

Automatic Segmentation and Labeling for Mandarin Chinese Speech Corpora for Concatenation-based TTS

Cheng-Yuan Lin*, Jyh-Shing Roger Jang* and Kuan-Ting Chen*

Abstract

Precise phone/syllable boundary labeling of the utterances in a speech corpus plays an important role in constructing a corpus-based TTS (text-to-speech) system. However, automatic labeling based on Viterbi forced alignment does not always produce satisfactory results. Moreover, a suitable labeling method for one language does not necessarily produce desirable results for another language. Hence in this paper, we propose a new procedure for refining the boundaries of utterances in a Mandarin speech corpus. This procedure employs different sets of acoustic features for four different phonetic categories. In addition, a new scheme is proposed to deal with the “periodic voiced + periodic voiced” case, which produced most of the segmentation errors in our experiment. Several experiments were conducted to demonstrate the feasibility of the proposed approach.

Keywords: speech assessment methods phonetic alphabet, speech corpus, sequential forward selection, k-nearest neighbor rule, leave-one-out, speaker-adapted model, context-dependent hidden Markov model (HMM).

1. INTRODUCTION

Corpus-based speech synthesis systems are becoming more and more popular due to the high degree of fluency achieved and the natural feel of the generated speech. However, such systems always require a significant amount of human effort in labeling the phonetic boundaries of the corresponding corpus [Van Erp *et al.* 1988] [Wang *et al.* 1999] [Cosi *et al.* 1991]. Therefore, a great deal of research on automatic phonetic labeling methods has been conducted over the past several years [Ljolje *et al.* 1993, 1994] [Demuyne *et al.* 2002]. In general, most of these methods involve the following two steps:

- (1) rough phonetic segmentation by means of Viterbi forced alignment using HMM (hidden Markov models) or other statistical methods;

* Multimedia Information Retrieval Laboratory, Dept. of Computer Science, National TsingHua University, Hsing-Chu, Taiwan, Tel: +88635715131-3506
E-Mail: {gavins, jang, marco}@wayne.cs.nthu.edu.tw

(2) high time-resolution analysis of the phonetic boundaries using boundary checking rules.

These HMM-based recognizers can be categorized in various ways. For example, some use context-dependent HMM, while others use context-independent HMM [Makashay *et al.* 2000]. Also, there are various types of HMM training methods, including speaker-dependent (SD), speaker-independent (SI), and speaker-adapted (SA) models. Although the HMM-based speech recognizer using MFCCs (mel-frequency cepstral coefficients) is well known for its excellent speech recognition, ability, its use of automatic phonetic segmentation and labeling does not always produce precise and satisfactory results necessary for the development of TTS. As a result, other acoustic features and refinement algorithms have been proposed in the literature to improve the phonetic labeling results obtained from HMM-based recognizers.

Several works have focused on automatic phonetic labeling, in the last few years. For example, in [Bonafonte *et al.* 1996], Bonafonte *et al.* took Gaussian probability density distribution as a similarity measure. In [van Santen *et al.* 1990], Jan P. H. van Santen *et al.* adopted broad-band and narrow-band edge detection. In [Torre Toledano *et al.* 1998], Toledano *et al.* tried to mimic human labeling using a set of fuzzy rules. In [Sethy *et al.* 2002], Sethy *et al.* employed adapted CDHMM (continuous density hidden Markov model) models [Lamel *et al.* 1993]. The main focus of all of these studies has been English speech, and they have seldom addressed the question of which phonetic class tends to be more error prone. Moreover, the methods proposed in the above papers may not perform equally well when dealing with another language. For example, most approaches for English utterance segmentation can be divided into two categories: rule-based [Torre Toledano *et al.* 1998] and statistics-based [Sethy *et al.* 2002] methods. For a rule-based approach, one needs to define a set of rules (crisp or fuzzy) for various phonetic transitions. For a statistics-based approach, one needs to collect a sample data set and label the set accordingly. Conceptually, the rule-based approaches for English corpora can be adapted for application to Chinese corpora. But in fact, it is hard to design such a system without the aid of human experts who have a thorough understanding of the similarities and differences between the phonetic sets of these two languages. It is our belief that the above two approaches should be used in a seamless, integrated manner. As a result, we have developed a hybrid approach, where most of the boundaries are identified via statistical pattern recognition (Sequential Forward Selection, K-Nearest Neighbor Rule and Leave-One-Out) [Whitney 1971] [Duda *et al.* 2001], while the most difficult cases (periodic voiced + periodic voiced) are handled using a rule-based approach.

Mandarin Chinese is a tonal language, and each character is associated with one or several syllables. A Chinese syllable is either composed of a CV (Consonant-Vowel or INITIAL-FINAL [Chou *et al.* 2002] [Lee 1997]) structure or a single V (Vowel) structure. Therefore, the primary effort in speech labeling focuses on precisely identifying the

boundaries of each syllable. Then the boundary between a consonant and a vowel within a syllable can be identified according to the type of a consonant. In most cases, the consonant is fricative, affricate, or plosive, and the consonant can easily be distinguished using several acoustic features other than MFCCs, such as zero-crossing rate or pitch, etc. If the consonant is periodic, as in the case of “ ㄉ ” (“l” in SAMPA (<http://www.phon.ucl.ac.uk/home/sampa/home.htm>), the acronym of ‘Speech Assessment Methods Phonetic Alphabet’), then the consonant does not need to be segmented, and the whole syllable should be treated as a single unit for TTS, since further operations involving pitch or time scale modification should be performed on both the consonant and the vowel.

In [Chou *et al.* 1998, 2002], Chou *et al.* proposed an SD-based HMM model plus simple boundary correction rules for Mandarin Chinese. However, to construct this system is time consuming because of the iterative procedure used for forced alignment, the correction rules and, re-training. In addition, it becomes particularly inefficient if the speech corpus is updated incrementally and regularly, such as by adding one hour of speech data per week. Furthermore, the SD-based HMM model may not outperform the SA-based HMM if the size of the training data is moderate; for example, there is one hour of data for the same speaker.

In this paper, we propose an SA-based HMM recognizer that performs a forced alignment first and then employ a refinement procedure to modify the identified boundaries. The proposed refinement procedure uses several innovative acoustic features to refine boundaries for various phonetic categories. These approaches and experimental results obtained using them will be described in the following sections.

This paper is organized as follows. Section 2 introduces our forced alignment procedure that uses an HMM recognizer to get initial estimations of all boundaries. Section 3 explains the refinement procedure specially designed for four phonetic categories and describes acoustic features are chosen by the SFS (Sequential Forward Selection) [Whitney 1971] algorithm. Section 4 describes the experiments conducted to demonstrate the performance of the proposed refinement procedure, and presents error analysis of irretrievable errors. Section 5 draws conclusions and discusses future work.

2. HMM BASED RECOGNIZER

2.1 From Orthographic Transcription to Phonetic Transcription

Forced alignment using the HMM-based recognizer relies on knowledge of the underlying phonetic transcription of a given utterance. In general, once the orthographic transcription and speech data are both available, we can employ forced alignment for automatic phonetic transcription. However, some commonly used Chinese characters have multiple syllables with different pronunciations, depending on the lexical contexts; For instance, the Chinese

character “重” (meaning “heavy”) is pronounced “ㄓㄨㄥˋ”(“TS-U-@N, 4th tone” in SAMPA) in “重要” (meaning “important”) and “ㄓㄨㄥˊ”(“TS_h-U-@N, 2nd tone” in SAMPA) in “重疊” (meaning “overlap”). As a result, word segmentation in the text sentence is necessary for correct phonetic transcription for the purpose of alignment. Commonly used approaches to word segmentation in Chinese NLP (natural language processing) include the forward or backward maximum word matching algorithm [Chen *et al.* 1992][Yeh *et al.* 1991], and the dynamic-programming-based statistic probability method [Sproat *et al.* 1990]. However, no word segmentation algorithm can guarantee perfect results for the following reasons:

- (1) Word segmentation relies on a collection of Chinese words in the form of a dictionary, which cannot cover all existing words since new words are constantly being created.
- (2) Even if the word dictionary were complete, some pronunciations could not be determined through dictionary lookup, especially for the case of Chinese poems. For instance, the first character of “朝辭白帝彩雲間” (meaning “leaving Baidi city in colored dawn”) is pronounced “ㄓㄠ” (“TS-au, 1st tone” in SAMPA, meaning “dawn”), not “ㄓㄠˊ” (“TS_h-au, 2nd tone” in SAMPA, meaning “to head for”). This error cannot be corrected through dictionary lookup since “朝” is a single-character word meaning “morning”.
- (3) Conflicts in word segmentation can lead to different results. For instance “老掌櫃順手把錢揣在懷裡” (meaning “the old shopkeeper smoothly slipped the money into his pocket”) will be labeled as “老 掌櫃 順手 把 錢 揣 在 懷裡” (meaning “the old + shopkeeper + smoothly + slipped + the money + into + his pocket”) if forward maximum word matching is used. On the other hand, it will be labeled as “老 掌櫃 順手把 錢 揣 在 懷裡” (meaning “The old + shopkeeper + smoothly + handle bar + the money + into + his pocket”) if the backward approach is adopted.

In order to avoid errors resulting from phonetic transcription, we perform the following two steps to achieve a better performance:

- (1) We perform word segmentation using forward and backward maximum matching based on a word dictionary containing around 90,000 entries. We keep the phonetic transcriptions as candidates for use in the next step. (If the result is the same, then we have only a single phonetic transcription.)
- (2) We expand the list of obtained phonetic transcription candidates by adding possible syllables for polyphonic characters that are not found in any of the words obtained through the above word segmentation process. We use these different phonetic transcription candidates to perform a forced alignment through Viterbi decoding. We accept the phonetic transcription that has the maximum log likelihood.

Speech Corpora for Concatenation-based TTS

The above steps combine both word segmentation in NLP and forced alignment in speech recognition to achieve better phonetic transcription performance. When the TTS-455 speech corpus with about 6,000 Chinese syllables was used, the syllable error rate was 2.1% and 1.9% for forward and backward maximum matching, respectively. With the addition of step 2, the error rate was reduced to 1.0%, which represents a significant reduction of 50% in the error rate. Some of the error cases are shown in Table 1.

Table 1. Labeling errors when orthographic transcription was transformed to phonetic transcription.

Text sentences of speech corpus.	Human transcription	Machine transcription
春風秋月何時『了』	ㄉㄨㄛˋ ("l-I-au, 3 rd tone")	ㄉㄛˊ ("l-@, 5 th tone")
他囊『括』七面金牌	ㄎㄨㄛˋ ("k h-U-o, 4 th tone")	ㄎㄨㄚˊ ("k-U-a, 1 st tone")
道『行』高深的老僧 掐指一算就知道對方的來意	ㄒㄩㄢˋ ("x-aN, 2 nd tone")	ㄒㄨㄢˋ ("6-I-@N, 2 nd tone")

Note: Symbols in parentheses are described in SAMPA.

The last character of the first sentence is a typical single character having multiple pronunciations that cannot be identified through word dictionary lookup. Unfortunately, forced alignment cannot find the correct phonetic transcription, either, because the utterance itself is ambiguous and unclear. The second sentence demonstrates the inadequacy of the word dictionary since “括” in “囊括” (meaning “to obtain”) is labeled “ㄎㄨㄚˊ” (“k-U-a, 1st tone” in SAMPA) in the dictionary, while it is also pronounced “ㄎㄨㄛˋ” (“k_h-U-o, 4th tone” in SAMPA) colloquially. The error from the third sentence indicates the inadequacy of the word dictionary; the word “道行” (meaning “capability” or “achievement”) should be in the word dictionary, but it is not.

2.2 Speech Corpus Introduction

Once a phonetic transcription is obtained, we can perform forced alignment by using a HMM recognizer. In this study, we used two Mandarin Chinese speech corpora:

- (1) TTS-455 speech corpus: This corpus contains 455 sentences spoken by one speaker and covers about 6,000 syllables. It is mainly for TTS. The details are as follows:
 - I. time duration: 30 minutes (66MB of disk space);
 - II. sampling rate and bit rate: 20,000 Hz, 16bits;
 - III. base syllables: 408;
 - IV. tonal syllables: 1196.

More information on this corpus can be found in (http://speech.cs.nthu.edu.tw/gavins/Research/SpeechSynthesis/content_hsf455.txt).

(2) TCC-300 speech corpus (http://rocling.iis.sinica.edu.tw/ROCLING/MAT/Tcc_300brief.htm): It contains sentences spoken by 300 subjects from National Taiwan University, Chiao Tung University, and Cheng Kung University in Taiwan. The recorded texts were selected from the “Academia Sinica Balanced Corpus” (<http://www.sinica.edu.tw/~tibe/2-words/modern-words>).

In order to perform a forced alignment on the TTS-455 speech corpus, we need to train an HMM-based recognizer. This recognizer will be described in Section 4.

3. DESIGN OF THE REFINEMENT PROCEDURE

A post-processing scheme must be used to refine the identified syllable boundaries. Specifically, since a forced alignment is based on MFCCs only, it makes sense to use other acoustic features to enhance precision. As mentioned in Section 1, using either a rule-based or a statistics-based approach alone is inadequate. Therefore, we combine these two methods to deal with a Mandarin Chinese speech corpus. First of all, we divide all Chinese phonemes into four categories. Then, we determine which set is suitable for which method (rule-based or statistics-based) by applying pattern recognition techniques. These steps will be described in detail in the following subsections.

3.1 Four Phonetic Categories

There are 37 distinct phonetic alphabets in Mandarin Chinese. This makes it difficult to develop a general method that can be used to refine labeling between all possible phonetic transitions. Hence, we divide all Chinese phonemes into four categories according to their acoustic characteristics. These four categories are fricative and affricate, unaspirated stop, aspirated stop, and periodic voiced [Lu 2002], as listed below in SAMPA format and in the MPA (Mandarin Phonetic Alphabet) format:

- Fricative and affricate: (consonants only)

(Fricative)

➤ SAMPA: f x ʃ S s

➤ MPA: ㄈ ㄨ ㄕ ㄙ ㄙ

(Affricate)

➤ SAMPA: tʃ tʃ_h TS TS_h ts ts_h

➤ MPA: ㄐ ㄑ ㄒ ㄕ ㄖ ㄗ

Speech Corpora for Concatenation-based TTS

- Unaspirated stop: (consonants only)
 - SAMPA: p t k
 - MPA: ㄅ ㄆ ㄇ
- Aspirated stop: (consonants only)
 - SAMPA: p_h t_h k_h
 - MPA: ㄅˊ ㄆˊ ㄇˊ
- Periodic voiced:
 - (Consonants)
 - SAMPA: m n l Z
 - MPA: ㄇㄣ ㄣㄣ ㄨㄣ ㄩㄣ
 - (Vowels)
 - SAMPA: a o @ e ai ei au ou an @n aN @N 2 I U y
 - MPA: ㄚ ㄛ ㄜ ㄝ ㄞ ㄟ ㄠ ㄡ ㄢ ㄣ ㄤ ㄨ ㄩ ㄩˊ ㄩˊˊ ㄩˊˊˊ

Fricative and affricate are combined in a single category is mainly because of the similarity of the acoustic characteristics. In particular, for any given syllable with an affricate or fricative consonant, according to our observations, the duration ratio between the aperiodic and periodic parts is almost constant; in addition, there usually exists a high zero-crossing rate at the aperiodic part. As for the periodic voiced category, we include both consonants and vowels since they both contain stable harmonic or pitch structures.

3.2 Feature Definition

In order to refine the boundaries identified by the HMM-based recognizer, we need to employ several acoustic features other than MFCCs. Some of these acoustic features are commonly used in speech processing; they include the zero-crossing rate, log energy, pitch, and entropy [Shen *et al.* 1998]. In addition, we also adopt two new acoustic features, the bisector frequency and the burst degree, to help identify boundaries more precisely.

3.2.1 Bisector Frequency

The bisector frequency is defined in equations (1) and (2):

$$freqIndex = \arg \min_{1 < k < N} \left| \sum_{f=1}^k A_f - \frac{\sum_{f=1}^N A_f}{2} \right|, \quad (1)$$

$$bisektorFreq = \frac{freqIndex}{N} \times sampleRate, \quad (2)$$

where A_f is the amplitude of the f^{th} frequency component and there are N distinct frequency components in the spectrum. The key characteristic of the bisector frequency is that its value is smaller for a voiced frame but larger for an unvoiced frame. Thus, we can use this feature to distinguish unvoiced from voiced patterns. Although the zero-crossing rate can also be used to detect unvoiced patterns, it is not sufficiently robust, especially when the mean amplitude of an unvoiced frame deviates from zero. For example, in Figure 1, the second unvoiced part of the waveform can be better detected by means of the bisector frequency than the zero-crossing rate.

In our implementation, we normalize the value of this feature to the range [0,1] according to equation (3):

$$bisektorfreq = \left(\frac{bisektorfreq - lowfreq}{highfreq - lowfreq} \right), \quad (3)$$

where the values of $highfreq$ and $lowfreq$ are empirically set to be $\frac{sampleRate}{2} \times 0.8$ and 100, respectively.

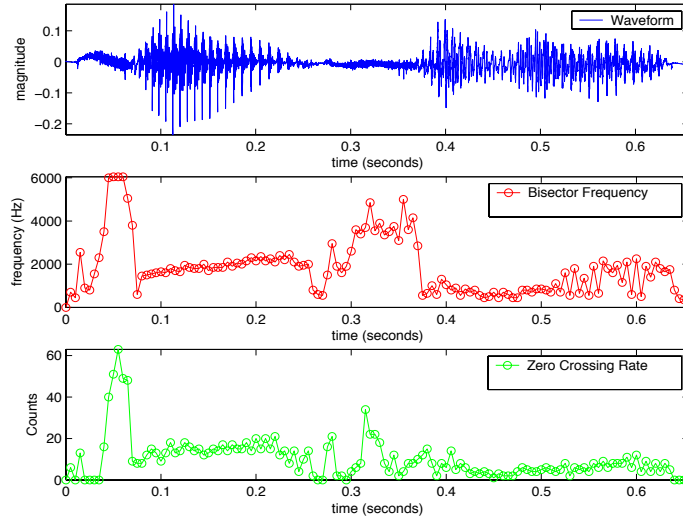


Figure 1. A comparison between the bisector frequency and the zero-crossing rate. The second unvoiced part of the waveform is better detected by means of the bisector frequency than the zero-crossing rate. The content of this waveform is “在視為” (“ts-ai, S, U-ei” in SAMPA).

3.2.2 Burst Degree

It is difficult to recognize a burst pattern in speech using the zero-crossing rate and/or pitch. This is a stable pitch structure does not exist, and the zero-crossing rate is relatively low. To deal with this situation, we adopt a new feature called the burst degree, which is a weighted average between the log energy and the reciprocal average distance between the local maxima, as shown in equation (4):

$$\text{burst degree} = \frac{\left(W_1 \times \frac{1}{\text{avg}(\text{local max Interval})} + W_2 \times \log \text{Energy} \right)}{(W_1 + W_2)}, \quad (4)$$

where W_1 and W_2 are two weighting factors with values of 4 and 1, respectively. The expression $\text{avg}(\text{local max Interval})$ is the average distance between the positions of neighboring local maxima of sample points. For instance, suppose that there are 4 local maxima located at positions 12, 52, 92 and 130 in a frame. Then, the intervals are 40, 40 and 38 and $\text{avg}(\text{local max Interval})$ is $(40+40+38)/3$. Figure 2 shows the result of the burst degree.

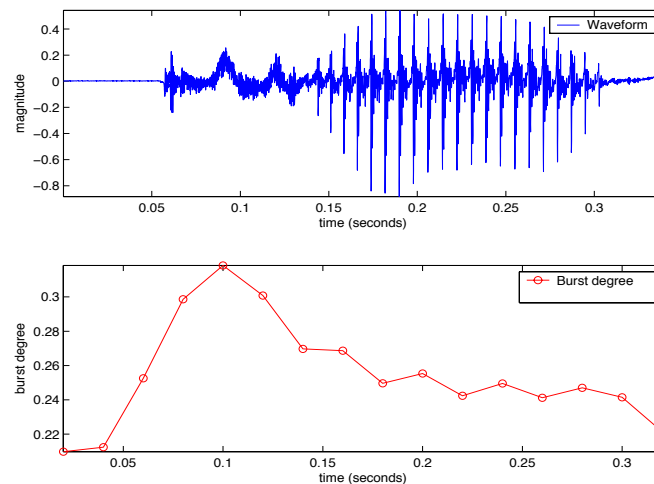


Figure 2. A speech waveform and its burst degree. The content of this waveform is “*咖*” (“*k_h-a*” in SAMPA).

3.3 Feature Selection Based on Phonetic Categories

In Section 3.1, we divided all phonemes into four phonetic categories. In this section, we divided boundaries into groups according to the transitions between phonetic categories. For instance, the boundaries of a given syllable with an aspirated stop consonant can be analyzed as follows:

- (1) Beginning boundary: “silence + aspirated stop” or “vowel + aspirated stop”.
- (2) Ending boundary: “vowel + silence” or “vowel + X”, where X is the consonant of the next syllable, which can be fricative and affricate, aspirated stop, unaspirated stop, or periodic voiced.
- (3) INITIAL/FINAL boundary: “aspirated stop + vowel”. (The INITIAL/FINAL boundary is the boundary between the consonant and the vowel within a syllable. In our experiments, we did not try to find these kinds of boundaries since they were not the focus of this study. However, we still discuss all three kinds of boundaries for the sake of completeness.)

Based on similar analysis, we constructed Table 2 which lists all possible transitions from the left side to the right side for beginning, ending, and INITIAL/FINAL boundaries.

Table 2. All possible category transitions of beginning, ending, and Initial/Final boundaries.

Left side	Right side	Beginning boundary	Ending boundary	Initial/Final boundary
Silence	Fricative and affricate	O	X	X
Silence	Aspirated stop	O	X	X
Silence	Unaspirated stop	O	X	X
Silence	Periodic voiced	O	X	X
Fricative and affricate	Periodic voiced	X	X	O
Aspirated stop	Periodic voiced	X	X	O
Unaspirated stop	Periodic voiced	X	X	O
Periodic voiced	Silence	X	O	X
Periodic voiced	Fricative and affricate	O	O	X
Periodic voiced	Aspirated stop	O	O	X
Periodic voiced	Unaspirated stop	O	O	X
Periodic voiced	Periodic voiced	O	O	O

O: possible transition; X: impossible transition.

It is evident that not all features work equally well for each phonetic group. Therefore we must design an efficient method to distinguish the most outstanding among all possible features. In our experiment, we collected a speech corpus that contained about 2,100 syllables from 20 long sentences from speech lasting a total of 10 minutes. This corpus was used for feature selection and was fully independent of our speech corpus mentioned in Section 2.2. The syllables covered every Mandarin Chinese phoneme. The beginnings and endings of the phonetic boundaries of these 2,100 syllables were manually labeled. In the following we describe the steps we performed to find the best combination of features for each of these phonetic category transitions.

Speech Corpora for Concatenation-based TTS

- (1) In order to find the most discriminative features, we had to create a set of training data. This was done by adding several candidate boundaries, 10 ms apart, located within ± 80 ms of a true (manually labeled) boundary. A candidate boundary was labeled “correct” if it was within ± 20 ms of the true boundary. (According to [Chou *et al.* 2002], manual labeling by two human experts can achieve about 90% consistency with 10 ms tolerance and 100% with 20 ms tolerance.) Therefore, we chose to use 5 correct candidates, all within 20 ms of the manually labeled one, in our experiments. If we had chosen only one, then the number of “correct” data might have been too small, leading to an unbalanced sample data set. In other words, for each true boundary, we created a set of 17 candidate boundaries (including the true one), with 5 labeled “correct” and 12 labeled “wrong” as the desired classification output as shown in Figure 3.

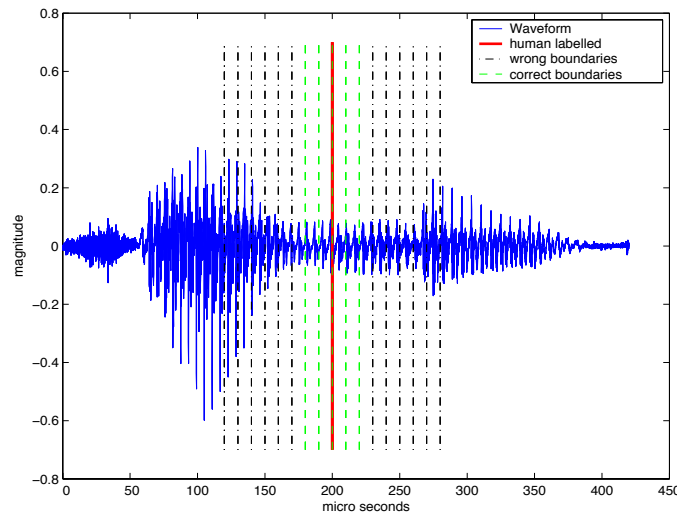


Figure 3. Training data of 5 correct boundaries and 12 wrong boundaries around the true boundary labeled by humans. The content of this waveform was “將離” (“t6-l-aN, l-l” in SAMPA).

- (2) For each candidate boundary, we evaluated the differences between all the acoustic features of its left and right frames. The size of each frame was 20 ms, and the “difference of acoustic features” was then used as a feature for designing a classifier.
- (3) In order to find the most influential acoustic features, we employed the method of sequential forward selection (SFS) [Whitney 1971] in the literature on pattern recognition. The idea behind SFS is to start with a single feature having the best classification rate. Then, we can keep the already selected features and try to identify a newly added feature that can increase

the classification rate the most. For instance, if features 2 and 5 are the currently selected features, then we will try to find another feature that, when combined with the selected features, can produce the best classification rate. This greedy step is repeated until the desired number of features has been selected or until there is no further improvement in the classification rate. In order to use SFS, we need to select a classifier together with its performance evaluation scheme. Here, we used KNNR (K-Nearest Neighbor Rule) as the classifier and LOO (Leave-One-Out) [Duda *et al.* 2001] as the performance criterion. The basic idea behind 1-NNR is to assign the class of a given test vector as the data point in the training data that is nearest to the given vector. In order to achieve better robustness, we can choose KNNR, where the K nearest neighbors are selected around the test vector and the assigned class is determined by means of a voting mechanism among these K points. Then, we performed a simple search to find the best value of K in KNNR is 9 in our experiment. To evaluate the performance of KNNR, we apply LOO, where a vector is selected as the test vector and all the other data as the training data. This process is repeated until each data point has served as the test vector. The final classification rate is the overall classification rate of these test vectors. KNNR with LOO is the most straightforward approach due to its simplicity, although other classifiers or performance criteria could also be used, too.

- (4) We applied the procedure described above to two parts of each syllable, that is, the beginning and ending boundaries.

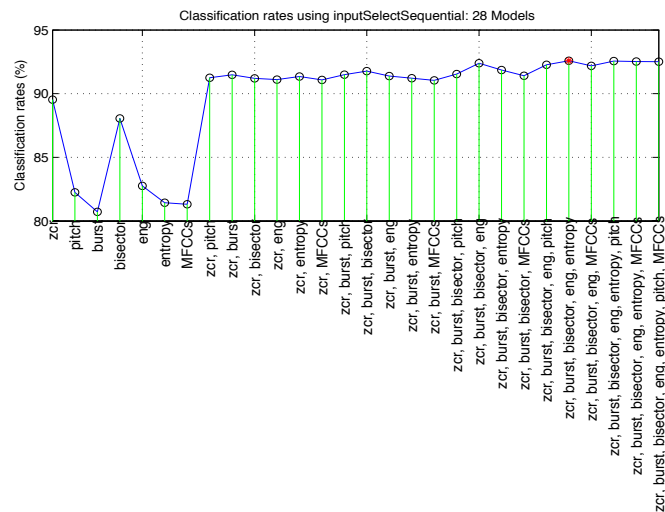


Figure 4. The LOO classification rates for different combinations of features for “silence + fricative and affricate” phonetic category at the beginning boundary.

Speech Corpora for Concatenation-based TTS

Figure 4 shows the SFS results for the “silence + fricative and affricate” phonetic category at the beginning boundary, where the x-axis is the selected features and the y-axis is the LOO classification rates. From Figure 4, it is evident that the most distinguishing features for the “silence + fricative and affricate” category at the beginning boundary are the zero-crossing rate, bisector frequency, log energy, entropy, and burst degree, with which a LOO classification rate of 92.6% could be achieved. By following this same procedure, we could identify the most distinguishing features and their corresponding LOO classification rates, as shown in Table 3.1 and Table 3.2.

Table 3.1 Classification rates of the beginning boundaries of syllables for four phonetic categories.

Phonetic category transitions		Classification rate	Selected features
Left side	Right side		
Silence	Fricative and affricate	92.6%	Zero-crossing rate, bisector frequency, log energy, entropy, and burst degree
Silence	Aspirated stop	89.0%	Zero-crossing rate, log energy, bisector frequency, and burst degree
Silence	Unaspirated stop	92.1%	Entropy, log energy, burst degree and bisector frequency, and MFCCs
Silence	Periodic voiced	89.1%	Log energy, pitch, and burst degree
Periodic voiced	Fricative and affricate	92.7%	Bisector frequency, log energy, zero-crossing rate, entropy, and burst degree
Periodic voiced	Aspirated stop	87.6%	Zero-crossing rate and bisector frequency
Periodic voiced	Unaspirated stop	89.2%	Zero-crossing rate, log energy, entropy, and bisector frequency
Periodic voiced	Periodic voiced	71.8%	Bisector frequency, log energy, zero-crossing rate, entropy, MFCCs, and burst degree

Table 3.2 Classification rates of the ending boundaries of syllables for four phonetic categories.

Phonetic category transitions		Classification rate	Selected features
Left side	Right side		
Periodic voiced	Silence	87.4%	Log energy, burst degree, entropy, and bisector frequency
Periodic voiced	Fricative and affricate	89.6%	Zero-crossing rate, bisector frequency, pitch, log energy, burst degree, and entropy
Periodic voiced	Aspirated stop	89.9%	Zero-crossing rate, bisector frequency, pitch, log energy, burst degree, and entropy.
Periodic voiced	Unaspirated stop	86.4%	Pitch and log energy
Periodic voiced	Periodic voiced	70.7%	Zero-crossing rate, bisector frequency, pitch, log energy, MFCCs, and entropy

The classification rates of “periodic voiced + periodic voiced” were only 71.8% at the beginning boundaries and 70.7% at the ending boundaries, respectively, which are comparatively low. This is mainly due to inseparable co-articulation. Later in this paper, we shall propose and detail other heuristic rules that can be applied to enhance the performance.

3.4 Further Improvement for “Periodic Voiced + Periodic Voiced” Cases

In our implementation, we first obtained an initial estimate of the beginning/ending boundaries based on the TCC-300 trained HMM with adaptation performed by means of a TTS-455 corpus. For every initial boundary, we selected candidate boundaries that were 2 ms apart and within 40 ms at both sides of this boundary. In other words, there were 41 candidate boundaries. The final boundary was determined by KNNR, where K was equal to 9, and the training data set is the one used for SFS and LOO mentioned above. The adopted features were those selected by the SFS as mentioned above.

However, for “periodic voiced + periodic voiced,” the performance was not good enough due to co-articulation. Hence, we devised a special scheme for this category. Specifically, we adopted only two features to determine the boundary. This approach is based on the observation that most boundaries labeled by humans are located in a region with lower log energy.

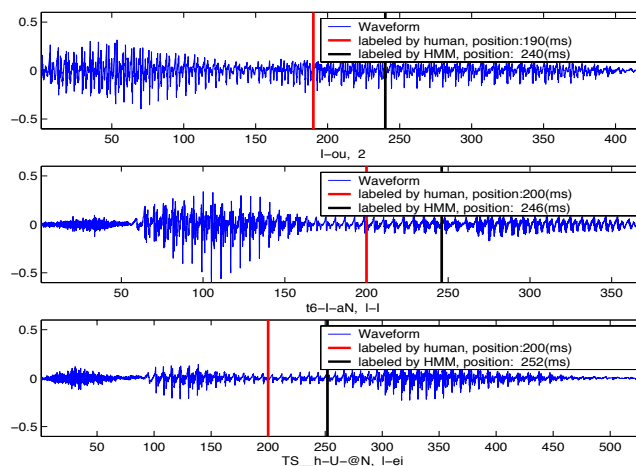


Figure 5. The three common errors in “periodic voiced + periodic voiced” cases. The content of the 1st waveform was “幼兒” (“I-ou, 2” in SAMPA), the content of the 2nd waveform was “將離” (“t6-l-aN, l-l” in SAMPA), and the content of the 3rd waveform was “蟲類” (“TS_h-U-@N, l-ei” in SAMPA).

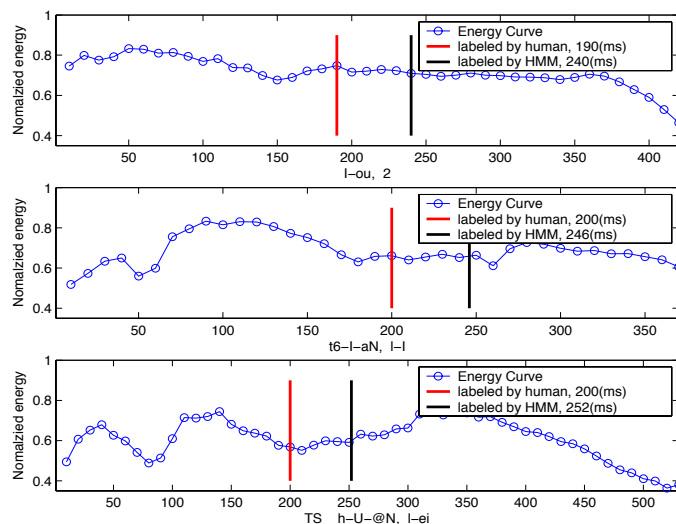


Figure 6. The corresponding log energy profiles for three “periodic voiced + periodic voiced” cases.

Figure 5 and Figure 6 show typical cases for 幼兒 (“I-ou” and “2” in SAMPA), 將離 (“t6-I-aN” and “l-l” in SAMPA), and 蟲類 (“TS_h-U-@N” and “l-ei” in SAMPA). Refining the boundary of this category is more complicated, and little related research has been reported in the literature. In this paper, we propose a new scheme to deal with this category using MFCCs and log energy, as described below:

- (1) The search region is increased from ± 40 ms to ± 80 ms since large deviations over 50 ms are common in the “periodic voiced + periodic voiced” category. The number of candidate boundaries is increased from 41 to 81.
- (2) We calculate the average log energy in the search region. We then set the new search region to be the one whose log energy is less than the log energy threshold, which is empirically defined as 0.9 times the average log energy.
- (3) Among the boundaries within the new search region, we select the one with the maximum distance between the MFCCs of its left and right frames.

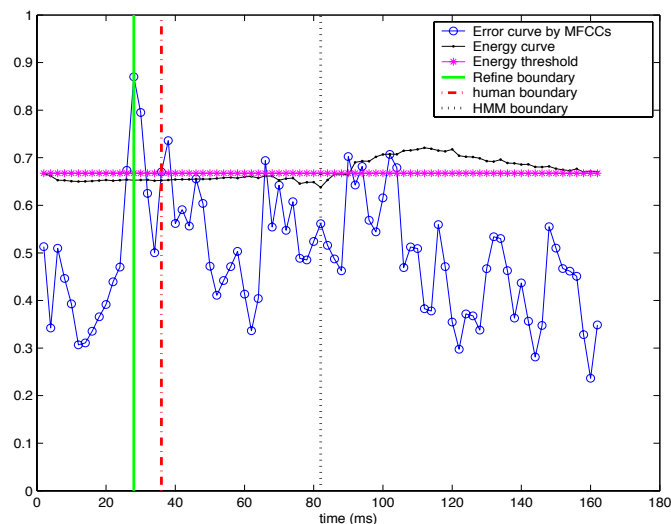


Figure 7. The refined results obtained based on MFCCs and log energy for the “periodic voiced + periodic voiced” case.

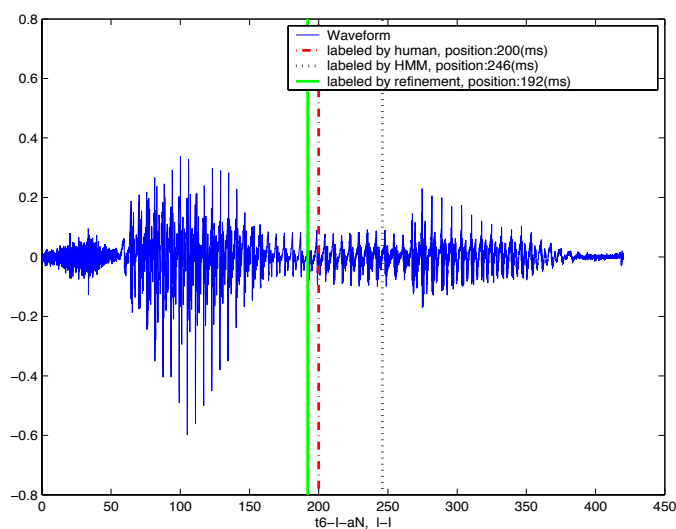


Figure 8. Typical results for the refined boundary of the “periodic voiced + periodic” case. The content of this waveform was “將離” (“t6-l-aN, l-l” in SAMPA).

Figure 7 shows typical results obtained by using the above refinement method. The refined boundary for the original waveform is shown in Figure 8. The experimental results and error analysis will be discussed in the next section.

4. EXPERIMENT RESULTS AND ERROR ANALYSIS

4.1 The Different Acoustic Models of HMM-based Recognizers

To evaluate the performance of our proposed system, we used the TTS-455 corpus to verify the segmentation results. First, we employed different types of model training to construct an HMM recognizer for forced alignment, as described below:

- (1) 1st model: Speaker-independent (SI) model constructed by using the TCC-300 corpus.
- (2) 2nd model: Speaker-dependent (SD) model constructed by using the TTS-455 corpus, with uniform segmentation.
- (3) 3rd model: Speaker-dependent (SD) model constructed by using the TTS-455 corpus, with initial segmentation performed by the model trained using the TCC-300 corpus.
- (4) 4th model: Speaker-independent (SI) model constructed by using the TCC-300 corpus first and then adapted by using the TTS-455 corpus.

Each of these four types of acoustic models was constructed based on context-dependent tri-phones. The MLLR method [Huang *et al.* 2001] used to construct the 4th model employs the regression class tree to estimate a set of linear transformations for the mean vectors and covariance matrices of a Gaussian mixture HMM system. The tree was constructed using a centroid-splitting algorithm based on the Euclidean distance measure. We applied a binary regression tree with thirty-two base classes to our adapted data. In order to speed up the adaptation process and preserve storage capacity, we used the diagonal transform matrix instead of the full transform matrix [Odell *et al.* 1995]. Hence, the 4th model can be regarded as a speaker-adapted (SA) model.

The difference between the 2nd and 3rd models lies in the initial segmentation for training. The 2nd model uses uniform segmentation, while the 3rd model uses the segmentation derived by the recognizer trained using the TCC-300 corpus. Both of them can be viewed as SD models derived from the TTS-455 corpus.

4.2 The Performance of Different Acoustic Modes for Labeling the TTS-455 Corpus

Table 4 summarizes the results obtained with different modeling methods. The acoustic model for HMM forced alignment is based on context-dependent triphone modeling. From Table 4, it is evident that the 4th model achieved the best performance.

Table 4. Segmentation results w.r.t. model training (including beginning and ending boundaries).

Model \ Errors	<=10ms	<=20ms	<=30ms	>50ms
1 st model	49.47%	70.58%	84.24%	4.90%
2 nd model	45.02%	69.83%	81.96%	7.64%
3 rd model	43.49%	65.55%	79.60%	8.49%
4 th model	46.09%	72.07%	87.40%	4.20%

4.3 A Comparison of the Segmentation Rate between Forced Alignment and Our Refinement Procedure

We have chosen the 4th model as our primary speech recognizer. However, its performance in segmentation is still not good enough for TTS application. The segmentation rate within 20 ms is only 72% when using the 4th model. It is probable that the system can be further improved. The following experiment was based on the initial boundaries identified by the 4th model. In our experiment, we divided the segmentation task according to groups of phonetic categories, as mentioned previously in Section 3.4. Table 5.1 shows the results for each phonetic category transition at the beginning boundary and Table 5.2 shows the results for each phonetic category transition at the ending boundary. Table 6 compares the overall segmentation rates obtained with the 4th model recognizer and our refinement procedure.

Table 5.1 Segmentation rates obtained with the HMM recognizer and the refinement procedure for all phonetic categories at the beginning boundaries of syllables.

Phonetic category Transitions		<=10 ms		<=20 m		<=30 ms		>50 ms	
Left side	Right side	H	R	H	R	H	R	H	R
Silence	Fricative and affricate	23.1	77.3	59.1	91.1	91.8	96.1	1.9	1.7
Silence	Aspirated Stop	13.7	81.9	54.5	94.3	93.3	98.7	0.3	0
Silence	Unaspirated stop	13.2	89.5	53.0	98.2	92.3	99.6	0.2	0
Silence	Periodic voiced	8.8	70.1	46.8	86.7	86.9	92.1	3.4	2.5
Periodic voiced	Fricative and affricate	59.4	84.7	83.6	94.7	95.3	97.8	0.7	0.7
Periodic voiced	Aspirated Stop	27.0	81.5	61.7	94.6	93.1	96.5	1.2	1.2
Periodic voiced	Unaspirated stop	30.0	85.2	67.0	95.8	91.4	98.4	1.3	0.5
Periodic voiced	Periodic voiced	45.0	66.3	60.0	75.3	71.9	79.2	10.9	6.7

Note: H: HMM results; R: Refined results; unit: %.

Table 5.2 Segmentation rates obtained with the HMM recognizer and the refinement procedure for all phonetic categories at the ending boundaries of syllables.

Phonetic category Transitions		<=10 ms		<=20 m		<=30 ms		>50 ms	
		H	R	H	R	H	R	H	R
Left side	Right side								
Periodic voiced	Silence	56.0	58.7	76.2	78.6	85.0	86.8	5.7	5.2
Periodic voiced	Fricative and affricate	58.3	75.0	88.2	92.8	97.2	97.3	0.5	0.3
Periodic voiced	Aspirated Stop	47.5	57.8	80.7	84.1	94.4	94.5	1.4	1.5
Periodic voiced	Unaspirated stop	57.0	73.1	91.5	91.6	98.1	97.8	0.1	0.3
Periodic voiced	Periodic voiced	42.9	63.5	60.8	72.8	70.9	79.6	11.5	8.4

Note: H: HMM results; R: Refined results; unit: %.

Table 6. The overall segmentation rates obtained with this system. (including beginning and ending boundaries).

	<=10ms	<=20ms	<=30ms	>50ms
HMM-based forced alignment	46.1%	72.1%	87.4%	4.2%
The proposed refinement method	69.1%	87.7%	94.2%	3.5%

4.4 Results and Discussions

From Table 5.1 and Table 5.2, we can observe that the performance for each phonetic category transition is satisfactory except for the category “periodic voiced + periodic voiced.” It may seem that our refinement method performed poorly for this category. We have carried out another experiment in which we applied the statistical method (just like the one applied to other phonetic categories) to this “periodic voiced + periodic voiced” category. The average segmentation rate of <=30ms for this “periodic voiced + periodic voiced” category was 60% lower. This clearly indicates that our refinement method (rule-based in this case) is definitely better. All in all, this category still poses a difficulty for automatic segmentation since there is usually very strong co-articulation between two neighboring syllables, such “第一” (meaning “number one”), “蘇武” (an ancient Chinese person’s name), and so on.

From Table 6, it is evident that our refinement approach leads to improvement in the overall segmentation rate. The segmentation rate within 20 ms is significantly increased by about 15.6%; and the segmentation rate within 30 ms after the refinement procedure is performed is 94.2%, which is acceptable for general TTS systems. Admittedly, however, there is still some room left for future improvement, as described in the following:

- (1) The size of our TTS-455 corpus is not large enough. A larger corpus will result in a better adapted model, which will reduce the segmentation errors that are larger than 50 ms.
- (2) Acoustic features other than MFCCs can potentially be used to obtain better segmentation rates. We are now in the process of identifying other more discriminative acoustic features for this purpose.

5. CONCLUSIONS

Correct phonetic labeling is very important for concatenation-based speech synthesis. Consequently, the application of automatic phonetic labeling and segmentation for corpora to be used in TTS has become a critical issue. In this paper, we have proposed a specific refinement procedure suitable for Mandarin Chinese. We divide all Chinese phonemes into four categories and employ the SFS algorithm to select the best features for each phonetic category. However, the proposed method does not work well in the “periodic voiced + periodic voiced” case. Hence, we have proposed an additional scheme to deal specifically with this case, using log energy and MFCCs. Several experiments have demonstrated the feasibility of the proposed approach.

In future work, we will focus on finding new features to improve the segmentation rate in the “periodic voiced + periodic voiced” case. We will also apply other classifiers, such as SVM (support vector machine), to further improve the classification results. Finally, we will apply other methods for feature extraction, such as linear discriminant analysis and principal component analysis.

Reference

- Bonafonte, A., A. Nogueiras and A. Rodriguez-Garrido, “Explicit segmentation of speech using Gaussian models,” *Proceedings of International Conference on Spoken Language Processing*, 1996, pp. 1269-1272.
- Chen, K. J. and S. H. Liu, “Word identification for mandarin Chinese sentences,” *Proceedings of the Fifteenth International Conference on Computational Linguistics*, 1992, pp. 101-107.
- Chou, F.-C., C.-Y. Tseng and L.-S. Lee, “Automatic Segmental and Prosodic Labeling of Mandarin Speech,” *Proceedings of International Conference on Spoken Language Processing*, 1998, pp. 1263-1266.
- Chou, F.-C., C.-Y. Tseng and L.-S. Lee, “A Set of Corpus-based Text-to-speech Synthesis Technologies for Mandarin Chinese,” *IEEE Transactions on Speech and Audio Processing*, 10(7), 2002, pp.481-494.
- Cosi, P., D. Falavigna and M. Omologo, “A Preliminary Statistical Evaluation of Manual and Automatic Segmentation Discrepancies,” *Proceedings of European Conference on Speech Communication and Technology*, 1991, pp. 693-696.

Speech Corpora for Concatenation-based TTS

- Demuyne, K. and T. Laureys, "A Comparison of Different Approaches to Automatic Speech Segmentation," *Proceedings of International Conference on Text, Speech and Dialogue*, 2002, pp. 277--284.
- Duda, R. D., P. E. Hart and D. G. Stork, *Pattern Classification*, 2nd ed., Wiley, New York, 2001.
- Huang, X., A. Acero and H. W. Hon, *Spoken language processing*, Prentice Hall, New Jersey, 2001.
- Lamel, L. F. and J. L. Gauvain, "High Performance Speaker-Independent Phone Recognition Using CDHMM," *Proceedings of European Conference on Speech Communication and Technology*, 1993, pp. 121-124.
- Lee, L.-S., "Voice Dictation of Mandarin Chinese," *IEEE Signal Processing Magazine*, 10(4), 1997, pp.63-101.
- Ljolje, A. and M. D. Riley, "Automatic segmentation of speech for TTS," *Proceedings of European Conference on Speech Communication and Technology*, 1993, pp. 1445-1448.
- Ljolje, A., J. Hirschberg and J. P. H. van Santen, "Automatic Speech Segmentation for Concatenative Inventory Selection," *Proceedings of ESCA/IEEE Workshop on speech synthesis*, 1994, pp. 93-96.
- Lu, H.-M., "An implementation and Analysis of Mandarin Speech Synthesis Technologies," MD thesis, National Chiao Tung University at Taiwan, 2002.
- Makashay, M. J., C. W. Wightman, A. K. Syrdal and A. Conkie, "Perceptual evaluation of automatic segmentation in text-to-speech synthesis," *Proceedings of International Conference on Spoken Language Processing*, 2000, pp. 431-434.
- Odell, J., D. Ollason, P. Woodland, S. Young and J. Jansen, *The HTK Book for HTK V2.0*, Cambridge University Press, Cambridge UK, 1995.
- Sethy, A. and S. Narayanan, "Refined Speech Segmentation for Concatenative Speech Synthesis," *Proceedings of International Conference on Spoken Language Processing*, 2002, pp. 149-152.
- Shen, J.-L., J.-W. Hung and L.-S. Lee, "Robust entropy-based endpoint detection for speech recognition in noisy environments," *Proceedings of International Conference on Spoken Language Processing*, 1998.
- Sproat, R. and C. Shih, "A statistical method for finding word boundaries in Chinese text," *Computer Processing of Chinese and Oriental Languages*, 1990, pp.336-351.
- Torre Toledano, D., M. A. Rodriguez Crespo and J. G. EscaladaSardina, "Trying to Mimic Human Segmentation of Speech Using HMM and Fuzzy Logic Post-correction Rules," *Proceedings of Third ESCA/COCOSDA Workshop on speech synthesis*, 1998, pp. 207-212.
- Van Erp, A. and L. Boves, "Manual segmentation and labelling of speech," *Proceedings of Speech*, 1988, pp. 1131-1138.

- van Santen, J. P. H. and R. Sproat, "High-accuracy automatic segmentation," *Proceedings of European Conference on Speech Communication and Technology*, 1990, pp. 2809–2812.
- Wang, H. C., R. L. Chiou, S. K. Chuang and Y. F. Huang, "A phonetic labeling method for MAT database processing," *Journal of the Chinese Institute of Engineers*, 22(5), 1999, pp. 529-534.
- Whitney, A., "A direct method of nonparametric measurement selection," *IEEE Transactions on Computers*, 20(9), 1971, pp.1100-1103.
- Yeh, C. L. and H. J. Lee, "Rule-based word identification for Mandarin Chinese sentences - A unification approach," *Computer Processing of Chinese and Oriental Languages*, 1991, pp. 97-118.

The Formosan Language Archive: Linguistic Analysis and Language Processing

Elizabeth Zeitoun* and Ching-Hua Yu*

Abstract

In this paper, we deal with the linguistic analysis approach adopted in the Formosan Language Corpora, one of the three main information databases included in the Formosan Language Archive, and the language processing programs that have been built upon it. We first discuss problems related to the transcription of different language corpora. We then deal with annotation rules and standards. We go on to explain the linguistic identification of clauses, sentences and paragraphs, and the computer programs used to obtain an alignment of words, glosses and sentences in Chinese and English. We finally show how we try to cope with analytic inconsistencies through programming. This paper is a complement to Zeitoun *et al.* [2003] in which we provided an overview of the whole architecture of the Formosan Language Archive.

Keywords: Formosan languages, Formosan Language Archive, corpora, linguistic analysis, language processing

1. Introduction¹

The Formosan Language Archive at Academia Sinica², Taipei, is part of the Language

* Institute of Linguistics, Academia Sinica, Taipei, Taiwan

E-Mail: {hsez, harryyu}@gate.sinica.edu.tw

¹ Earlier drafts of this manuscript were presented in different occasions and among others, at The Fourth Workshop on Asian Language Resources, March 25, 2004, Sanya, Hainan Island, China and at the Tuesday Seminar at the Institute of Linguistics, University of Hawai'i at Mānoa on Feb.1, 2005. We are thankful to all the participants for their helpful suggestions and comments. We are also grateful to two reviewers for their insightful comments that helped us revise an earlier version of this manuscript.

² The Formosan Language Archive is located at: <http://formosan.sinica.edu.tw/>. The project work team (headed by E. Zeitoun) includes/included* the following assistants and members.

- Language analysis: E. Zeitoun, Hui-chuan Lin, *Tien-hsin Hsin (Rukai)
Tai-hwa Chu, E. Zeitoun (Saisiyat)
Yu-ting Yeh, E. Zeitoun
*Cui-wei Lin (Amis)
Jia-jing Hua, E. Zeitoun (Paiwan)

Archives Project, developed within the five-year National Digital Archives Program (NDAP) and launched in 2002 under the auspices of the National Science Council of Taiwan. A pilot study was conducted in 2001.

The main purpose of our project is to collect, preserve, edit and disseminate via the World Wide Web a virtual library of language and linguistic resources permitting access to recorded and transcribed Formosan text collections, comparative data and related references. Its goal is two-fold: (i) to provide a platform upon which research on various linguistic phenomena can be done through search in the Language Archive and (ii) to develop a pedagogical tool. The first goal has been partially achieved, as will be demonstrated below, but we need more funding and help from the linguistic and non-linguistic community to reach the second goal.

The Formosan Language Archive includes both Chinese and English browsing display on the Internet, and contains three main types of information databases: (1) the corpora of nine Formosan languages with annotated texts³ (see Table 1 for a list of the digitized texts, as of June 2005) and audio files if available, (2) a geographic information system and (3) four bibliographical databases. The Formosan Language Corpora consist of a trilingual platform, with Formosan texts glossed and translated into Chinese and English. Texts are assigned to four categories: (i) folktales, (ii) narratives, (iii) conversations and (iv) songs with audio files

E. Zeitoun, Qiu-yun Liu, Bukun Ismahasan (Bunun)

Lin Zhi-xian (tapest)

- GIS: Jia-jing Hua & *Bai Bing-ling
- Engineering: Ching-hua Yu
- Metadata: *Weng Cui-xia, E. Zeitoun, Ching-hua Yu
- References: E. Zeitoun, Qiu-yun Liu, Jia-jing Hua

Most of the assistants working on language analysis (Hui-chuan Lin, Tai-hwa Chu, Yu-ting Yeh, Cui-wei Lin, Jia-jing Hua, Bukun Ismahasan) are aboriginal and have been trained for years (since 1997-1998) in recording and analyzing their own language, i.e., they know how to transcribe and annotate a corpus. All the analyses are supervised by the project director.

³ The Formosan languages belong to the Austronesian language family, which includes a diversity of languages stretching west to east from Madagascar to Easter Island and north to south from Taiwan to New Zealand. There are still fourteen extant Formosan languages, five of which are moribund and are preceded with an asterisk in the list that follows. While population statistics are available, it is rather difficult to identify the number of speakers for each community. The languages include: Atayal, Amis, Bunun, *Kanakanavu (about a hundred speakers left), *Kavalan, Paiwan, *Pazih (one speaker left), Puyuma, Rukai, *Saaroa (about a hundred speakers left), Saisiyat, Seediq, *Thao (about twenty speakers left), and Tsou. Yami, spoken on Orchid Island (politically part of Taiwan) is genetically closer to the Philippine languages (the Batanic subgroup).

transcribed as faithfully as possible⁴. The Formosan Language Corpora provide different types of search systems -- sentence-based, paragraph-based, concordance-based, keyword-based, affix-based and lexical category-based -- and preserve the original work recorded by earlier scholars by providing two kinds of display, *cf.*, “original data” and “re-edited data,” which can be viewed separately or conjointly. The geographic search system permits users to determine the geographical distribution of each language/dialect. It is hoped that in the future, we will be able to further develop this system so that it will be possible to observe the expansion/decrease of a particular linguistic community over the last hundred years. Another goal is to provide mappings of phonemes, lexical items (arranged in different semantic fields) and grammatical words to allow users to see the distribution of cognates within the Formosan languages and identify areal features. The search system for the four bibliographical databases allows access to the latest information in publications about Formosan languages pertaining to linguistics, language teaching, literature and music. The display of the Archive will not be further discussed in this paper, as it has been reported in more detail elsewhere (see Zeitoun *et al.* [2003]).

Table 1. Digitized texts in Chinese and English, as of June 2005

Language	Dialect	Fieldworker and/or analyst	Texts (Stories)	Words	Sentences	Voice file (mp3)	Web Display available
Rukai	Mantauran	1) E. Zeitoun & Hui-chuan Lin [2003]	14	6598	764	60MB	✓
		2) E. Zeitoun & Hui-chuan Lin [1999-2004]	21	7000	1200	65MB	
	Maga	Tien-hsin Hsin [2002]	24	3945	419	50MB	✓
	Tona	1) E. Zeitoun [1993-2001]	12	11281	899	60MB	✓
		2) E. Zeitoun [2003-2004]	8	3400	500	35MB	
	Labuan	E. Zeitoun [2003]	9	650	200	14MB	
Tanan	Paul Li [1975]	26	10656	1237	--		
Saisiyat	Tunggho	1) Chu Tai-hwa [2003], supervised by E. Zeitoun	14	4479	374	30MB	✓
		2) Chu Tai-hwa [2004~], supervised by E. Zeitoun	3	800	250	15MB	
Atayal	Squliq	Ye Yu-ting [2003] supervised by E. Zeitoun	20	10439	1476	80MB	✓
Tsou	Tfuya		48	9088	1362	70MB	✓
	Tapangu	Tung <i>et al.</i> [1964]	57	8334	1003	66MB	✓
	Duhtu		29	5589	661	43MB	✓

⁴ Texts are recorded in the villages where the informants live (usually either inside or outside their houses). Texts recorded in the Paiwan language have also been video-taped. The informants are free to record narratives, folktales or songs. Conversations only include two speakers.

Amis	Central	Fey <i>et al.</i> [1993]	25	50000	1780	200MB	✓
Bunun	Isbukun	Tseng <i>et al.</i> [1998]	49	35089	1265	--	✓
Kanakanavu	--	Tsuchida [2003]	10	5961	781	--	✓
Pazih	--	Li & Tsuchida [2001]	31	7590	991	--	✓
Paiwan	Southern	Hua Jia-jing [2004~2005] supervised by E. Zeitoun	20	12000	800	55MB	

The goal of the present paper is to discuss the linguistic analysis approach adopted in the Formosan Language Corpora and the processing programs that have been developed for it. Indeed, the digitization of various Formosan languages and dialects has posed numerous challenges on both the linguistic and computational levels. We have had to develop not only a uniform annotation system to account for language variation and typology but also processing tools for annotating the growing corpus and retrieving and displaying the data from/on the Internet.

This paper is organized as follows. In section 2, we discuss problems related to the transcription of different corpora. In section 3, we deal with annotation rules and standards. In section 4, we turn to the notion of text structure. In section 5, we discuss problems related to analytic and programming consistency. Conclusions are drawn in section 6.

2. Transcriptions

In this section, we first deal with the orthographic system adopted in the Archive and then discuss IPA conversions from one operating environment (Word) to another (Web).

2.1 Orthographic system adopted in the Archive

We first outline the phonemic inventory of the Formosan languages. We then provide an overview of the diverse writing systems that have been used to transcribe the Formosan languages. Finally, we deal with the problems raised by these writing systems, and explain our preference for using IPA for standardized transcription.

2.1.1 Outline of the phonemic inventory of the Formosan languages

The Formosan languages exhibit fairly simple phonemic inventory systems consisting usually of no more than twenty consonants and four vowels, which typically include a series of voiceless and voiced stops: /p, t, k, q, ʔ, b, d, g/; an affricate: /ts/; fricatives: /s, z/; a series of nasals: /m, n, ŋ/; liquids: /l, r/; and four vowels: /a, i, u, ə/. Of course, there is great variation among these languages which has arisen through phonological changes. They will not be detailed in the present paper. Most noticeably, Paiwan has developed a series of palatals: /c, ɟ, ʎ/; Rukai, Paiwan and Puyuma exhibit a partial/full series of retroflexes: /ɬ, ɗ, ʁ/. Atayal, Seediq, Bunun, Paiwan and Thao distinguish between velar and pharyngeal sounds, while

Amis differentiates glottal and epi-glottal sounds [Li 1999]. A few languages such as Sguliq Atayal, Tsou, Maga Rukai and Saisiyat have developed more complex vocalic systems. All the consonants and vowels found in the Formosan languages are given in Table 2 below.

Table 2. The phonemic inventory of the Formosan languages

【 CONSONANTS 】

		labial	Dental	palatal	retroflex	velar	pharyngeal	epi-glottal	glottal
stop	-vd	p	t	c	ʈ	k	q	ʔ	ʔ
	+vd	b ɸ	d d'	ɟ	ɖ	g			
affricate			ts						
fricative	-vd	f ɸ	θ s	ʃ	ʂ	x	χ	h	h
	+vd	β v	ð z	ʒ	ʐ	ɣ	ʁ		
nasal		m	n			ŋ			
liquid			l ɭ [lh]	ʎ	ɭ				
trill/flap			r r [r]						
glide		w		y					

【 VOWELS 】

	front	central	back
high	i	ɨ ʉ	u
mid	e	ə, œ	o
low	æ	a	

The basic syllable structure in most languages is CVC, though both Rukai and Tsou now exhibit a CV syllable structure. Consonant clusters occur in only a few languages (e.g., Tsou, Maga Rukai, Thao and Atayal). Stress is usually non-phonemic.

2.1.2 Writing systems adopted to transcribe the Formosan languages during the past four hundred years

Different writing systems (alphabetic, syllabic and logographic) have been adopted to transcribe the Formosan languages during the past four hundred years. Four stages can be distinguished that reflect the history of Taiwan. The last of them is the most complex.

Dutch colonization (1629-1661):

The Roman alphabet was first used in Taiwan in the 17th century by Dutch missionaries to record Siraya and Favorlang. They devised a Romanization system based on the Dutch spelling, which at the time had not yet been standardized.

Chinese colonization (1661-1895):

With the colonization of Taiwan by the Chinese, many land contracts, songs, place or family names and reports were transcribed with Chinese characters. The phonetic value of these Chinese characters is somewhat complex, sometimes referring to Mandarin Chinese and at other times to the Minnan pronunciation.

Japanese colonization (1895-1945):

From 1895 to 1945, Taiwan was a Japanese colony. Aboriginal children were enrolled in schools (up to the age of 12) and learnt Japanese, so they were able, in later years, to transcribe their own language in katakana.

Post-1945:

With the arrival of the Nationalist Chinese under the leadership of Chiang Kai-shek, the Chinese government imposed Mandarin Chinese as the sole official language. The Zhuyin fuhao system more popularly known as Bopomofo, was introduced and used in textbooks, dictionaries etc. At one time, it was also used to transcribe the Formosan languages (e.g., the Bible, songs and textbooks). Bopomofo consists of 37 symbols derived from Chinese characters, and some of these symbols were slightly altered to convey sounds recorded in the Formosan languages that are not found in Chinese. Different writing systems (all Romanized) were devised by the Catholic and the Protestant Church and used during the same period. The lack of adherence to common principles had the unfortunate consequence of producing different writing systems for different tribes. Diacritics were introduced: in Amis, for instance, ^ is used to represent a glottal stop.

In 1991, Prof. Li Jen-kuei [Li 1992] was asked by the Ministry of Education of Taiwan to devise writing systems for the Formosan languages and proposed different solutions (e.g., replacing IPA symbols such as *ŋ* with a capital letter *N* or with two symbols, *ng*).

In 2002, linguists were asked by the Council of [Taiwan] Indigenous Peoples, Executive Yuan, to work in collaboration with each tribe according to their individual expertise and finalize the orthographic system(s) of all the aboriginal languages of Taiwan. This has also led to a variety of Romanized systems that try to improve on the Romanization systems of the Catholic and Protestant Church while taking into account Li's [Li 1992] recommendations.

2.1.3 Problems raised by more recent Romanized writing systems

We will not discuss problems with earlier writing systems (the Dutch-based transcription system and the use of Chinese characters and symbols) as these have been addressed elsewhere (see, for instance, Adelaar [1999] and Rau [1995]). We will, rather, focus on the inconsistencies in the Romanization systems, devised either by missionaries or linguists.

The various Romanization systems devised by missionaries were not usually based on a phonemic representation of the language being transcribed. This had, in many cases, an unfortunate consequence: a relevant phonemic contrast was not represented while other non-distinctive features were taken into account. Early *et al.* [2003:15] showed, for instance, that in Paiwan the orthography used by two Swiss Catholic missionaries did not distinguish between the two phonemes /tj/ and /dj/ but represented both phonemes with a single graph, cf., *tj*. Li [1992:21] also noted that an orthographic system was devised for Paiwan whereby a distinction between /θ/ and /s/ was made, but such a contrast does not exist in that language.

No common principles have been applied to the Formosan languages nor have they been consistently adopted among linguists. In Amis, for instance, *d* represents the lateral fricative /ʎ/, but in all the other Formosan languages, it refers to a dental /d/. Blust (see Blust [2003]) transcribes [θ] as *c* (to show phonological change, PAN *C > Thao θ), while most other linguists transcribe [θ] as *th*. Table 3 provides a comparison of the various symbols used to transcribe the Formosan languages along with their IPA equivalents.

Table 3. Comparison of the various graphs used to transcribe the Formosan languages along with ipa equivalents

【VOWELS】	
IPA	GRAPH
i	i
ɨ	ɨ
ʉ	ʉ ɨ U
u	u
e	e, é ⁵
œ	oe
ə	e
o	o
æ	ae
a	a

⁵ This graph is used for Maga Rukai, which has /ə/ and /e/ as distinctive vowels.

【 CONSONANTS 】

IPA	GRAPH	IPA	GRAPH
p	p	s	s
t	t	ʃ	sh
ts	ts	ʃ	sh
	c		S
t̥	tr	x	x
	T	χ	h
c	tj	h	h
	t̥	h̥	
	t̥	v	v
k	k	ð	dh
q	q	z	z
	ʻ		z
ʔ	ʻ		rh
ʔ	ʻ	z̥	z
	^	ʁ	R
b	b	m	m
ḃ		n	n
β		ŋ	ng
d	d	ɬ	d
d̥			l
d̥	dr		l
	D	l	
	rh	lr	
	r̥	L	
j	dj	ɮ	ɬ
	d̥		l̥
	d̥		l
g	g	r	lj
ɣ			
f	f	r	r
ϕ		w	w
θ	th	j	y
	c		

To overcome the problem of non-standardization in the current writing systems, we decided to record or re-edit texts in IPA, a recommended standard used in many Archive projects (e.g., the Rosetta Project). However, to preserve the integrity of earlier recorded data, we keep intact original materials recorded with certain Romanization systems and produce

new versions of these based on our own standardized format.

It became necessary for us to make changes in our corpus, as we were including more and more languages. The first languages we started to digitize and to annotate were Rukai, Atayal and Tsou. The commonly accepted use of *c* in Formosan linguistics as a replacement for [ts] seemed at the time the best choice⁶, as there are consonant clusters in three of these languages, cf., Maga Rukai, Squliq Atayal and Tsou. However, the introduction of Paiwan, in which there is a distinction between palatalized and non-palatalized sounds, forced us to change our writing policy though, as *c* is the standard IPA symbol used to represent a palatal stop. We thus changed the earlier *c* to *ts*, to distinguish the affricate [ts] from the palatal [c].

Other changes may be needed in the future as we include more languages, but we plan to keep them to a minimum.

2.2 Using Unicode IPA symbols

To convert IPA symbols from Word documents (in which texts are typed) to the Web, we make use of the Unicode encoding system, which offers the possibility of displaying symbols uniformly across browser platforms. In Unicode, each IPA character is assigned a standardized encoding number so as to avoid using the same code for two different symbols. In theory, Unicode represents the best way to display IPA characters on the Web. In practice, it requires an initial configuration. Displaying IPA symbols on certain platforms is sometimes difficult as will be shown below (Webster [2002]).

This section discusses how we use IPA in our two working environments (Word and the Web) and how we convert IPA symbols into a computer-readable form.

Three things are required to convert IPA symbols from word processing documents into HTML files:

1. An operating system that supports Unicode (e.g., Windows 2000/XP).
2. An installed Unicode font that includes IPA (e.g., *SILDoulosIPA* for Microsoft Word, and *Lucida Sans Unicode* for the Web).
3. A Unicode-compliant application (e.g., Microsoft Word or Internet Explorer).

2.2.1 Creating Word processing documents

All the texts included in the Formosan Language Corpora contain different kinds of information: metadata information, utterance identifications, orthographic transcriptions, interlinear word-glosses and free translations. Specific IPA symbols are introduced in the files whenever necessary. We make use of the Unicode-compliant font *SILDoulosIPA*, made

⁶ At the same time, we started to analyze data on Saisiyat, a language that has no affricate.

available through SIL. The data is typed as follows:

(1) **ʔinaʔi vaha-nai ʔi ʔoponoho toramoro ka ma-kotsiŋai.**

這 話-我們.屬格 * 萬山 很 * 狀態.虛擬式-難

this language-IPE.Gen * Mantauran very * Stat.Subj-difficult

我們萬山話很難(學)。

Our language is very difficult (to learn). (Zeitoun and Lin [2003, ex. 01-002-a])

Strictly speaking, a Word document is not an ASCII text file, as it may contain formatting code (e.g., indenting, italics, etc.) and IPA symbols, which are challenging for computer processing. It is thus necessary to convert these phonetic symbols into computer-readable forms. Thus the interoperability can be achieved on another application or platform. A macro can be used to transliterate IPA symbols as decimal numeric entities. For instance, the **ɔ̃** character is rendered by the HTML code **ð**. Each IPA symbol is automatically converted into its corresponding numeric reference entity throughout a document. When this operation is finished, we import these alphanumeric characters into the textual database. Once the database has been established, the query operation can be performed as desired.

2.2.2 Creating Unicode IPA Web pages

To display IPA symbols in Web pages, some preliminary work must be done by the user, i.e., his/her computer must be configured with a Unicode IPA font and a Unicode-compliant browser for viewing IPA symbols on the Web. Internet Explorer automatically views web pages encoded with UTF-8, an encoding standard, provided that an appropriate font is installed. As for the font, most Windows 2000/XP machines make use of the *Lucida Sans Unicode* font, which contains the Unicode IPA symbols.

In order to display Unicode IPA Web pages, we declare that the HTML page is using:

```
(2) <head>
    <meta http-equiv="Content-Type" content="text/html; charset=utf-8">
    ...
</head>
```


Then, we need to either specify the name of the font locally, e.g., *Lucida Sans Unicode* as:

```
(3) <font face="Lucida Sans Unicode">⊂</font>
```

or declare it globally in the <head> element of the HTML page, for example:

```
(4) <head>
    ...
    <style type="text/css">
        BODY {font-family: Lucida Sans Unicode; font-size: 10pt;}
    </style>
    ...
</head>
```

Our database keeps track of all the graphs and symbols in each corpus. In other words, not only the Romanization systems but also the numeric reference entities for IPA symbols are stored. This means that when a query is issued from the user's machine, the request is then sent to the server application, which sends the query command to the back-end database, producing a query result that satisfies the initial criteria. The result is then sent back to the server program, which finally produces the HTML output for the user. Our web application is oriented to both browsing and searching the corpus. Either method displays the HTML output, including the IPA codes (if any), and finally displays it in the client browser. If the client computer has the appropriate font installed, e.g., *Lucida Sans Unicode*, then the IPA symbols are guaranteed to be displayed correctly; if not, the user's web browser will display "?????" or empty boxes .

2.2.3 Keyword search with IPA symbols

As briefly outlined above (see section 1), the Formosan Language Archive not only permits the browsing of texts, but also allows for searching based on (i) keywords, (ii) list of affixes and (iii) lexical categories. While the search through affixes and lexical categories is rather simple, as the user browses a separate database⁷, keyword search is one of the most important features of the Formosan Language Archive. The search can be made by typing a word in any

⁷ These two databases can be cross-referenced, i.e., if a user intends to look for the distribution of a particular affix, then examples will be drawn from the main text archive.

of the Formosan languages included in the corpora, its Chinese or English translation or glosses. Of interest for us is searching performed by typing a word in a specific Formosan language. Since each corpus includes IPA symbols, the type of search must also handle these.

The two applications we are using, Microsoft Word and Internet Explorer, do not allow the automatic insertion of Unicode IPA. However, it is easier to insert manually IPA symbols in Word than in Internet Explorer. The insertion of IPA symbols will first be discussed here with respect to these two environments. We will then explain how we devised a keyboard mapping mechanism that allows the insertion of IPA symbols on the Web.

In Microsoft Word (e.g., 2000/XP), there are several ways to insert a Unicode IPA symbol. The first is the well-known *Insert...Symbol* menu command. After *Insert...Symbol* is chosen, a Unicode font is then selected, the pull-down list on the right displays all of the Unicode code points (such as “IPA Extension”) included in that font. The second method consists of using the AutoCorrect feature, which is designed to replace mnemonic abbreviations with their Unicode IPA equivalents. This method is handy, but a constraint is placed on codes. They must all begin and end with a non-alphabetic character (see Webster [2002]). A third method consists of inserting IPA symbols using the find/replace function.

It is extremely difficult, if not impossible to insert Unicode IPA symbols when using Web browsers like Internet Explorer. Such symbols, if inserted, usually become empty boxes in the field. To display such symbols, we decided to design a keyboard in which all occurring IPA symbols (so far, 15) along with their numeric equivalents could be displayed (Figure 1). When the user clicks on one of the IPA buttons, the reference code is inserted into the field automatically, and the code is enclosed by “less than” and “greater than” marks (e.g., <660>). The reason for not inserting the typical reference entity (e.g., ʔ) directly is that the ampersand character is significant for Web processing. When the field data is posted onto the server, the Web application can manipulate it due to its computer-readability. In the server, each of the posted IPA symbols is converted back into the standard entity (e.g., \$#660;). During this process, we can guarantee that the search string is kept undistorted when sent to the server. It should be noted, however, that a few IPA symbols are able to appear AS IS in the field. Even so, these symbols would be urlencoded⁸ into unexpected character strings which would be hard for the program to parse.

When we started digitizing data on the Formosan languages and were confronted with the insertion of IPA symbols on the Web, we found the above method most acceptable. The sole limitation is that users must have installed Unicode IPA symbols beforehand to take advantage of this type of input mechanism.

⁸ This method is normally used when the browser sends form data to a Web server. It replaces spaces with "+" signs, and unsafe ASCII characters with "%" signs followed by their hexadecimal equivalents.

Keyword (original):

Keyword (English):

Lexical Category:

Personal Pronoun:

Usage:

1. To enter an IPA character using original language, press the code button instead:

IPA	ð	ŋ	ɖ	ə	ɭ	ʔ	ɨ	θ
Code	240	331	598	601	621	660	616	952

IPA	∅	ʃ	æ	œ	β	ɰ	ʏ
Code	216	643	230	339	946	649	404

2. To enter the English keyword, please type any word that may occur (such as book, wine, etc).

Figure 1. IPA Keyboard Mapping

3. Annotation rules and standards

3.1 Ontology of different Formosan languages

The use of language codes is necessary when constructing the ontology of different Formosan language families included in the corpora. Our coding system is actually based on the latest version of Ethnologue, which was developed by the Summer Institute of Linguistics and is available on the Internet (e.g., DRU for Rukai and BNN for Bunun). As the SIL website does not provide abbreviated names for dialects. We use a two-letter code based on the dialect name itself (e.g., Mn from *Mantauran* Rukai). Thus, the language and the dialect codes form distinct entries in our database.

The codes used for the Formosan languages (along with the dialects they include) that are being archived are shown in Table 4.

Table 4. The code system used in the Formosan language archive

Language	SIL Code	Dialect	Code
Rukai	DRU	Mantauran	Mn
		Maga	Mg
		Tona	To
		Budai	Bu
		Tanan	Ta
		Labuan	La
Saisiyat	SAI	Taai	Ta
		Tungho	Tu
Atayal	TAY	Squliq	Sq
		C'uli'	Cu
Tsou	TSY	Duhtu	Du
		Tfuya	Tf
		Tapangu	Ta

Amis	ALV	Sakizaya Northern Tavalong-Vata'an Central Southern	Sa No Ta Ce So
Bunun	BNN	Takituduh Takibakha Takbanuaz Takivatan Isbukun	Td Th Tb Tn Is
Kanakanavu	QNB	Kanakanavu	Ka
Puyuma	PYU	Nanwang Kapitul	Na Ka
Paiwan	PWN	t-dialect tj-dialect	Td Tj
Pazih	PZH	Pazih Kaxabu	Pzh Kx

3.2 Rules for annotating the corpus in English and Chinese

The Formosan languages are morphosyntactically heterogeneous, and though the literature on a number of Formosan languages is now much more abundant than it used to be, many grammatical phenomena have yet to be clarified or need to be further investigated. This poses a challenge for the analysis of each Formosan language corpus that we deal with, as will be explained below.

As pointed out by Zeitoun *et al.* [2003], each text is annotated based on linguistic annotation standards. The transcription of a text in the original language is divided into utterances, sentences and clauses. Words are glossed, and sentences are given free translations. Glosses (or tagset) can be provided at two different levels: the word level (stems) and at the morphemic level (roots and affixes). The major difference between these two types of annotations lies in the fact that glosses at the word level might provide only a vague interpretation of a word and render its word formation opaque. In the texts that have been collected for the Formosan languages (e.g., Tung [1964], Li [1975]), we find that this interpretation is most often context-based (i.e., subject to the context of the whole sentence). At the morphemic level, on the other hand, roots and affixes as well as morphological alternations must be identified and further analyzed.

Since we started our research in 2001, we have applied a morphemic analysis to annotate all the texts that have been recorded or re-analyzed by ourselves. This method has many advantages in spite of its shortcomings (see below). First, the linguist can annotate the corpus consistently, i.e., words are not “contextually” glossed but their “core” meaning is sought. Second, it helps to determine the distribution and meaning of nearly each affix, thus allowing

construction of an affix database. Third, it deepens one's understanding of the grammar of a specific language, making it easier to identify major lexical and syntactic categories (also included in a database).

The first corpus was annotated in 2001 and focused on only one dialect of Rukai, Mantauran. Over the past four years, as different languages have been annotated, we have been obliged to add more abbreviations to our original list, taking into account morphosyntactic distinctions that exist in these languages. This does not pose a problem, as far as linguistic analysis is concerned, because we know that the Formosan languages exhibit much typological variation. As our abbreviation list was discussed in Zeitoun *et al.* [2003], we will only deal in this section with problems that have arisen due to inclusion of more languages in our corpora.

The addition of new abbreviations has had two different consequences: (i) the use of particular glosses for a single language, and (ii) the insertion of new symbols to distinguish different types of affixes. We will discuss these two consequences in turn below.

Some of these abbreviations are (so far) only used for one language. In Atayal, for instance, there is a distinction between the immediate progressive and remote progressive (cf., *nyux* vs. *cyux*). As progressive auxiliary verbs have grammaticalized from earlier existential verbs that still co-occur productively in this language, the same immediate/remote distinction is also found in these existential verbs. This dichotomy has been reported in Seediq, a language from which collections of texts ready for digitization have not yet been retrieved. Atayal is, thus, the only language in our corpora that makes use of these four abbreviations. Other abbreviations, e.g., AF, PF, Red and LocNmz, are much more common and widely spread cross-linguistically.

One of the most important changes we have had to make has been the insertion of brackets $\langle \rangle$, commonly used to delimit infixes and recommended by the Max Planck Institute, Leipzig⁹. Initially, that symbol was not used in our glosses because in the languages that we were annotating (Rukai, Tsou, Atayal and Saisiyat), two infixes barely co-occur simultaneously. In Saisiyat, for instance, though the combination $f\langle om \rangle \langle in \rangle \beta \alpha t$ [beat<AF><Perf>beat] 'beat' is grammatically correct, it was not found in our corpus. Originally, if we had a word like $fom\beta\alpha t$ 'beat' to annotate, we would use hyphens to show its word formation, cf., $f-om-\beta\alpha t$ [beat-AF-beat], following a common practice among Formosanists. The introduction of two new languages, Bunun and Paiwan, forced us to use brackets instead, as the occurrence of two infixes in these languages is quite productive.

⁹ Abbreviations and glosses recommended by the Max Planck Institute, Leipzig (www.eva.mpg.de/lingua/files/morpheme.html) were made available to the public following the creation of our own archive.

Our newest abbreviation list is shown in Table 5. Abbreviations are given both in English and in Chinese, as one of the major goals of the Formosan Language Archive is to build a multilingual corpora in which the original orthography and Chinese-English translations co-exist.

Table 5. Abbreviations used in the Corpora

ABBREVIATION	CHINESE	ENGLISH
ActNmz	動態名物化	Action nominalization
AF	主事焦點	Agent Focus
Asp	時貌 (或 動貌)	Aspect
Caus	使役	Causative
ClsNmz	分句名物化	Clausal nominalization
Cnc	讓步	Concessive
Cntrfct	違反事實	Counterfactual
Dyn	動態	Dynamic
E	排除式 (= 我們)	Exclusive
EP	強調助詞	Emphatic Particle
Excl	驚嘆語	Exclamation
Ext.Imm	存在.近距	Existential Immediate
Ext.Rem	存在.遠距	Existential Remote
Fill	填充語	Filler
Fin	限定	Finite
FP	語尾助詞	Final Particle
Fut	未來	Futute
Gen	屬格	Genitive Case
Hab	習慣	Habitual
HP	勸建助詞	Hortative Particle
I	包含式 (=咱們)	Inclusive
IF	工具焦點	Instrumental Focus
Imp	命令	Imperative
Imprs	無人稱	Impersonal pronoun
InstNmz	工具名物化	Instrument nominalization
Irr	非實現	Irrealis
LF	處所焦點	Locative Focus
LF.Hort	處所焦點.勸建	Locative Focus Hortative
Lig	連繫詞	Ligature
Loc	處所格	Locative Case
LocNmz	處所名物化	Locative nominalization
NAgPass	非主事被動	Non agentive passive
Neg	否定	Negation
NegImp	否定命令	Negative Imperative
NFin	非限定	Non-Finite

NSpec	未指定	Non-specific
Nom	主格	Nominative Case
ObjNmz	受事名物化	Objective Nominalization
Obl	斜格	Oblique
Pass	被動	Passive
P, plur	複數	plural
Perf	完成貌	Perfective
PF	受事焦點	Patient Focus
PF.Hort	受事焦點.勸建	Patient Focus Hortative
Prfct	完成進行	Perfect
Prog.Imm	進行.近距	Progressive Immediate
Prog.Rem	進行.遠距	Progressive Remote
QP	引述助詞	Quotative Particle
Real	實現	Realis
Ref	反身	Reflexive
Rec	相互	Reciprocal
Red	重疊	Reduplication
S	單數	Singular
Stat	狀態	Stative
StatNmz	狀態名物化	State nominalization
Spec	指定	Specific
Subj	虛擬式	Subjunctive
SubjNmz	主語名物化	Subjective nominalization
Sup	最高級	Superlative
TempNmz	時間名物化	Temporal nominalization
Top	主題	Topic
1	我(們)	1 st Person
2	你(們)	2 nd Person
3	他(們)	3 rd Person
.	帶著兩種功能之詞素	Portmanteau Morpheme
:	(可區分之)詞綴	(Divisible) Affix
-	接詞	Affix or Clitic
<	中綴	Infix
*	無法確定構詞語法功能	Morphosyntactic function undetermined

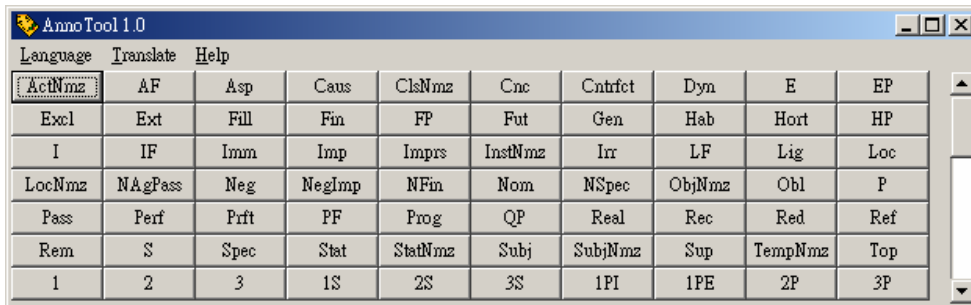
Our annotation system is not without shortcomings which we are well aware of. First, though our morphemic analysis allows for the development of different (re)search tools (e.g., keywords, list of affixes and lexical categories), the reading of a word without a whole translation of the sentence is nearly impossible for someone not familiar with Formosan languages. To cope with this problem, lists of lexical categories have been made for each

corpus that allow the user to search for a word, to determine its word formation, to check for related words and to understand its meaning. Second, morphemic analysis can be performed only if a language is well understood by the analyst. Though the project leader trained for many years aboriginal assistants in linguistic analysis, and is supervising the development of each corpus to help make the consistency rate higher through the use of the same terminology, it has become clear that to overcome analytical problems, the participation of more language specialists in the development of each different corpus is crucial. Third, while users can cross-reference rather easily both “original” and “linguistically re-annotated/re-edited” data files, our system can not display the phonetic/phonemic transcriptions of languages, as in the case of Maga Rukai for instance, where morphophonemic alternations render systematic morphemic analysis opaque. This inoperability of our system results from the fact that only a few languages exhibit such dense internal variation so that it is hard to generalize a program for the whole corpora. But this limitation has been solved by adding columns pertaining to morphophonemic alternations in the databases for lexical categories.

Other shortcomings (e.g., inconsistencies in glosses or “wrong” analyses) have been either remedied through the development of new programs or can be resolved through follow-up revisions and corrections of earlier corpora.

3.3 AnnoTool: An Annotation Tool for Formosan Languages

To help with annotation of the corpora, a program called **AnnoTool** (see Figure 2) has been developed. It has two main functions: it facilitates the tagging of texts and the translation of the linguistic terminology from English to Chinese or vice versa.



The screenshot shows a window titled 'AnnoTool 1.0' with a menu bar containing 'Language', 'Translate', and 'Help'. Below the menu bar is a table with 10 columns and 7 rows of morphosyntactic abbreviations. The first row is highlighted with a mouse cursor. The table contains the following data:

ActNmz	AF	Asp	Caus	ClsNmz	Cnc	Cntrfct	Dyn	E	EP
Excl	Ext	Fill	Fin	FP	Fut	Gen	Hab	Hort	HP
I	IF	Imm	Imp	Imprs	InstNmz	Irr	LF	Lig	Loc
LocNmz	NAgPass	Neg	NegImp	NFin	Nom	NSpec	ObjNmz	Obl	P
Pass	Perf	Pfct	PF	Prog	QP	Real	Rec	Red	Ref
Rem	S	Spec	Stat	StatNmz	Subj	SubjNmz	Sup	TempNmz	Top
1	2	3	1S	2S	3S	1PI	1PE	2P	3P

Figure 2. A screenshot of AnnoTool

When launched, the program pops up the list of morphosyntactic abbreviations used to annotate each corpus. To satisfy the requirements outlined in section 3.1, **AnnoTool** allows the expansion of abbreviations used by linguists. Its second major function is translating annotation tags from English into Chinese – or vice versa – in order to reduce the work involved in glossing each text. The above two facets of the program are explained below.

AnnoTool has been designed to work with Microsoft Word. The user can have both programs running concurrently. However, it is necessary to arrange the desktop so that the two windows do not overlap each other. As shown in Figure 3, **AnnoTool** usually occupies one-third of the screen, and Word two-thirds. When the user clicks on one of the buttons in **AnnoTool**, a tag is inserted into the Word document automatically. This method makes linguistic analysis more efficient and more accurate. It is more efficient because the linguist can view the on-screen list and stick to pre-defined terminology. It is more accurate because the likelihood of introducing typos is kept to a minimum.

Labels can be translated from English into Chinese, or vice versa. To do so, the user must first select a single term or an entire line from a document and then switch to **AnnoTool** and click on English→Chinese (or Chinese→English) in the Translate menu. Accordingly, the selected sequence in Word can be translated into one of these two languages.

We are conscious that one limitation of **AnnoTool** is that it has been programmed to handle a specific terminological set. It does not deal with the literal translation of lexical words or phrases. Nevertheless, using this tool makes our linguistic analysis easier than it used to be.

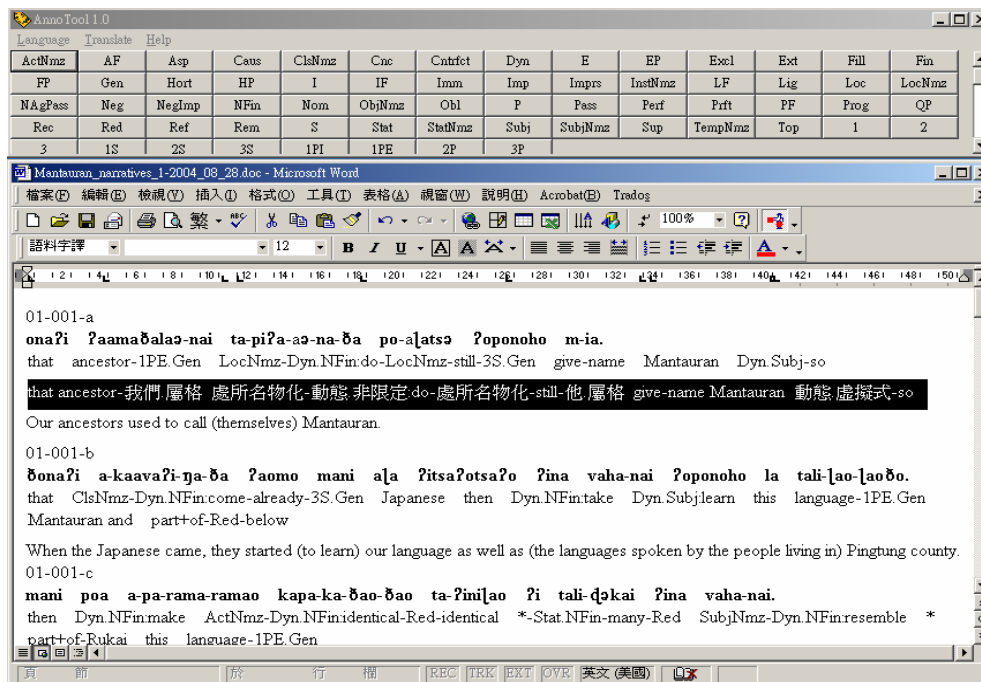


Figure 3. Using AnnoTool with Word

3.4 Affixes and lexical categories

For the tagging of major lexical categories, we follow – though with some reservations – the standardization established by CKIP in charge of the Academia Sinica Chinese Corpora. Not all of the lexical categories devised by CKIP are found in the Formosan languages, and conversely, some lexical categories not listed by CKIP are necessary to describe the Formosan languages, as illustrated in Tables 6 and 7. The set of lexical categories has been improved since two more languages (Atayal and Saisiyat) other than Rukai¹⁰ were tagged.

4. Text structure

In this section, we deal with linguistic “recognition” of clause/sentence and paragraph boundaries, and the programs that have been developed to obtain from the Internet a parallel alignment of words, glosses and sentences both in Chinese and in English.

Table 6. A comparison of existing lexical categories in Chinese and in Formosan languages

✓: lexical category found in Rukai or in other Formosan languages

(+): rare

—: non-existent

Abbreviated basic lexical categories for Chinese	Non-abbreviated basic lexical categories for Chinese	Basic lexical categories in Rukai	Basic lexical categories in other Formosan languages
1. A	Adjective	—	(+)
2. C	Conjunction	✓	✓
3. ADV	Adverb	—	(+)
4. ASP	Aspect	✓	✓
5. N	Noun	✓	✓
6. DET	Determiner	✓	✓
7. M	Measure	✓	✓
8. T	Particle	✓	✓
9. P	Preposition	✓	✓
10. VI	Intransitive Verb	✓	✓
11. VT	Transitive Verb	✓	✓
12. POST	Postposition	—	(✓)
13. FW	Foreign Words	✓	✓
14. U	Undecided	✓	✓

¹⁰ We are grateful to the two assistants, Yu-ting Yeh and Tai-hwa Chu, in charge of the Atayal and Saisiyat corpora for their help in improving the databases for lexical categories.

Table 7. Unlisted lexical categories for Chinese that must be included in our description of the Formosan languages

Other basic lexical categories not listed for Chinese	Non-abbreviated basic lexical categories	Basic lexical categories in Rukai	Basic lexical categories in other Formosan languages
1. AUX	Auxiliary	—	(✓)
2. NEG	Negator	✓	✓
3. TOP	Topic	✓	✓
4. TNS	Tense	—	(✓)
5. MOD	Mood	✓	✓
6. CM	Case marker	✓	✓
7. INT	Interrogative word	✓	✓
8. LIG	Ligature	✓	✓
9. EXC	Exclamation word	✓	✓
10. ONOM	Onomatopoeia	✓	✓

4.1 Clause/sentence and paragraph boundaries

As far as linguistic data is concerned, two major factors help in the recognition of clauses/sentences: intonation and syntactic structure. We transcribe every text based on voice files that are recorded and digitized. Though we have not taken into account nor have we tried to provide the duration of each word, intonation plays quite an important role in the detection of clause/sentence boundaries. The analyst’s knowledge of the language also helps him/her determine the beginning and the end of a clause vs. that of a sentence. To give but one example, in Tona Rukai, *si* ‘and’ can appear at the end of a sentence or between two nouns or two clauses. Syntactically speaking, it thus functions as a phrasal or causal coordinator/conjunction. In terms of discursive practices, it is used to mark a pause. That pause can be perceived as “long” (as in (5)), in which case the analyst has to treat the clause as a full sentence, or as short (as in (6)), in which case, two clauses will be treated as being coordinated and forming a longer sentence.

- (5) Tona Rukai
la ʔabəə m-wa nakay baivi si...(where ...= pause)
 then Dyn.NFin:return Dyn.Subj-go this village and
 ‘They returned to the village and...’ (Zeitoun [2004, ex. 01-002-e])

(6) Tona Rukai

la wa waməcə na bəkəʔə na caŋacaŋə
 then Dyn.NFin:go Dyn.Subj:take * pig * white and black
la paowa po-ʔaɗiŋi si la so
 then Caus:Dyn.NFin:go Caus:to-inside and then just
doo ki paŋətəɗə ʔaboalə si...
 Dyn.NFin:can * person name Dyn.Subj:come out and
 ‘Then they brought a black flecked with white pig, put it inside (the hole) and
 Pangetede could come out.’ (Zeitoun [2004, ex. 01-004-b])

4.2 Design of programs to recognize words, sentences and paragraphs

In accordance with annotating conventions, the transcription of a text is divided into sentences, which are further segmented into space-delimited words. There are two types of translations: glosses at the word level and free translations at the sentence level. Sentences are numbered for reference purposes. The encoded format of the reference number is xx-xxx-x, where the first part refers to the text id, the second indicates the paragraph id, and the third corresponds to the sentence id¹¹.

Each utterance or sentence contributes to the concept of “one block.” A block thus includes: (i) the reference information, (ii) the original utterance or sentence, (iii) word glosses and (iv) free sentential translations.

The annotated data has a three-level hierarchy. It includes the “text,” the “word” and the “sentence.” Transcriptions, glosses and translations are associated with one of these three levels. Metadata is associated at the text level. The structure is hierarchical in that a text contains sentences and words. Based on this hierarchical structure, it is easy for a computer to handle a text as an object (see Jacobson *et al.* [2000]).

A parse program was written to extract sentence and word objects from each corpus. The location of each sentence, their translations and other related information are stored in the sentence-level database (Figure 4). The locations of words, their transcriptions, their word order, Chinese and English word glosses, and punctuation are stored into the word-level database (Figure 5). The *location* field, as a primary key, is used to relate one database to another.

¹¹ In the corpus files, we simply use the three-part encoding format to represent a sentence location. In the implementation of the database, however, we prefix it with a language id. It then has the final format of xxxxx-xx-xxx-x, in which the first part stands for a language and its dialect.

翻譯: 資料表		
location	c_freetrans	e_freetrans
DRUMn_01_001_a	我們的祖先萬山自稱是萬山人。	Our ancestors used to call (themselves) Mantauran.
DRUMn_01_001_b	日本人來了以後就(開始)學我們	When the Japanese came, they started (to learn) our language as we
DRUMn_01_001_c	他們比較這兩種語言後, (就發現	They compared (our) languages and (discovered that) there were n
DRUMn_01_001_d	然後他們告訴我們說: 「你們是同	Then they told us: "You share the same ancestry."
DRUMn_01_001_e	我們萬山人(才)知道「原來咱們	(That's how) we, Mantauran, learnt that actually we were Rukai.
DRUMn_01_001_f	日本人還沒來之前, 我們不知道我	Before the Japanese came, we did not know we were related to the

Figure 4. Sentence-level database

原文與註解: 資料表							
location	wordorde	orthog	punct	pul	pur	cgl	egls
DRUMn_01_001_a	0	onaʔi				那	that
DRUMn_01_001_a	1	ʔaamaðala&				祖先-我們屬格	ancestor-1PE.Gen
DRUMn_01_001_a	2	ta-piʔa-aə ;n				處所名物化-動態	LocNmz-Dyn.NFin.do-
DRUMn_01_001_a	3	po-aɭatsə ;				取-名	give-name
DRUMn_01_001_a	4	ʔpponoho				萬山	Mantauran
DRUMn_01_001_a	5	m-ia				動態-虛擬式-這樣	Dyn.Subj-so

Figure 5. Word-level database

There is no translation at a higher level than the sentence, so there is no need for a paragraph-level table. The free, sentence-level translations can be strung together and arranged in the original order, and they serve as intelligible, if not always smooth or elegant, translations of the whole text.

4.3 Bilingual translation and alignment

Our project consists of multilingual parallel corpora, which in turn consist of Formosan utterances and bilingual translations. At the sentence level, a source segment and two translations of this source are included. At the word level, each lexical unit and their bilingual glosses are included. From the typographic format, two linking correspondences can be inferred from the text: sentence alignment and word alignment. We developed a morphological program to convert the implicit structure of the text into the XML format, which is now the commonly used standard for corpus encoding (Figure 6), as well as a database format (Figures 4 & 5), which can be utilized by the relational method and is accessed by using the structured query language (SQL). When corpus information has been encoded in such formats, it is easier to handle the alignment problems.

```
<?XML version="1.0" encoding="BIG5" ?>
<TEXT id="01" code="DRUMn" lang="Rukai" dial="Mantauran">
<HEAD>
```

```

<TITLE>Our language</TITLE>
...
</HEAD>
<BODY>
<S id="01-001-a">
  <TRANSCR>
    <W><FORM WO="0">ona&#660;i</FORM><CGLS>那</CGLS><EGLS>that</EGLS>
      </W>
    <W><FORM WO="1">&#660;aama&#240;ala&#601;-nai</FORM><CGLS>祖先-我們.屬格
      </CGLS><EGLS>ancestor-1PE.Gen</EGLS></w>
    <W><FORM WO="2">ta-pi&#660;a-a&#601;-na-&#240;a</FORM>
      <CGLS>處所名物化-動態.非限定:做-處所名物化-還-他.屬格</CGLS>
      <EGLS>LocNmz-Dyn.NFin.do-LocNmz-still-3S.Gen</EGLS></W>
    <W><FORM WO="3">po-a&#621;ac&#601;</FORM><CGLS>取-名</CGLS>
      <EGLS>give-name</EGLS> </W>
    <W><FORM WO="4">&#660;oponoho</FORM><CGLS>萬山</CGLS>
      <EGLS>Mantauran</EGLS></W>
    <W><FORM WO="5">m-ia</FORM><CGLS>動態.虛擬式-這樣</CGLS>
      <EGLS>Dyn.Subj-so</EGLS></W>
    <PUNCT>.</PUNCT>
  </TRANSCR>
  <FREETRAN lang="Chinese">我們的祖先自稱萬山人。</FREETRAN>
  <FREETRAN lang="English">Our ancestors used to call (themselves) Mantauran.
  </FREETRAN>
</S>
...
</BODY>
</TEXT>
</XML>

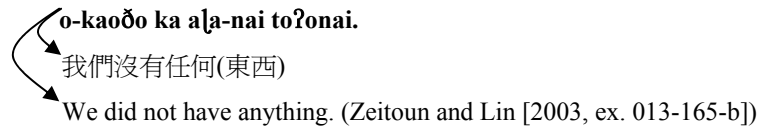
```

Figure 6. XML markup of a linguistic text

4.3.1 Sentence alignment

According to our conventional notations, sentences have been aligned since the first corpus (that for Mantauran Rukai) was initially built on a sentence-by-sentence basis. Then the Chinese and English translations were appended. They are clearly distinguishable for distinct line position in the file:

(7) Mantaaran Rukai

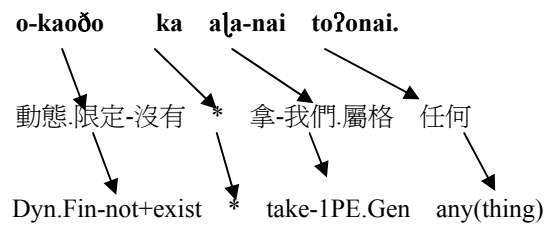


To keep the sentences aligned, our approach maps the linking relationships of the sentence segments and stores or encodes them in a standard format. In other words, the sentential information is stored in the individual fields of a record or in certain XML node elements.

4.3.2 Word alignment

In the Formosan Language Corpora, each uttered word is space-delimited and owns its bilingual glosses appear below it. If no gloss is available, then an asterisk * replaces it:

(8) Mantaaran Rukai



Interlinear morpheme-by-morpheme glosses provide most of the information necessary to build a word alignment database. In the database design, each record is based on a transcribed word. This lexical unit includes important pieces of information, such as a unique identifier (here called a *location*), a spelled form, a specific word order, and glosses. Word order plays a major role in word arrangement. It allows words (along with their glosses) to be pieced together and to reappear in the same order as in the original format.

Word alignment provides a basis for the extraction of bilingual lexicons. Using the alignment database, we can get the full index of a particular language. However, as each word is deliberately cut into pieces corresponding to morphemes rather than being given a literal meaning, it is impractical to put them together in the order of the source language since the result would be incomprehensible gibberish. That is why we provide a lexical category search, which allows the user to browse the meaning of each word, and a reference to its word formation (see section 3.1).

Our aligning strategy thus consists quite simply of arranging words with their

corresponding glosses and attaching bilingual translations in a correct order. It will be possible to enhance this approach in the future so that it can be used for other processing tasks.

5. Consistency

The consistency of transcribed words and aligned glosses is one of our major concerns in the construction of each corpus. Even though the various types of corpora collected follow standard notations (IPA transcriptions, interlinear glosses and sentence translations), a certain degree of inconsistency can be found in each corpus.

Inconsistencies can be found at different levels: lexical (i.e., inconsistency in transcriptions; word glossing problems), morphological (incorrect identification of morpheme boundaries, hesitation regarding the distribution of certain affixes or roots), and syntactic (differences in syntactic structures between different dialects or languages that may yield incorrect interpretations of the data at hand). We provide examples of these three types of inconsistencies below and show how we are able to deal with them.

5.1 Transcriptions and word glosses

Certain incorrect transcriptions are easily “repaired,” e.g., in Tanan Rukai, Li [1975] recorded ‘very’ as *aramor* and *?aramor*, but later fieldwork showed that the second instance is the correct one. Other discrepancies are more difficult to account for. In Mantauran Rukai, the term *ivoko* ‘male friend’ contrasts with *la-?ivoko* ‘male friends’. We checked both words many times, and both forms are correct (the first without a glottal, and the second with a glottal).

Word glosses pose another challenge to the linguist, who, for one thing, must be familiar with the culture of the language in question. We are confronted with two interrelated problems: (i) the analyst must decide on the “core meaning” of a word, but at the same time be aware of instances of polysemy or homophony; (ii) there must be a concordance between Chinese and English, but that concordance will sometimes be difficult to reach. To give one example, we were confronted with a series of words in Mantauran Rukai that have to do with social organization, cf., *va|ova|o* ‘young (between 15 and 30 years old), maiden woman’, *savarə* ‘young (between 15 and 30 years old), unmarried man’, *titina* ‘young or middle-aged (between 15 and 45 years old) married woman with children’ (also referring to one’s aunt or a woman of the same age as one’s mother), *tamatama* ‘young or middle-aged (between 15 and 45 years old) married father with children’ (also referring to one’s uncle or a man of the same age as one’s father). When glossing these terms, we had to make decisions about the most linguistically meaningful and culturally relevant aspects of these words and also be able to find the equivalence between English and Chinese. We finally decided to use such glosses as ‘young woman’, ‘young man’, ‘middle-aged woman’ and ‘middle-aged man’, which have

been/are being adopted for other Rukai dialects and other languages whenever necessary.

5.2 Morphology

Morphology plays a crucial role in understanding the Formosan languages, and the morphemic method we have adopted to annotate each corpus has forced us to deal even more carefully with word formation. The analyst is confronted with two major problems, (i) the incorrect identification of morpheme boundaries, and (ii) the restricted distribution of certain affixes or roots that might render their use and functions opaque.

5.2.1 Morpheme boundaries

Blust [Forthcoming] states that “most AN languages can be characterized as agglutinative-synthetic.” Our assumption is based on the fact that morphemes can either be free or bound and can include roots, function words, clitics and affixes and on the fact that morpheme boundaries are usually clear. However, morpheme boundaries might also be difficult to identify, and linguists sometimes propose different approaches to analyzing for the same words. The first problem that has to be settled is whether a word is composed of one or two morphemes. It happens that in some languages/dialects, certain words are no longer divisible, though historically, an affix could be identified. That is the case with the word *ʔoponoho* ‘name of a tribe (Mantauran) or the place they inhabit’, which derives from the prefixation of *ʔo-* (<*swa-* from) to *ponoho* (<*ponogo* ‘name place’).

Different analyses from ours are found in the literature, and we must take them into consideration. In Saisyat (Chu [2003]), for instance, we analyze *ʔi/ʔik* as a ligature, i.e., a grammatical word that carries no lexical meaning. These two morphemes occur in complementary distribution and must be glossed slightly differently, *ʔi* as ‘Lig’ and *ʔik* as *ʔi-k* ‘Lig-Stat’. The first occurs before dynamic verbs and the second before stative verbs. Li [1999], on the other hand, has analyzed both morphemes as sometimes bound and sometimes free, and translated them as ‘not’.

5.2.2 Distribution of affixes and roots

Some morphemes are invariable. Because their distribution is very much restricted and their morphophonemic/morphemic alternations are nonexistent, it might be difficult to determine their roots, their origins, their lexical categories. This is the case with Mantauran Rukai *tila!* which translates as ‘Leave/Go away’ but is actually formed with a first person plural pronoun *t(a)-* adjoined to what was originally the root *ila*. This type of analysis can only be drawn on external evidence, and as mentioned above, necessitates a good understanding of the language being investigated.

Likewise, some affixes are very non-productive, and it might be difficult to determine their meaning. This is the case with Mantauran Rukai *taʔa ʔa ʔanə* ‘house warning’ (< *ʔa ʔanə* ‘house’); the meaning of *taʔa* is still poorly understood.

5.2.3 Syntactic structures

The major problem that the linguist must be aware of regarding syntactic structures has to do with typological diversity. For instance, in Mantauran and Labuan Rukai, though subordinate temporal clauses are superficially identical, in the former, the subject is marked by the genitive, and in the latter, it is marked by the nominative.

(9) Mantauran Rukai

onaʔi	ʔiʔa	a-paka-kanə-ŋa-li
that	yesterday	Cl:Nmz-Dyn.NFin:finish-eat-already-1S.Gen
(ʔa)	o-ʔavacə-ŋa-ʔao.	
Top	Dyn.Fin-leave-already 1S.Nom	

‘Yesterday, after I had eaten, I left.’

(10) Labuan Rukai

sa	maka-kanə-ŋ-ako	ko	aga	ka
when	Dyn.Fin:finish-eat-already-1S.Nom	Acc	rice	Top
w-a-davac-ako.				
Dyn.Fin-Real-leave-1S.Nom				

‘Yesterday, after I had eaten, I left.’

5.3 Programs developed to remedy analytic inconsistencies

From the processing perspective, a hyphen is used as a morpheme boundary and as such provides morphemic information that can be used to parse word tokens (e.g., *om-ia-nai* ‘Dyn.Fin-so-1PE.Nom’) without difficulty. To remedy inconsistencies in transcriptions and glosses, all the words can be extracted from the corpus data to create an index. This index list (or finderlist) enables the analyst to compare all the words in order to minimize incorrect spelling or glosses. This program can also output a frequency list of morphemes (Hockey [1998]).

Initially, the design of each database had to take punctuation into account. We treat a

space between two words as a punctuation mark, so every word can be said to have an associated punctuation mark. Although this mark indicates a boundary between a group of words, in practice it is connected to the preceding word. Following this approach, we can treat punctuation as a field of the preceding word, as shown in Figure 5 (Leech *et al.* [1995]).

5.3.1 Word-by-word alignment consistency checker

At a very early stage in the development of the Formosan Language Archive, a program called **Chkgloss** was designed to verify the rigid structure of the corpus by comparing the number of orthographic words with that of their glosses (see Figure 7). In each corpus, transcribed words are aligned vertically with their interlinear glosses. For example:

- (11) **o-kaodɔ** **ka** **a|a-nai toʔonai.**
 動態.限定-沒有 * 動態.非限定:拿-我們.屬格 任何
 Dyn.Fin-not+exist * Dyn.NFin:take-1PE.Gen any(thing)
 我們沒有任何（東西）。
 We did not have anything.

As mentioned above, to guarantee that transcribed words are the same in number as their glosses, an asterisk is used to represent an empty word (whose meaning or morphosyntactic function remains opaque). It is only after the verification process is completed that a text can undergo whole transformation and be displayed on the Internet.

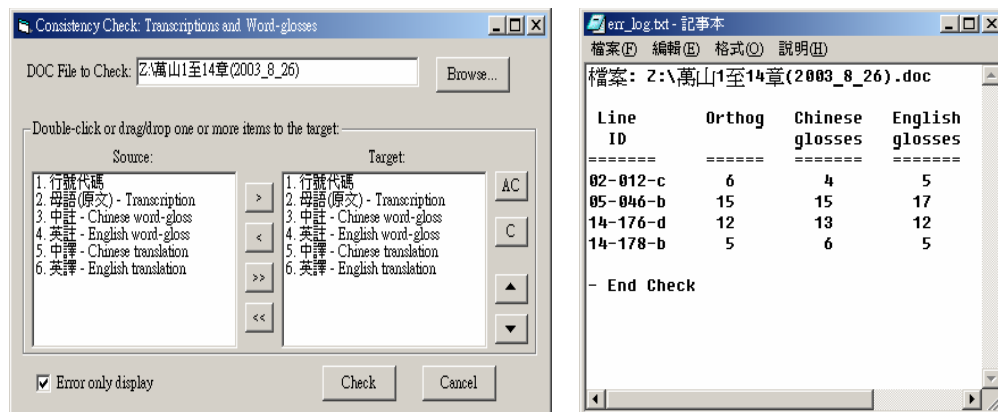


Figure 7. A Screenshot of Chkgloss

Chkgloss is helpful for identifying errors because it provides the consistency rate between (i) each tagged word and its gloss and (ii) each sentence and its bilingual translation. In most

cases, a corpus has to undergo back-and-forth processing several times before it can be deemed to be valid (Figure 8).

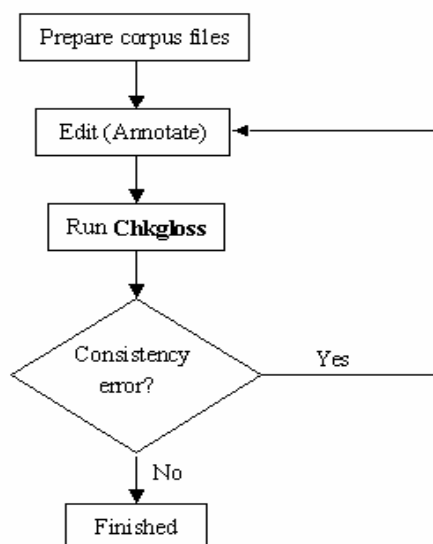


Figure 8. The workflow of using *Chkgloss*

6. Conclusion

The Formosan Language Archive is a useful tool for conducting research on the Formosan languages. The multilingual comparable corpora have begun to find their way in linguistic applications and natural language processing. As far as linguistic applications are concerned, each corpus features well-analyzed data that can serve as a basis for more in-depth studies. There are a number of advantages in providing word alignment, sentence alignment, linguistic annotations and bilingual translations. Computer-aided linguistic research is being carried out using tools and techniques that improve the work of the analyst. Applications that were developed for the Formosan Language Archive include Unicode IPA symbols, **AnnoTool**, **Chkgloss** and **Indexer**.

Drawbacks are inevitable, however. If suitable electronic text versions had been available, progress would have been more rapid. Admittedly, a lot of time has been spent on reformatting the legacy data to make it computer readable. In addition, electronic versions of earlier published materials have to be made from scratch, since there were previously no electronic files (e.g., Li [1975], Tung *et al.* [1964], Fey *et al.* [1993] etc.).

It is the purpose of our project to collect, analyze and digitize data on many, if not all, of the Formosan languages for which texts are available, but more corpora need to be included to refine the original architecture of the archive. On the other hand, we also need to think about

how to develop new tools, make use of existing tools described in the literature (cf., Szakos *et al.* [2004]) and process the voice files for further research (e.g., phonetic and discursive studies)¹². We might also be able at some point to conduct an experiment on natural language processing (e.g., corpus-based machine translation).

References

- Adelaar, K.A., "Retrieving Siraya phonology: a new spelling for a dead language," In *Selected Papers from the Eighth International Conference of Austronesian Linguistics*, ed. by E. Zeitoun and P. J.-K. Li, Symposium Series of the Institute of Linguistics (Preparatory Office), No. 1, Institute of Linguistics, (Preparatory Office), Academia Sinica, Taipei, 1999, pp. 313-354.
- Bird, S. and G. Simons, "Seven dimensions of portability for language documentation and description," *Language*, 79(3), 2003, pp.557-582.
- Blust, R., *Thao dictionary*, Language and Linguistics Monograph Series, No. A5 Institute of Linguistics, (Preparatory Office), Academia Sinica, Taipei, 2003.
- Blust, R., *The Austronesian Languages*, Ms., Forthcoming.
- Bow, C., B. Hughes and S. Bird, "Towards a general model of interlinear text," In *Proceeding of EMELD 2004: the Workshop on Linguistic databases and best practice*, Detroit, Michigan, July 15-18, 2004.
- Chu, T.-H., *Saisyat texts*, ms., 2003.
- Early, R. and J. Whitehorn, *One hundred Paiwan texts*. Pacific Linguistics, 542, Research School of Pacific and Asian Studies, The Australian National University, Canberra, 2003.
- Fey, V., *et al.*, *O'Orip no' - Amis Ameizu wenhua - Amis Culture*. Taiwan Bible Society, Taipei, 1993.
- Hockey, S., "Textual Database," *Using Computers in Linguistics: A Practical Guide*, ed. by J. Lawler and H. Aristar-Dry, Routledge, London, 1998, pp.101-133.
- Hua J.-J., *Southern Paiwan texts*, ms., 2005.
- Hsin, T.-H., *Maga (Rukai) texts*, ms., 2002.
- Jacobson, M., B. Michailovsky and B. Lowe, "Linguistic documents synchronizing sound and text," *Speech Communication*, 33, 2000, pp.79-96.
- Leech, G., G. Myers and J. Thomas, *Spoken English on Computer*, Longman Publishing, New York, 1995, pp. 208-219.
- Li, P. J.-K., *Rukai Texts*, Institute of History and Philology, Special Publication No. 64-2, Academia Sinica, Taipei, 1975.

¹² In the case of phonetic studies, the manual re-segmentation of the files would be necessary.

- Li, P. J.-K., *Orthographic systems for Formosan languages*, Ministry of Education, Taipei, 1992. [In Chinese]
- Li, P. J.-K., *The history of Formosan aborigines: Linguistic, Nantou*, The Historical Research Commission of Taiwan Province, Taiwan, 1999. [In Chinese]
- Li, P. J.-K. and S. Tsuchida, *Pazih texts and songs*, Language and Linguistics Monograph Series, No. A2-2, Institute of Linguistics, Preparatory Office, Academia Sinica, Taipei, 2002.
- Rau, V. D., "The scientific and social principles of the orthographic symbols of the aboriginal languages of Taiwan: a case study of Atayal," *The Languages of the Austronesian tribes of Taiwan*, ed. by P. J.-K. Li and Y.-C. Lin, Ministry of Education, ROC, Taipei, 1995, pp.31-47. [In Chinese]
- Szakos, J. and U. Glavitsch, "Portability, modularity and seamless speech-corpus indexing and retrieval: a new software for documenting (not only) the endangered Formosan aboriginal languages. Paper read at the Workshop on Linguistic databases and best practice," In *Proceeding of EMELD 2004: the Workshop on Linguistic databases and best practice*, Detroit, Michigan, July 15-18, 2004.
- Tseng, S.-Q. *et al.*, *zouguo shikong de yueliang*, Chen-hsing Publ. Co., Taipei, 1998. [In Chinese]
- Tsuchida, S., *Kanakanavu Texts (Austronesian Formosan)*, Endangered Language of the Pacific Rim. ELPR Publication Series A3-104, Nakanishi Printing Co., Kyoto, 2003.
- Tung, M.-N. and V. Rau, *Yami text.*, ms., 2002.
- Tung, T.-H. *et al.*, *A descriptive study of the Tsou Language, Formosa*, Institute of History and Philology, Special Publication 48, Academia Sinica, Taipei, 1964.
- Venezky, R. L., "Principles for the design of practical writing systems," *Anthropological Linguistics*, 12(7), 1970, pp.256-270.
- Webster, G., "Using Unicode IPA on the web and in word processing," University of Washington Language Learning Center, (paper available at : <http://depts.washington.edu/llc/help/presentations/index.php>)2002.
- Ye, Y.-T., *Atayal texts*, ms., 2003.
- Yu, C.-H., "Discussion on the digitization of the Formosan Language Archive – building up of the architecture of the archive." In *Proceeding of the First workshop on the Digital Library Projects*, July 25-26, 2002, Taipei.
- Zeitoun, E., *Tona texts*, ms., 2004.
- Zeitoun, E., C.-H. Yu and C.-X. Weng, "The Formosan Language Archive: development of a multimedia tool to salvage the languages and oral traditions of the indigenous tribes of Taiwan," *Oceanic Linguistics*, 42(1), 2003, pp.218-232.
- Zeitoun, E. and C.-H. Yu, "The Formosan Language Archive: Language Processing and Linguistic Analysis," In *Proceeding of 1st International Joint Conference on Language Language Processing (IJCNLP-04) Fourth Workshop on Asian Language Resources*, March 25, 2004, Sanya, Hainan Island, China.

Zeitoun, E. and H.-C. Lin, *We should not forget the stories of the Mantauran : Memories of the past*, Language and Linguistics Monograph Series, No. A4, Institute of Linguistics, Preparatory Office, Taipei, 2003.

Zeitoun, E. and H.-C. Lin, *We should not forget the stories of the Mantauran: Traditional folktales*, ms., 2004.

Zeitoun, E., *The Formosan Language Archive: Linguistic analysis and language processing*, Invited talk at Tuesday Seminar of the Institute of Linguistics, University of Hawai'i at Mānoa, January 25, 2005.

Related web pages:

<http://www.ailla.org/pc/mainindex.html>

<http://lacito.archivage.vjf.cnrs.fr/>

<http://www.emeld.org/workshop/2004/paper.html>

<http://www.language-archives.org>

<http://www.rosettaproject.org:8080/live/>

<http://www.sinica.edu.tw/SinicaCorpus>

<http://www.talana.linguist.jussieu.fr>

<http://sino-tibetan.cityu.edu.hk/rda/> -- no longer available, later moved to:

<http://victoria.linguistlist.org/~lapolla/RDA/index.html>

<http://paradisec.org.au>

Mandarin Topic-oriented Conversations

Shu-Chuan Tseng*

Abstract

This paper describes the collection and processing of a pilot speech corpus annotated in dialogue acts. The Mandarin Topic-oriented Conversational Corpus (MTCC) consists of annotated transcripts and sound files of conversations between two familiar persons. Particular features of spoken Mandarin, such as discourse particles and paralinguistic sounds, are taken into account in the orthographical transcription. In addition, the dialogue structure is annotated using an annotation scheme developed for topic-specific conversations. Using the annotated materials, we present the results of a preliminary analysis of dialogue structure and dialogue acts. Related transcription tools and web query applications are also introduced in this paper.

Keywords: Taiwan Mandarin, dialogue act, speech corpus

1. Introduction

A number of large scale corpora have been collected, processed, and made available for public use over the decades, for instance, the British National Corpus [Leech 1994] and the American National Corpus [Ide and Macleod 2001]. However, most of these corpora contain written language data only. For modern Mandarin, the Sinica Balanced Corpus contains a small percentage of transcripts of spoken data [Chen and Huang 1996]. Duanmu *et al.* (1998) also published the Taiwanese Putonghua Corpus (TWPTH) via LDC (Linguistic Data Consortium), and it includes five dialogues and thirty monologues. The above two corpora contain materials of Mandarin which is used in Taiwan. In addition, the Chinese Academy of Social Sciences (CASS) has created a national corpus of phonetically and prosodically labelled speech data for the purpose of speech synthesis [Li *et al.* 2000]. The focus of the CASS corpus is the phonetic variations of spontaneous Mandarin spoken in Mainland China. Nevertheless, because free conversations have no domain specification in the topics, it leads to greatly diverse vocabulary types and sentence varieties. It is sometimes disadvantageous to use free conversations for linguistic analysis or as engineering training data because the individual tokens available in the data are not enough for statistical analysis.

* Institute of Linguistics, Academia Sinica, Taipei, Taiwan
E-Mail: tsengsc@gate.sinica.edu.tw

To resolve this problem, several pilot spoken corpora which collected natural dialogues, such as the ATIS Corpus [Kowtko and Price 1989], the TRAINS Corpus [Heeman and Allen 1995], and the Map Task Corpus [Anderson *et al.* 1991], are all recorded in specific situations or tasks. They have been available to the research community for more than ten years. As stated by Zheng (2004), proper design of a speech corpus before the actual collection process takes place influences the value of the corpora to a great degree. Research results on spoken language processing obtained by applying the above pilot corpora illustrate the importance of spoken corpora. For linguists, corpora are more than merely data. They enable researchers to gain a different understanding of human language use. The enormous, automatic calculation power now available through modern information technology, including software and hardware, facilitates data analysis and summarization for speech engineers. However, well-designed method for the collection and processing of speech corpora will produce more information on research topics, because they will give considerations to the properties and structures of the to-be-collected corpora. This must be done beforehand by humans, and it is not a trivial task. Through the Mandarin speech corpus presented in this paper, we hope to make substantial contributions to linguistic analysis, automatic speech processing, and dialogue structure research.

2. Data Collection and Processing

The Mandarin Topic-oriented Conversational Corpus (MTCC) is part of the National Digital Archives Project (2002-2006). Its special focus is on the collection of spoken Mandarin data which reflect synchronic language use and document sociolinguistic properties in Taiwan. Our first aim in data collection is to archive conversations between familiar persons. The topics should be freely chosen by the dialogue participants to a certain degree, but restricted to contemporary events. Therefore, thirty speakers who participated in a previous project that involved collecting dialogues between strangers in 2001¹ were invited to join the MTCC project [Tseng 2004b]. They were required to come to Academia Sinica with a person with whom they were familiar. Before recordings were made, an instruction sheet was given to the speakers. It indicated that the speakers should choose one piece of news from the year 2001 and carry on a conversation about that topic. The length of the conversation was limited to twenty minutes. When the conversation time had nearly reached twenty minutes, the lab assistants signalled to the speakers to end the conversation naturally. Because the well-known Switchboard corpus is also a topic-specific corpus, we compared it with the MTCC corpus and found three main differences. (1) We collected conversations between familiar persons; the Switchboard Corpus contains conversations between strangers. (2) We recorded conversations

¹ Mandarin Conversational Dialogue Corpus (MCDC).

in an ordinary room so that the conversation partners would have visual contact with each other during the conversation. The Switchboard conversations were recorded over the telephone without visual contact. (3) We only allowed the participants to choose one event to discuss in depth (for approximately twenty minutes); the Switchboard participants chose from very general topics, such as sports, and the conversations were, in general, shorter than ours (lasting a maximum of ten minutes).

2.1 Goals of MTCC Collection

Our goal in collecting the MTCC is threefold. (1) Because this is part of the national digital archive project, the collected data must reflect the synchronic use of Mandarin in Taiwan, possibly covering lexical use, communication habits, and contemporary topics and events. (2) We intend to develop an infrastructure for building a spoken corpus that includes transcription tools, formats, database management, and tools such as a web querying system. (3) In order to undertake linguistic analysis of communication habits, annotated dialogue act data are to be produced.

2.2 Digital Recording and Subjects

The dialogues were recorded by a SONY TCD-D10 Pro II DAT tape recorder with Audio-Technica ATM 75 headset microphones at a sampling rate of 48 kHz. Each subject was recorded on a separate channel on a DAT tape. All recordings on DAT tapes were transformed into digitized audio files (.wav format) via the Tascam US224 interface. The process of collecting approximately 11 hours (6.8 GB) of conversations for the MTCC was completed at the Institute of Linguistics, Academia Sinica, in 2002.

In total, thirty-three female and twenty-seven subjects were recorded. Their ages ranged from 14 to 63. The pairs were siblings, friends, spouses, relatives, or mothers and daughters. The details of the corpus statistics are summarized in Table 1.

2.3 Transcription

Except for one dialogue in which the participants mainly spoke Southern-Min², all the dialogues were orthographically transcribed. The transcription process was performed with the assistance of TransList, which was developed specifically for collecting Mandarin spoken corpora. We decided not to use Transcriber [Barras *et al.* 2001] to process our data for two reasons. Transcriber was developed specifically for broadcast news data, so the terminology used in the programme does not fit our data type well. Second, we needed two ways of transcribing content (romanization and characters) to be input to the database conversion

² Southern-Min is the main dialect spoken in Taiwan.

programme. Therefore, it was much easier to use a working interface specifically designed for our purposes.

Table 1. Subjects in the MTCC

Dialogue Length	Subjects' Relationship	Subject's Sex: age	Dialogue Length	Subjects' Relationship	Subject's Sex: age	Dialogue Length	Subjects' Relationship	Subject's Sex: age
d-01 19 min.	Siblings	M: 40 M: 45	d-11 22 min.	Spouses	F: 34 M: 43	d-21 17 min.	Spouses	F: 36 M: 36
d-02 23 min.	Friends	F: 30 M: 36	d-12 21 min.	Friends	F: 23 M: 23	d-22 21 min.	Siblings	M: 45 F: 40
d-03 17 min.	Siblings	F: 37 F: 39	d-13 26 min.	Relatives	M: 47 F: 43	d-23 22 min.	Mother-daughter	F: 46 F: 21
d-04 11 min.	Friends	M: 26 F: 22	d-14 23 min.	Friends	F: 35 F: 43	d-24 16 min.	Spouses	M: 45 F: 45
d-05 19 min.	Friends	M: 29 F: 26	d-15 20 min.	Friends	M: 42 M: 40	d-25 22 min.	Friends	M: 22 M: 23
d-06 21 min.	Siblings	F: 33 M: 36	d-16 17 min.	Mother-daughter	F: 43 F: 15	d-26 21 min.	Siblings	F: 17 F: 14
d-07 23 min.	Spouses	F: 47 M: 46	d-17 21 min.	Mother-daughter	F: 36 F: 63	d-27 22 min.	Friends	M: 26 M: 28
d-08 18 min.	Spouses	F: 37 M: 42	d-18 21 min.	Relatives	M: 40 M: 24	d-28 20 min.	Friends	M: 29 M: 28
d-09 20 min.	Mother-daughter	F: 20 F: 49	d-19 23 min.	Friends	F: 27 F: 27	d-29 22 min.	Friends	M: 21 M: 23
d-10 25 min.	Mother-daughter	F: 48 F: 23	d-20 22 min.	Friends	F: 45 M: 53	d-30 23 min.	Friends	F: 42 F: 35

TransList provides two transcription methods: Pinyin transcription (using Latin alphabet) and character transcription. TransList automatically converts characters to Pinyin and checks the consistency of the character counts and syllable counts. Paralinguistic sounds, such as laughing and coughing, are marked in parentheses in the sentence position where they are produced. Phonetically reduced word forms are transcribed in the form of SAMPA-M in the Pinyin transcription component [Tseng 2004b]. In the character transcription component, we transcribe the full word form in characters. This follows the guidelines given by Gibbon *et al.* (1997), but is different from the transcription approach proposed by Zheng (2004), where Pinyin, characters, surface forms, and paralinguistic sounds are all documented in individual layers. In addition, two Mandarin dictionaries are used for checking standard pronunciation and mispronunciation: the Modern Mandarin Dictionary (2001) and Mandarin Dictionary (1995). Moreover, we do not segment data into sentences because the data is produced spontaneously and therefore contains a wide range of grammatical variations, e.g., ill-formed, incomplete sentences, repairs, and so on. Our solution is to arrange the dialogue content in terms of speaker turns. Furthermore, TransList provides next-phase database construction, which transforms the transcribed texts into a syllable-based database. Due to the lack of space, we will not go into the details of the transcription interface and conventions here; they can be found in [Tseng 2004a]. Details about the database construction and integration process as

well as the programmes and tools can be found in [Tseng 2004b].

2.4 Preliminary Results of Transcription

Except for the first conversation in which Southern-Min was spoken, all conversations were transcribed one by one by a transcriber. The transcription precision was relatively high, as we will see from the statistics given below. Nevertheless, before we release the MTCC corpus, a second check of the transcribed data will be necessary to ensure inter-transcriber consistency. In this phase, we will introduce a preliminary version of the corpus.

In total, 134,868 characters and 50,312 paralinguistic sounds and unclear syllables were transcribed. They were segmented by applying the CKIP³ automatic word segmentation system developed for written Mandarin (Academia Sinica). The resulting transcribed characters consist of 1,527 distinct, monosyllabic words (52,285 word tokens in total), 4,404 disyllabic words (35,296 tokens), 803 trisyllabic words (3,356 tokens), and 267 words with more than three syllables (471 tokens). In the transcribed data, a few utterances are spoken in Southern-Min. Our solution was to transcribe them in the form of Mandarin sentences while trying to keep the original meaning as much as possible. In total, 189 characters were used to transcribe Southern-Min sentences, making up approximately 0.14 % of the total transcribed characters. Five hundred and fifty-nine syllables were regarded as uncertain, and their phonetic forms were transcribed without characters. They make up about 0.4% of the total number of syllables.

3. Annotation of the MTCC

Among the twenty-nine transcribed conversations, sixteen are annotated as dialogue acts. Different from the traditional pragmatic speech act research approach [Levinson 1993] which emphasizes the function of speech acts, we focus on macro-structure annotation. Our idea is to sketch a global dialogue structure from a top-down perspective. Local phenomena, such as repairs of single words within sentences, are not considered in the annotation system. Only repairs in the form of complete propositions are annotated. Referring to the Verbmobil annotation schema for appointment scheduling and travel planning [Alexandersson *et al.* 1998], we designed an annotation system for our topic-specific dialogues in the MTCC. The Verbmobil system is based on task-specific information management, which is different from the MTCC conversations, where no concrete information is required to fulfil the task-specific

³ CKIP signifies the Chinese Knowledge Information Processing Group at the Institute of Information Sciences and the Institute of Linguistics at Academia Sinica. Because the CKIP automatic word segmentation system in principle segments compound words into smaller units, we did not experience significant problems when using the system for spoken data. However, we encountered greater problems using the part-of-speech tagging system for spontaneous spoken data.

goals. A number of particular tags, such as *politeness_formula*, *thanks*, *bye* etc., in the Verbmobil system are irrelevant to the MTCC and not considered in the MTCC annotation convention. In Sections 3.1 and 3.2, we will introduce the dialogue structure and the system of dialogue acts and in Sections 3.4 and 3.5, we will present the annotated results and a preliminary analysis of those results.

3.1 Dialogue Structure

In general, a text, whether written or spoken, consists of three components: the opening, the main body, and the closing. We are concerned with a specific type of topic-oriented conversation which resembles a formal discussion of a topic. Thus, in addition to the opening, the main body, and the closing of the conversation, such conversation components as negotiation of a topic and introduction of a topic are also relevant to the conversation style of our data. We, therefore, propose to divide the dialogue acts into five main categories: (1) *opening*: dialogues that start conversations, (2) *topic-negotiation*: dialogues that negotiate topics, (3) *topic-introduction*: dialogues that introduce topics, (4) *main discussion*: dialogues about topics and (5) *closing*: dialogues that end conversations. Furthermore, we need one category of marked up sentences for which the annotators were unable to choose suitable dialogue acts from among the available ones: (6) *sentential fragments*.

Based on the above dialogue structure, we propose a linear system for annotating dialogue acts for the MTCC. As shown in Table 2, we use thirty-seven annotation tags to mark up the discourse functions of the utterances. Unlike the sequential dialogue structure, the main discussion of a topic here is rather dynamic. The conversation participants may raise issues, exchange opinions, give examples, express different point of views, hesitate, and so on. The interaction is spontaneous, so we expect to observe a variety of discourse functions. Basically, the discourse functions of the main interaction are divided into eight types: those for managing sub-topics, expressing opinions, adding supplemental information, signalling feedback, requesting further actions and information, completing unfinished sentences, expressing exclamation, and hesitating. An overview of all thirty-seven annotation tags for dialogue acts is given in Table 2.

3.2 Dialogue Acts

Based on Table 2, this section presents a brief introduction to the annotation tags without giving explicit examples due to the lack of space⁴. **To start a conversation** contains only one annotation tag. *Opening* marks utterances used by the conversation participants to express

⁴The operational definitions and examples of the individual annotation tags are available at <http://mmc.sinica.edu.tw> (currently only in Chinese).

Table 2. Overview of dialogue acts

To start a conversation	To negotiate a topic	To introduce a topic	To talk about a topic	To end the conversation	Sentential fragments
<i>opening</i>	<i>suggest_topic</i>	<i>introduce_topic</i>	<ul style="list-style-type: none"> • dialogue acts marking sub-topic management • dialogue acts marking opinion expression • dialogue acts marking sentential supplementation • dialogue acts marking feedback • dialogue acts marking action/info requests • dialogue acts marking sentential completion • dialogue acts marking exclamation • dialogue acts marking hesitation 	<i>conclude</i>	<i>not_classified</i>
	<i>accept_topic</i>			<i>closing</i>	
	<i>reject_topic</i>				
	<i>comment_topic</i>				

Dialogue act categorization	Dialogue act annotation
• sub-topics management	<i>begin_statement, connect_statement, explain, give_example</i>
• opinion expression	<i>agree, agree_part, oppose, oppose_part, comment_by_self, comment_by_other</i>
• sentential supplementation	<i>confirm, correct, rephrase, repeat</i>
• feedback	<i>feedback, feedback_understanding, feedback_non_understanding, backchannel</i>
• action/info requests	<i>request, question, answer, question_request_answer, rhetorical_question, rhetorical_question_answered</i>
• sentential completion	<i>completion_by_self, completion_by_other</i>
• exclamation	<i>Exclamation</i>
• hesitation	<i>hesitation</i>

their readiness to begin a conversation. **To negotiate a topic** contains four annotation tags used to mark up different stages in which the conversation partners agree on a specific topic. It includes suggesting a topic (*suggest_topic*), accepting a topic suggestion (*accept_topic*), opposing a topic suggestion (*reject_topic*), and commenting on a topic suggestion (*comment_topic*). **To introduce a topic** contains only one tag used to annotate utterances that officially begin the main discussion (*introduce_topic*).

To talk about a topic contains twenty-eight tags used to annotate the main discussion between the conversation participants. A speaker makes a statement related to the topic (*begin_statement*). This can be his/her opinion on certain events related to the topic. For different sub-topics, some participants may prefer to use conjunctions or fixed expressions, e.g., “in fact” or “to be honest”, to bridge a topic shift (*connect_statement*). Within a statement, further clarifications can be made to explain the content of the statement (*explain*) or to provide examples (*give_example*). Another conversation participant can either express complete agreement (*agree*), partial agreement (*agree_part*), complete opposition (*oppose*), or partial opposition (*oppose_part*) with regard to the statement made by the other conversation

partner. Comments on statements can be made by the speaker him/herself (*comment_by_self*) or by the listener (*comment_by_other*). The listener can confirm the content of the previous statements made by the partner (*confirm*). This is not agreement on a certain opinion but simply confirmation that the stated information is correct. Utterances can be corrected (*correct*), rephrased (*rephrase*), or repeated (*repeat*).

The listener can give explicit signals through overt utterances to express that he/she is considering/processing the statement made by the speaker (*feedback*). The listener can produce simple sounds or words to show that he/she understands the message (*feedback_understanding*) or does not understand the message delivered by the speaker (*feedback_non_understanding*). Or the listener can also give simple signals such as “uh hm” to inform the speaker that the delivered message has been received (*backchannel*). Sometimes, the speaker may require answers or actions from the listener (*request*). Speakers can raise questions (*question*) or ask questions which explicitly require an answer from the listener (*question_request_answer*). Questions are answered (*answer*). Some questions are asked for rhetorical reasons (*rhetorical_question*). They are real question, but are used to trigger a new topic or a new thought. Often, these kinds of questions are answered by the speakers themselves (*rhetorical_question_answered*). Unfinished utterances can sometimes be completed by the speaker (*completion_by_self*) or by the listener (*completion_by_other*). The speaker can express exclamation (*exclamation*) or hesitate while he/she is planning the next utterance or when he/she has doubts about the content of the statement just made (*hesitation*).

To end the conversation contains two annotation tags used to close a conversation. The conversation participants draw conclusions about the topic (*conclude*) or express their readiness to end the conversation in general (*closing*). **Sentential fragments** for which the transcribers do not know the intended content are marked *not_classified*.

3.3 Annotation Example

The example given below illustrates our transcription and annotation format. Upper-case Latin letters are used to transcribe discourse particles of Mandarin. Paralinguistic sounds and pauses are enclosed in parentheses. Annotation tags begin with <b tagname> and end with </b tagname>. Word strings that fulfil the discourse function defined for a given dialogue act are annotated. They can be a single word, a single utterance, or a complete speaker turn. The length of an annotated word string is dependent on the discourse function, not the syntactic units. All word strings can only be annotated once; no cross-marking is allowed. It is also not possible to use multiple acts to annotate a single word string, either.

MISC-97 : <b connect_statement>(breathe)當然 LA 你知道(short break)當然還是我還是覺得美國人把人命看得比較值錢</b connect_statement>(short break)<b comment_by_self>因為是因爲富有 LA</b comment_by_self> (extracted from DA-2002-15.WAV , record 76/165 , 0688160-0698644)⁵

MISC-98 : <b hesitation>E</b hesitation>(extracted from DA-2002-15.WAV , record 77/165 , 0694203-0696533)

MISC-98 : <b agree_part>美國可能對他有拿美可能是美國公民他們(inhale)可能會比較重視 BA 外國的恐怕不見得</b agree_part> (extracted from DA-2002-15.WAV , record 78/165 , 0699154-0708400)⁶

MISC-97 : <b agree>對對對(pause)O 對應該是這樣子對你如果說是我的公民的話我就特別照顧你 A EN EN EN (short break)他一看情形不對在世界各地一樣 A 他就馬上撒僑 A (inhale)他也派專機 A 派專船 A 去接 A (pause)</b agree><b comment_by_self>這種事要是落到我們中國人頭上恐怕沒有沒有沒有這等好事了(short break)自己先跑了</b comment_by_self> (extracted from DA-2002-15.WAV , record 79/165 , 0704278-0428490)⁷

3.4 Annotation of Dialogue Acts

Applying the above annotation system to dialogue acts, we completed the annotation of sixteen dialogues. The results are shown in Table 3. The overall distribution of annotated dialogue acts in percentages can be found in Appendix A. The most frequently produced dialogue acts are summarized in three groups. The first contains backchannels and signals for understanding feedback. They are important for keeping a conversation going. The listener has no substantial issues to address, so it is necessary for the listener to acknowledge that the message delivered by the speaker has arrived, and that he/she is listening. The second group contains dialogue acts used to begin a sub-topic or to explain the content of a sub-topic. Both *begin_statement* and *explain* make up the essential part of a discussion. They build up the framework of the whole discussion. The third frequent group contains questions and answers. It is difficult in practice for communication to be fluent if the conversation partners simply

⁵ <b connect_statement>(breathe) of course LA you know (short break) of course still I still think that the Americans esteem human lives </b connect_statement>(short break)<b comment_by_self> because it is because they are rich LA</b comment_by_self>

⁶ <b agree_part>The United States possibly to its, have- , possibly American citizens, they esteem more (inhale) for foreigners not necessarily </b agree_part>

⁷ <b agree> yeah yeah yeah (pause) O yeah it should be so. if you are citizen, I will take care of you A EN EN EN (short break) when the situation is urgent A, the States will evacuate their citizens immediately A (inhale). They will send airplanes A ships A to pick them up A (pause)</b agree><b comment_by_self> If this happens to Chinese, I am afraid that this will not be done. (short break) you have to escape the trouble by yourself.</b comment_by_self>

express their opinions without interacting with each other. By asking questions and getting answers, the conversation partners construct natural communication.

Table 3. Annotation results (in alphabetical order)

Annotation tag dialogue	d-02	d-03	d-04	d-05	d-06	d-07	d-08	d-09	d-10	d-11	d-12	d-13	d-14	d-15	d-16	d-17	Total
accept_topic	1	0	1	1	1	0	1	1	0	0	1	0	0	1	0	0	8
agree	0	7	8	6	15	18	10	27	14	3	7	5	20	18	4	15	177
agree_part	7	4	1	0	2	6	3	2	0	1	0	5	0	5	0	2	38
answer	1	2	27	12	33	25	16	37	1	6	18	13	3	1	8	87	290
backchannel	18	19	18	63	32	52	60	28	73	9	117	117	200	79	45	80	1,010
begin_statement	37	38	17	36	28	31	27	33	12	9	3	23	9	14	8	6	331
closing	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	2
comment_ by_other	1	18	9	23	18	8	18	16	4	0	3	2	1	3	0	7	131
comment_ by_self	7	6	2	1	5	11	0	3	12	7	6	4	8	14	7	5	98
comment_topic	0	1	1	0	0	0	0	0	0	0	0	0	1	0	0	0	3
completion_ by_other	0	17	7	4	6	11	2	6	13	5	14	20	15	12	8	13	153
completion_ by_self	2	3	2	0	1	10	3	1	15	2	10	24	12	15	7	18	125
conclude	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
confirm	1	17	6	8	19	16	9	14	1	7	5	9	0	1	0	0	113
connect_ statement	3	1	0	3	1	3	1	0	2	1	1	0	3	0	3	7	29
correct	1	0	3	0	5	1	0	1	0	0	0	0	0	0	1	0	12
exclamation	0	0	1	3	0	4	0	1	0	0	0	1	3	3	0	5	21
explain	17	16	17	17	28	12	24	66	9	6	12	11	4	9	0	12	260
feedback	0	0	0	0	0	6	0	0	1	0	0	9	4	5	0	1	26
feedback_ understanding	37	15	18	43	68	29	36	40	3	0	7	15	16	5	1	28	361
feedback_non_ understanding	0	0	2	2	0	2	0	10	0	0	2	0	0	0	1	4	23
give_example	12	4	5	12	16	7	12	13	5	3	4	8	7	4	3	1	116
hesitation	1	0	5	1	9	0	0	1	3	2	1	10	1	13	4	0	51
introduce_topic	0	0	6	0	0	1	0	0	1	1	1	1	0	1	0	0	12
not_classified	11	18	7	6	9	17	7	29	0	2	4	20	1	1	0	8	140
opening	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1
oppose	23	6	13	0	4	13	6	44	9	0	4	11	0	4	1	14	152
oppose_part	2	1	5	1	1	0	3	7	1	0	1	1	0	3	0	1	27
question	10	3	25	6	12	16	8	39	0	7	12	11	1	0	5	78	233
question_ request_answer	4	5	11	9	33	6	17	13	1	4	6	5	0	1	2	10	127

reject_topic	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
repeat	7	2	0	0	0	1	3	4	0	0	1	1	0	0	0	0	0	19
rephrase	2	0	2	2	0	2	2	2	1	1	0	0	0	0	0	0	0	14
request	0	0	1	0	0	0	0	1	2	0	6	3	0	0	0	0	14	27
rhetorical_question_answered	8	0	0	1	0	8	0	1	2	2	0	3	0	3	0	3	0	31
rhetorical_question	2	0	0	0	0	0	0	0	0	0	0	0	2	1	0	0	0	5
suggest_topic	1	1	0	1	1	0	1	1	0	2	1	0	0	1	0	0	0	10
Total	216	204	221	261	347	317	269	441	186	80	247	332	311	217	108	419	4,176	

3.5 Preliminary Analysis

The macro-structure of the topic-oriented dialogues is observed in Figure 1. Annotation tags used for the main interaction make up more than ninety percent of the overall data across all the speakers. Sentential fragments for which the human annotators could not identify the dialogue acts make up approximately three percent of all the dialogue acts. Because the subjects were familiar with each other, in general, they did not need opening or closing dialogue acts. Also, they did not need a lot of time to negotiate a topic or introduce a topic. Thus, our proposal to divide dialogues into five parts as described in Section 3.1 may be revised for conversations between familiar speakers. In the corpus, most of the speakers went directly into the main discussion on the chosen topic.

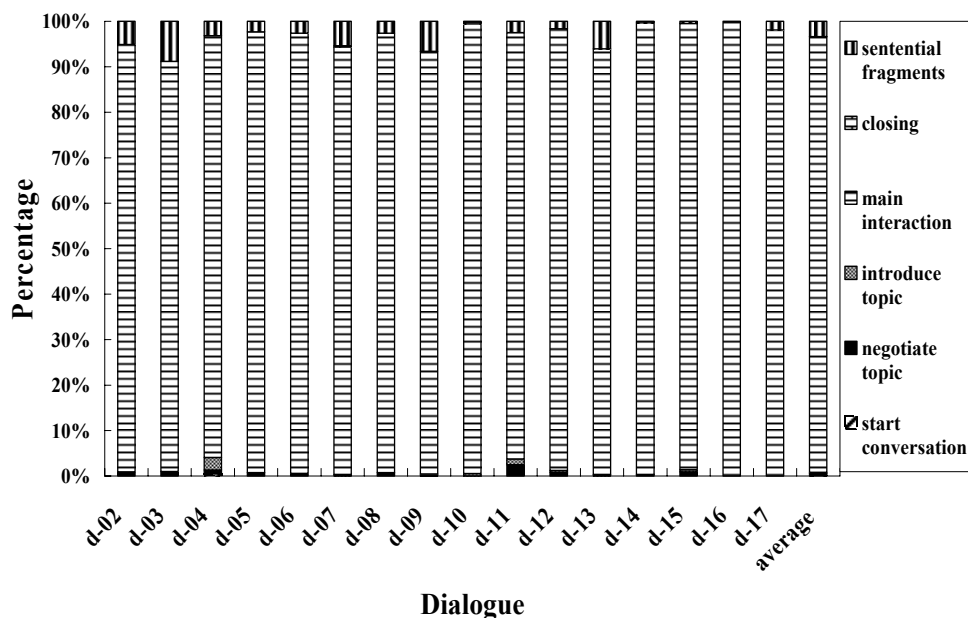


Figure 1. Dialogue structure

Furthermore, we investigated the main interaction in the conversations which made up more than ninety percent of the total annotated dialogue acts. Figure 2 shows that dialogue acts related to feedback, sub-topic management, and requests were more frequently annotated than the other dialogue act types. Detailed results in percentages can be found in Appendix B. Exclamatory expressions were seldom used, perhaps because the situation at the time of recording was formal. Although the subjects knew each other well, their behaviour was relatively conservative. Some of the subjects supplemented or completed utterances produced by themselves or their conversation partners relatively often. Some of the conversations in the corpus are long, but very few dialogue acts are used. And some speakers show differing preference for certain dialogue acts, for instance questions requesting further information. As part of the next analysis step, we are currently analyzing the cross-effect between speaking rate (fast vs. slow talkers), gender (female vs. male speakers), subject relationship (friends and siblings vs. parents-children), and annotated dialogue acts. From the linguistic point of view, the above preliminary results can be investigated in a more sophisticated way, for instance, to determine what Mandarin sentences in spoken use may look like and what their discourse functions may be [Chao 1968; Li and Thompson 1981].

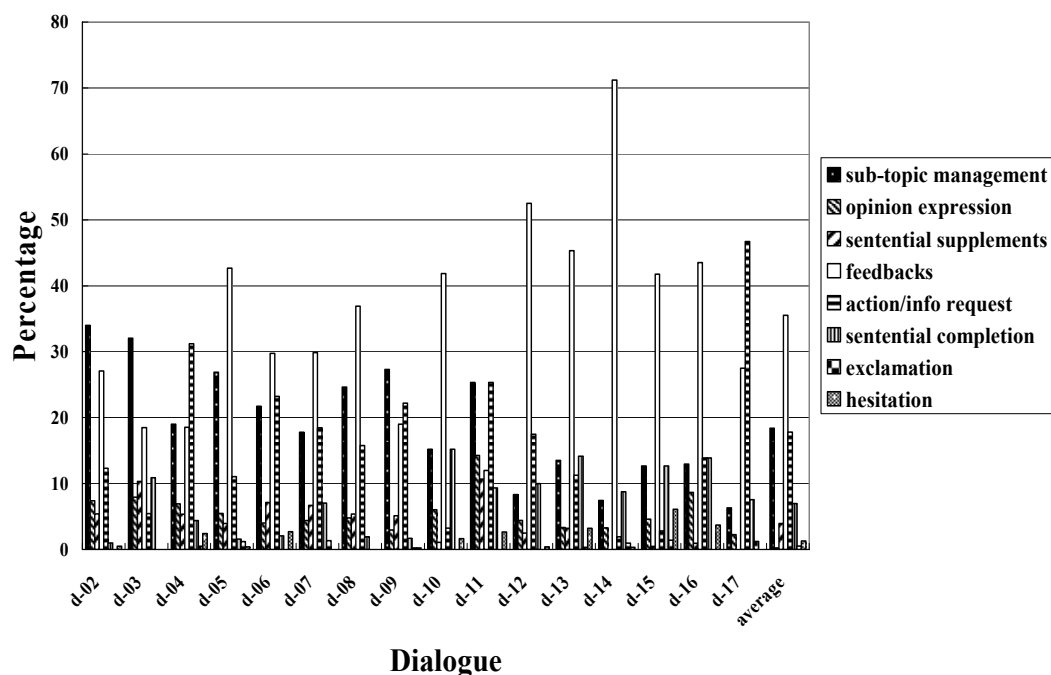


Figure 2. Major interaction of topic-oriented conversations

4. Web Query

For written corpus data, on-line query systems for key word search have been available for a long time. However, in spite of the high demand for empirical spoken data, no such application tool is provided on the Web yet. As our corpus has a database format, the MTCC transcripts can be transformed into a clear, syllable-based database of transcribed and annotated spoken Mandarin data. With such a database, we have already developed a Web search tool for querying keywords. The results are presented in the form of a concordance with information about the subjects. Sound files can be downloaded to check the transcribed content. Furthermore, we also enable the user to search only for sequences annotated with any given tag or to search for keywords and given annotation tags together.

4.1 Query Tool for Spoken Mandarin

The web query system provides four variable settings: search type, corpus, subject, and search content. The tool is shown in Figure 3. The search type can be keyword search only, annotation tag search only, or a combination search for keywords and annotation tags. The user can choose one or more corpora from among our spoken corpora. The gender and age of the subjects can be selected by the user. A keyword search is entered in the form of characters. It also includes pauses or paralinguistic sounds (they should be given in parentheses, as stated in Section 3.3). If a search involves annotation tags, a complete list of annotation tags is automatically made available to the user. For instance, for the case shown in Figure 4, we want to search for the keyword “不” occurring in the annotated tag *agree_part* produced in the MTCC by all male subjects aged from 20 to 40.

語料庫檢索

| 標記系統說明 | 使用說明 |

設定檢索功能	
<input type="radio"/> 關鍵字 <input type="radio"/> 標記 <input checked="" type="radio"/> 標記&關鍵字	
設定語料庫	
<input type="radio"/> MCDC(現代漢語連續口語對話語音語料庫) <input type="radio"/> MMTCC(現代漢語地圖導引語音語料庫) <input checked="" type="radio"/> MTCC(現代漢語主題對話語音語料庫)	
設定發音人性別及年齡	
性別	<input type="radio"/> 全選 <input checked="" type="radio"/> 男 <input type="radio"/> 女
年齡	<input type="radio"/> 全選 <input checked="" type="radio"/> 20 歲 - <input type="radio"/> 40 歲
設定語料檢索項目	
關鍵字	<input type="text"/> <input type="checkbox"/> AND <input type="text"/> <input type="button" value="開始檢索"/>
標記	<input type="text" value="標記"/> <input type="button" value="開始檢索"/>
標記&關鍵字	標記 <input type="text" value="agree_part"/> 關鍵字 <input type="text" value="不"/> <input type="button" value="開始檢索"/>
備註：	

Figure 3. Web Query Tool

4.2 Query Result Illustration

The results for the above query are shown in Figure 4. Four items are found. The presented results contain information about the dialogue coding, the complete speaker turn containing the annotated content, the subject, the gender and age of the subject, and the audio file, which can be listened on-line or using a video file, a feature which is not yet available. The results can be saved and downloaded for further analysis.

檢索標記 : agree_part

關鍵字 : 不

總共在MTCC尋找到4筆的資料

[重新檢索](#)

【檢索結果】

號碼	檔名	內容	發音人	性別	年齡	聲檔	影像檔
1	d-02	就很難很難去(short_break)抓住哪個的輕重(pause)<b agree_part>好的(inhale)或許它就是它的好的沒有錯(inhale)(short_break)或者是他可能不具新聞點</b agree_part>(pause)新聞(pause)EI(exhale)	MISC-72	male	36	DA-2002-02_R0722489.WAV	
2	d-02	會做專題(swallow)(inhale)會這樣做專題NA我(short_break)<b agree_part>你看我也會承認A現在新聞素質(short_break)並不是那麼高因為需求量大太大了</b agree_part>(inhale)所以從業人員們可能沒有受到	MISC-72	male	36	DA-2002-02_R0886848.WAV	
3	d-02	(inhale)他們有可能有他們的革新方法(short_break)<b agree_part>他們也是可能也是跟你一樣看不下去了(inhale)(short_break)所以有想說(short_break)自己跳下去做(short_break)做理想</b agree_part>	MISC-72	male	36	DA-2002-02_R1335231.WAV	
4	d-04	<b agree_part>EN(pause)恐怖那就算不要有病毒就好了</b agree_part>	MISC-75	male	26	DA-2002-04_L0445260.WAV	

page 1 / 1 pages 1

儲存結果 : [save result](#)

Figure 4. Query result

5. Conclusion

This paper has presented preliminary results of an annotated Mandarin conversational corpus and an analysis of annotated dialogue acts. It is well known that spontaneous speech is difficult to deal with, no matter what aspects are considered and that the basic task in research on spontaneous speech is the construction of well-defined data. We have collected a situation-specific spoken corpus and annotated it in dialogue acts. The size of data can definitely be extended, and the annotation scheme improved. The aim of this paper was to illustrate the importance of such a pilot corpus. For instance, we have shown in the analysis presented in this paper that topic-specific dialogues have similar dialogue structures. In addition to dialogue acts, more research topics can be studied with the available spoken corpora. From the linguistic point of view, pronunciation variations, sentence types, and discourse functions are interesting issues. From the speech engineering point of view,

interesting subjects of research on spontaneous speech are pronunciation modelling, parsing algorithms and the intentions of dialogue acts. Hopefully, our annotated MTCC corpus will be useful for research on the above-mentioned issues in Mandarin.

Acknowledgements

The author gratefully acknowledges financial support for the National Digital Archives Project provided by the National Science Council of Taiwan and thanks the three anonymous reviewers for *the International Journal of Computational Linguistics and Chinese Language Processing*. The author expresses sincere thanks to all the assistants who contributed to this project: Ya-Fang He, Vincent Liu, Kah-Lai Chen, Zhe-Ming Chen, and Hong-Da Shi.

References

- Alexandersson, J., B. Buschbeck-Wolf, T. Fujinami and M. Kipp, "Dialogue Acts in VERBMOBIL-2," Report no. 226, July 1998, DfKI.
- Anderson, A., M. Bader, E. Bard and E. Boyle, "The HCRC Map Task Corpus," *Language and Speech*, vol. 34, 1991, p. 351-366.
- Barras, C., E. Geoffrois, Z. Wu and M. Liberman, "Transcriber: Development and Use of a Tool for Assisting Speech Corpora Production," *Speech Communication*, vol. 33, 2001, p. 5-22.
- Chao, Y.-R., "A Grammar of Spoken Chinese," University of California Press, 1968.
- Chen, K.-J. and C.-R. Huang, "SINICA CORPUS: Design Methodology for Balanced Corpora," *Proceedings of the Eleventh Pacific Asia Conference on Language, Information and Computation*, 1996, p. 167-176.
- Chinese Academy of Social Sciences, "Modern Mandarin Dictionary," Xiandaihanyu Cidian, Beijing, 2001.
- Duanmu, S., G. H. Wakefield, Y. P. Hsu, G. Cristina and S. P. Qiu, "Taiwanese Putonghua Speech and Transcript Corpus," *Linguistic Data Consortium*, 1998.
- Gibbon, D., R. Moore and R. Winski (Eds.), "Handbook of Standards and Resources for Spoken Language Systems," Berlin, Mouton de Gruyter, 1997.
- Heeman, P. and J. Allen, "The TRAINS 93 Dialogues," 94-2, Technical Report, Department of Computer Science, University of Rochester, 1995.
- Ide, N. and C. Macleod, "The American National Corpus: A Standardized Resource for American English," *Proceedings of Corpus Linguistics 2001*, 2001, Lancaster, p. 274-280.
- Kowtko, J. and P. Price, "Data Collection and Analysis in the Air Planning Domain," *Proceedings of the DARPA Speech and Natural Language Workshop*, 1989, p. 119-125.
- Leech, G., "100 Million Words of English: The British National Corpus", *English Today*, 9(1):9-15, 1994.

- Levinson, S., *“Pragmatics,”* Cambridge University Press, 1993.
- Li, C. and S. Thompson, *“Mandarin Chinese. A Functional Reference Grammar,”* University of California Press, 1981.
- Li, A.-J., F. Zheng, W. Byrne, P. Fung, T. Kamm, Y. Liu, Z. Song, U. Ruhi, V. Venkataramani and X.-X. Chen, “CASS: A Phonetically Transcribed Corpus of Mandarin Spontaneous Speech,” *Proceedings of the International Conference on Spoken Language Processing (ICSLP 2000)*, 2000, vol. I, p. 485-488.
- Ministry of Education, *“Mandarin Dictionary (Revised version),”* Guoyucidian, Taipei, 1995.
- Tseng, S.-C., “Processing Mandarin Spoken Corpora,” *Traitement Automatique des Langues*. Special Issue: Spoken Corpus Processing. 45(2): 89-108. 2004a.
- Tseng, S.-C., “Mandarin Conversational Dialogue Corpus,” *Post-Conference Proceedings for the International Symposium of Spontaneous Speech Processing: Data and Analysis*, National Institute for Japanese Language, 2004b, Tokyo, p. 73-86.
- Zheng, F., “Making Full Use of Chinese Speech Corpora,” *Proceedings of the Oriental-COCOSDA*, 2004, Singapore, p. 9-23.

request	0.00	0.00	0.45	0.00	0.00	0.00	0.00	0.23	1.08	0.00	2.43	0.90	0.00	0.00	0.00	3.34	0.65
rhetorical_ question answered	3.70	0.00	0.00	0.38	0.00	2.52	0.00	0.23	1.08	2.50	0.00	0.90	0.00	1.38	0.00	0.72	0.74
rhetorical_question	0.93	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.64	0.46	0.00	0.00	0.12
suggest_topic	0.46	0.49	0.00	0.38	0.29	0.00	0.35	0.23	0.00	2.50	0.40	0.00	0.00	0.46	0.00	0.00	0.24

Appendix B: Main Interaction Types (in Percentages)

main interaction types	d-02	d-03	d-04	d-05	d-06	d-07	d-08	d-09	d-10	d-11	d-12	d-13	d-14	d-15	d-16	d-17	mean
sub-topic management	34.0	32.1	19.0	26.9	21.7	17.8	24.6	27.3	15.2	25.3	8.3	13.5	7.4	12.7	13.0	6.3	18.4
opinion expression	7.4	7.9	6.9	5.5	4.0	4.4	4.8	2.9	6.0	14.3	4.4	3.3	3.3	4.6	8.6	2.2	0.2
sentential supplementation	5.4	10.3	5.4	4.0	7.1	6.7	5.4	5.1	1.1	10.7	2.5	3.2	0.0	0.5	0.9	0.0	4.0
feedback	27.1	18.5	18.5	42.7	29.8	29.9	36.9	19.0	41.8	12.0	52.5	45.3	71.2	41.8	43.5	27.5	35.5
action/info request	12.3	5.4	31.2	11.1	23.2	18.5	15.8	22.2	3.3	25.3	17.5	11.3	1.9	2.8	13.9	46.7	17.8
sentential completion	1.0	10.9	4.4	1.6	2.1	7.0	1.9	1.7	15.2	9.3	10.0	14.1	8.7	12.7	13.9	7.5	7.0
exclamation	0.0	0.0	0.5	1.2	0.0	1.3	0.0	0.2	0.0	0.0	0.0	0.3	1.0	1.4	0.0	1.2	0.5
hesitation	0.5	0.0	2.4	0.4	2.7	0.0	0.0	0.2	1.6	2.7	0.4	3.2	0.3	6.1	3.7	0.0	1.3

MATBN: A Mandarin Chinese Broadcast News Corpus

Hsin-Min Wang*, Berlin Chen[†], Jen-Wei Kuo[†] and Shih-Sian Cheng*

Abstract

The MATBN Mandarin Chinese broadcast news corpus contains a total of 198 hours of broadcast news from the Public Television Service Foundation (Taiwan) with corresponding transcripts. The primary purpose of this collection is to provide training and testing data for continuous speech recognition evaluation in the broadcast news domain. In this paper, we briefly introduce the speech corpus and report on some preliminary statistical analysis and speech recognition evaluation results.

Keywords: broadcast news, corpus, speech recognition, Mandarin Chinese, transcription, annotation

1. Introduction

Starting in 1995, the Defense Advanced Research Projects Agency of the United States (DARPA) directed its research program for continuous speech recognition to focus on automatic transcription of broadcast news [Stern 1997]. Since then, many research groups worldwide have paid attention to this challenging task and put much effort into collecting broadcast news corpora of various languages [Matsuoka *et al.* 1997, Federico *et al.* 2000, Graff 2002]. Though some Mandarin Chinese broadcast news corpora are available from LDC (Linguistic Data Consortium, USA)¹, they all exhibit the Mainland China accent, and the wording is quite different to that used in the Taiwan area. To support researchers and technology developers who are interested in studying Mandarin Chinese spoken in the Taiwan area, we have collected Mandarin Chinese news broadcast in Taiwan.

Due to the success of a previous project which collected Mandarin speech data across Taiwan (MAT) [Wang 1997] and was completed by a group of researchers from several universities and research institutes in Taiwan, the same group of people decided to collaborate again on collecting spontaneous speech data in 2001. The first MAT project spanned the

* Institute of Information Science, Academia Sinica, Taipei, Taiwan
E-Mail: whm@iis.sinica.edu.tw

[†] Graduate Institute of Computer Science and Information Engineering,
National Taiwan Normal University, Taipei, Taiwan

¹ Linguistic Data Consortium: <http://www ldc.upenn.edu>

period August 1995 through July 1998. Speech files were collected through telephone networks. The content included read speech (numbers, Mandarin syllables, words of 2 to 4 syllables, and phonetically balanced sentences) and a small amount of spontaneous speech (short answering statements). The second MAT project spanned the period August 2001 through July 2004. We expected to collect both dialogue speech and broadcast news speech. In the broadcast news part, we expected to transcribe 220 hours of broadcast news in 3 years. The first 40-hour corpus was scheduled to be completed by the end of the first year (July 2002), the other 80 hours were due to be completed in July 2003, while the remaining 100 hours were planned to be ready for testing in July 2004. However, we were able to process only 198 hours of broadcast news because our transcribers spent much time correcting errors reported by our colleagues who participated in this project during the second and third years. The 198-hour corpus spanned the period November 17, 2001 through April 3, 2003.

The transcripts are in Big5-encoded form, with SGML tagging to annotate acoustic conditions, background conditions, story boundaries, speaker turn boundaries, and audible acoustic events, such as hesitations, repetitions, vocal non-speech events, external noises, etc. These tags include time stamps that are used to align the text with the speech data. In the 198-hour broadcast news corpus, based on hand-segmentation results, there are 4,100 stories, 581 headlines, 197 weather forecasts, and 197 ending sections. Around 150 hours of speech from 10 weather forecasts and all the stories, headlines, and ending sections were carefully transcribed, while the remaining weather forecasts and segments containing advertising or pure music were just annotated with time stamps without orthographic transcripts. The transcripts contain around 2.3 million Chinese characters in total.

We have established a webpage for the corpus. On this webpage, there are tools that users can employ to query the corpus. Though the project is finished, we will continue to correct errors reported by users. Also, we have selected five one-hour shows as a development set and five more one-hour shows as an evaluation set, and conducted speech recognition experiments on them. The rest of this paper is organized as follows: The data collection procedures and the details of transcription and annotation are presented in sections 2 and 3, respectively. Then, a preliminary assessment of the 198-hour Mandarin Chinese broadcast news corpus is given in section 4. The corpus webpage and corpus tools are introduced in section 5. Speech recognition evaluation results are discussed in section 6. Finally, conclusions are drawn in section 7.

2. Data Collection

The Public Television Service Foundation (Taiwan)² kindly agreed to share their broadcast

² The Public Television Service Foundation (Taiwan): <http://www.pts.org.tw>

news with us. The recordings spanned the period November 7, 2001 through June 30, 2003. A Digital Audio Tape (DAT) recorder, which was connected to a broadcasting machine using an XLR balanced cable, was set up in the TV broadcasting studio. That is, the broadcast news speech was recorded at the same time it was broadcasted to avoid any modulation effect. Recordings are made in stereo with a 44.1kHz sampling rate and 16 bit resolution. Each recording consists of a broadcast news episode 60 minutes long.

Each DAT was manually processed to convert the digital speech samples into a single Microsoft Windows wave file and stored in a hard disk. Then, the signal was down-sampled to 16kHz with a resolution of 16 bits. During this operation, only the left channel was selected. Thus, broadcast news speech in mono, down-sampled to 16kHz with 16 bit resolution was used for further transcription and annotation. More than 250 one-hour broadcast news shows were recorded in this way. However, we were able to transcribe only 198 of them.

Since video can provide visual clues to facilitate transcription and annotation, video recordings were also made simultaneously with the audio recordings. The recordings were made on VHS video tapes. Because we did not have enough space to store several hundred video tapes, each recording was first converted into an MPEG1 file and then stored on a CD-ROM. After conversion was completed, the video tape was reused again. With the video recording, the broadcast news speech corpus can be expanded into a video broadcast news corpus, though at this stage, we are only focusing on the audio track.

3. Transcription and Annotation

The corpus has been segmented, labeled, and transcribed manually using a tool developed by DGA (Délégation Générale pour l'Armement, France) and LDC, called “Transcriber” [Barras *et al.* 2001]. Two full-time transcribers were engaged in this project. They were educated native Mandarin speakers. They worked next to each other on separate machines, so they could easily share their experiences and discuss the problems they encountered on the job. Every sound file was transcribed by one transcriber. When the transcription of a sound file was completed, an additional verification step was performed by the other transcriber to eliminate errors and to ensure consistency of the data. In addition, the first author of this paper and the two transcribers held a regular meeting every week, at which further checking of the transcription and annotation work was performed, and some specific problems encountered by the transcribers were discussed and solved.

Sometimes it was hard to correctly identify the speakers or background conditions by only listening to the audio recording. In each such case, the transcribers would look for clues in the corresponding video file. In addition to the original conventions used in the DGA&LDC Transcriber, we also included the other two sets of annotation tags. The first of these was designed by Dr. Shu-chuan Tseng for annotating Mandarin conversational dialogue speech

[Tseng 2004], and the second one was provided by Dr. Chiu-yu Tseng, which was originally designed for annotating spontaneous monologue speech. All the annotation tags frequently used in this corpus are listed in the Appendix.

The studio anchor speech always exhibits a high standard of fluency, good pronunciation, and good acoustic quality. Most of the field reporter speech also exhibits a high standard of fluency and good pronunciation, but sometimes the acoustic quality is low. Some of the interviewee speech is of very low quality and intelligibility with background speech and noises of various types, and the speech itself sometimes contains mispronunciations, particles, repetitions, repairs, etc. As a result, much more time was required to transcribe and annotate the interviewee speech. The segments containing dialects or foreign languages were annotated with the language identity and time stamps without orthographic transcripts.

3.1 SGML structure of transcriptions

Owing to the complexity and hierarchical nature of the additional information needed in the transcripts, SGML was chosen by the DGA&LDC Transcriber as a suitable framework for formatting the text. The document structure used in all the transcripts is as follows [Barras et al. 2001].

For each waveform file (a full 60-minute program here), there is one accompanying transcript file, containing a single "Episode" element; the Episode has attributes to identify the file name, transcriber, and release version.

Each Episode contains a series of "Section" elements, which correspond to the topical units (stories, advertising, etc.) in the Episode; the Section attributes identify the type of unit, and the points in time at which the Section begins and ends in the corresponding waveform file.

Within each Section containing material to be transcribed, there are one or more "Segment" elements, corresponding to speaker turns within the Section; the Segment attributes identify the speaker, the speaking mode, the channel fidelity, and the points in time at which the speaker turn begins and ends.

At any point within an Episode, a Section or a Segment where there is a change in the presence of music, background voices, or other noises, a "Background" element is inserted to mark the change; the Background attributes identify the type of background (music, speech, other, and shh) and the point in time at which the change occurs.

3.2 Automatic generation of initial transcriptions

After we examined the anchor scripts on the website of the TV broadcaster, we found that they matched the content of the studio anchor speech rather well. This observation led us to design

an automatic tool to generate initial transcription files for our transcribers to start with. The flowchart of the tool for generating initial transcriptions is depicted in Figure 1. It primarily consists of a story segmentation module, a speech recognition module, an anchor script alignment module, and an output module.

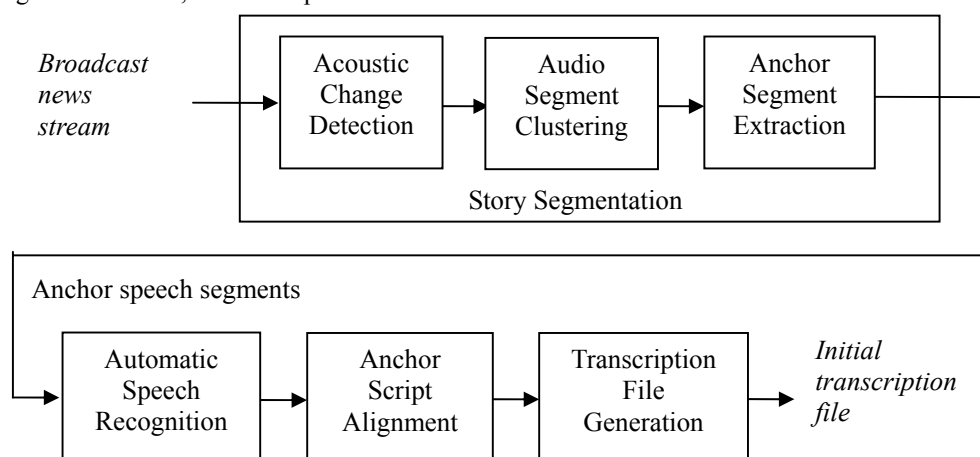


Figure 1. The flowchart of our approach to automatically generating initial transcription files

The story segmentation module first performs speaker and environment change detection, followed by hierarchical clustering of audio segments. We assume that the largest cluster is the studio anchor cluster, and that every studio anchor speech segment is the first segment of a story. Thus, the number of studio anchor speech segments corresponds to the number of stories in the audio stream, and the starting time of a story is the starting time of its studio anchor speech segment. The speech recognition module transcribes all the studio anchor speech segments, and the anchor script alignment module aligns the anchor scripts with the recognition output. Here, we use a vector-space-model-based information retrieval approach for alignment. For each studio anchor speech segment, the recognition output is converted into feature vectors, where the weight of an indexing term (which can be a syllable, a character, a word, or an overlapping N-gram combination) is represented as $tf \times idf$, the term frequency multiplied by the inverse document frequency. Each anchor script is also represented by feature vectors in a similar way. The Cosine measure is used to estimate the relevance between the recognition output and the anchor script. The anchor script with the largest relevance value is aligned with the studio anchor speech segment. Details about the story segmentation approach and the information retrieval approach to story alignment can be found in [Wang *et al.* 2004]. Finally, the output module generates an initial transcription file

consisting of time stamps, topical descriptions³, and anchor scripts of the stories that correspond to the audio stream.

There is much room for improvement in the present tool; e.g., the IDs of the studio anchor and field reporters could be identified by applying speaker identification techniques. Nevertheless, with this initial transcription, the efficiency was improved to some extent.

4. Corpus Assessment

4.1 A brief description

Each one-hour news show usually has two or three parts separated by advertisements. Sometimes, however, there is no advertising at all within a show. Each part starts with headlines with background music, followed by a number of stories. Because the news shows were collected from a non-profit public TV station, the advertising was composed of public service announcements and previews rather than commercials. Generally, a one-hour news show contains one to three headline sections, zero to two advertising sections, depending on the number of headline sections, a number of news stories, a weather forecast section, and an ending section.

Figure 2 depicts a partial transcription of a broadcast news show. The transcription has three hierarchically embedded segmentation layers (orthographic transcription, speaker turns, and sections (stories)), plus a fourth segmentation layer (acoustic background conditions), which is independent of the other three. Frequently, the non-speech part between speech segments produced by two different speakers is chopped into several distinct short segments according to their acoustic foreground and background conditions. Moreover, a speaker turn may be separated into several segments by short silence segments.

4.2 Preliminary statistical analysis

Some preliminary assessments of the 198-hour Mandarin Chinese broadcast news corpus have been conducted. There are seven distinct studio anchors; three are male, and four are female. The distribution of studio anchors is summarized in Table 1. It is obvious that the distribution is quite unbalanced; one female anchored 83.84% of the news shows, while two males each anchored only one show. According to the hand-segmentation results, there are 4,100 news stories in total. The total length is around 143 hours, and the average length per story is 2.1 minutes. Some other brief statistical information about these 4,100 stories is summarized as follows:

³ The topical descriptions were copied from the website of the TV broadcaster because every anchor script was associated with a manually-generated topical description there.

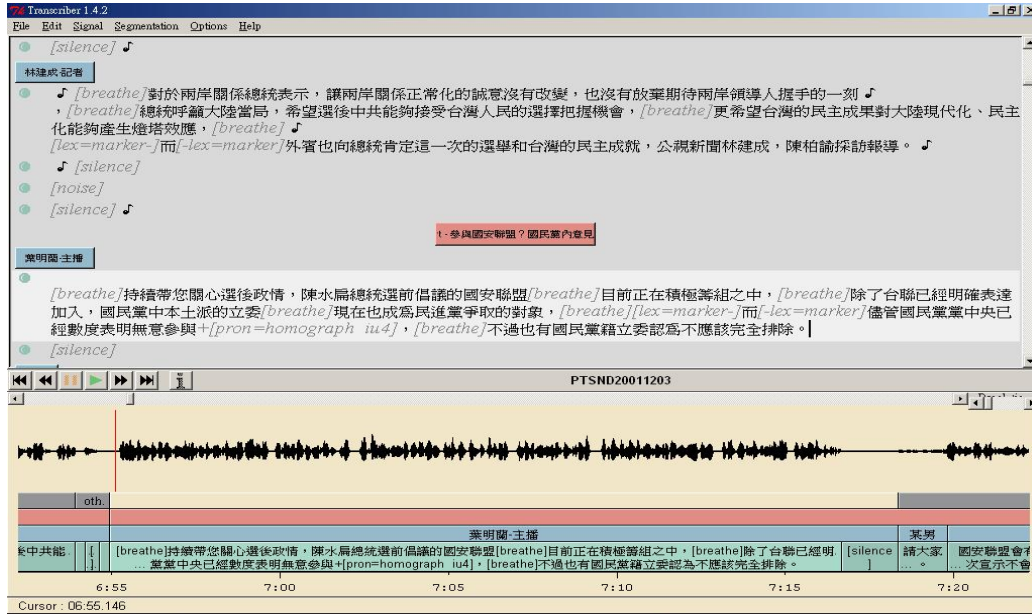


Figure 2. A partial transcription of a broadcast news show

Table 1. The distribution of studio anchors in the 198-hour broadcast news corpus

Gender	Identity	Number of shows	Percentage
Male	Male 1	15	7.58%
	Male 2	1	0.51%
	Male 3	1	0.51%
Female	Female 1	166	83.84%
	Female 2	7	3.54%
	Female 3	5	2.53%
	Female 4	3	1.52%
Sum		198	100%

- (1) There are 386 distinct field reporters. Of these, 77 field reporters are male. The identities of 22 male and 70 female field reporters can be identified. The true identities of 294 field reporters are undetermined even though our transcribers have referred to the video recording for clues. It is very likely that some of the unidentified field reporters in different stories in fact correspond to the same reporter, while some correspond to the 92

identified reporters. Therefore, the exact number of distinct field reporters may be much lower than 386 but slightly higher than 100.

- (2) There are around 5,900 distinct interviewees. Of these, around 4,000 interviewees are male. It is interesting to find that, unlike the situation with the field reporters, the percentage for male speakers is relatively high. Moreover, even though the identities of some interviewees are also unknown, it is very likely that the unknown interviewees in different stories are in fact different people.
- (3) The total lengths of the studio anchor speech, field reporter speech, and interviewee speech are around 27 hours, 71 hours, and 45 hours, respectively. The numbers of characters transcribed from them are around 493,000, 1,167,000, and 616,000, respectively. The speaking rates are 4.9, 5.1, and 4.5 syllables per second, respectively. Some segments contain pure music or noise. In addition, some speech segments contain foreign languages, dialects, or aboriginal languages. If these segments are omitted, then the total lengths of the studio anchor speech, field reporter speech, and interviewee speech are around 25 hours, 58 hours, and 35 hours, respectively. Some speech segments contain overlapping speech.
- (4) The frequency counts of the most frequently used tags in the corpus are shown in Table 2. The overall top 5 most frequently used tags are “breathe,” “pause,” “mispronunciation,” “particle,” and “discourse marker,” respectively. The “breathe” and “pause” tags are common to different types of speech, but very high percentages of the “mispronunciation,” “particle,” and “discourse marker” tags are found in the interviewee speech. This is because the studio anchor speech and field reporter speech mostly consist of planned speech, while the interviewee speech mostly consists of spontaneous speech. It is interesting that the 9th most frequently used tag in the field reporter speech is “Formosan,” though the count number is not very high. This is because some field reporters were aborigines and they pronounced their own names in aboriginal languages. We also found that the interviewees spoke in dialects more often than the studio anchors and field reporters did. Moreover, in different types of speech, some speech segments contain English terms.

In addition to the 4,100 news stories, there are 581 headline sections (~5.5 hours), 652 advertising sections (~23.5 hours), 197 weather forecast sections (~10.3 hours), and 197 ending sections (~0.8 hours). All the headline sections and ending sections were carefully transcribed. The weather forecasts in 10 shows were also carefully transcribed, but the remaining 187 weather forecasts and all the advertising sections were simply annotated with time stamps without orthographic transcripts. The transcripts contain around 2.3 million Chinese characters in total.

Table 2. The frequency counts of the most frequently used tags in the corpus. Min-Nan is a common dialect and Formosan denotes all the aboriginal languages used in the Taiwan area

Studio anchor speech		Field reporter speech		Interviewee speech		Overall	
Tags	Count	Tags	Count	Tags	Count	Tags	Count
Breathe	20780	Breathe	47125	pause	33987	breathe	89352
pause	5881	Pause	26070	breathe	21447	pause	65938
mispronunciation	780	mispronunciation	6845	particle	21153	mispronunciation	26134
Particle	327	English	992	mispronunciation	18509	particle	22428
English	306	Particle	948	discourse marker	5258	discourse marker	6090
discourse marker	232	discourse marker	600	restart	3695	restart	3911
Restart	160	Min-Nan	263	syllable contraction	1407	English	2626
repair	58	syllable contraction	219	English	1328	syllable contraction	1667
repetition	51	Formosan	79	repetition	1197	Min-Nan	1424
Syllable contraction	41	restart	56	Min-Nan	1131	repetition	1263

5. Corpus Webpage and Tool

Though the project is finished, we will continue to correct errors reported by users and post the most up-to-date versions of transcriptions on the corpus webpage, <http://sovideo.iis.sinica.edu.tw/SLG/corpus/MATBN-corpus.htm>. Currently, on this webpage, two tools are available for querying the corpus. Figure 3 shows a snapshot of the tool for querying sections. The tool shows statistical information, time stamps, and orthographic transcripts corresponding to sections such as news stories, headlines, endings, weather forecasts, etc. Figure 4 shows a snapshot of the tool for querying speakers. With this tool, users can easily find statistical information, time stamps, and orthographic transcripts corresponding to the speakers they specify.

6. Speech Recognition Evaluation

6.1 The development and evaluation sets

Ten one-hour shows were selected from the 198-hour carefully transcribed broadcast news database to evaluate the speech recognition performance. That is, we spot-checked about 5% ($10/198 * 100\% = 5.05\%$) of the database. We divided the shows into a development set and an evaluation set. The development set consisted of five shows recorded on 2003/01/24, 2003/01/27, 2003/02/07, 2003/03/05, and 2003/03/06, while the evaluation set consisted of five shows recorded on 2003/01/28, 2003/01/29, 2003/02/11, 2003/03/07, and 2003/04/03. The basic guidelines for making selections were as follows: First, we wanted to include as many studio anchors as possible. Second, the test shows had to be broadcast after January 1st, 2003 so that we could use the newswire text before January 1st, 2003 to train the language models. In this section, we will report the speech recognition results obtained for the development set and evaluation set. These results are meant to serve as a reference for future

studies that use this corpus.

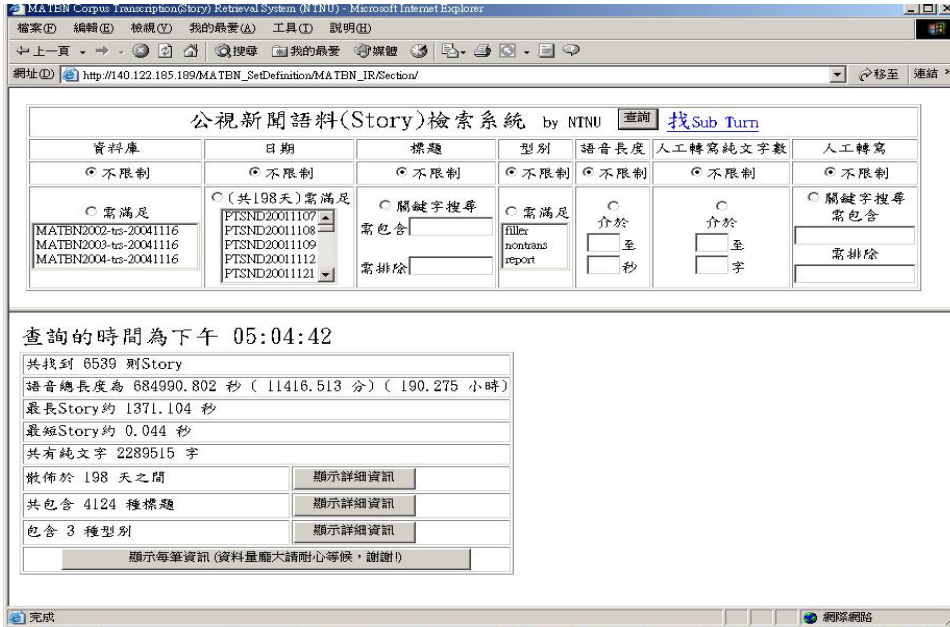


Figure 3. The tool for querying sections

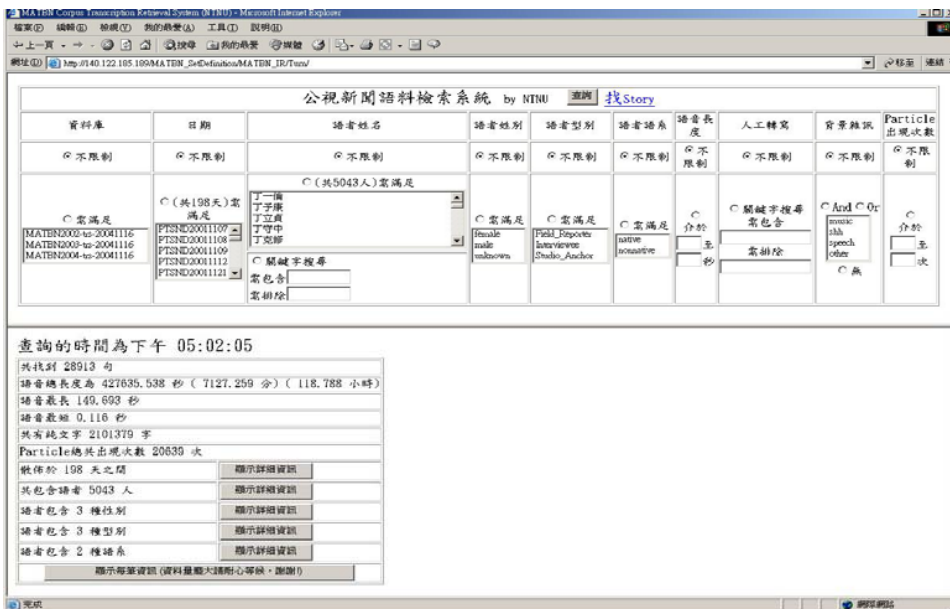


Figure 4. The tool for querying speakers

6.2 The NTNU broadcast news transcription system

The NTNU broadcast news transcription system [Chen *et al.* 2004] was employed here for speech recognition evaluation. Some features of the speech recognizer are reviewed below.

6.2.1 Front-end signal processing

In the speech recognizer, spectral analysis is applied to a 20 ms frame of speech waveform every 10 ms. For each speech frame, 12 mel-frequency cepstral coefficients (MFCCs) and the logarithmic energy are extracted, and these coefficients and their first and second time derivatives are spliced together to form a 39-dimensional feature vector. In addition, to compensate for channel noise, utterance-based cepstral mean subtraction (CMS) is applied to the training and testing speech.

6.2.2 Acoustic modeling

The acoustic models were trained with a database of 16 hours of broadcast news speech collected from several radio stations located in Taipei. The broadcast news data were recorded using a wizard FM radio connected to a PC and digitized at a sampling rate of 16 kHz with 16 bit resolution. The data collection period was from December 1998 to July 1999. The training database is a combination of two corpora: The first corpus contains two hours of field reporter and interviewee speech, and four hours of studio anchor speech. The manual transcripts have been time-aligned to the phrasal level. The second corpus contains ten hours of studio anchor speech. Each audio file is a short news abstract (lasting 50 seconds on average) produced by a studio anchor. Unlike the first corpus, only the orthographic transcripts were available for each audio file; detailed time alignment was unavailable.

Due to the monosyllabic structure of the Chinese language, where each syllable can be decomposed into an INITIAL and a FINAL, the acoustic units used in the speech recognizer are intra-syllable right-context-dependent INITIAL/FINALs, including 112 context-dependent INITIALs and 38 context-independent FINALs. Each INITIAL or FINAL is represented by a Continuous Density Hidden Markov Model (CDHMM) with two to four states. The Gaussian mixture number per state ranges from 1 to 64, depending on the amount of corresponding training data available. In addition, the silence model is a 1-state CDHMM with 128 Gaussian mixtures trained with the non-speech segments. A total of 12,419 mixtures were obtained.

6.2.3 Lexicon and language modeling

The lexicon contains 71,694 words, including 66,290 words selected manually from the CKIP lexicon and 5,404 new words or compound words extracted automatically from the language model training corpus. The word-based unigram, bigram, and trigram language models were trained using the newswire text corpus, consisting of 170 million Chinese characters collected

from Central News Agency (CNA) in 2001 and 2002 (the Chinese Gigaword Corpus released by LDC). The language models were further processed with Katz backoff smoothing [Katz 1987] using the SRI Language Modeling Toolkit (SRILM)⁴.

6.2.4 The speech recognizer

The speech recognizer was implemented with a left-to-right frame-synchronous tree search as well as a lexical prefix tree organization of the lexicon. At each speech frame, the so-called word-conditioned method was used to group path hypotheses that shared the same history of predecessor words into the same copies of the lexical tree, and to expand and recombine them according to the tree structure until a possible word ending was reached. At word boundaries, the path hypotheses among the tree copies that had the same search history were recombined and then propagated to the existing tree copies or used to start new ones if none yet existed. At each speech frame, a beam pruning technique, which considered the decoding scores of path hypotheses together with their unigram language model look-ahead and syllable-level acoustic look-ahead scores [Chen *et al.* 2004], was used to select the most promising path hypotheses. Moreover, if the word hypotheses ending at each speech frame had scores that were higher than a predefined threshold, then their associated decoding information, such as the word start and end frames, the identities of current and predecessor words, and the acoustic score, were kept in order to build a word graph for further language model rescoring. Once the word graph had been built, the Viterbi beam search with higher language model weighting [Wessel *et al.* 2001] was performed on it to generate the most likely word sequence. In this study, the word-based trigram language model was used in both the tree search procedure and the word graph rescoring procedure.

6.3 Speech recognition experiments

We conducted speech recognition experiments on the development set and the evaluation set both separately and jointly. Table 3 provides some statistical information about the development set and evaluation set, and the speech recognition results obtained from them. The data corresponding to the development set, the evaluation set, and the union of the two sets are, respectively, depicted in Table 3 (a), (b), and (c). To facilitate the calculation of recognition rates, utterances containing foreign languages, dialects, or aboriginal languages were discarded prior to analysis. It is obvious from Table 3 that both the statistical information and speech recognition results for the development set and evaluation set are very similar to each other. The recognition accuracy for the interviewee speech is extremely poor, but the accuracy for the studio anchor speech and the field reporter speech is quite reasonable. This is

⁴ The SRI Language Modeling Toolkit: <http://www.speech.sri.com/projects/srilm/>

obviously because most of the anchor speech and field reporter speech exhibits a high level of fluency, good pronunciation, and good acoustic quality, while most of the interviewee speech is of very low quality and intelligibility with background sounds of various types, and the speech itself sometimes contain mispronunciations, particles, repetitions, repairs, etc.

Table 3. Speech recognition evaluation results

(a) The 5-hour development set

	Number of speakers	Number of utterances	Speech length (sec.)	Syllable accuracy	Character accuracy	Word accuracy
Studio anchor	3	168	2438.709	79.18%	74.05%	65.74%
Field reporter	36	345	5608.170	65.43%	58.04%	48.74%
Interviewee	142	191	2463.509	26.57%	19.71%	15.55%
Overall	181	704	10510.388	60.57%	53.82%	45.34%

(b) The 5-hour evaluation set

	Number of speakers	Number of utterances	Speech length (sec.)	Syllable accuracy	Character accuracy	Word accuracy
Studio anchor	3	160	2347.630	78.98%	73.72%	66.42%
Field reporter	32	317	5215.351	67.42%	59.78%	50.27%
Interviewee	151	197	2300.656	26.75%	19.53%	14.66%
Overall	186	674	9863.637	61.46%	54.58%	46.06%

(c) Overall estimation (the union of the development and evaluation sets)

	Number of speakers	Number of utterances	Speech length (sec.)	Syllable accuracy	Character accuracy	Word accuracy
Studio anchor	3	328	4786.339	79.08%	73.89%	66.07%
Field reporter	45	662	10823.522	66.31%	58.88%	49.48%
Interviewee	279	388	4764.165	26.66%	19.62%	15.11%
Overall	327	1378	20374.026	61.00%	54.19%	45.69%

6.4 Discussion

Notice that, in a benchmark test, the model parameters optimally tuned for the development set are applied to the evaluation set directly. Further tuning conducted on the evaluation set is not allowed. In this study, we aimed to select development and evaluation data from the corpus so that users can experiment on it with the same setting. Therefore, we have only reported the baseline recognition accuracy of the development set and evaluation set. Furthermore, in this corpus, it is obvious that most of the anchor reporters and field reporters are female, while most of the interviewees are male. Therefore, it is difficult to design a benchmark test with a balanced number of female and male speakers using different types of

speech. However, this may simply reflect the real situation in Taiwan. In that case, we do not need to worry about the gender issue very much.

7. Concluding Remarks

Speech resources are crucially important for research and development in speech technology. But the development of a speech corpus used to be very tedious and time consuming. In August 2001, we started a 3-year project aimed at collecting a Mandarin Chinese broadcast news corpus in Taiwan. At the end of the project, we had labeled 198 one-hour news shows using the DGA&LDC Transcriber. In this paper, we have discussed the development and evaluation of the corpus. The speech corpus will soon be available through the Association for Computational Linguistics and Chinese Language Processing (ACLCLP). In the future, we will try to collect data from other TV broadcasters. We will also try to collect data from various programs.

Acknowledgements

This project was funded by the National Science Council of the Republic of China under grants NSC 90-2213-E-009-109, NSC-91-2219-E-009-039, and NSC-92-2213-E-009-021. The authors would like to thank the Public Television Service Foundation (Taiwan) for sharing their broadcast news with us and their employees for helping us to set up the recording machines in their broadcasting studio and operating them regularly. Acknowledgements go to Dr. Chiu-yu Tseng and Dr. Shu-chuan Tseng for their valuable assistance and comments on the transcription and annotation, Prof. Sadaoki Furui and his colleagues for sharing their experience with us, Ms. Kuan-jung Chen, Ms. Mei-li Chang, and Ms. Tzau-fang Yan for their hard work on transcription and annotation, Mr. Guo-hsien Wang and Mr. Yi-hsiang Chao for cloning speech data on the DAT to PC, Mr. Tzan-hwei Chen for developing the tool for automatically generating initial transcriptions, and all the colleagues from universities and research institutes that participated in this project. Acknowledgements also go to three anonymous reviewers for their helpful comments.

References

- Barras, C., E. Geoffrois, Z. B. Wu and M. Liberman, "Transcriber: Development and Use of a Tool for Assisting Speech Corpora Production," *Speech Communication*, 33, 2001, pp. 5-22.
- Chen, B., J. W. Kuo and W. H. Tsai, "Lightly Supervised and Data-driven Approaches to Mandarin Broadcast News Transcription," *Proceedings of the 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2004.

- Federico, M., D. Giordani and P. Coletti, "Development and Evaluation of an Italian Broadcast News Corpus," *Proceedings of the 2nd International Conference on Language Resources and Evaluation*, 2000.
- Graff, D., "An Overview of Broadcast News Corpora," *Speech Communication*, 37,2002, pp. 15-26.
- Katz, S. M., "Estimation of Probabilities from Sparse Data for Other Language Component of a Speech Recognizer," *IEEE Transactions on Acoustics, Speech and Signal Processing*, 35(3), 1987, pp. 400-401.
- Matsuoka, T., Y. Taguchi, K. Ohtsuki, S. Furui, and K. Shirai, "Toward Automatic Transcription of Japanese Broadcast News," *Proceedings of the 5th European Conference on Speech Communication and Technology*, 1997.
- Stern, R. M., "Specification of the 1996 Hub 4 Broadcast News Evaluation," *Proceedings of the 1997 DARPA Speech Recognition Workshop*, 1997.
- Tseng, S.-C., "Processing Spoken Mandarin Corpora," *Traitement automatique des langues, Special Issue: Spoken Corpus Processing*, 45(2), 2004, pp. 89-108.
- Wang, H. C., "MAT - A Project to Collect Mandarin Speech Data through Telephone Networks in Taiwan," *Computational Linguistics and Chinese Language Processing*, 2(1), 1997, pp. 73-89.
- Wang, H. M., S. S. Cheng and Y. C. Chen, "The SoVideo Mandarin Chinese Broadcast News Retrieval System," *International Journal of Speech Technology*, 7(2-3), 2004, pp. 189-202.
- Wessel, F., R. Schluter, K. Macherey and H. Ney, "Confidence Measures for Large Vocabulary Continuous Speech Recognition," *IEEE Transactions on Speech and Audio Processing*, 9(3), 2001, pp. 288-298.

Appendix

In the DGA&LDC Transcriber, the annotation tags are divided into 4 categories, namely, noise, pronounce, language, and lexical. We list them in Table A1. An event can be instantaneous (e.g., 我從來沒想過 <Event desc="laugh" type="noise" extent="instantaneous"/> 會來參加這個勞工遊行，), sustained (e.g., 這個倒 <Event desc="laugh" type="noise" extent="begin"/> 沒什麼 <Event desc="laugh" type="noise" extent="end"/>。), or associated with the previous character (e.g., 所以根據不同可能發 <Event desc="mispronunciation hua1" type="pronounce" extent="previous"/> 生的狀況，)。

Table A1. The tags used in the corpus

(a) noise

description	example
breathe, clear throat, click, cough, cry, hiccup, laugh, noise, pause, sigh, silence, smack, sneeze, sniffle, swallow, yawn, etc.	我從來沒想過 <Event desc="laugh" type="noise" extent="instantaneous"/> 會來參加這個勞工遊行，
cough, cry, laugh, yawn, etc.	這個倒 <Event desc="laugh" type="noise" extent="begin"/> 沒什麼 <Event desc="laugh" type="noise" extent="end"/> 。
particle	所以各位 <Event desc="particle" type="noise" extent="begin"/> NE <Event desc="particle" type="noise" extent="end"/> 在我們各自的工作崗位上，
unrecognizable non-speech sound	我揣摩 <Event desc="unrecognizable non-speech sound" type="noise" extent="begin"/> ... <Event desc="unrecognizable non-speech sound" type="noise" extent="end"/> 笛子的

(b) pronounce

description	example
alternative	也就是待 <Event desc="alternative dai1" type="pronounce" extent="previous"/> 在家裡
mispronunciation	所以根據不同可能發 <Event desc="mispronunciation hua1" type="pronounce" extent="previous"/> 生的狀況，
stutter, syllable contraction, uncertain, unrecognizable speech sound, etc.	尤其是派系比較嚴重的 <Event desc="syllable contraction" type="pronounce" extent="begin"/> 這些 <Event desc="syllable contraction" type="pronounce" extent="end"/> 地方，
zhuyin	就連 <Event desc="zhuyin" type="pronounce" extent="begin"/> ㄅㄆㄇ <Event desc="zhuyin" type="pronounce" extent="end"/> 他也是看不懂。

(c) lexical

description	example
abridged, cut, editing term, error, interrupted, discourse marker, new word, repair, repetition, restart, etc.	因為肌瘤切除術我們曉得 <Event desc="repetition" type="lexical" extent="begin"/> 它的它的 <Event desc="repetition" type="lexical" extent="end"/> 一些缺點

(d) language

description	example
English, Formosan, Hakka, Japanese, Min-Nan, Unknown, etc.	經濟部長林信義今天已經率領 <Event desc="English" type="language" extent="begin"/> WTO <Event desc="English" type="language" extent="end"/> 代表團出發前往卡達。

TAICAR - The Collection and Annotation of an In-Car Speech Database Created in Taiwan

**Hsien-Chang Wang*, Chung-Hsien Yang+, Jhing-Fa Wang+,
Chung-Hsien Wu** and Jen-Tzung Chien****

Abstract

This paper describes a project that aims to create a Mandarin speech database for the automobile setting (TAICAR). A group of researchers from several universities and research institutes in Taiwan have participated in the project. The goal is to generate a corpus for the development and testing of various speech-processing techniques. There are six recording sites in this project. Various words, sentences, and spontaneously queries uttered in the vehicular navigation setting have been collected in this project. A preliminary corpus of utterances from 192 speakers was created from utterances generated in different vehicles. The database contains more than 163,000 files, occupying 16.8 gigabytes of disk space.

Keywords: TAICAR, in-car speech, speech database, multi-channel recording, corpus collection and annotation

1. Introduction

1.1 In-car speech corpora review

Driver information systems are becoming increasingly complex as more and more functions are integrated into modern cars. Speech-enabled functions will enhance the safety and convenience of operating for future vehicles. To realize such functions, in-car speech processing techniques need to be built and tested first. Thus, it is necessary to collect an in-car speech database. Although many speech corpora [Tapisa *et al.* 1994], [Roach *et al.* 1996], [Kudo *et al.* 1994], [Bernstein *et al.* 1994] have been created to improve speech-processing effectiveness, few in-car speech databases have been reported.

* Department of Information Management, Chang Jung Christian University,
396 Chang Jung Road, Sec.1, Kway Jen, Tainan, Taiwan, R.O.C.
Tel: 886-6-2785123 ext. 2071; Fax: 886-6-2785657
E-Mail: wangbb@mail.cju.edu.tw

+ Department of Electrical Engineering, National Cheng Kung University

** Department of Computer Science and Information Engineering, National Cheng Kung University

Researches on speech processing in the vehicular environment, including works on speech recognition, noise reduction and speaker adaptation, have been published at numerous conferences, for example, the *International Workshop on Hand-Free Speech Communication*, which was held in 2001 in Kyoto, Japan; the biannual *European Conference on Speech Communication and Technology (EuroSpeech)*; and the *International Conference on Spoken Language Processing (ICSLP)*. To our knowledge, several research organizations have carried out in-car speech database collection. In Japan, professor Itakura at CIAIR collected multimedia data, such as audio, video, and auxiliary vehicle information, from dialogues spoken in moving cars [Itakura 2001]. The system was built in a Data Collection Vehicle (DCV) supporting the synchronous recording of multi-channel audio and video data through microphones and cameras. In Europe, researchers in countries such as France, Germany, Britain, and Spain joined in a cooperative project, SpeechDat [Heuvel *et al.* 1999] to collect an in-car speech database for multi-lingual speech processing purposes. The resulting SpeechDat-Car database contains speech data recorded from three microphones and one cellular phone. A similar project has also been reported by Langmann and his colleagues. [Langmann *et al.* 1998]. Researchers at the University of Illinois in Champaign-Urbana designed a project whose purpose was to collect multi-channel database consisting of both speech and video data. One hundred speakers participated in the project, and a total of 59,000 utterances were collected [Lee *et al.* 2004]. Table 1 shows a brief comparison of some existing in-car speech corpora and the TAICAR corpus.

Table 1. Survey of several in-car speech corpora

Corpus name (year)	CSDC-MoTiV (1998)	SpeechDat-Car (1999)	CU-Move (2000)	CIAIR-HCC (2001)	CMU (2001)	AVICAR (2004)	TAICAR (2004)
Country	Germany	Europe	USA	Japan	USA	USA	Taiwan
# of People	641	N/A	N/A	ongoing	43	100	192
Microphone	Array	Array	Array	Mesh	Array	Array	Array
Content	Digits; Commands	Multi-lingual	Digits; Commands	Digits; Words; Sentences	Short words	Digits; Letters; Sentences	Digits; Words; FAQs
Need Specific Car	No	No	No	Yes	No	No	No

1.2 Motivation and Setup

A group of researchers in the field of speech processing in Taiwan initiated an in-car speech collection project called TAICAR (Taiwan in-CAR speech database). The goal is to generate an in-car speech database to be applied to various noisy speech processing researches. In order to generate the corpus rapidly and usefully, some considerations with regard to setting up the data collection procedure were deemed important. These considerations are described below.

1.3 Setup of the TaiCar project

The philosophy behind the TaiCar corpus collection procedure is to use convenient and readily available equipment to collect speech and environmental noise in various vehicles. The following are the ten considerations deemed important.

1. The platform for in-car speech collection should be a notebook PC.
2. The resulting speech database should follow the Microsoft file format for audio waveforms.
3. Multiple channels of microphone signals should be recorded.
4. A channel of clean speech signal should be recorded simultaneously for reference purposes.
5. The recording devices (microphones, recording card, etc.) should be readily available.
6. The speakers should reflect Taiwan's demographics in terms of gender, dialect, education, age, and population.
7. The database should cover all the phonetic properties of Mandarin.
8. In addition to the speech data, the corpus should also include environmental noises.
9. The database should reflect two real-world road conditions of the real world. The vehicle should be routed through a downtown area and along a highway during a recording session.
10. The database should contain some spontaneous sentences to facilitate research on mobile dialogue systems.

This paper is organized as follows. Section 2 describes the recording procedure. Section 3 presents the annotating procedure. Preliminary results of the TaiCar project are given in Section 4, and Section 5 gives a conclusion.

2. Recording Procedures

2.1 Data collection system

In this project, six recording sites at universities and research institutions have been set up so far across Taiwan. Each site uses a notebook PC equipped with a PCMCIA multi-channel signal-recording card as the recording platform. A pre-amplification circuit amplifies the input signals, which go to the recording card from the microphones. Six microphones are placed in the vehicle. A microphone array with four omni-microphones is placed on the sun visors. The distance between the microphones is 30 cm. Another microphone is bound above the notebook PC placed on the lap of the speaker. Due to safety considerations, the speaker should be the navigator instead of the driver. The last microphone, a unidirectional anti-noise one, is worn

on the head of the speaker. The reason for using such a good microphone is to provide nearly clean speech for reference purposes. The hardware elements are described in detail below:

1. A DAQP PCMCIA multi-channel signal recording card capable of recording up to 16 channels of signal is plugged into the notebook PC as the recording interface.
2. Four omni-directional microphones form a linear microphone array (channels 0-3).
3. One omni-directional microphone is placed in front of the speaker (channel 4).
4. One unidirectional microphone is worn on the head of the speaker (channel 5).
5. A pre-amplification circuit is utilized before the speech signal is fed to the PCMCIA card.

Figure 1 shows the configuration and the positioning of the microphone array, the navigator, and the pre-amplification circuit.

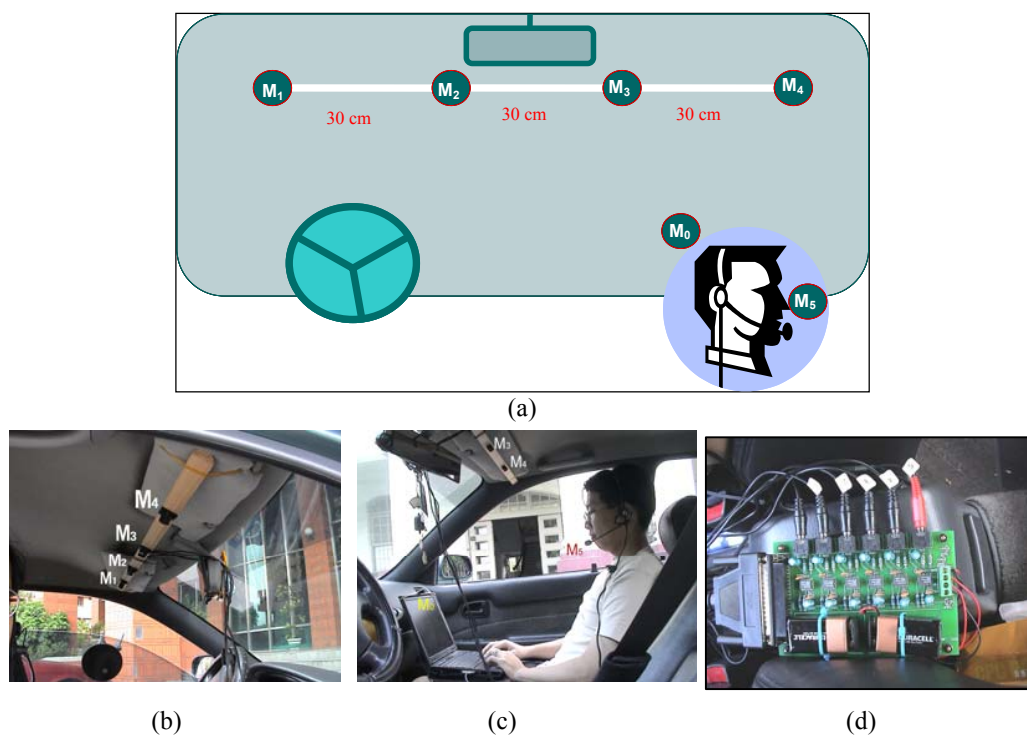


Figure 1. (a) The configuration of TAICAR recording system. The distance between the microphones in the array is 30 cm. (b) The microphone array attached to the sun visor above; (c) the positions of the speaker and recording notebook PC; (d) the amplification circuit board for multi-channel recording.

During the recording process, the notebook PC is placed on the lap of the navigator. The material to be uttered is shown on the screen in prompts so that the speaker can follow. A sample screenshot captured during the recording procedure is shown in Figure 2.

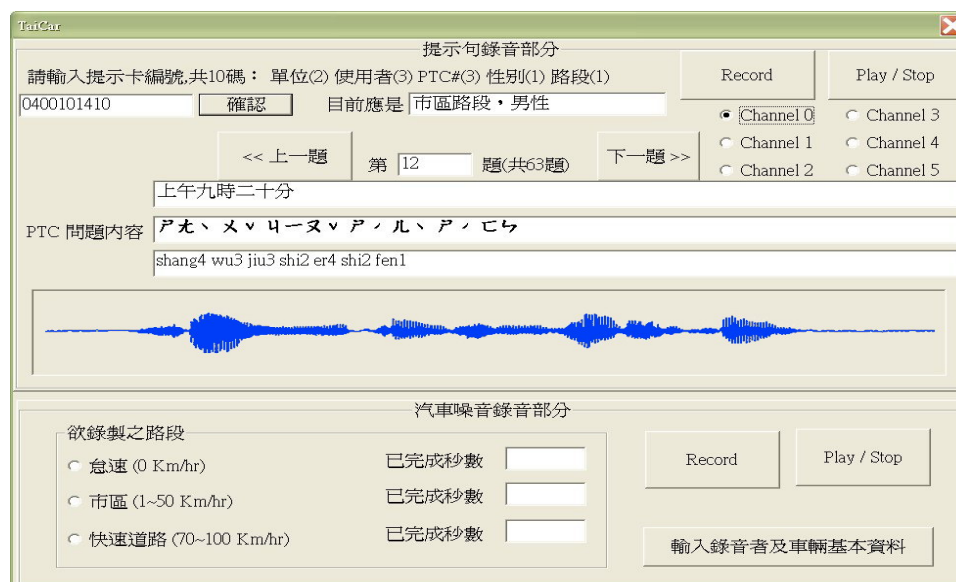


Figure 2. A screenshot from the TAICAR database recording procedure

2.2 Speech Files Format

For each utterance, six speech files are recorded. The files, saved in the MS-Windows file format for audio waveforms, are composed of two parts: a file header and sampled data. The file header contains the following information about the speech: 1) the number of channels, which indicates whether the speech was recorded in mono or stereo; 2) the number of samples recorded per second; 3) the number of bits per sample; and 4) the size of the speech data. The sampled data of speech signals are in the binary format. The files retain the waveforms of the recorded utterances as well as the preceding and following silence.

Unlike several existing speech databases, for example, MAT [Wang 1997], the transcribed Chinese characters are stored in separate files using Big-5 code. This makes it convenient to preview these files using common text processing programs under most operating systems.

2.3 Corpus Design

The TAICAR database material contains two parts. The first part is used to collect the reading speech of the speakers. It is generated by following the philosophy of the creation of

MAT-2400 database material [Wang 1997]. The framework for this material was created by Dr. Tseng of Academia Sinica [Tseng 1995]. The materials were extracted from two text corpora consisting of 77,324 lexical entries and 5,353 sentences. The material contains 407 base-syllables in Mandarin Chinese without tones; 1,062 words with two to four syllables; and 200 numbers in five different contexts, including digital sequences, dates, time, prices, and car license plate numbers.

The second part consists of spontaneous FAQ's (Frequently Asked Questions) collected from the general public in Taiwan. This material was generated by asking them several questions. The scenario questions were given to ordinary citizens, and their answers were transcribed and used as the material for the spontaneous FAQ's. The scenario questions include a description of seven query domains containing questions which are usually asked while driving a car. The seven query domains and some collected FAQ's are listed in Table 2.

Table 2. Some example of FAQ's

Domain	Scenario	Collected FAQ
Food	◆ You are hungry and looking for a restaurant.	– Where is the nearest fast-food restaurant? – Guide me to the nearest restaurant.
Clothes	◆ You want to buy some clothes.	– Where is the nearest Hang Ten? – I want to go to Far Eastern Department Store.
Lodging	◆ You are looking for a place to stay.	– I would like to know the location of the Hilton Hotel. – Show me the nearby hotels.
Navigation	◆ You want to know how to get to a destination.	– How do I go to CKS airport? – Where is City Hall?
Entertainment	◆ You want to have fun.	– How do I get to the nearest theater?
Others	◆ Weather conditions.	– What's the temperature in Taipei?
	◆ Other information one wants to know while driving.	– Is there any museum nearby? – Turn the CD player on.

The collected FAQ's were randomly chosen to be included in the TAICAR prompt sheets. Each prompt sheet contains 10 FAQ's that the speaker utters spontaneously.

2.4 Prompt Sheet

The prompt sheets are designed to serve as guides for the speaker to follow while uttering speech. The prompt sheet contains two parts of the aforementioned materials--the spontaneous speech and FAQ's. A total of 72 items are listed on a prompt sheet. The items are:

- ◆ 5 numbers spoken in different ways (No's. 10-14);
- ◆ 12 isolated Mandarin syllables (No's. 15-26);
- ◆ 45 isolated words (No's. 27-56, 67-82);
- ◆ 10 FAQ sentences (No's. 57-66).

The prompt sheet is designed to contain as many syllable and phonetic combinations as possible. The FAQ's are also included on the prompt sheet since they are useful for research of vehicular dialogue systems. An example of a prompt sheet is shown in Appendix A.

3. The Annotating Process

For a speech corpus to be useful, various phenomena of speaker behaviour and the deficiencies of the speech files should be annotated correctly. Since the annotation of a speech database is a labour consuming task, the tagging procedure for the TAICAR database was designed to be as convenient as possible. In the annotation phase, the annotators check whether the speech files are intelligible and whether the auto-transcribed syllables match the speaker's utterances, and they mark the starting and ending points of the speech. Figure 3 shows a screenshot of the annotation process.



Figure 3. A screenshot of TAICAR database tagging

- (1) Prompt sheet number.
- (2) & (3) Text and phonetic transcript of the current speech.
- (4) & (5) Back/Proceed to other speech.
- (6) Mini-keyboard for modifying the phonetic syllables.
- (7) Click this area to select one of the six channels. The speech waveform shown will change correspondingly.
- (8) Play/Stop-playing this speech.
- (9) & (10) Left/Right Click on the waveform to mark the starting/ending point of the speech.
- (11) Update the database when tagging is finished.

If the starting or ending point of an utterance does not match the syllables, the annotator should mark another boundary of the utterance and correct both the text content and phonetic syllables in the database.

4. Preliminary Data Collection Result

The TAICAR project was carried out between 2002 and 2003. According to the initial plan, researchers at each recording site would record the speech of 40 speakers and annotate the utterances. However, for technical and financial reasons, researchers at some sites did not complete these tasks. In all, 192 speakers at the six recording sites participated in this project. The result was an in-car speech database consisting of utterances recorded in both downtown and highway environments. Since it is hard to accurately read long sentences on a screen while driving, utterances consisting of FAQ sentences were collected at only one site. Some statistics for the resulting database are shown in Table 3. Note that the number of files or contents is for 192 speakers driving along two different routes with six recording-channels.

Table 3. Some brief statistics for the preliminary result of TAICAR database collection

Items	Description
Speakers	Total: 192
	Male: 115 (59.8%)
	Female: 77 (40.2%)
	Age: from 19 to 58, mostly 20~30 (71.3%)
	Education: most has BS degrees (89.6%)
Daily Language: Taiwanese (64.6%)	
Car	Type: mostly sedans (71.3%)
	Engine capacity: below 2.0L: (57.8%); 2.0~3.0L (37.5%)
Speech data amount	6 DVDs 163,890 files 16.8 gigabytes 145.8 hours
Database content	16,128 digits 4,608 English letters 27,648 Isolated syllables 101,376 Words with two-four characters 960 FAQ's

Figure 4 shows the waveforms of the utterance “EQ7673” from channel 0 to channel 5.

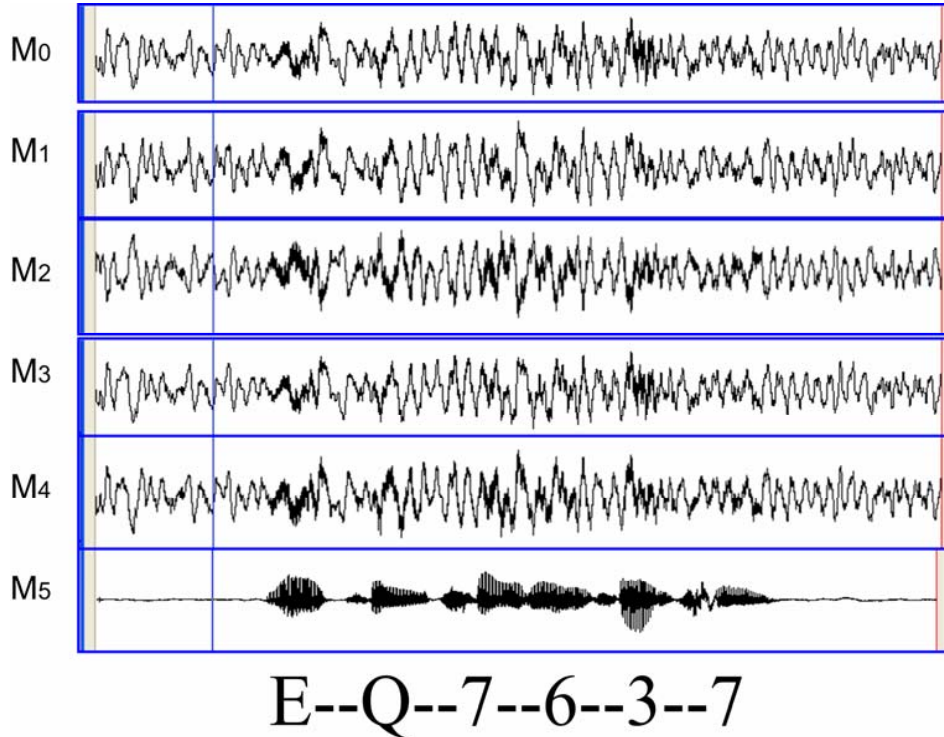


Figure 4. Speech waveforms of the utterance “EQ7673” (in Mandarin), from channel 0 to channel 5.

As mentioned in Section 2, the microphone for channel 5 is unidirectional and anti-noise. It is adopted to record the reference signal for calculating the signal-to-noise ratio (SNR) and the time shift for other channels. The SNR can be computed as follows:

$$SNR = 10 \cdot \log_{10} \frac{E[speech]}{E[noise]}, \quad (\text{in dB})$$

where $E[x]$ stands for the energy of signal x .

To estimate the SNR for M_5 , the speech region is detected first. Then the noise can be estimated from the non-speech part. Based on the estimated noise level, the average SNR in the speech region can be determined. By aligning the signal of M_5 with the signals of $M_0 \sim M_4$, one can locate the speech regions in $M_0 \sim M_4$. Then the noise level and SNRs for $M_0 \sim M_4$ can be computed. The SNRs for different routes measured in the downtown and highway environments are reported in Table 4.

Table 4. SNRs (in dB) for microphones M_0 – M_5 measured in the highway and downtown environments

	Channel 0	Channel 1	Channel 2	Channel 3	Channel 4	Channel 5
Highway	-2.8550	-2.4770	-2.7171	-2.5318	-2.6763	11.4040
Downtown	-2.6187	-2.2637	-2.0714	-2.4655	-2.5261	11.2245

To calculate the time shift between channel k ($0 \leq k \leq 4$) and channel 5, the configuration of all six microphones should be considered first, as shown in Figure 5. The microphone for channel k is M_k . The distance between M_i and M_j is D_{ij} . The distances for $D_{3,5}$ and $D_{0,5}$ are predefined as 40 cm and 60 cm, respectively. The distance between the microphones in the microphone array is 30 cm, i.e. $D_{1,2} = D_{2,3} = D_{3,4} = 30$ cm. Applying the Pythagorean Theorem, we can calculate the distances $D_{1,5}$, $D_{2,5}$, and $D_{4,5}$ obtaining 72, 50, and 50 cm, respectively. Because the speed of sound is 32,000 cm/sec, the time shift between M_i and M_j ($0 \leq i, j \leq 4$), denoted as $T_{i,j}$, can be determined. Since M_5 is placed in front of the mouth of the speaker, it can be regarded as the original source of the utterance. The time shift for each channel can be determined as $T_{0,5} = 0.00187$, $T_{1,5} = 0.00156$, $T_{2,5} = 0.00125$, $T_{3,5} = 0.00156$, and $T_{4,5} = 0.00221$.

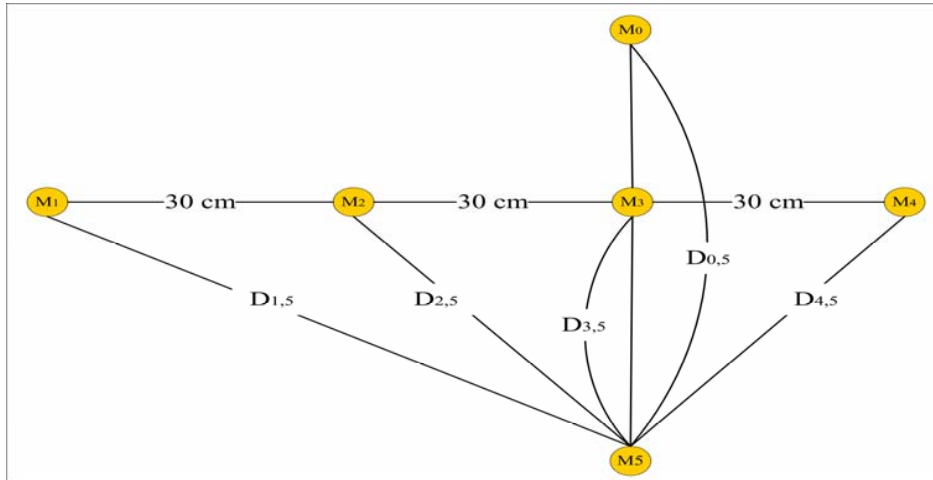


Figure 5. The detailed configuration and distances between the six microphones

5. Conclusion

This paper has described the TAICAR project that aims to create a Mandarin Chinese speech database based on the in-car environment in Taiwan. The preliminary result is a 192-speaker speech database containing 145.8 hours of utterances and environmental noises recorded in various types of automobiles. So far, two works have adopted the TaiCar corpus studies on speech enhancement in car noise environment [Yang et al. 2004], [Wang et al. 2004] and have

shown that the use of this corpus is of fundamental importance for the testing of in-car noise reduction technology. The database can also be used to develop various in-car speech processing techniques, such as speech source separation, active speech detection, channel equalization, and robust noisy speech recognition.

Acknowledgement

The authors would like to express their appreciation to the researchers from National Taiwan University, National Chiao-Tung University, National Tsing-Hua University, National Cheng-Kung University, Industrial Technology Research Institute (ITRI), and Chungwa Telecom Laboratories (CTL). Without their help, the TaiCar project would not have been possible.

References

- Bernstein, J., K. Taussig and J. Godfrey, "MACROPHONE: An American English Telephone Speech Corpus for Polyphone Project," In *Proceedings of the IEEE International Conference on Acoustics Speech and Signal Processing*, 1994, Adelaide, Australia, Vol. I, pp. 81-84.
- Heuvel, H., A. Bonafonte, J. Boudy, S. Dufour, P. Lockwood, A. Moreno and G. Richard, "SpeechDat-Car: Towards a collection of speech databases for automotive environments," In *Proceedings of the Nokia-COST249 Workshop on Robust Methods for Speech Recognition in Adverse Conditions*, 1999, Tampere, Finland, pp. 135-138.
- Heuvel, H., J. Boudy, R. Comeyne, S. Euler, A. Moreno and G. Richard, "The SpeechDat-Car multilingual speech databases for in-car applications: Some first validation results," In *Proceedings of 6th European Conference on Speech Communication and Technology*, 1999, Budapest, Hungary, Vol.5, pp. 2279-2282.
- Itakura, F., "Multi-Media Data Collection for In-Car Speech Communication - Ongoing Data Collection and Preliminary Results," In *Proceedings of International Workshop on Hand-Free Speech communication*, 2001, Kyoto, Japan, pp. 1-5.
- Kudo, I., T. Nakama, N. Arai and N. Fujimura, "The Database Collection of Voice Across Japan (VAJ) Project," In *Proceedings of 2nd International Conference on Spoken Language Processing*, 1994, Yokohama, Japan, pp.1799-1802.
- Langmann, D., H.R. Pfitzinger, T. Schneider, R. Grudszus, A. Fischer, M. Westphal, T. Crull and U. Jekosch, "CSDC, the MoTiV Car Speech Data Collection," In *Proceedings of 1st International Conference on Language, Resources and Evaluation*, 1998, Granada, Spain, pp. 1107-1110.
- Lee, B., H.-J. Mark, C. Goudeseune, S. Kamdar, S. Borys, M. Liu and T. Huang, "AVICAR: Audio-Visual Speech Corpus in a Car Environment," In *Proceedings of 8th International Conference on Spoken Language Processing*, 2004, Jeju Island, Korea.

- Roach, P., S. Arnfield, W. Barry, J. Baltova, M. Boldea, A. Fourcin, W. Gonet, R. Gubrynowicz, E. Hallum, L. Lamel, K. Marasek, A. Marchal, E. Meister and K. Vicsi, "BABEL: An Eastern European Multi-language Database," In *Proceedings of 4th International Conference on Spoken Language Processing*, 1996, Philadelphia, USA, pp. 1892-1893.
- Tapias, D., A. Acero, J. Esteve and J.C. Torrecilia, "The VESTEL Telephone Speech Database," In *Proceedings of 3rd International Conference on Spoken Language Processing*, 1994, Yokohama, Japan, pp.1811-1814.
- Tseng, C.Y. "A Phonetically Oriented Speech Database for Mandarin Chinese," In *Proceedings of the 13th International Congress on Phonetic Sciences*, 1995, Stockholm, Sweden, Vol. 3, pp.326-329.
- Wang, H.C., "MAT - A Project to Collect Mandarin Speech Data Through Telephone Networks in Taiwan," *Computational Linguistics and Chinese Language Processing*, 2(1), 1997, pp. 73-90.
- Wang, J.-F., C.-H. Yang and K.-H. Chang, "Using Perceptual Wavelet Decomposition and Subspace Tracking for Noise Removal in Car Environment," In *Proceedings of ROCLING XVI: Conference on Computational Linguistics and Speech Processing*, 2004, Taipei, Taiwan.
- Yang, C.-H., J.-F. Wang and K.-H. Chang, "Subspace Tracking for Speech Enhancement in Car Noise Environments," In *Proceedings of International Conference on Acoustic, Speech, and Signal Processing*, 2004, Quebec, Canada.

Appendix A. A Sample of the Prompting Sheet

- | | |
|---------------|-----------------------|
| (10) 七三八 零四零八 | (47) 改革派 |
| (11) 十二月十七日 | (48) 閩南語 |
| (12) 上午九時二十分 | (49) 消防車 |
| (13) 四千五百二十二元 | (50) 療養院 |
| (14) EQ 七六三七 | (51) 風風雨雨 |
| (15) 該 | (52) 未雨綢繆 |
| (16) 案 | (53) 布魯塞爾 |
| (17) 鎮 | (54) 訓導主任 |
| (18) 榮 | (55) 血本無歸 |
| (19) 轉 | (56) 莫名其妙 |
| (20) 卡 | (57) 我要找吃飯的地方 |
| (21) 推 | (58) 哪裡有餐廳 |
| (22) 桌 | (59) 最近的加油站在哪裡 |
| (23) 怒 | (60) 附近有沒有加油站 |
| (24) 跄 | (61) 加油站還有多遠 |
| (25) 說 | (62) 最近的餐廳在哪裡 |
| (26) 的 | (63) 附近有沒有麥當勞速食店 |
| (27) 予以 | (64) 我想找肯得基，哪裡有？ |
| (28) 財務 | (65) 車子快沒油了，最近的加油站在哪裡 |
| (29) 喜愛 | (66) 台南車站要怎麼走 |
| (30) 案由 | (67) 七五九五 |
| (31) 給予 | (68) 四五七一八七九 |
| (32) 爲要 | (69) 泰國 |
| (33) 台銀 | (70) 人事行政局 |
| (34) 掃蕩 | (71) 聯華電子公司 |
| (35) 手腕 | (72) 彰化商業銀行 |
| (36) 搬運 | (73) 無法 |
| (37) 合約 | (74) 代表 |
| (38) 應用 | (75) 不過 |
| (39) 黨外 | (76) 報導 |
| (40) 加速 | (77) 方式 |
| (41) 花園 | (78) 調查 |
| (42) 去年 | (79) 工業區 |
| (43) 佛像 | (80) 台中市 |
| (44) 尊重 | (81) 戡亂時期 |
| (45) 狀況 | (82) 金融機構 |
| (46) 內閣制 | |

Design and Development of a Bilingual Reading Comprehension Corpus

Kui Xu* and Helen Meng*

Abstract

This paper describes our initial attempt to design and develop a bilingual reading comprehension corpus (BRCC). RC is a task that conventionally evaluates the reading ability of an individual. An RC system can automatically analyze a passage of natural language text and generate an answer for each question based on information in the passage. The RC task can be used to drive advancements of natural language processing (NLP) technologies imparted in automatic RC systems. Furthermore, an RC system presents a novel paradigm of information search, when compared to the predominant paradigm of text retrieval in search engines on the Web. Previous works on automatic RC typically involved English-only language learning materials (Remedia and CBC4Kids) designed for children/students, which included stories, human-authored questions, and answer keys. These corpora are important for supporting empirical evaluation of RC performance. In the present work, we attempted to utilize RC as a driver for NLP techniques in both English and Chinese. We sought parallel English, and Chinese learning materials and incorporated annotations deemed relevant to the RC task. We measured the comparative levels of difficulty among the three corpora by means of the baseline bag-of-words (BOW) approach. Our results show that the BOW approach achieves better RC performance in BRCC (67%) when compared to Remedia (29%) and CBC4Kids (63%). This reveals that BRCC has the highest degree of word overlap between questions and passages among the three corpora, which artificially simplifies the RC task. This result suggests that additional effort should be devoted to authoring questions with a various grades of difficulty in order for BRCC to better support RC research across the English and Chinese languages.

Keywords: bilingual, reading comprehension, corpus.

* Human-Computer Communications Laboratory, Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong, Hong Kong SAR, China
E-Mail: {kxu, hmmeng}@se.cuhk.edu.hk

1. Introduction

RC is a task that conventionally evaluates the reading ability of an individual, especially during language learning. Typically, the subject is presented with a passage of natural language text and asked to read and comprehend the passage. He or she is then presented with a series of questions about the passage and asked to answer each question based on information understood from the passage.

Recently, research efforts have been devoted to the development of *automatic* RC systems [Anand *et al.* 2000; Charniak *et al.* 2000; Hirschman *et al.* 1999; Ng *et al.* 2000; Riloff and Thelen 2000]. An RC system can automatically analyze a passage of natural language text. When the system is then presented with a series of (human-generated) questions, it is expected to automatically generate an answer for each question, based on information it extracted or retrieved from the passage. While the conventional RC task can evaluate the reading capability of a human, the task can also be used to drive advancements in natural language processing (NLP) technologies incorporated into automatic RC systems. Furthermore, an RC system presents a novel paradigm of information search, when compared to the predominant paradigm of text retrieval in search engines on the Web. Several comparisons are made below:

1. An RC system has only a *limited amount* of direct context upon which to draw in order to answer questions, whereas a Web-based search engine has *huge amounts* of direct context available on the Internet.
2. An RC system aims to generate *specific, precise answers* to user-posed questions based on the given passage, thus eliminating the need for the user to read the entire passage, whereas a Web-based search engine presents a list of text documents that closely match the user's query so that the user to *further browse for potential answers*.
3. A cross-language RC system is analogous to a cross-language text retrieval system, in that a user-posed question/query may be expressed in a different language from that used for passages or archived documents.

The first two points suggest that in-depth syntactic and semantic analyses are needed to facilitate the automatic RC task. Hence, the RC task has been a driver of NLP development.

Previous works on automatic RC, as cited above, typically involved *language learning materials* designed for children/students, which included stories, human-authored questions, and answer keys. These materials included the Remedia corpus [Hirschman *et al.* 1999] and the CBC4Kids corpus [Anand *et al.* 2000; Dalmás *et al.* 2003]. These corpora are important because they support empirical evaluation of RC performance. An analytical review of the passages, questions, and answer keys in the Remedia and CBC4Kids training sets reveals that a suite of natural language processing and information extraction technologies are

indispensable for achieving comprehension. Hence, an RC corpus needs to be annotated to support the development of such technologies, as explained below:

1. If questions and answers have equivalent meanings in term of the same frame structures, such as the same predicates, logical subjects, and logical objects, then RC systems should be able to identify the same frame structures.
2. Inference may be based on a common ontology (e.g., the synonymy/antinomy, is-a, part-of, causality, and entailment relations in WordNet), human feelings (e.g., what is strange, happy, sad, etc.) or semantically equivalent descriptions (e.g., “*the man is a farmer*” means “*the man makes a living by farming*”).
3. Inference may be based on the context knowledge in a passage. For example, the two sentences “*A merry-go-round has wooden animals on it*” and “*The weather damages the animals*” imply that “*The weather damages the merry-go-round.*” An RC system should be able to find inferences by using context knowledge to identify equivalent meanings.
4. An RC system should be able to perform summarization to answer such questions as “*What can we draw from this story?*”
5. An RC system should be able to perform calculations in order to answer such questions as “*How many boroughs are there in New York City?*”
6. An RC system should be able to resolve anaphora in documents in order to identify equivalent meanings.
7. When questions ask for specific persons, times, locations, or numbers, an RC system should be able to identify different named entity types in order to identify equivalent meanings.
8. For definition questions, the answers follow some patterns. An RC system should be able to perform pattern matching to identify equivalent meanings. An example question-answer pair is “*Who is Christopher Robin?*” and “*He is the same person that you read about in the book, Winnie the Pooh.*”

In addition, previous corpora involved *English-only language learning materials*. This paper describes our initial attempt to design and develop a bilingual corpus for reading comprehension (RC). In the current work, we attempted to utilize RC as a driver for NLP techniques in both English and Chinese. As an initial step, we sought parallel English and Chinese learning materials for language learning and incorporated annotations deemed relevant to the RC task. We refer to this corpus as the bilingual reading comprehension corpus (BRCC).

The rest of this paper is organized as follows. Section 2 reviews related works.

Section 3 presents considerations for designing a bilingual corpus. Section 4 describes the development of the BRCC. Section 5 discusses BOW matching results on the BRCC, and section 6 draws conclusions.

2. Related Work

Remedia is the first corpus developed for the evaluation of automated RC systems [Hirschman *et al.* 1999]. This corpus consists of remedial reading materials for grades 3 to 6 and was annotated by the MITRE Corporation [Hirschman *et al.* 1999]. An example passage with questions and answer keys is shown in Table 1. An answer key is the answer to a given question as provided by the publisher. It may not be an extract from the passage itself. In each story, as exemplified in Table 1, the first line is the title; the second line is the dateline; the others are story sentences. This corpus was used as a test-bed in 2000 in an ANLP-NAACL workshop on “Reading Comprehension Tests as Evaluation for Computer-Based Language Understanding Systems” [Charniak *et al.* 2000; Hirschman *et al.* 1999; Ng *et al.* 2000; Riloff and Thelen 2000].

Table 1. Sample story and questions in the Remedia corpus

Passage	Pony Express Makes Final Run (JOPLIN, MISSOURI, October 26, 1861) From now on, mail will be sent a new, faster way. It is called the telegraph. It uses wires to send messages. Now there will be no need for the Pony Express. Since April, 1860, mail has been sent this way. The Last Pony Express rider leaves town today....
Questions	Who left Joplin on October 26, 1861? What did the Pony Express riders do? When did the Pony Express start?
Answer keys	the last Pony Express rider they carried mail April, 1860

To evaluate RC systems on Remedia, three evaluation metrics were proposed in [Hirschman *et al.* 1999], namely, P&R, HumSent, and AutSent. P&R directly uses answer keys to compute precision and recall, while both HumSent and AutSent use answer sentences which are obtained according to answer keys. The difference between HumSent and AutSent is that answer sentences in HumSent are marked by humans, while those in AutSent are generated by an automated routine based on which sentence has the highest recall compared with the answer key [Hirschman *et al.* 1999]. HumSent accuracy is calculated by comparing the system answers with the human-marked answers, scoring one point to an answer from the system that is identical to the human marked answer and zero points otherwise. The average score across all the questions is the HumSent accuracy. Riloff and Thelen [2000] believed that

HumSent was more reliable than P&R and AutSent, since human-marked answer sentences were involved.

In 2000, the CBC4Kids corpus was developed based on the Canadian Broadcasting Corporation web page¹ for kids [Anand *et al.* 2000; Dalmás *et al.* 2003]. Stories in CBC4Kids are all news articles and cover 12 domains: politics, health, education, science, human interest, disaster, sports, business, crime, war, entertainment, and the environment [Anand *et al.* 2000]. For each story, Ferro and Bevins from the MITRE Corporation added between eight and twelve questions and answer keys [Anand *et al.* 2000]. According to the answer keys, the answer sentences were also annotated. In 2000, CBC4Kids was used to develop and evaluate reading comprehension technologies in a summer workshop on reading comprehension held at Johns Hopkins University.²

Details about Remedia and CBC4Kids³ are given in Table 2. The distributions of different types of questions⁴ in the Remedia training set and CBC4Kids training set are shown in Table 3 and Table 4, respectively. In Table 3 and Table 4, we divide *what* questions into four sub-types, since this question type asks for a variety of information, ranging from definitions to reasons, numbers, events, time, locations, person names, etc. *What-DEF* questions ask for definitions; *What-VP* questions ask about actions that the question subjects performed; *What-NP* questions ask for noun phrases which are subjects or objects of question predicates; *What-OTH* questions ask for reasons, numbers, times, locations, person names, organizations, etc.

Table 2. Details of Remedia and CBC4Kids

	Remedia	CBC4Kids
Publisher	Remedia Publications	Canadian Broadcasting Corporation
Training set	55 stories	73 stories
Test set	60 stories	52 stories
Corpus size	20K words	35K words
# questions	575	1232
Annotated information	named entities, anaphora co-references, and answer sentences	part-of-speech tags, base forms of words, named entity tags, anaphora co-references, parse trees, and answer sentences

¹ <http://www.cbc4kids.ca>

² <http://www.clsp.jhu.edu/ws2000/groups/reading/>

³ We obtained CBC4Kids from Lisa Ferro.

⁴ The question type “others” in Tables 3, Tables 4, 5 and 7 refers to *how many, how much, how long, how often, how far, how tall, etc.*

Table 3. Question distributions in the Remedia training set

Question type	# questions	Example
Who	43	Who is Christopher Robin?
What-DEF	13	What is the stock market?
What-VP	8	What did Jackie Cochran do?
What-NP	15	What did Alex eat on the island?
What-OTH	8	What causes these lights? What is the baby's name? What is the name of our national library?
When	43	When was Winnie the Pooh written?
Where	42	Where did young Chris live?
Why	42	Why did Chris write two books of his own?
Other	1	How high did she take her plane?
Overall	215	

Table 4. Question distributions in the CBC4Kids training set

Question type	# questions	Example
Who	100	Who runs the club?
What-DEF	37	What is Bill 101?
What-VP	39	What did Meiorin do to get her job black?
What-NP	69	What does the round goby eat?
What-OTH	29	What causes a solar eclipse? What company runs YNN? What is the capital of Turkey?
Which	10	Which leader is the premier of Ontario?
When	68	When did the metal shop close?
Where	79	Where is Brasilia?
Why	92	Why is the school opening the club?
How	53	How can the satellite help farmers?
Others	75	How big is the club? How many people live in La Ronge? How much was Babe paid to play basketball?
Overall	651	

In the above, we have summarized the common characteristics of two English corpora, Remedia and CBC4Kids, which can be used to guide the design of a bilingual reading comprehension corpus. Passages in these corpora are taken from language learning materials designed for children. They consist of stories that cover a variety of domains, which we refer to as open domains. As shown in Table 2, the two corpora contain on the order of a hundred passages, several tens of thousands of words, and several hundred questions. In addition, these two corpora provide experimental materials to support the development of automatic syntactic and semantic analysis techniques that can contribute to reading comprehension. Hence, they include annotations such as named entities, anaphora co-references, part-of-speech tags, and parse trees in order to support automatic named entity filtering, pronoun resolution, and syntactic and semantic analysis.

To estimate the reading difficulty of the passages in Remedia and CBC4Kids, we use the Dale-Chall readability formula⁵ [Dale and Chall 1948]. It is believed that this formula is more accurate than other existing formulas when it is applied to the passages for grades four and above [Klare 1963]. To measure the readability of a passage, we first compute a raw score according to the following formula:

$$\text{Raw score} = (0.0496 \times \text{average sentence length}) + (0.1579 \times \text{percent of words in passage not found on Dale Word List}^6) + 3.6365$$

We then consult the Grade Equivalent Conversion Chart [Dale and Chall 1948] with the raw score to obtain a grade equivalent reading score. For example, a raw score in the range from 7.0 to 7.9 implies suitability for grades 9–10. According to the Dale-Chall formula, the maximum readability grade of Remedia is 8; the minimum grade of Remedia is below grade 4. The Dale-Chall formula cannot determine the exact minimum readability score when the grade is below 4. CBC4Kids is at readability grades 7–15.

3. Design Considerations for the Bilingual Reading Comprehension Corpus (BRCC)

We referenced the English corpora, Remedia and CBC4Kids, in formulating the design considerations for the BRCC. We searched for language learning materials that have both English and Chinese versions. These materials needed to contain sufficient numbers of passages and corresponding English and Chinese questions as well as answer keys. The passages also needed to cover a range of readability, as measured using methods such as Dale Chall's formula. We summarize the design considerations for passages, questions, and answer keys, respectively, in the following:

- Parallel Chinese-English passages should be provided for cross-lingual evaluation.
- Passages should be open domain.
- Passages with different grades of difficulty should be selected in order to challenge RC systems to achieve higher levels of performance.

Our design considerations for *questions* were:

- Parallel Chinese-English questions should be provided.
- Questions should cover different types (see Table 4) that correspond to key information in the passages, and they should be authored with various levels of difficulty.

⁵ <http://www.interventioncentral.org/htmldocs/tools/okapi/okapi.shtml>

⁶ <http://www.interventioncentral.org/htmldocs/tools/okapi/okapimanual/dalechalllist.shtml>

Our design considerations for *answers* were:

- Parallel Chinese-English answer keys should be provided.
- Sentences that contain the same meanings as answer keys should be provided for the HumSent evaluation metric.

Annotations in the corpora should be based on solid principles. Inter-annotator agreement should also be enforced whenever possible. In order to evaluate different NLP technologies used in RC tasks, we propose to annotate linguistic knowledge about the structure and the meaning of language at different levels [Allen 1995]. In the initial version of the BRCC, we annotated such linguistic knowledge as the boundaries of noun phrases (e.g., “*the English language*” is a noun phrase), several types of named entities (e.g., person names, locations, etc.), and anaphoric references (e.g., “*Edison*” may be the referent of the pronoun “*he*”).

4. Development of the BRCC

According to our design considerations discussed in the above section, we selected a bilingual RC book as raw data to develop the BRCC. The book “*英語閱讀100天*,” published by the Chung Hwa Book Company (Hong Kong) Limited, supports English learning by Chinese readers. In total, there are 100 parallel Chinese-English reading comprehension passages. The corpus size is about 18K English words and 17K Chinese characters. There are 414 questions with corresponding answer keys in English. We manually translate the English questions and answers into Chinese. In addition, the passages cover the following domains: the English language, tourism, culture, society, sports, history, geography, arts, literature, economy, business, science, and technology. These attributes of the bilingual book are comparable to those of Remedia and CBC4Kids. We reserved 50 passages as the training set and the other 50 passages as the test set.

According to the Dale-Chall formula, the readability levels of the English training passages in the BRCC range from grades 7 to 12. Hence, compared to the passages in Remedia, those in the BRCC are more difficult. Compared to the passages in CBC4Kids, those in the BRCC are easier.

There are four questions on average for each passage. The distribution of different types of questions in the training set is listed in Table 5. We found that the corpus provides twelve types of questions that ask for key information from the passages, namely, *Who*, *What-DEF*, *What-VP*, *What-NP*, *What-OTH*, *Which*, *When*, *Where*, *Why*, *Yes/No*, *How* and *Others* (see Table 5). Table 6 shows a sample passage with questions and answer keys from the BRCC in both English and Chinese. In this table, the Chinese questions and answer keys are translated from English.

Table 5. Question type distribution in the BRCC training set

Question type	# questions	Example
Who	23	Who was William the Conqueror?
What-DEF	37	What is shocking story?
What-VP	4	What was Abraham Lincoln determined to do?
What-NP	35	What were well developed?
What-OTH	14	What is the size of Disney World? What is the nickname of Franz Beckenbauer? What is the main cause for those tragedies in America at schools?
Which	14	Which language was the other choice?
When	9	When did Disney World come into being?
Where	16	Where is New York City mainly located?
Why	15	Why do people have such a worry?
Yes/No	16	Did Thomas Edison like talking much?
How	15	How is an American nicknamed today?
Others	22	How many World Cups have there been? How long will the improvement take? How often do people use "hello"?
Overall	220	

Table 6. A sample passage with questions and answer keys in both Chinese and English in the BRCC

English passage	Imagine this: you have just won a competition, and the prize is an English language course at a famous school in Britain or the United States. You can either take a 30-week course for four hours a week, or a four-week course for 30 hours a week. Which one should you choose?...
English questions	1. If you win a competition, what may be the prize? 2. What may be the two kinds of courses?
English answer keys	1. The prize may be an English language course at a famous school in Britain or the United States. 2. They are either a 30-week course for four hours a week or a four-week course for 30 hours a week.
Chinese passage	想像一下：你刚赢得一场比赛，其奖赏是在英国或美国的一所名牌大学学习一门英语语言课程。你可以选一门为30周的课程，每周学习4小时，或者选一门为期4周的课程，每周30小时。你将作何选择？...
Chinese questions	1. 如果你赢得一场比赛，奖赏或许是什么？ 2. 两类课程会是什么？
Chinese answer keys	1. 奖赏会是在英国或美国的一所名牌大学学习一门英语语言课程。 2. 它们分别是一门为期30周的课程，每周学习4小时，或者选一门为期4周的课程，每周30小时。

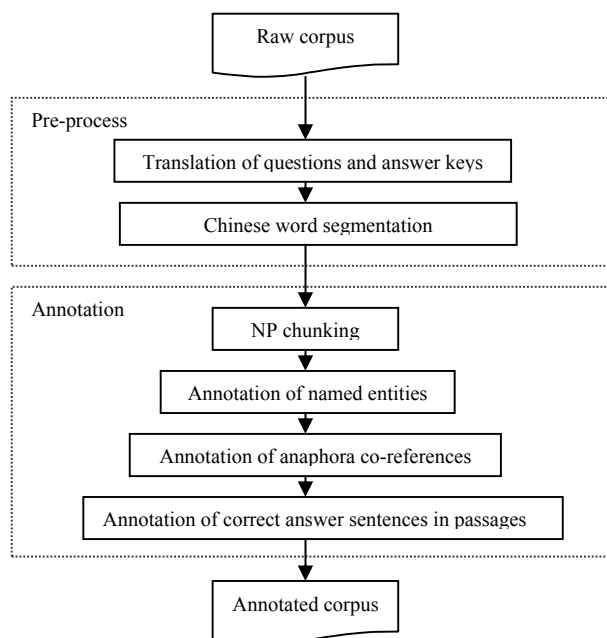


Figure 1. The development flow of the BRCC

The development flow is shown in Figure 1. We pre-processed the raw data in two steps:

1. We translated English questions and answer keys into Chinese manually.
2. We segmented Chinese words in Chinese passages and questions using the Chinese segmenter⁷ provided by the LDC. The Chinese language is written without word delimiters. In order to process the text based on words, we needed to tokenize the Chinese character strings into Chinese words. In addition, we annotated noun phrases, named entities, anaphora co-references, and correct answer sentences for passages and questions in both languages.

4.1 Translation of Questions and Answer Keys

The original book contains English passages, English questions, English answer keys, and Chinese passages, but no Chinese questions or answer keys. We arranged for two individual translators with high proficiency in English to translate the English questions and answer keys into Chinese. They worked separately without any communication between them. The two translators then cross-checked the translations and labelled identical translations (exact string matching) as correct. Disagreements in translations were reviewed individually and resolved

⁷ http://www ldc.upenn.edu/Projects/Chinese/LDC_ch.htm

by both translators together. In total, they found that 48.1% of the translations were exact matches. The distribution of miss-matches is explained in the following.

1. Using synonymous words (36.2%): For example, the translator either used “喝” (meaning: drink) or “饮” (meaning: drink) in the translation. Either translation is acceptable.
2. Using semantic equivalent descriptions (8.2%): The two translators express the same meaning in different ways. For example, they used “北方军” (meaning: northern army) or “北方的士兵” (meaning: northern soldiers) in their translation. The meaning of the second translation is equivalent to the English question, “*How were the northern soldiers called by the southern army during the Civil War?*” Therefore, the second translation was adopted.
3. Using active or passive voice (2.7%): For example, the two translators used “什么被播出” (in passive voice) or “播出什么” (in active voice) to translate “*what is shown.*” Since the English question uses the passive voice, the Chinese translation with passive voice was adopted.
4. Using anaphora versus direct reference (1.9%): The two translators represented the same entity using its name or its anaphor. For example, they used “他” (meaning: he) or “李克” (meaning: Rick) to refer to “*Rick*” in the question. Since “*Rick*” (i.e., direct reference) was used in the question, the second translation was adopted.
5. Different meanings (2.9%): For example, the English question was “*What did the music center on?*” The two translations were “该音乐用在什么地方” (meaning: what did the music use for) and “该音乐把什么作为中心” (meaning: what did the music center on). The first translation is wrong. The second translation is correct. The translation with the equivalent meaning to the English question was adopted.

4.2 Chinese Word Segmentation

Chinese word segmentation is a process in which word boundaries are identified. Typically, a space is inserted as a delimiter between Chinese words during segmentation. In this process, we manually corrected the outputs of an automated segmenter to produce correct Chinese words boundaries. We began by using the LDC Chinese segmenter to process all the passages, questions, and answer keys in order to maintain consistent segmentation across the board. This segmenter applies a dynamic programming approach to find the word segmentation path which has the highest multiple of word probabilities. In addition, this segmenter includes a lexicon in GB code, which contains 44,405 entries. Each entry of the lexicon contains a word, the occurrence frequency of the word, and the pinyin spelling (indicating the Mandarin pronunciation) of the word. Then, two linguistic annotators corrected and cross-checked the output of the Chinese segmenter so that the meanings of the Chinese sentences agreed with

those of the corresponding English sentences.

For example, “*Hingis*” is a person name in the English sentence, “*Hingis was so frustrated after the game.*” The corresponding Chinese translation is “赛后，轩芝丝感到灰心丧气。”“轩芝丝” is the Chinese transliteration of “*Hingis*.” It should be a Chinese word. However, the segmenter did not recognize the person name and mistakenly segmented it into three single-character words: “轩” (meaning: room), “芝” (meaning: a kind of herb), and “丝” (meaning: silk). These meanings are not found in the corresponding English sentence. Therefore, “轩”, “芝”, and “丝” was merged into a transliterated name, “轩芝丝”.

As another example, consider the English sentence, “*the USA’s population increased 3.3%.*” The segments in the output, “美国人 口,” included “美国人” (meaning: American) and “口” (meaning: mouth), which do not agree with the meaning of the original meaning of the English sentence. Hence, the correct segments should be “美国(meaning: USA) 人口 (meaning: population).”

4.3 Noun Phrase Chunking

This section describes the annotation of noun phrase (NP) boundaries for all the passages as well questions in both English and Chinese. This step was important because we assumed that the named entities, anaphors, and antecedent boundaries were consistent with the NP boundaries. Named entities and anaphora co-references were annotated based on the NP boundaries.

Noun phrase chunking is the process that segments sentences into non-recursive portions of noun phrases. After NP chunking, the NP boundaries (denoted by square brackets) are marked in the sentence. For example, note the following sentence before and after NP chunking:

The government has other agencies and instruments for pursuing these other objectives.

[*The government*] has [*other agencies and instruments*] for pursuing
[*these other objectives*] .

We first annotated English NP boundaries and referenced these to annotate the Chinese NP boundaries. Hence, the annotations of the NP boundaries in both English and Chinese were consistent. The annotation procedure is described below.

1. We applied Brill's part-of-speech transformational tagger⁸ to annotate each word in the English text (passages, questions, and answer keys) with its part-of-speech tag [Brill 1994].
2. We applied the BaseNP Chunker⁹ to identify English noun phrases. The input of the BaseNP Chunker was the output of the Brill's part-of-speech transformational tagger. With heuristic transformational rules trained on Wall Street Journal text from the UPenn Treebank, the BaseNP Chunker inserted square brackets marking the contained baseNP structures [Ramshaw and Marcus 1995].
3. We automatically replaced the left bracket "[" with "<NP>" and right bracket "]" with "</NP>". Thus, the annotation tags of named entities and anaphora co-references could be inserted within the brackets of the noun phrase, "<NP>".
4. Two human annotators corrected and cross-checked the outputs of the BaseNP Chunker. Each noun phrase should have had a noun or pronoun as the head. Other words in the noun phrase were modifiers of the head. Normally, a noun phrase should have ended with a noun/pronoun/adjective. In addition, a noun phrase should not have overlapped other noun phrases since we only considered non-recursive baseNP structures.

For example, consider the input sentence:

Some say it came from the French, "ho" and "la" – "Ho, there!"

The output of the BaseNP chunker was:

[*Some say*] [*it*] *came from* [*the French*] , "*ho*" and "*la*" – "*Ho, there!*"

After replacing the left bracket "[" with "<NP>" and the right bracket "]" with "</NP>", the output became:

<NP> *Some say* </NP> <NP> *it* </NP> *came from* <NP> *the French*
</NP> , "*ho*" and "*la*" – "*Ho, there!*"

⁸ http://www.cs.jhu.edu/~brill/RBT1_14.tar.Z

⁹ ftp://ftp.cis.upenn.edu/pub/chunker/basenp_chunker_1/

Actually, the NP boundary behind “Some say” is wrong because “say” is a verb in the sentence. The valid boundary should be inserted behind “Some”. Therefore, after the annotator corrected the error, the annotated sentence became:

< NP > *Some* < /NP > *say* < NP > *it* < /NP > *came from* < NP > *the French*
 < /NP > , “*ho*” and “*la*” – “*Ho, there!*”

In the process of annotating Chinese NP boundaries, we referenced the annotations of English NP and attempted to follow these to achieve cross-language consistency. For example, the corresponding Chinese sentence with word segmentations for the previous example is:

有人 认为 他 来自 於 法语 的 “*ho*” 和 “*la*” – 意思 是 : 「喂 , 那边 的 ! 」

The corresponding annotation of Chinese NP chunks is:

< NP > 有人 < /NP > 认为 < NP > 他 < /NP > 来自 於 < NP > 法语 < /NP > 的
 “*ho*” 和 “*la*” – 意思 是 : 「喂 , 那边 的 ! 」

In this example, the English counterpart of “有人” (meaning: somebody) is “*some*”; the English counterpart of “他” (meaning: he or it) is “*it*”; the English counterpart of “法语” (meaning: French) is “*the French*.”

However, due to differences between the two languages, occasions may arise in which a strict correspondence between the NP in English and the NP in Chinese cannot be maintained. Consider the following example:

English sentence: *It was once predicted that British and American English would draw so far apart that eventually they would become separate languages.*

Chinese sentence: 曾经 有人 预言 英国 英语 和 美国 英语 的 区别 会 越来越 大 直至 成为 两种 不同 的 语言。

“*It was predicted*” is translated as “有人 (meaning: somebody) 预言 (meaning: predict)” in the Chinese sentence. “*Somebody*” (有人) cannot match “*it*” because the subject of “*predict*” is not “*somebody*” in the English sentence but in the Chinese sentence. In addition, the translation of the Chinese word “区别” (meaning: difference) “*draw so far apart*,” which is

not a noun phrase. For Chinese NPs whose corresponding English NPs could not be found, the annotators still annotated them in the Chinese sentences. Therefore, “有人,” “区别,” “英国 英语,” “美国 英语,” and “两种 不同的 语言” in the previous example were annotated as noun phrases. The annotated NP boundaries in English and Chinese were:

English sentence: < NP > It < /NP > was once predicted that < NP > British < /NP > and < NP > American English < /NP > would draw so far apart that eventually < NP > they < /NP > would become < NP > separate languages < /NP >.

Chinese sentence: 曾经<NP>有人</NP>预言<NP>英国 英语</NP>和<NP>美国 英语</NP>的<NP>区别</NP>会 越来越大 直至 成为<NP>两种 不同的 语言</NP>。

4.4 Annotation of Named Entities

This section describes the annotation of named entities. Knowledge about named entities has been applied to develop RC systems [Hirschman *et al.* 1999]. Here, we used a filtering module to rank sentences higher if they contained appropriate named entities. Positive results were achieved by applying named entity filtering.

Table 7. Named entity types for different question types

Question type	Named entity type	examples
Who	PERSON	Thomas Alva Edison, soldier, etc.
What	-	
Which	-	
When	TIME	100 years, weekend, etc.
Where	LOCATION	the US, England, etc.
Why	-	
Yes/No	-	
How	-	
Others	NUM	Millions, 20 stories, etc.

According to the question types listed in Table 5, we annotated four types¹⁰ of named entities: PERSON, TIME, LOCATION, and NUM (see Table 7). Each annotator examined all the noun phrases and marked named entity tags in the left NP boundaries, “<NP>”. The format used was “<NP NE=value>”. The expected named entity type was assigned as the value to

¹⁰ ORGANIZATION was not annotated because there were no questions that asked about organizations, even though organizations (e.g., government names, company names, etc.) do appear in the BRCC passages. We will include ORGANIZATION in the BRCC in the future if necessary.

“NE”. The guidelines for identifying the four types of name entities were as follows:

- If the NP contains a person name or an occupation of person, the value should be PERSON.
- If the NP contains a year, month, week, date, duration, specific hour, or minute, the value should be TIME.
- If the NP contains a street name, a park name, a building name, a city name, or a country name, the value should be LOCATION.
- If the NP contains a specific number, e.g., for money, frequency, length, height, distance, width, area, weight, or age, the value should be NUM.

For example, after named entity annotation, the English sentence

This greeting may have arrived in England during the Norman Conquest in the year 1066

became

This greeting may have arrived in <NP NE=LOCATION> England </NP> during the Norman Conquest in <NP NE=TIME> the year 1066 </NP>.

The corresponding Chinese sentence

这个打招呼的用语也许是在1066年诺曼第人征服英国时带去的，

became

这个打招呼的用语也许是在<NP NE=TIME>1066年</NP>诺曼第人征服<NP NE=LOCATION>英国</NP>时带去的。

In this example, the noun phrase “*England*” (“英国” in Chinese) is tagged LOCATION; “*the year 1066*” (“1066年” in Chinese) is tagged TIME.

4.5 Annotation of Anaphora Co-references

Anaphora co-references show the relationships between anaphors and their antecedents. Both an anaphor and the corresponding antecedent refer to the same referent. We believe

anaphora co-references are important in RC tasks because anaphors can be matched with their antecedents when anaphora resolution is applied.

In this process, noun phrases that contain pronouns are annotated as anaphors; their corresponding antecedents are noun phrases that contain the same entities to which the anaphors refer. If multiple antecedents exist, the nearest prior one is used. For example, consider the following sentences:

The American inventor, Thomas Alva Edison, is believed to be the first person to use “hello” in the late 1800’s, soon after the invention of the telephone.

...

Thomas Edison was a man of few words.

He wasted no time.

“*He*” is the anaphor and refers to “*Thomas Edison*” and “*Thomas Alva Edison.*” The annotator chooses “*Thomas Edison*” as the antecedent because it is the nearest one.

We annotate anaphora co-references in the left NP tag “<NP>”. The annotation format of the anaphor is “< NP REF=value >”. The format of the antecedent is “< NP REFID=value >”. Each antecedent has a unique value, which is based on a counter that starts counting at the beginning of the passage. The value of an anaphor is identical to that of its antecedent. In other words, the annotator marks the co-reference relationship between an anaphor and its antecedent by assigning them the same value.

Refer again to the previous example; the last two sentences (in both English and Chinese) have the following co-reference annotations:

<NP NE=PERSON REFID=7> *Thomas Edison* </NP> *was a man of few words.*

<NP REF=7> *He* </NP> *wasted no time.*

<NP NE=PERSON REFID=7> 托马斯·爱迪生</NP> 是个 沉默寡言 的 人。

<NP REF=7>他</NP> 从 不 浪 费 时 间。

In this example, “*Thomas Edison*” (托马斯·爱迪生 in Chinese) is assigned a “REFID” of 7. “*He*” (他 in Chinese) refers to “*Thomas Edison*,” and its “REF” is assigned a value of 7.

4.6 Annotation of Correct Answer Sentences

In order to evaluate RC systems with the HumSent evaluation metric, we annotate answer sentences according to published answer keys, which are written by hand. Answer sentences are passage sentences that contain published answer keys or have the same meaning of published answer keys. Consider the following example:

Question 1: *What word is used most often in the world?*

Published answer key: *The word “hello” is used most often.*

Answer sentence: *The word “hello” is used more often than any other one in the English language.*

In this example, the answer key is an “extract” from the passage. Consider another example:

Question 2: *Did Thomas Edison like talking much?*

Published answer key: *He didn’t.*

Answer sentence: *Thomas Edison was a man of few words.*

In this example, the answer key is not an “extract” from the passage. However, the answer sentence is equivalent to the published answer key, because the statement that “Thomas Edison” did not like talking much means he was a man of few words.

We mark each answer sentence with left and right boundaries: “<ANSQ_{*i*}>” and “</ANSQ_{*i*}>”, where *i* is the sequence of the question. After the previous example is annotated, the correct answer sentences in English and Chinese are as follows:

<ANSQ2> <NP NE=PERSON REFID=7> Thomas Edison </NP> was a man of few words. </ANSQ2>

<ANSQ2> <NP NE=PERSON REFID=7> 托马斯·爱迪生</NP> 是个 沉默寡言的人。 </ANSQ2>

We list the distributions of different annotations among the 100 passages in Table 8. After noun phrases, named entities, anaphora co-references, and correct answer sentences are annotated, the annotated passage and questions in Table 6 are as shown in Table 9.

Table 8. The distribution of different annotations among the 100 passages

Annotation type	# annotations
Noun phrase	6877
Named entity-PERSON	1504
Named entity-LOCATION	558
Named entity-TIME	307
Named entity-NUM	416
Anaphora co-reference	379

Table 9. The annotated sample passage and questions from Table 6. This sample includes annotation tags for NP boundaries, named entities, anaphora co-references, and answer sentences.

English passage	<ANSQ1>Image <NP>this</NP>: you have just won <NP>a competition</NP>, and <NP>the prize</NP> is <NP REFID=1>an English language course</NP> at <NP>a famous school</NP> in <NP NE=LOCATION>Britain</NP> or <NP NE=LOCATION>the United States</NP>. </ANSQ1> <ANSQ2>You can either take <NP NE=TIME>a 30-week course</NP> for <NP NE=TIME>four hours a week</NP>, or <NP NE=TIME>a four-week course</NP> for <NP NE=TIME>30 hours a week</NP>. </ANSQ2> <NP REF=1>Which one</NP> should you choose?...
English questions	1. If you win <NP>a competition</NP>, what may be <NP>the prize</NP>? 2. What may be <NP>the two kinds</NP> of <NP>courses</NP>?
Chinese passage	<ANSQ1>想像一下：你 刚 赢得 <NP>一 场 比 赛</NP>，其 <NP>奖 赏 </NP> 是 在 <NP NE=LOCATION>英 国 </NP> 或 <NP NE=LOCATION>美 国 </NP> 的 <NP>一 所 名 牌 大 学 </NP> 学 习 <NP REFID=1>一 门 英 语 语 言 课 程 </NP>。 </ANSQ1> <ANSQ2>你 可 以 选 <NP>一 门 为 期 30 周 的 课 程 </NP>， <NP>每 周 </NP> 学 习 <NP>4 小 时 </NP>， 或 者 选 <NP>一 门 为 期 4 周 的 课 程 </NP>， <NP>每 周 </NP> <NP>30 小 时 </NP>。 </ANSQ2>你 将 作 <NP REF=1>何 选 择 </NP>？ ...
Chinese questions	1. 如 果 你 赢 得 <NP>一 场 比 赛 </NP>， <NP>奖 赏 </NP> 或 许 是 什 么 ？ 2. <NP>两 类 </NP> <NP>课 程 </NP> 会 是 什 么 ？

5. Benchmark Experiments and Discussion

In order to measure the comparative levels of difficulty among the BRCC, Remedia, and CBC4Kids, we applied the baseline bag-of-words (BOW) approach in our experiments. The same baseline has been previously applied to both Remedia and CBC4Kids. The RC system applied to Remedia is called Deep Read [Hirschman *et al.* 1999]. The input sentence of the

BOW matching approach is represented by a set of words, and the output is the first occurrence of the sentence that has the maximum number of matching words between the word set of the sentence and that of the question. The answer sentences are used to obtain HumSent results with the BOW matching approach.

Three pre-processes are performed prior to BOW matching:

1. The stemmed nouns and verbs are used to replace the original words¹¹.
2. English stop-word removal: We use the same stop-words list used in the Deep Read system [Hirschman *et al.* 1999]. They are forms of *be, have, do*, personal and possessive pronouns, *and, or, to, in, at, of, a, the, this, that*, and *which*.
3. Chinese stopword removal: The stop-words are the Chinese translations of the English personal/possessive pronouns, 和, 或, 到, 在, 中, 的, 这, and 那. We use the Chinese word segmentations in the BRCC directly.

In addition, we used named entity filtering (NEF) and pronoun resolution (PR) [Hirschman *et al.* 1999] to investigate the annotations of named entities and anaphora co-references. Both approaches have been applied in Deep Read [Hirschman *et al.* 1999]. The results obtained with both approaches showed significant improvements [Hirschman *et al.* 1999]. We applied these two approaches and repeated the experiments on the BRCC. For NEF, three named entity types (PERSON, TIME, and LOCATION) were used to perform answer filtering for three types of questions (*who, when and where*). The relationships are listed in the following [Hirschman *et al.* 1999]:

- For *who* questions, a candidate sentence that contains PERSON is assigned higher priority.
- For *where* questions, a candidate sentence that contains LOCATION is assigned higher priority.
- For *when* questions, a candidate sentence that contains TIME is assigned higher priority.

For Chinese questions, the question types refer to the corresponding English questions.

The Deep Read system uses a very simplistic approach to match five pronouns (*he, him, his, she and her*) to the nearest prior person name [Hirschman *et al.* 1999]. In addition, a different module uses the hand-tagged reference resolution of these five pronouns. In our experiment, we automatically resolved these five pronouns based on our hand-tagged references. For Chinese passages, we replaced the four pronouns 他, 她, 他的 and 她的 with their hand-tagged references. The detailed results obtained by applying BOW, NEF, and PR to the BRCC test set are listed in Table 10.

¹¹ A C function (morphstr) provided by WordNet is used to obtain the base forms of words [Miller *et al.* 1990].

Table 10. The detailed results obtained by applying bag-of-words (BOW), named entity filtering (NEF), and pronoun resolution (PR) to the BRCC test set

Corpus	BOW	BOW+NEF	BOW+PR
BRCC English test set	67%	68%	68%
BRCC Chinese test set	68%	69%	69%

The BOW approach achieved 29% HumSent accuracy when applied to the Remedia test set and 63% HumSent accuracy when applied to the CBC4Kids test set. As shown in Table 10, the BOW matching approach seemed to achieve especially good results when applied to the BRCC in comparison with the other corpora, as the improvement achieved by applying NEF and PR were not significant. A possible reason is that the questions tended to use the same words that were used in their answers in the BRCC. In the following example, we list a question and its correct answer in the BRCC training set. The corresponding word sets (i.e., bags of words) were obtained following stemming and stop-word removal:

Question: *Where did many sports played all over the world grow up to their present-day form?*

BOW: {*where many sport play all over world grow up present-day form*}

Correct answer: *Many sports which nowadays are played all over the world grew up to their present-day form in Britain.*

BOW: {*many sport nowadays play all over world grow up present-day form Britain*}

In this example, the intersection between the two word sets contains 10 words, 91% of which are in the question. We further calculated the overlap ratios for all the question-answer pairs in the English parts of the BRCC, Remedia, and CBC4Kids, and show the results in Table 11. We used the following formula:

$$\frac{\text{\# matching words between a question and its correct answer sentence}}{\text{\# words in the question}}$$

Table 11. The word overlap ratios for the English parts of the BRCC, Remedia, and CBC4Kids¹²

	BRCC(English)	Remedia	CBC4kids
Training set	71.7%	39.3%	46.3%
Test set	62.1%	37.8%	-

¹² Since the human-marked answers were not provided in the test set of our CBC4Kids copy, we were not able to compute the overlap ratio for the test set.

A high degree of word overlap between questions and correct answers could result in good BOW matching performance, which may mislead us to think that BOW is a sufficient approach for RC. Such overlap will artificially ease the task of RC. The difficulty levels of RC tests depend not only on the overlap between questions and correct answers but also on the world knowledge, domain ontology, etc. Questions may ask for information that is not provided in the passage, or for information that resides in different parts of the passage. Human beings can perform reasoning based on their world knowledge and domain ontology, but this process is really a challenge for machine performing automatic reading comprehension. For example, consider the following question and candidate answers:

Question: *Who owned the Negroes in the Southern States?*

Candidate sentence 1: *The blacks were brought to the Southern States as slaves.*

Candidate sentence 2: *They were sold to the plantation owners and forced to work long hours in the cotton and tobacco fields.*

If an RC system can infer that “*Negroes are sold to the plantation owners*” means “*they own Negroes*,” then it will be easy to know that candidate sentence 2 is the correct answer. In this paper, we present RC performance measured using the bag-of-words (BOW) approach in order to use it as a baseline performance benchmark. The BOW approach relies heavily upon the degree of word overlap between questions and their corresponding answer sentences. Improvement beyond this benchmark requires the use of more sophisticated techniques for passage analysis, question understanding, and answer generation. It also requires further work in authoring questions that cover various grades difficulty in order to challenge techniques used in automatic natural language processing.

In addition, it is insufficient to only consider the word overlap in the BOW matching approach. The inter-word relationships, such as lexical dependencies among concepts in syntactic parsing, are also important for developing RC systems. In the following example, we list a question, two candidate sentences, and their corresponding word stems:

Question: *What is the new machine called?*

BOW: {*what new machine call*}

Candidate sentence 1: *A new machine has been made.*

BOW: {*new machine make*}

Candidate sentence 2: *The machine is called a typewriter.*

BOW: {*machine call typewriter*}

In this example, both candidate sentences have two matching words. The BOW matching approach cannot distinguish between them. The matching words between candidate sentence 1 and the question are: “new” and “machine,” where “new” is the modifier of “machine.” The matching words between candidate sentence 2 and the question are: “machine” and “call,” where “machine” is the object of “call.” Candidate sentence 2 and the question share a dependency with respect to the verb “call.” But candidate sentence 1 and the question do not share any dependency with respect to any verb. Actually, candidate sentence 2 is the correct answer sentence. Based on this example, we believe that syntactic structures, such as verb dependencies between words, can be applied to improve the performance of RC systems.

6. Conclusions

In this paper, we have presented the design and development of a bilingual reading comprehension corpus (BRCC). The reading comprehension (RC) task has been widely used to evaluate human reading ability. Recently, this task has also been used to evaluate automatic RC systems [Anand *et al.* 2000; Charniak *et al.* 2000; Hirschman *et al.* 1999; Ng *et al.* 2000; Riloff and Thelen 2000]. An RC system can automatically analyze a passage and generate an answer for each question from the given passage. Hence, the RC task can be used to assess the state of the art of natural language understanding. Furthermore, an RC system presents a novel paradigm of information retrieval and complements existing search engines used on the Web.

So far, two English RC corpora, Remedia and CBC4Kids, have been developed. These corpora include stories, human-authored questions, answer keys, and linguistic annotations, which provide important support for the empirical evaluation of RC performance. In the current work, we developed an RC corpus to drive research of NLP techniques in both English and Chinese. As an initial step, we selected a bilingual RC book as the raw data, which contained English passages, questions, answer keys, and Chinese passages. We then manually translated the English questions and answer keys into Chinese and segmented the Chinese words. We also annotated the noun phrases, named entities, anaphora co-references, and correct answer sentences for the passages.

We gauged the comparative readability levels of the English passages by applying the Dale-Chall formula to the BRCC, Remedia, and CBC4Kids. We also measured the comparative levels of difficulty among the three corpora in terms of question answering using the baseline bag-of-words (BOW) approach. Our results show that the readability level of the BRCC is higher than that of Remedia and lower than that of CBC4Kids. We also observed that the BOW approach attains a better RC performance when applied to the BRCC (67%) than that it does when applied to Remedia (29%) and CBC4Kids (63%). The measured overlap values were 71.7% (training set) and 62.1% (test set) for the BRCC, compared with 39.3% (training set) and 37.8% (test set) for Remedia. This indicates that there is a higher degree of

word overlap which artificially simplifies the RC task with the BRCC. This strongly suggests that more effort must be made to author questions at various difficulty levels in order for the BRCC to better support RC research across the English and Chinese languages.

Acknowledgements

This project has been partially supported by a grant from the Area of Excellence in Information Technology of the Hong Kong SAR Government. In addition, we would like to thank the anonymous reviewers for their comments.

References

- Allen, J., *Natural Language Understanding*, The Benjamin/Cummings Publishing Company, Menlo Park, CA, 1995.
- Anand, P., E. Breck, B. Brown, M. Light, G. Mann, E. Riloff, M. Rooth, and M. Thelen, "Fun with Reading Comprehension," Final Report of the Workshop 2000 of Language Engineering for Students and Professionals Integrating Research and Education, Reading Comprehension, in Johns Hopkins University, 2000.
- Brill, E., "Some advances in rule-based part of speech tagging," In *Proceedings of the Twelfth National Conference on Artificial Intelligence (AAAI-94)*, 1994, pp. 722–727.
- Buchholz, S., "Using Grammatical Relations, Answer Frequencies and the World Wide Web for TREC Question Answering," In *Proceedings of the tenth Text Retrieval Conference (TREC 10)*, 2001, pp. 502–509.
- Chall, J. S. and E. Dale, *Readability revisited: The new Dale-Chall readability formula*, Cambridge, MA: Brookline Books, 1995.
- Charniak, E., *Towards a Model of Children's Story Comprehension*, Ph.D. thesis, Massachusetts Institute of Technology, 1972.
- Charniak, E., Y. Altun, R. D. S. Braz, B. Garrett, M. Kosmala, T. Moscovich, L. Pang, C. Pyo, Y. Sun, W. Wy, Z. Yang, S. Zeller, and L. Zorn, "Reading Comprehension Programs In a Statistical-Language-Processing Class," In *ANLP-NAACL 2000 Workshop: Reading Comprehension Tests as Evaluation for Computer-Based Language Understanding Systems*, 2000, pp. 1–5.
- Collins, M., *Head-Driven Statistical Models for Natural Language Parsing*, PhD thesis, University of Pennsylvania, 1999.
- Dale, E. and J. S. Chall, "A Formula for Predicting Readability: Instructions," *Educational Research Bulletin*, 1948, pp. 37–54.
- Dalmas, T., J. L. Leidner, B. Webber, C. Grover, and J. Bos, "Generating Annotated Corpora for Reading Comprehension and Question Answering Evaluation," In *Proceedings of the Workshop on Question Answering held at the Tenth Annual Meeting of the European Chapter of the Association for Computational Linguistics 2003 (EACL'03)*, 2003, pp. 13–19.

- Hirschman, L., M. Light, E. Breck, and J. Burger, "Deep Read: A Reading Comprehension System," In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, 1999, pp. 325–332.
- Klare, G., "The Measurement of Readability," In *Iowa State University Press, Ames, Iowa*, 1963.
- Light, M., G. S. Mann, E. Riloff, and E. Breck, "Analyses for Elucidating Current Question Answering Technology," In *Journal of Natural Language Engineering*, 4(7), 2001, pp. 1351–3249.
- Miller, G. A., R. Beckwith, C. Fellbaum, D. Gross, and K. J. Miller, "Introduction to WordNet: An on-line lexical database," In *International Journal of Lexicography*, 1990, pp. 235–312.
- Ng, H. T., L. H. Teo, and L. P. Kwan, "A Machine Learning Approach to Answering Questions for Reading Comprehension Tests," In *Proceedings of the 2000 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, 2000, pp. 124–132.
- Ramshaw, L. and M. Marcus, "Text Chunking Using Transformation-Based Learning," In *Proceedings of the Third ACL Workshop on Very Large Corpora*, 1995, pp. 82–94.
- Riloff, E. and M. Thelen, "A Rule-based Question Answering System for Reading Comprehension Test," In *ANLP/NAACL-2000 Workshop on Reading Comprehension Tests as Evaluation for Computer-Based Language Understanding Systems*, 2000, pp. 13–19.
- Voorhees, E. M., "Overview of the TREC 2001 Question Answering Track," In *Proceeding of the NIST Special Publication 500-250: The Tenth Text REtrieval Conference (TREC 2001)*, 2001, pp. 1–15.

A Chinese Term Clustering Mechanism for Generating Semantic Concepts of a News Ontology

Chang-Shing Lee*, Yau-Hwang Kuo⁺, Chia-Hsin Liao⁺ and Zhi-Wei Jian⁺

Abstract

In order to efficiently manage and use knowledge, ontology technologies are widely applied to various kinds of domain knowledge. This paper proposes a Chinese term clustering mechanism for generating semantic concepts of a news ontology. We utilize the parallel fuzzy inference mechanism to infer the *conceptual resonance strength* of a Chinese term pair. There are four input fuzzy variables, consisting of a *Part-of-Speech (POS)* fuzzy variable, *Term Vocabulary (TV)* fuzzy variable, *Term Association (TA)* fuzzy variable, and *Common Term Association (CTA)* fuzzy variable, and one output fuzzy variable, the *Conceptual Resonance Strength (CRS)*, in the mechanism. In addition, the *CKIP* tool is used in Chinese natural language processing tasks, including POS tagging, refining tagging, and stop word filtering. The *fuzzy compatibility relation* approach to the semantic concept clustering is also proposed. Simulation results show that our approach can effectively cluster Chinese terms to generate the semantic concepts of a news ontology.

Keywords: Ontology, Chinese Natural Language Processing, Fuzzy Inference, Feature Selection, Concept Clustering

1. Introduction

An ontology is an explicit, machine-readable specification of a shared conceptualization [Studer *et al.* 1998]. It is an essential element in many applications, including agent systems, knowledge management systems, and e-commerce platforms. It can help generate natural language, integrate intelligent information, provide semantic-based access to the Internet, and extract information from texts [Gomez-Perez *et al.* 2002] [Fensel 2002] [Schreiber *et al.* 2001]. Soo *et al.* [2001] considered an ontology to be a collection of key concepts and their inter-relationships, collectively providing an abstract view of an application domain. With the

* Department of Information Management, Chang Jung Christian University, Tainan, Taiwan
E-Mail: leecs@mail.cju.edu.tw; leecs@cad.csie.ncku.edu.tw

⁺ CREDIT Research Center, National Cheng Kung University, Tainan, Taiwan

support of an ontology, a user and a system can communicate with each other through their shared and common understanding of a domain. M. MissiKoff *et al.* [2002] proposed an integrated approach to web ontology learning and engineering that can build and access a domain ontology for intelligent information integration within a virtual user community. The proposed approach involves automatic concept learning, machine-supported concept validation, and management. Embley *et al.* [1998] presented a method of extracting information from unstructured documents based on an ontology. Alani *et al.* [2003] proposed the Artequakt, which automatically extracts knowledge about artists from the Web based on a domain ontology. It can generate biographies that are tailored to a user's interests and requirements. Navigli *et al.* [2003] proposed OntoLearn with ontology learning capability to extract relevant domain terms from a corpus of text. OntoSeek [Guarino *et al.* 1999] is a system designed for content-based information retrieval. It combines an ontology-driven content-matching mechanism with moderately expressive representation formalism. Lee *et al.* [2004] proposed an ontology-based fuzzy event extraction agent for Chinese news summarization. The summarization agent can generate a sentence set for each piece of Chinese news.

In this paper, we propose a Chinese term clustering mechanism for generating the semantic concepts of a news ontology. The parallel fuzzy inference mechanism is adopted to infer the conceptual resonance strength for any two Chinese terms. The *CKIP* tool [Academia Sinica 1993] is used in Chinese natural language processing, including POS tagging, refining tagging, and stop word filtering. The remainder of this paper is structured as follows. Section 2 introduces the structure of the Chinese term clustering mechanism. Semantic concept analysis for Chinese term clustering is presented in Section 3. Section 4 introduces the parallel fuzzy inference mechanism for semantic concept generation. Section 5 presents experimental results. Finally, some conclusions are drawn in Section 6.

2. The Structure of the Chinese Term Clustering Mechanism

An ontology is defined as a set of representational terms called concepts. The inter-relationships among these concepts describe a target world. Here, we will briefly describe the structure of the object-oriented ontology [Lee *et al.* 2003]. An object-oriented ontology consists of several basic components: (1) *Domain*: The top layer of the ontology is the name of the domain knowledge. In this study, an ontology was constructed for Chinese news, so its domain name is Chinese news. (2) *Category*: The second layer contains the categories of the domain ontology. Each category is composed of some concepts with various inter-relationships. There are seven categories for our Chinese news ontology. They are "Political" (政治焦點), "International" (國際要聞), "Finance" (股市財經), "Cross-Strait" (兩岸風雲), "Societal" (社會地方), "Entertainment" (運動娛樂), and "Life" (生活新知). (3)

Concept Set: The *Concept Set* is composed of various concepts and relations. We treat each concept in the ontology as a class, so the structure of the *Concept Set* can be treated as a class diagram. Figure 1 shows an example for our Chinese Political news domain ontology [Lee et al. 2003].

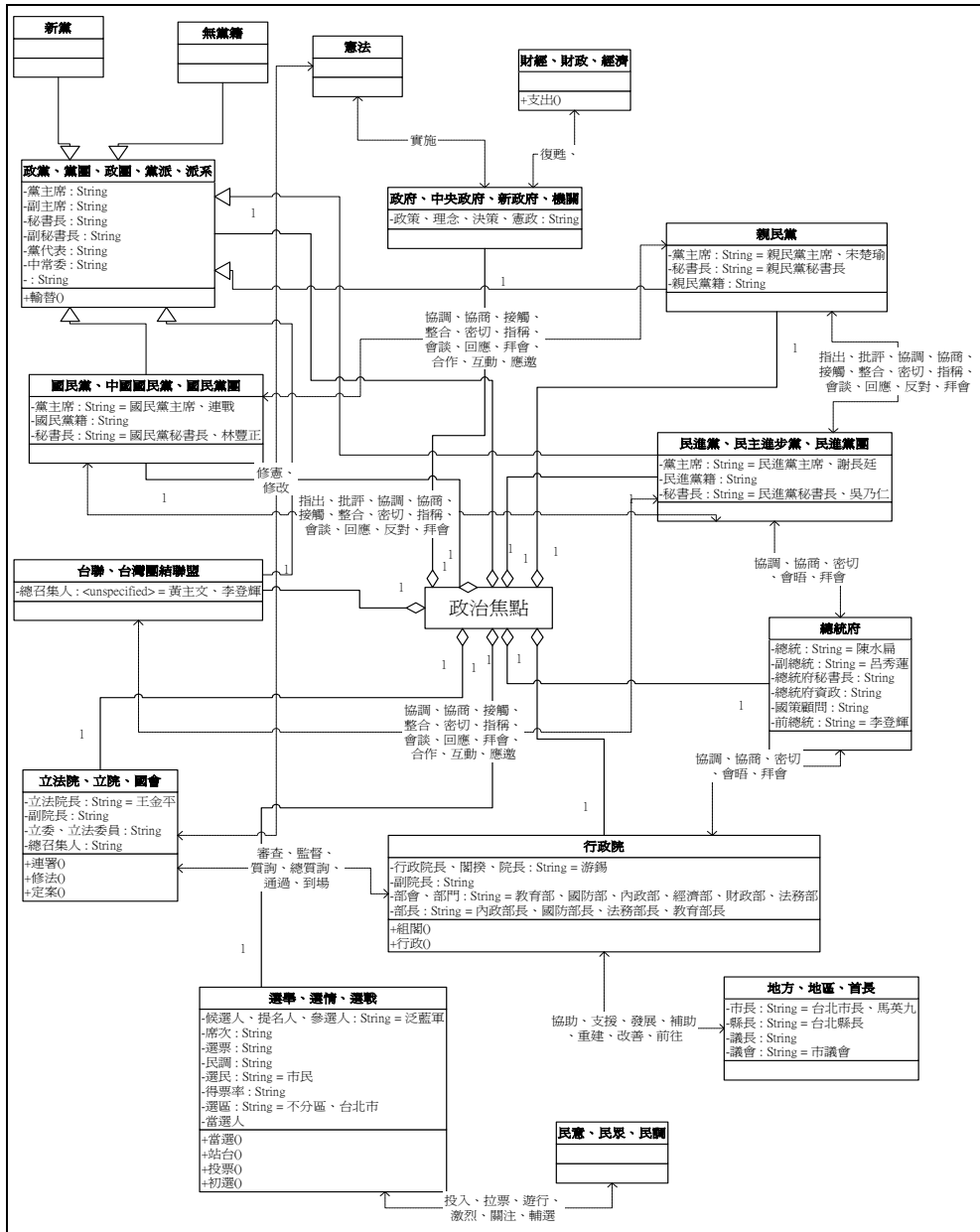


Figure 1. The domain ontology for the news category "Political"

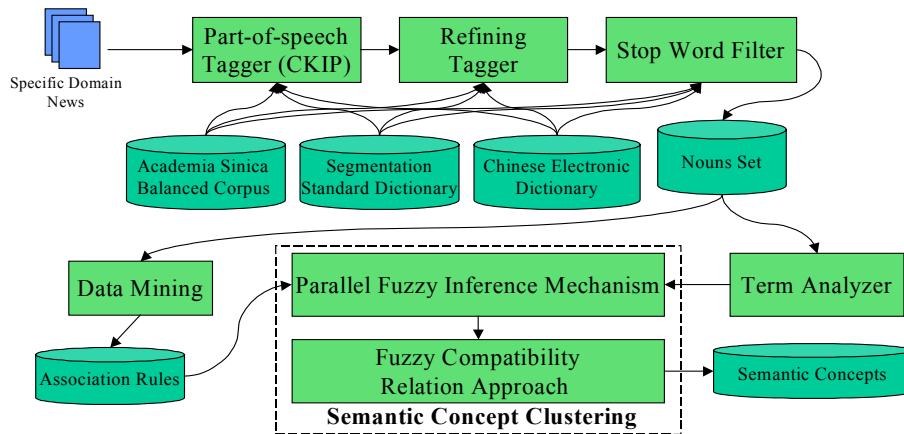


Figure 2. The structure of the Chinese term clustering mechanism

In this section, we will propose a Chinese term clustering mechanism for generating the semantic concepts of a news ontology. Figure 2 shows the structure of the Chinese term clustering mechanism. Natural language processing technologies were utilized to deal with the Chinese news that we gathered from the China Times website (<http://www.chinatimes.com.tw>). Several technologies, including a part-of-speech tagger, refining tagger, stop word filter, and term analyzer, were adopted for document pre-processing. Chinese language processing tools, such as *CKIP* [Academia Sinica 1993], the *Academia Sinica Balanced Corpus*, *Segmentation Standard Dictionary* [Academia Sinica 1998], and *Chinese Electronic Dictionary* [Academia Sinica 1993] provided by Academia Sinica, were used to deal with the Chinese news. In addition, the data mining technique and the concept clustering approach based on the *fuzzy compatibility relation* were employed. We will briefly describe these technologies in the following.

First, the *CKIP* is used to tag each word with its POS tag for the Chinese news. The refining tagger then refers to the *Academia Sinica Balanced Corpus* and *Chinese Electronic Dictionary* to refine the POS tags. With the aid of the corpus and the dictionary, we have sufficient Chinese POS knowledge to analyze the features of the terms for semantic concept clustering. The *stop word filter* is used to select terms with useful POS tags as candidate features. Table 1 shows unmeaning tags as stop words. Then, the *term analyzer* analyzes the term frequency of the news to select the important terms from a specific class of news. For example, the terms with the POS tags Na (普通名詞), Nb (專有名詞), Nc (地方名詞), and Nd (時間名詞) are preserved and sent to the *Parallel Fuzzy Inference Mechanism* for further processing. The *Data Mining* mechanism adopts the *Apriori Algorithm* to generate association rules, which are used in the *Parallel Fuzzy Inference Mechanism*. The *Apriori Algorithm* [Jacobes 1993] is described as follows.

Table 1. The stop word list used in the stop word filter [Academia Sinica 1993]

Part-of-Speech Tag	Meaning	Examples
Ca	並列連接詞	和、或者
Cb	關聯連接詞	雖然、不但
Da	數量副詞	一共、恰好
Db	法相副詞	一定、也許
Dbb,Dbc	評價副詞	居然、果然
Dc	否定副詞	沒有、未
Dd	時間副詞	隨即、稍後
Df	程度副詞	非常、更
Dg	地方副詞	到處、遍地
Dh	方式副詞	如此、從中
Di	標誌副詞	過、起來
Dj	疑問副詞	為何、何故
Dk	句副詞	據報、據了解
I	感歎詞	哦、哇
P	介詞	經過、遭受
T	語助詞	了、的

Apriori Algorithm:

Find frequent itemsets using an iterative level-wise approach based on candidate generation.

Input:

Database D of transactions and minimum support threshold min_sup .

Output:

L , frequent itemsets in D .

Method:

Step 1: $L_1 = \text{find_frequent_1-itemsets}(D, min_sup)$; //find_frequent_1-itemsets denotes to find frequent 1-itemsets in D

Step 2: For ($k=2; L_{k-1} \neq \phi; k++$) {

Step 2.1: $C_k = \text{apriori_gen}(L_{k-1}, min_sup)$;

Step 2.2: For each transaction $t \in D$ { //scan D for counts

$C_t = \text{subset}(C_k, t)$; //get the subsets of t that are candidates

Step 2.3: For each candidate $c \in C_t$

$c.count++$;

}

Step 3: $L_k = \{c \in C_k | c.count \geq min_sup\}$

}

Step 4: Return $L = \cup_k L_k$

Step 5: End.

Procedure find_frequent_1-itemsets(D, min_sup)

Step 1: Get C_1 from D // C_1 denotes candidate 1-itemsets

```

Step 2: For each transaction  $t \in D$  { //scan D for counts
   $C_t = \text{subset}(C_k, t)$ ; //get the subsets of  $t$  that are candidates
}
Step 3: For each candidate  $c \in C_1$ 
   $c.\text{count}++$ ;
}
Step 4:  $L_1 = \{c \in C_k | c.\text{count} \geq \text{min\_sup}\}$ 
Step 5: Return  $L_1$ 
Step 6: End.
Procedure apriori_gen( $L_{k-1}; \text{min\_sup}$ )
Step 1: For each itemset  $l_1 \in L_{k-1}$ 
  Step 1.1: For each itemset  $l_2 \in L_{k-1}$ 
    Step 1.2: If  $(l_1[1] = l_2[1]) \wedge (l_1[2] = l_2[2]) \wedge \dots \wedge (l_1[k-1] < l_2[k-1])$  then {
       $c = l_1 \cup l_2$ ; //join step: generate candidates
    }
    Step 1.3: If Has_infrequent_subset( $c, L_{k-1}$ ) then
      delete  $c$ ; //prune step: remove unfruitful candidate
    Else add  $c$  to  $C_k$ 
  }
Step 2: Return  $C_k$ ;
Step 3: End.
Procedure Has_infrequent_subset ( $c; L_{k-1}$ ) //use priori knowledge
Step 1: For each  $(k-1)$ -subset  $s$  of  $c$ 
  Step 1.1: If  $s \notin L_{k-1}$  then return TRUE;
  Step 1.2: Else return FALSE;
Step 2: End.

```

3. Semantic Concept Analysis for Chinese Term Clustering

In this paper, we propose the *Conceptual Resonance Strength (CRS)* fuzzy variable for Chinese term clustering. The *CRS* is the similar degree for any term pair in the same concept. Hence, any Chinese term pair with a strong *CRS* will be classified as the same concept. We use four fuzzy variables, the *resonance strength in Part-of-Speech (POS)*, *resonance strength in Term Vocabulary (TV)*, *resonance strength in Term Association (TA)*, and *resonance strength in Common Term Association (CTA)*, to compute the *CRS* of the Chinese term pair. We will describe these variables in the following.

A. Resonance Strength in Part-of-Speech (POS)

The first fuzzy variable for *CRS* is the *resonance strength in Part-of-Speech (POS)*. Figure 3 shows the structure of the tagging tree that is used to compute the *resonance strength of POS* for any Chinese term pair. Table 2 shows the refining POS tags of Chinese noun terms.

Table 2. The refining POS tags of Chinese noun terms

Part-of-Speech Tag		Meaning	Examples
粗詞類	細詞類		
Na		普通名詞	
	Naa	物質名詞	泥土、水
	Nab	個體名詞	桌子、杯子
	Nac	可數抽象名詞	夢、符號
	Nad	抽象名詞	風度、香氣
	Nae	集合名詞	車輛、船隻
Nb		專有名詞	
	Nba	正式專有名詞	雙魚座、余光中
	Nbc	姓氏	張、王
Nc		地方名詞	
	Nca	專有地方名詞	西班牙、台北
	Ncb	普通地方名詞	郵局、市場
	Ncc	名方式地方名詞	海外、身上
	Ncd	表事物相對位置的地方詞	上頭、中間
	Nce	定名式地方名詞	四海、當地
Nd		時間名詞	
	Nda	時間名詞(歷史性、循環重複)	唐朝、春、夏、秋、冬
	Ndc	名方式時間名詞	年底、週末
	Ndd	副詞性時間名詞	現在、當今
Ne		定詞	這、哪、少許
Nf		量詞	
	Nfa	個體量詞	一"張"桌子、一"個"杯子
	Nfb	跟述賓式合用的量詞	寫一"手"好字、下一"盤"棋
	Nfc	群體量詞	一"雙"筷子、一"副"耳環
	Nfd	部分量詞	一"節"甘蔗、一"段"文章
	Nfe	容器量詞	一"箱"書、一"碗"飯
	Nff	暫時量詞	一"頭"秀髮、一"地"落葉
	Nfg	標準量詞	公斤、法郎
	Nfh	準量詞	國、面
	Nfi	述詞用量詞	看一"遍"、摸一"下"
	Nfzz	零量詞	"三萬"人口
Ng		方位詞	接"上"、屋"後"、睡覺"之前"
Nh		代名詞	
	Nha	人稱代名詞	你、我、他、自己
	Nhb	疑問代名詞	誰、什麼
	Nhc	泛指代名詞	之、其

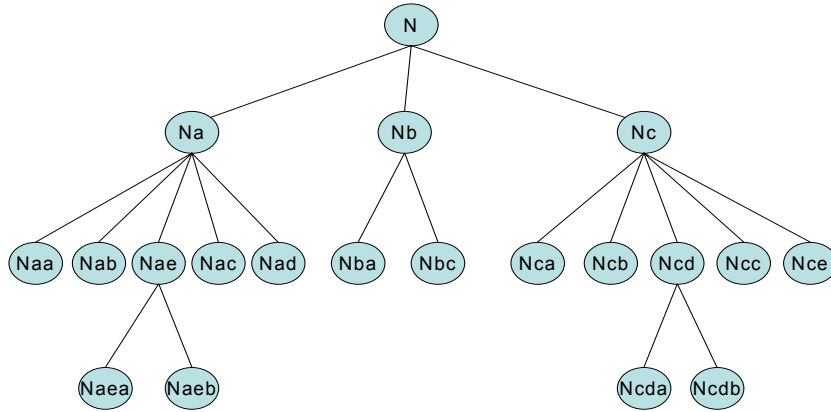


Figure 3. The structure of the tagging tree derived using CKIP

The resonance will be strong when the path distance of any Chinese term pair is short. For example, the two terms “電腦(computer)” and “軟體(software)” with their POS are “電腦(computer) (Nab)” and “軟體(software) (Nac),” respectively. Hence, the path distance of the term pair (“電腦(computer)”, “軟體(software)”) is 2 (Nab -> Na -> Nac).

B. Resonance Strength in Term Vocabulary (TV)

From the viewpoint of Chinese language characteristics, any term pair with common words will be similar in semantic meaning. For example, the Chinese terms in the term set {民進黨, 民進黨團, 民主進步黨} are similar in semantic meaning since they are composed of the common words “民”, “進”, and “黨”. We also consider another characteristic of Chinese terms with respect to term vocabulary. This assumes that terms having the same starting or ending word will share some common linguistic properties [Yang *et al.* 1994][Gao *et al.* 2001]. Good examples of starting and ending words are as follows: {星期一 (Monday), 星期六 (Saturday), 星期日 (Sunday)} and {昨天 (yesterday), 明天 (tomorrow), 今天 (today), 每天 (everyday)}. The first term set has the same starting word “星,” and the second term set has the same ending word “天.” The algorithm for computing resonance strength in *TV* [Lee *et al.* 2003] is shown below.

Algorithm for computing the resonance strength in TV **Input:** All terms (t_1, t_2, \dots, t_n) selected from Term Analyzer**Output:** Resonance strength in TV between any two Chinese terms**Method:****Step1:** For all terms (t_1, t_2, \dots, t_n) **Step1.1:** For all terms (t_1, t_2, \dots, t_n) **Step1.1.1:** Generate a term pair (t_a, t_b) $1 \leq a < b \leq n$ **Step1.1.2:** $TV(t_a, t_b) = N$ /* N represents identical words between the Chinese term pair (t_a, t_b) */**Step1.1.3:** If the starting word of two Chinese terms is the same then $TV(t_a, t_b) = TV(t_a, t_b) + 0.5$.**Step1.1.4:** If the ending word of two Chinese terms is the same then $TV(t_a, t_b) = TV(t_a, t_b) + 0.5$.**Step2:** Max_{TV} = maximum $TV(t_i, t_j)$ value of all term pairs.**Step3:** Min_{TV} = minimum $TV(t_i, t_j)$ value of all term pairs.**Step4:** End.

For example, the two terms “民進黨團” and “民主進步黨” have three common words, “民,” “進,” and “黨,” and the same starting word, “民,” so the total strength is 3.5.

C. Resonance Strength in Term Association (TA)

A large amount of previous research has focused on how to best cluster similar terms together. The proposed methods can be roughly grouped into two categories: knowledge-based clustering and data-driven clustering [Gao *et al.* 2001]. However, the obtained term knowledge is not sufficient for concept clustering, because a term pair is sometimes similar in meaning but lacks common properties of knowledge. Therefore, the confidence value derived using the *Apriori Algorithm* for the term pair can be applied to decide the strength of term relation. A term pair with a high confidence value consists of two terms that have a strong relationship and can be classified as the same concept. For example, the term set {總統 (President) (Nab), 總統府(The Office of the President) (Nca), 陳水扁(President Chen) (Nb)} represents similar concepts, so they will be clustered into the same concept. But from the viewpoint of term knowledge, only the two terms “總統(President)” and “總統府(The Office of the President)” will be clustered into the same concept. The term “陳水扁(President Chen)” will not be clustered into the concept {總統(President), 總統府(The Office of the President)}. Therefore, the *resonance strength in TA* is necessary for concept clustering. In addition, the resonance strength is decided by the confidence value of the two terms, so we adopt the average of the two confidence values as the resonance strength in *TA*. For example, the term pair {總統 (Nab), 陳水扁 (Nb)} for the “Political(政治焦點)” category (<http://www.chinatimes.com.tw>) with (總統 -> 陳水扁) has a confidence value of 0.84, and the confidence value of (陳水扁 -> 總統) is 0.80, so the resonance strength in *TA* is 0.82 $((0.84+0.80)/2)$.

D. Resonance Strength in Common Term Association (CTA)

Any two Chinese terms with the same common words or starting/ending words may not have the similar meaning. For example, consider the three Chinese terms “美國(U.S.A.),” “美方(U.S.A.),” and “警方(police)”. The Chinese terms “美國(U.S.A.)” and “美方(U.S.A.)” have the common starting word “美”; meanwhile, the Chinese terms “美方(U.S.A.)” and “警方(police)” have the common ending word “方”. But the common terms with a specific threshold of confidence for “美國,” “美方,” and “警方” are as follows:

- 美國(U.S.A.) -> {白宮(White House), 布希(Bush), 紐約(New York)};
- 美方(U.S.A.) -> {白宮(White House), 布希(Bush), 五角大廈(Pentagon)};
- 警方(police) -> {警員(policeman), 刑事組(criminal investigation), 分局(police station)}.

Hence, the common term set for {美國(U.S.A.), 美方(U.S.A.)} is {白宮(White House), 布希(Bush)}, and for {警方(police), 美方(U.S.A.)} is *Null*. Therefore, the term pair {美國(U.S.A.), 美方(U.S.A.)} has stronger resonance in *CTA* than the term pair {警方(police), 美方(U.S.A.)} does.

4. The Parallel Fuzzy Inference Mechanism for Semantic Concept Generating

We adopt the parallel fuzzy inference mechanism for semantic concept clustering. The fuzzy variables for computing the *CRS* of any Chinese term pair are adopted in the mechanism.

4.1 Aggregate Term Resonance with the Parallel Fuzzy Inference Mechanism

In this subsection, we will explain how four input fuzzy variables can be aggregated into one output fuzzy variable to compute the *CRS* of each Chinese term pair. There are four input fuzzy variables, consisting of *Part-of-Speech Similarity (POS)*, *Term-Vocabulary Similarity (TV)*, *Term-Association Strength (TA)*, and *Common Term-Association Strength (CTA)*, and one output fuzzy variable, *Conceptual Resonance Strength (CRS)*, in the Parallel Fuzzy Inference Mechanism. Two linguistic terms, *POS_Low* and *POS_High*, are defined in the *POS* fuzzy variables. Figure 4 shows the membership functions of the fuzzy sets {*POS_Low*,

POS_High} for the fuzzy variable *POS similarity*, where $p = \frac{\max_{POS} - \min_{POS}}{100}$.

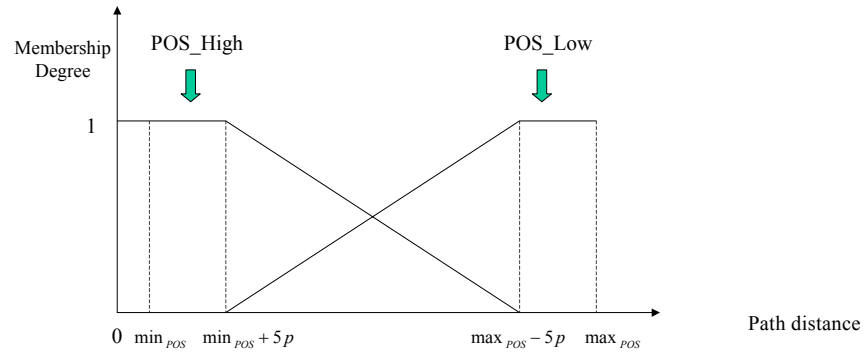


Figure 4. The membership functions of the POS fuzzy variable

Two linguistic terms, *TV_Low* and *TV_High*, are defined in the *TV* fuzzy variable. Figure 5 shows the membership functions of the fuzzy sets $\{TV_Low, TV_High\}$ for the fuzzy variable *TV similarity*, where $p = \frac{\max_{TV} - \min_{TV}}{100}$.

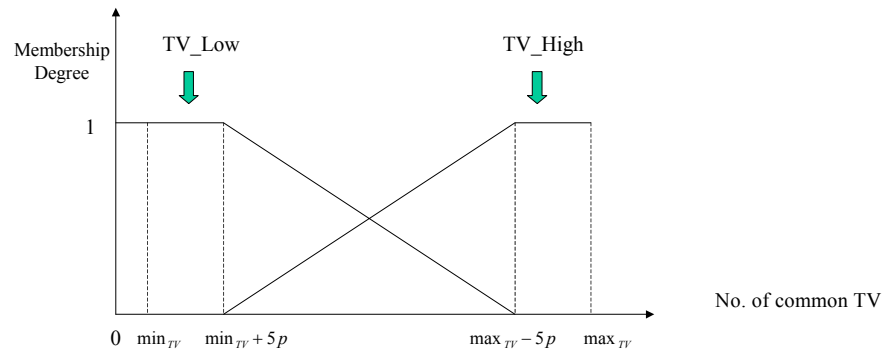


Figure 5. The membership functions of the TV fuzzy variable

Three linguistic terms, consisting of *TA_Low*, *TA_Medium*, and *TA_High*, are defined in the *TA* fuzzy variable. The membership functions of the fuzzy sets $\{TA_Low, TA_Medium, TA_High\}$ for the fuzzy variable *TA strength* are shown in Figure 6, where

$$p = \frac{\max_{TA} - \min_{TA}}{100}.$$

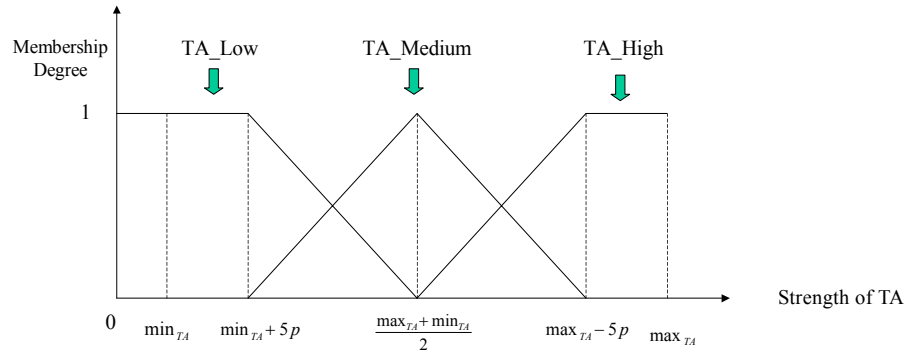


Figure 6. The membership functions of the TA fuzzy variable

Three linguistic terms, consisting of *CTA_Low*, *CTA_Medium*, and *CTA_High*, are defined in the *CTA* fuzzy variable. Figure 7 shows the membership functions of the fuzzy sets $\{CTA_Low, CTA_Medium, CTA_High\}$ for the fuzzy variable *CTA strength*, where

$$p = \frac{\max_{CTA} - \min_{CTA}}{100}.$$

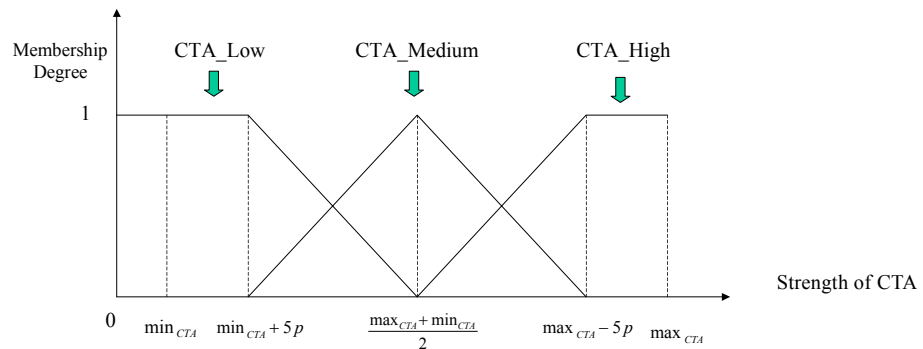


Figure 7. The membership functions of the CTA fuzzy variable

Five linguistic terms, consisting of *CRS_Very_Low*, *CRS_Low*, *CRS_Medium*, *CRS_High*, and *CRS_Very_High*, are defined in the *CRS* fuzzy variable. Figure 8 shows the membership functions of the fuzzy sets $\{CRS_Very_Low, CRS_Low, CRS_Medium, CRS_High,$

*CRS_Very_High\} for the fuzzy variable *CRS strength*, where $p = \frac{\max_{CRS} - \min_{CRS}}{100}$.*

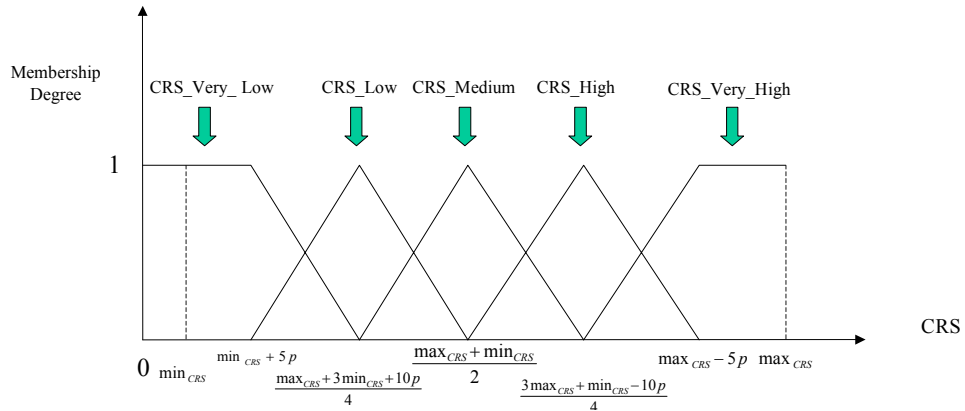


Figure 8. The membership functions of the CRS fuzzy variable

Having described the fuzzy variables used to compute the CRS of a Chinese term pair, we will next explain how the parallel fuzzy inference mechanism proposed by Kuo *et al.* [1998] and Lin [1991] is used to perform semantic concept clustering.

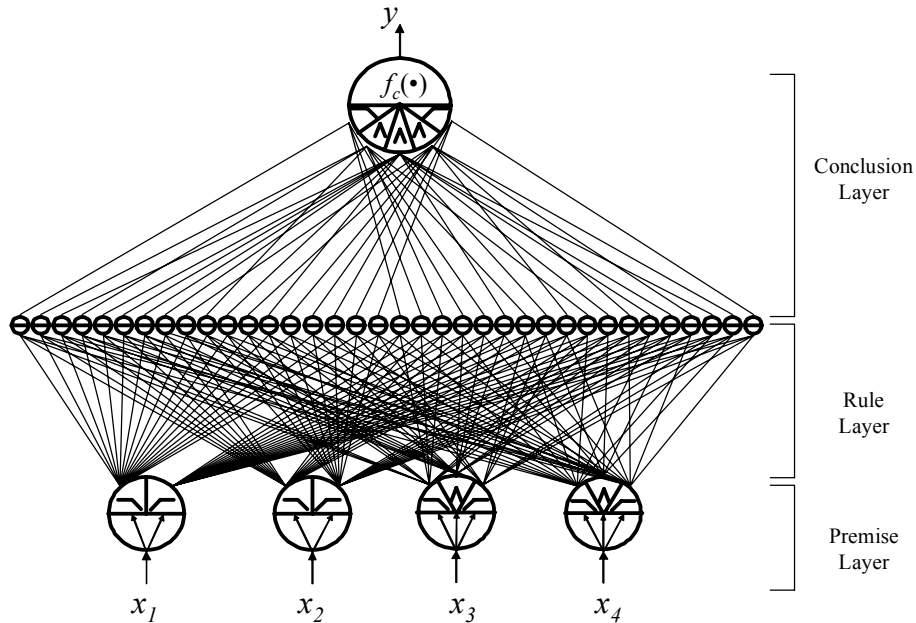


Figure 9. The structure of the parallel fuzzy inference mechanism for semantic concept clustering

Figure 9 shows the structure of the parallel fuzzy inference mechanism. It is a three-layered network which can be constructed by directly mapping from a set of specific fuzzy rules, or can be learned incrementally from a set of training patterns. In our approach,

the rules are defined by the domain expert. The structure consists of a *premise layer*, *rule layer*, and *conclusion layer*. There are two kinds of nodes, *fuzzy linguistic nodes* and *rule nodes*, in this model. A *fuzzy linguistic node* represents a fuzzy variable and manipulates the information related to that linguistic variable. A *rule node* represents a rule and determines the final firing strength of that rule during the inferring process. The *premise layer* performs the first inference step to compute matching degrees. The *conclusion layer* is responsible for drawing conclusions and defuzzification. We will describe each layer in the following.

A. Premise layer:

As shown in Figure 9, the first layer is called the *premise layer* and is used to represent the premise part of the fuzzy system. Each fuzzy variable appearing in the premise part is represented with a condition node. Each of the outputs of the condition node is connected to some nodes in the second layer to constitute a condition specified in some rules. Note that the output links must be emitted from proper linguistic terms as specified in the fuzzy rules. In other words, a linguistic node is a polymorphic object that can be viewed differently by different fuzzy rules. Figure 10 shows the fuzzy linguistic node for the *TA* fuzzy variable.

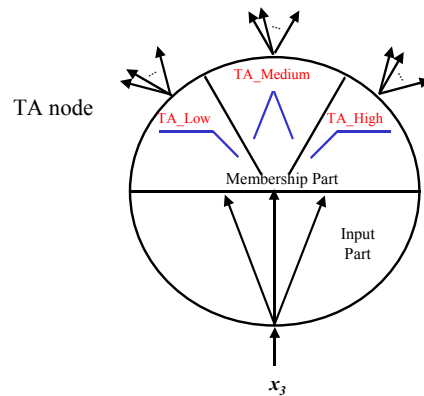


Figure 10. The structure of the fuzzy linguistic node for the *TA* fuzzy variable

The *premise layer* performs the first inference step to compute matching degrees. The input vector is $x = (x_1, x_2, \dots, x_n)$, where x_i is denoted as the input value of the i th linguistic node. Thus, the output vector of the premise layer is $\mu^1 = ((u_{11}^1, u_{21}^1, \dots, u_{N_1 1}^1), (u_{12}^1, u_{22}^1, \dots, u_{N_2 2}^1), \dots, (u_{1n}^1, u_{2n}^1, \dots, u_{N_n n}^1))$, where u_{ij}^1 is the matching degree of the j -th linguistic term in the i -th condition node. In our approach, the triangular function and trapezoidal function are adopted as the membership functions for the linguistic terms. Equation 1 and 2 show the triangular and trapezoidal membership functions, respectively:

$$f_{triangle}(x : a, b, c) = \begin{cases} 0 & x < a \\ (x-a)/(b-a) & a \leq x \leq b \\ (c-x)/(c-b) & b \leq x \leq c \\ 0 & x > c \end{cases} \quad (1)$$

$$f_{trapezoidal}(x : a, b, c) = \begin{cases} 0 & x < a \\ (x-a)/(b-a) & a \leq x \leq b \\ 1 & b \leq x \leq c \\ (d-x)/(d-c) & c \leq x < d \\ 0 & x \geq d \end{cases} \quad (2)$$

$$f_{ij}^1 = \begin{cases} f_{triangular} & j \neq 1 \text{ or } n \\ f_{trapezoidal} & j = 1 \text{ or } n \end{cases} \quad (3)$$

where n is the number of linguistic terms for the i -th linguistic node. Therefore, $\mu_{ij}^1 = f_{ij}^1(x)$.

B. Rule layer:

The second layer is called the *rule layer*. In it, each node is a rule node and is used to represent a fuzzy rule. The links in this layer are used to perform precondition matching of fuzzy logic rules. The output of a rule node in the rule layer is linked to associated linguistic nodes in the third layer. In our model, the rules are previously defined by domain experts. Table 3 shows the fuzzy inference rules for the parallel fuzzy inference mechanism. Figure 11 shows the structure of the rule node.

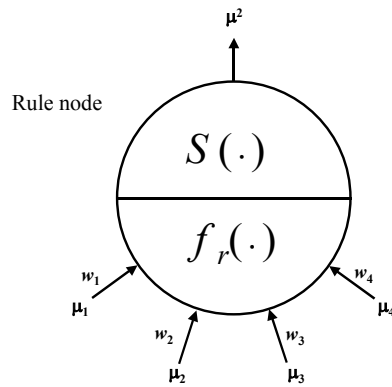


Figure 11. The structure of the rule node

Table 3. The fuzzy inference rule for the parallel fuzzy inference mechanism

Rule	POS	TV	TA	CTA	CRS		Rule	POS	TV	TA	CTA	CRS
1	L	L	L	L	VL		19	H	L	L	L	VL
2	L	L	L	M	VL		20	H	L	L	M	VL
3	L	L	L	H	L		21	H	L	L	H	L
4	L	L	M	L	L		22	H	L	M	L	L
5	L	L	M	M	L		23	H	L	M	M	M
6	L	L	M	H	M		24	H	L	M	H	M
7	L	L	H	L	M		25	H	L	H	L	H
8	L	L	H	M	H		26	H	L	H	M	H
9	L	L	H	H	H		27	H	L	H	H	H
10	L	H	L	L	M		28	H	H	L	L	M
11	L	H	L	M	M		29	H	H	L	M	H
12	L	H	L	H	H		30	H	H	L	H	H
13	L	H	M	L	H		31	H	H	M	L	H
14	L	H	M	M	H		32	H	H	M	M	H
15	L	H	M	H	H		33	H	H	M	H	VH
16	L	H	H	L	H		34	H	H	H	L	VH
17	L	H	H	M	VH		35	H	H	H	M	VH
18	L	H	H	H	VH		36	H	H	H	H	VH

The f_r function in Figure 11 provides the net input for this node and is defined as $f_r = \sum_{i=1}^p w_i \mu_i$.

The S function is used to normalize the f_r function and is defined in Eq. 4:

$$S(x : a, b) = \begin{cases} 0, & x < a \\ 2\left(\frac{x-a}{b-a}\right)^2, & a \leq x \leq \frac{a+b}{2} \\ 1-2\left(\frac{x-b}{b-a}\right)^2, & \frac{a+b}{2} \leq x < b \\ 1, & x \geq b \end{cases} \quad (4)$$

In our case, the rule node has four inputs, and each input value is between 0 and 1.

C. Conclusion layer:

The third layer is called the *conclusion layer*. This layer is also composed of a set of fuzzy linguistic nodes. A fuzzy linguistic node can also operate in a reverse mode, called a conclusion node. In the reverse mode, fuzzy linguistic nodes are responsible for drawing conclusions and defuzzification. Figure 12 shows the structure of a linguistic node in the reverse mode, and shows that it is also an output node.

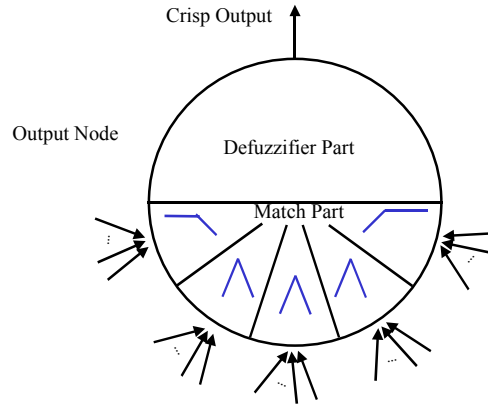


Figure 12. The structure of a fuzzy linguistic node for the conclusion layer

In our model, the final output y is the crisp value that is produced by combining all the inference results with their firing strengths. The defuzzification process is defined in Eq. 5:

$$CrispOutput = \frac{\sum_{i=1}^r \sum_{j=1}^c y_{ij}^k w_{ij}^k V_{ij}}{\sum_{i=1}^r \sum_{j=1}^c y_{ij}^k w_{ij}^k}, \quad (5)$$

where $w^k = \frac{\sum_{i=1}^n \mu_i^1}{n}$, V_{ij} is the center of gravity, r is the number of corresponding rule nodes,

c is the number of linguistic terms of the output node, n is the number of fuzzy variables in the premise layer, and k represents the k -th layer. The values of r , c , n , and k adopted here are 36, 5, 4 and 2, respectively.

4.2 Fuzzy Compatibility Relation Approach to Semantic Concept Clustering

The conceptual resonance of terms pair can be treated as a fuzzy compatibility relation, because it satisfies the properties of reflexivity and symmetricalness. Therefore, the problem of concept clustering is that of finding all the classes of maximal α -compatibles with fuzzy compatibility relations. In this model, the value of α represents a specified membership degree of the fuzzy compatibility relation. The semantic concept clustering algorithm based on the fuzzy compatibility relation approach is described as follows.

Semantic Concept Clustering Algorithm based on the Fuzzy Compatibility Relation Approach
Input:

1. Fuzzy Compatibility Membership Degree α
2. The term set $X = \{Term[1], Term[2], \dots, Term[n]\}$ with n terms for the specific category News, and its corresponding fuzzy conceptual resonance matrix $A = [\alpha_{ij}]_{n \times n}$.

Output:

The *Final_Concept_Set*, which is the set of Domain Ontology Concepts.

Method:

Step 1: For $i \leftarrow 1$ to n

Step 1.1: $Set_i \leftarrow \Phi$ /* Set_i denotes the Term Set regarding $Term[i]$, and all the compatibility membership degrees α_{ij} 's of the terms in Set_i are not less than α */

Step 1.2: $S_i \leftarrow 0$ /* S_i denotes the cardinality of Set_i */

Step 1.3: $Set_i \leftarrow Set_i \cup \{Term[i]\}$

Step 1.4: $Temp_Set \leftarrow \Phi$, /* $Temp_Set$ denotes the set of existing concept subsets */

Step 1.5: For $j \leftarrow i$ to n

Step 1.5.1:

If $\alpha_{ij} \geq \alpha$ Then

Step 1.5.1.1: $Set_i = Set_i \cup \{Term[j]\}$

Step 1.5.1.2: $S_i \leftarrow S_i + 1$

Step 1.6: Determine the power set p_k of Set_i .

Step 1.6.1: $S_{p_k} \leftarrow |p_k|$, where $k = 1, \dots, 2^{S_i}$
/* S_{p_k} Denotes the cardinality of p_k */

Step 1.7: For $k \leftarrow 1$ to 2^{S_i}

Step 1.7.1: If $p_k \in Temp_Set$
Continue

Step 1.7.2: $flag \leftarrow 0$

Step 1.7.3: For $l \leftarrow 1$ to $S_{p_k} - 1$

Step 1.7.3.1: For $m \leftarrow l + 1$ to S_{p_k}

Step 1.7.3.1.1:

$n \leftarrow$ Index of $p_k[l]$ in X

$q \leftarrow$ Index of $p_k[m]$ in X

Step 1.7.3.1.2:

If $\alpha_{nq} < \alpha$

Then $flag \leftarrow 1$ and Break

Step 1.7.3.2: If $flag = 1$

Then Break

Step 1.7.4: If $flag = 0$ Then

Step 1.7.4.1: $Final_Concept_Set \leftarrow Final_Concept_Set \cup p_k$

Step 1.7.4.2: $Temp_Set \leftarrow Temp_Set \cup \{P(p_k) - p_k - \Phi\}$

/* $P(p_k)$ Denotes the power set of p_k */

Step 2: End.

The method used to determine α is very important for semantic concept clustering, because it will influence the number of concepts and the degree of compatibility of Chinese terms. The value of α may vary for different domain documents, because their properties may be different. Now, we use an example to explain the relation of concept clustering for different value of α . In Figure 13, the terms are clustered based on a specific value of α , and they point to the same concept if their CRS values are greater than α .

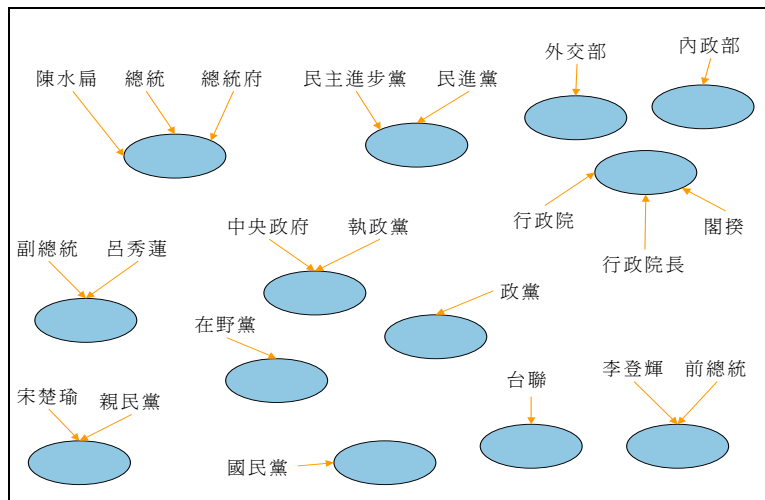


Figure 13. The concepts clustered based on a specific value of α .

If we reduce the value of α , then the terms will be clustered with high compatibility. Figure 14 shows the concepts clustered based on a lower value of α corresponding to Figure 13.

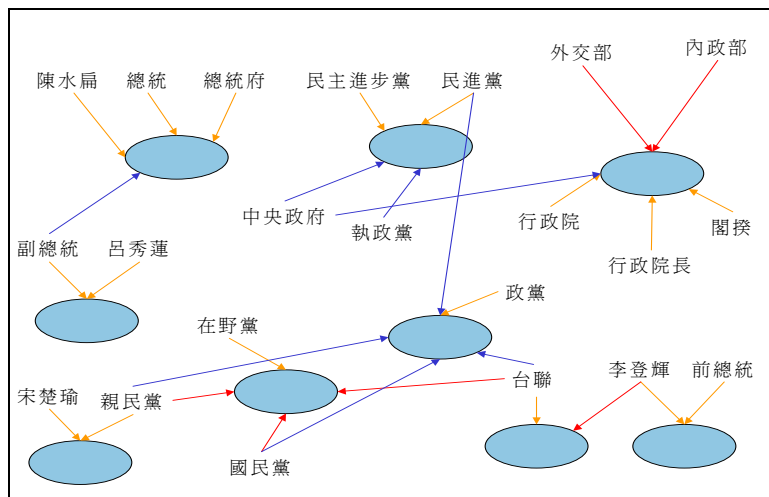


Figure 14. The concepts clustered based on a lower value of α .

The lower α value will result in the formation of the more concepts and strengthen the compatibility degree of terms for a specific concept. The α decision algorithm for semantic concept clustering can be described as shown below. The prune-and-search strategy will be applied to solve this problem.

The α Decision Algorithm for semantic concept clustering

Input:

CRS of terms for a specific news category
The interval $[a, b]$ for the scope of the concept number

Output:

Fuzzy Membership Degree α

Method:

Step 1: Read all the values of conceptual resonance into a double array $R = double[n]$, where n is the number of values.

Step 2: Sort all the elements of R , where $R[i] \leq R[j]$ for $0 \leq i < j \leq n$.

Step 3: $p \leftarrow \frac{n}{2}$

Step 4: $count \leftarrow 1$

Step 5: $\alpha \leftarrow R[p]$, and let the number of classes of maximal α -compatibles be c .

Step 5.1: $count \leftarrow count + 1$

Step 5.2: If $c > b$ Then

Step 5.2.1: $p \leftarrow p + \frac{n}{2^{count}}$

Step 5.2.2: Go to Step 5.

Step 5.3: If $c < a$ Then

Step 5.3.1: $p \leftarrow p - \frac{n}{2^{count}}$

Step 5.3.2: Go to Step 5.

Step 5.4: If $a \leq c \leq b$

Then go to Step 6.

Step 6: End.

5. Experimental Results

In this section, some experiments obtained using the proposed approach will be presented. The news corpus was gathered between May 2001 and March 2002 from the ChinaTimes website (<http://www.chinatimes.com.tw>). Seven categories of news, consisting of “Political” (政治焦點), “International” (國際要聞), “Finance” (股市財經), “Cross-Strait” (兩岸風雲), “Societal” (社會地方), “Entertainment” (運動娛樂) and “Life” (生活新知), were used in the experiments. Table 4 lists the number of documents for each news category, the Chinese terms produced by the refining tagger, the remaining terms produced by the stop word filter, and the filtering percentages for the Chinese terms and remaining terms.

Table 4. The experimental results obtained using the proposed filter

News Category	政治焦點 (Political)	國際要聞 (International)	股市財經 (Finance)	兩岸風雲 (Cross-Strait)	社會地方 (Societal)	運動娛樂 (Entertainment)	生活新知 (Life)
Number of Doc.	11277	13542	22756	6040	13441	5974	9279
Chinese Terms	25448	25484	18960	22856	35846	24178	35932
Remaining Terms	17091	15367	11346	15085	24813	16543	24287
Filter Percent	32.84%	39.70%	40.16%	34.00%	30.78%	31.58%	32.41%

Next, we will analyze the results of *CRS* for any Chinese term pair. Table 5 shows the partial results of *CRS* for the “Political” (政治焦點) category with the highest values. Notice that each term pair not only exhibits strong similarity in term knowledge for the *POS* fuzzy variable and *TV* fuzzy variable but also high strength for the *TA* fuzzy variable and *CTA* fuzzy variable. The term pairs marked with asterisks (*) exhibited strong *TA* and *CTA* but weak *POS* and *TV*.

Table 5. Partial conceptual resonance results for the “Political” category with the highest values

Chinese Term Pair	CRS	Chinese Term Pair	CRS
(民主進步黨,民進黨)	0.595481543	* (國民黨主席,連戰)	0.534303235
(親民黨主席,親民黨)	0.594101448	(國民黨,親民黨)	0.534003082
(國民黨主席,國民黨)	0.587348993	(民進黨籍,民進黨團)	0.5338768
(民進黨主席,民進黨)	0.581987023	(副秘書長,秘書長)	0.531714804
(親民黨主席,國民黨主席)	0.577182427	* (親民黨主席,宋楚瑜)	0.530113977
(行政院長,行政院)	0.571720293	(立委,立法委員)	0.52974897
(民進黨主席,民主進步黨)	0.571574542	(國防部,國防)	0.529134478
(立院,立法院)	0.56774317	(馬英九,台北市長)	0.522650369
(立法院,立法院長)	0.558938037	(親民黨,民進黨)	0.522529795
(民進黨團,民進黨)	0.558824035	(委員會,委員)	0.522146828
(台北市,台北市長)	0.557260479	(縣市長,縣市)	0.520918138
(民進黨籍,民進黨)	0.555527542	(政府,中央政府)	0.518259497
(民進黨主席,國民黨主席)	0.554741061	* (呂秀蓮,副總統)	0.51637767
(高雄市,高雄)	0.553642597	(民主進步黨,民進黨團)	0.516346569
(台聯,台灣團結聯盟)	0.553602148	* (李登輝,前總統)	0.515308468
(國民黨,民進黨)	0.551060166	* (陳水扁,總統)	0.513920884
(國民黨,國民黨籍)	0.547553789	(副院長,立法院長)	0.512207578
(親民黨主席,民進黨主席)	0.54062093	(民主進步黨,國民黨)	0.511493131
(民進黨籍,國民黨籍)	0.53928488	(總統府,總統)	0.507264737
(民進黨主席,黨主席)	0.538797148	* (王金平,立法院長)	0.504385767

(a)

(b)

The next experiment was conducted to obtain semantic concept clustering results under various α values. In this experiment, the number of concepts varied between 500 and 1,000 for each news category. Table 6 shows that the different values of α produced various numbers of concepts containing different terms. The experimental results show that the semantic concept clustering results were influenced by the values of α .

Table 6. Analysis of various α values for each news category

News Category	政治焦點 (Political)	國際要聞 (International)	股市財經 (Finance)	兩岸風雲 (Cross-Strait)	社會地方 (Societal)	運動娛樂 (Entertainment)	生活新知 (Life)
α	0.40	0.40	0.42	0.41	0.40	0.39	0.40
Number of Concepts	971	948	543	791	783	640	880
Number of Average Terms per Concept	3.45	3.56	3.13	3.39	3.08	3.65	3.64

Table 7 shows the concept clustering results under various values of α for the “Life” (生活新知) category. Table 8 shows a partial listing of the concepts, including concept names, attributes and operations, in the golden standard ontology of the “Life” (生活新知) category.

Table 7. Analysis of various α values for the “Life” (生活新知) category

$\alpha = 0.44$	Concept No. 1	教育 教師 教授 教育部長
	Concept No. 2	學校 學生 學術 大學 台灣大學
	Concept No. 3	學者 學術 大學 教授
$\alpha = 0.42$	Concept No. 1	教育 教師 教授 教育部長 學生 學校 教育部
	Concept No. 2	學校 學生 學術 大學 台灣大學 校長
	Concept No. 3	學者 學術 大學 教授 研究
	Concept No. 4	學校 家長 老師 學生
	Concept No. 5	科學 學術 大學
	Concept No. 6	學者 科學家 專家
$\alpha = 0.40$	Concept No. 1	教育 教師 教授 教育部長 學生 學校 教育部 大學 課程 資源
	Concept No. 2	學校 學生 學術 大學 台灣大學 校長 院長 教授
	Concept No. 3	學者 學術 大學 教授 研究 成果 領域
	Concept No. 4	學校 家長 老師 學生 教育部長 教育部 高中 大學
	Concept No. 5	科學 學術 大學 研究所 研究
	Concept No. 6	學者 專家 科學家 科學
	Concept No. 7	學者 科學 學生 生物 領域 學術 大學
	Concept No. 8	成果 研究 科學 領域 學術
	Concept No. 9	技術 研究 領域 應用 產業

Table 8. An example of the gold-standard concepts for the “Life” (生活新知) category

Concept name	Attribute	Operation
教育、教育部	教育部長：String 教師、教授、教師：String 學生：String	Null
學校、校園、校方	小學：String 中學：String=國中、高中 大學、學院：String=台灣大學 研究所：String 課程：String=考試、暑假、寒假	Null
學術、研究	學者：String=教授、科學家、專家 領域：String=生物、醫學、電腦科學 研討會：String	發展、研發、研究、實驗
科技、科學	領域：String=電機、電子、資訊、 通訊、半導體、光電、網路 產品：String=手機、硬體、軟體、 電腦、處理器、液晶螢幕	Null
醫院、醫界	醫師、醫生：String	手術、檢驗
衛生局、衛生署	健保：String=全民健保、健保卡	Null
氣象、天氣、氣候	雨量、雨勢、雨：String=大雨、豪雨、 雷雨、陣雨、雷陣雨 鋒面：String 氣溫：String 季風：String=東北季風	變化

Notice that the concepts with higher values of α are subsets of the concepts with lower values of α . That is, a lower value of α generated a concept with more Chinese terms. In the final experiment, we tested the performance measures Precision and Recall. We choose four students who were currently working toward the M.S. degree in Computer Science and Information Management, and let them to evaluate the values obtained using Eq.(6) and Eq.(7) for precision and recall. Figure 15-20 show the average precision and recall curves based on the evaluations performed by the four experts. Table 8 shows an example of the gold-standard concepts for the “Life” (生活新知) category. The Precision and Recall measure formulas used in this study are as follows:

$$Precision = \frac{\text{The number of relevant common terms in gold-standard concept and the automatically generated semantic concept}}{\text{The number of terms in the automatically generated semantic concept}}, \quad (6)$$

$$Recall = \frac{\text{The number of relevant common terms in gold - standard concept and the automatically generated semantic concept}}{\text{The number of terms in the gold - standard concept}}. \quad (7)$$

Figure 15-17 show the average precision results obtained based on the evaluations performed by the four domain experts for various α values. Figure 18-20 show the average recall results obtained based on the evaluations performed by the four domain experts for various α values.

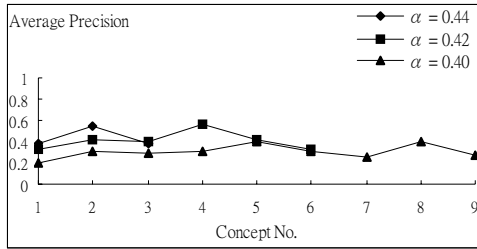


Figure 15. The average precision results for different α values (Concept name)

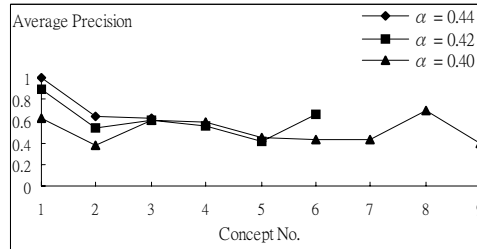


Figure 16. The average precision results for different α values (Concept name + Attribute)

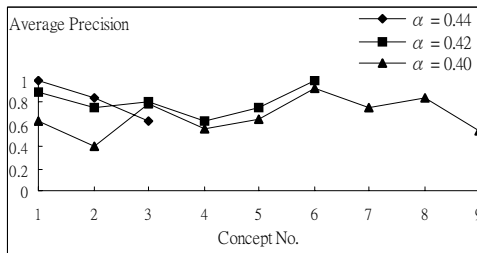


Figure 17. The average precision results for different α values (Concept name + Attribute + Attribute value)

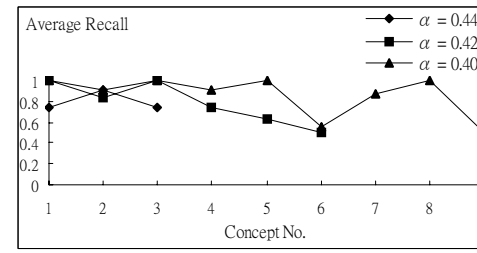


Figure 18. The average recall results for different α values (Concept name)

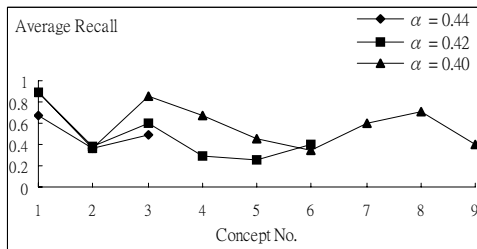


Figure 19. The average recall results for different α values (Concept name + Attribute)

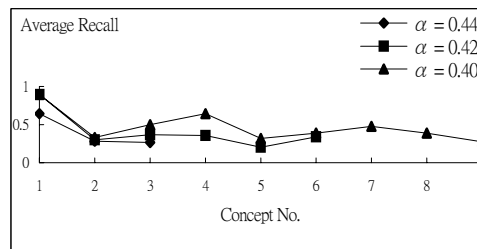


Figure 20. The average recall results for different α values (Concept name + Attribute + Attribute value)

6. Conclusions

This paper has presented a Chinese term clustering mechanism for generating semantic concepts of a news ontology. We utilize the parallel fuzzy inference mechanism to infer the conceptual resonance strength of any two Chinese terms. In addition, the *CKIP* tool is used in Chinese natural language processing, including part-of-speech tagging, Chinese term analysis, and Chinese term feature selection. A fuzzy compatibility relation approach to semantic concept clustering has also been proposed. Simulation results show that our approach can effectively cluster Chinese terms to generate the semantic concepts of a news ontology. In the future, we will extend use our approach to help construct a domain ontology more efficiently. Moreover, we will adopt the genetic learning mechanism to learn the membership functions of fuzzy inference rules for the parallel fuzzy inference mechanism. Finally, mixed Chinese/English documents will also be employed to construct more a complex domain ontology.

Acknowledge

The authors would like to express their gratitude to the anonymous reviewers for their comments, which improved the quality of this paper. This work was partially supported by the Ministry of Economic Affairs in Taiwan under grant 93-EC-17-A-02-S1-029 and partially sponsored by the National Science Council of Taiwan (R. O. C.) under grant NSC-93-2213-E-309-003.

Reference

- Alani, H., S. Kim, D. E. Millard, M. J. Weal, W. Hall, P. H. Lewis and N. R. Shadbolt, "Automatic Ontology-Based Knowledge Extraction from Web Documents," *IEEE Intelligent Systems*, 18 (1) 2003, pp. 14-21.
- CKIP, "Academia Sinica Balanced Corpus," Technical Report, No. 95-02/98-04, Academia Sinica, Taiwan, 1998.
- CKIP, "Chinese Electronic Dictionary," Technical Report, No. 93-05, Academia Sinica, Taiwan, 1993.
- Embley, D. W., D. M. Campbell, R. D. Smith and S. W. Liddles, "Ontology-based extraction and structuring of information from data-rich unstructured documents," *Proceeding Of ACM Conference on Information and Knowledge Management*, USA, 1998, pp. 52-59.
- Fensel, D., "Ontology-based Knowledge Management," *IEEE Computer*, 35 (11) 2002, pp. 56-59.
- Gao, J., J. T. Goodman and J. Miao, "The Use of Clustering Techniques for Language Modeling – Application to Asian Language," *Computational Linguistics and Chinese Language Processing*, 6 (1) 2001, pp. 27-60.

- Gomez-Perez, A and O. Corcho, "Ontology languages for the semantic web," *IEEE Intelligent Systems*, 17 (1), 2002, pp. 54-60.
- Guarino, N., C. Masolo and G. Vetere, "OntoSeek: Content-based access to the web," *IEEE Intelligent Systems*, 14 (3) 1999, pp. 70-80.
- Jacobes, P. S., "Using Statistical Methods to Improve Knowledge-Based News Categorization," *IEEE Expert*, 8 (2) 1993, pp. 13-23.
- Kuo, Y. H., J. P. Hsu and C. W. Wang, "A Parallel Fuzzy Inference Model with Distributed Prediction Scheme for Reinforcement Learning," *IEEE Trans. on Systems, Man, and Cybernetics*, 28 (2) 1998, pp. 160-172.
- Lammari, N. and E. Metais, "Building and maintaining ontologies: a set of algorithm," *Data & Knowledge Engineering*, 48 (2) 2004, pp. 155-176.
- Lee, C. S., Y. J. Chen and Z. W. Jian, "Ontology-based Fuzzy Event Extraction Agent for Chinese e-news Summarization," *Expert Systems with Applications*, 25 (3) 2003, pp. 431-447
- Lee, C. S., S. M. Guo and Z. W. Jian, "Weighted Fuzzy Ontology for Chinese e-News Summarization," *2004 IEEE International Conference on Fuzzy Systems*, USA, 2004.
- Lee, R. C. T., R. C. Chang, S. S. Tseng and Y. T. Tsai, "Introduction to the Design and Analysis of Algorithms," Unalis co., Taipei, 1999.
- Lin, C. T. and C. S. G. Lee, "Neural-Network-Based Fuzzy Logic Control and Decision System," *IEEE Trans. Computers*, 40 (12) 1991, pp. 1320-1336.
- Missikoff, M., R. Navigli and P. Velardi, "Integrated approach to web ontology learning and engineering," *IEEE Computer*, 35 (11) 2002, pp. 60-63.
- Navigli, R. and P. Velardi, "Ontology learning and its application to automated terminology translation," *IEEE Intelligent Systems*, 18 (1) 2003, pp. 22-31.
- Schreiber, A.T., B. Dubbeldam, J. Wielemaker and B. Wielinga, "Ontology-based photo annotation," *IEEE Intelligent Systems*, 16 (3) 2001, pp. 66-74.
- Soo, V. W. and C. Y. Lin, "Ontology-based information retrieval in a multi-agent system for digital library," *Proceeding Of the sixth conference on artificial intelligence and applications*, Taiwan, 2001, pp. 241-246.
- Studer, R., R. Benjamins and D. Fensel, "Knowledge engineering: principles and methods," *Data and Knowledge Engineering*, 25 (1) 1998, pp. 161-197.
- Van Der Vet, P.E. and N.J.I. Mars, "Bottom-Up Construction Ontologies," *IEEE Trans. on Knowledge and data Engineering*, 10 (4) 1998, pp. 513-526.
- Yang, Y. J., S. C. Lin, L. F. Chien, K. J. Chen and L. S. Lee, "An intelligent and efficient word-class-based Chinese language model for Mandarin speech recognition with very large vocabulary," *Proceeding of ICSLP-94*, Yokohama, Japan, 1994, pp. 1371-1374.