>，in1<in1> 感覺<kam2kak4> 講<kong2>，a0<a0>，咱 <lan2> 就是<to7si7> [m 日行一善] ，hoonn01<hoonn0> 咱<lan2> 本底<pun2te2> ne0<ne0>，就是<to7si7> 欲<beh4> 來<lai5> 幫助<pang1coo7> 別人<pat8lang5> e0<e0> hoonn0<hoonn0>。所以<soo2i2> 講<kong2>，伊<i1> 是<si7> xxx<xxx> 真 <cin1> 好意<ho2i3>，真<cin1> 善心<sian7sim1>，a01<a0> 伊<i1> 嘛<ma7> 無 1<bo5> 求<kiu5> 對方<tui3hong1> 來<lai5> 回報<hue5po3>。

Passage C (Duration: 40sec)

Participants: 001 (hostess 1)

Filename: RK006

001: 做陣<co3tin7> 收聽<siu1thiann1> 幸福<hing7hok4> 萬事通<ban7su7thong1>。

001: 我<gua2> 是<si7> [m 幸福] [m 妹妹] e0<e0> 淑芬<siok4hun1>。

001: 來<lai5> 今仔日<kin1a2jit8> 幸福<hing7hok4> 銀行<gin5hang5> 咱<lan2> 來<lai5> 儉 1<khiam7>，o0<o0> 兩 1<nng7> 个<e5> 朋友<ping5iu2> e0<e0> 故事 <koo3su7>。

001: 咱<lan2> 講<kong2> a02<a0>，人生<jin5sing1> 旅途<lu2too5> oo02<oo0>，有<u7> 朋友<ping5iu2> hoonn0<hoonn0>，m0<m0> 咱<lan2> 會 1<e7> 感覺<kam2kak4> 誠<ciann5> 幸福<hing7hok4>。

001: 因爲<in1ui7> 朋友<ping5iu2> e0<e0> 好處<ho2chu3> 就是<to7si7> 講<kong2> 會當<e7tang3> 佮<kah4> 你<li2> 分擔<hun1tam1> 你<li2> o0<o0> 心內<sim1lai7>，你<li2> 歡喜<huann1hi2> e0<e0> 事志<tai7ci3>，o0<o0> 你<li2> 感覺<kam2kak4> m0<m0> 悲傷<pi1siong1> e0<e0> 事志<tai7ci3> 攏<long2> 會當<e7tang3> 佮<kah4> 對方<tui3hong1> 講<kong2> la0<la0> hoonn0<hoonn0>。

001: a01<a0> 咱<lan2> 今仔 2<cim2a2> 講著<kong2tioh8> 這 2<cit4> 兩 1<nng7> 个<e5> 朋友<ping5iu2> a02<a0>，in1<in1> 就是<to7si7> 相招<sio1cio1> 去<khi3> chit4tho5<chit4tho5>，hoonn01<hoonn0>。

001: a01<a0> in1<in1> 去<khi3> chit4tho5<chit4tho5> 這 2<cit4> 个<e5> 所在<soo2cai7> hoonn0<hoonn0>，ai0ioo0<ai0ioo0> 去<khi3> [m 沙漠] [m 旅行] ne0<ne0>，hoonn01<hoonn0>。

001: 但是<tan7si7> 咱<lan2> 講<kong2> a02<a0>，閣<koh4> 較 1<khah4> 好<ho2> e0<e0> 人 1<lang5> hoonn0<hoonn0> 嘛<ma7> 有<u7> 可能<kho2ling5> 會 1<e7> 冤家<uan1ke1> hoonn0<hoonn0>，e01<e0> 翁仔某<ang1a2boo2> 較 1<khah4> 好<ho2> 嘛<ma7> 會 1<e7> 相觸<sio1tak4> le02<le0> hoonn0<hoonn0>。

Passage D (Duration: 39sec)

Participants: 001 (hostess 1)

Filename: RK007

001: 來<lai5> 共<ka7> 聽眾<thiann1ciong3> 朋友<ping5iu2> 講<kong2> 一 1<cit8> 个<e5> 鳥仔<ciau2a2> e0<e0> 故事<koo3su7> hoonn0<hoonn0>，onn0<onn0> 有<u7> 一 1<cit8> 个<e5> 拍獵<phah4lah8> e0<e0> 人 1<lang5> a02<a0> hoonn0<hoonn0>，a01<a0> 伊<i1> 掠著<liah8tioh8> 一 1<cit8> 隻<ciah4> 鳥仔<ciau2a2>，hoonn01<hoonn0>，這 2<cit4> 隻<ciah4> 鳥仔<ciau2a2> 足<ciok4> 水 2<sui2> 足<ciok4> 水 2<sui2> e02<e0> hoonn0<hoonn0>，[是<si7> 一 1<cit8> 隻<ciah4> 真<cin1> i] ，e01<e0> 恰若<kah4na2> 彩色<chai2sik4> e0<e0> 就<to7> 著 1<tioh8>，好親像<ho2chin1chiunn7> 咱<lan2> 彼 1<he1> 南部<lam5poo7> e0<e0>，onn0<onn0> 彼 2le0<hit4le0> 彩色<chai2sik4> 鳥<ciau2> 共款<kang7khuan2>，m0<m0> [m 確實] [m 很] [m 美] ，

001: 但是<tan7si7> 這 2<cit4> 隻<ciah4> 鳥仔<ciau2a2> ne0<ne0>，e01<e0> 予 1<hoo7> 伊<i1> 掠著<liah8tioh8> 了後<liau2au7> a02<a0>，這 2<cit4> 隻<ciah4> 鳥仔<ciau2a2> 講<kong2> 會 1<e7> 講話<kong2ue7> la0<la0> hoonn0<hoonn0>，

001: a01<a0> 這 1<ce1> 講話<kong2ue7> 是講<si7kong2> 啥物<siann2mih8> 話<ue7> ne0<ne0>？伊<i1> 就<to7> 共<ka7> 這 1<ce1> 个<e5> 拍獵<phah4lah8> e0<e0> 人 1<lang5> 講<kong2> a02<a0>，enn0<enn0> 爲著<ui7tioh8> 欲<beh4> 感謝<kam2sia7> 你<li2> 會 1<e7> 共<ka7> 我<gua2> 放開<pang3khui1>，所以<soo2i2> 我<gua2> 送<sang3> 你<li2> 三 1<sann1> 項<hang7> 寶<po2>，這 1<ce1> 三 1<sann1> 項<hang7> 寶<po2> ne0<ne0>，是<si7> 三 1<sann1> 句<ku3> 話<ue7>。