# SPOKEN CORPORA AND ANALYSIS OF NATURAL SPEECH[*]

Shu-Chuan Tseng

**ABSRACT**

This paper introduces spoken corpora of Taiwan Mandarin created at Academia Sinica and gives an overview of some recent studies carried out utilizing the spoken data. Spoken language resources of Taiwan Mandarin have been collected and processed at Academia Sinica since 2001. As a result, spoken data, which are useful not only for language archives purpose, but also for linguistic studies, has been made available. In addition to creation of the corpus, two lines of research are discussed in which theoretical and empirical studies are connected by using the aforementioned language resources: 1) language variation and change and 2) spoken discourse analysis. Phonetic reduction is one of the main reasons for changes within a language and it is important to take into account different levels of variations in spontaneous speech. For this purpose, we studied syllable contraction/merger, vowel reduction, and phonetic reduction in directional complements. Discourse items also play an essential part, because they add specific implications to sentences and their use is mainly marked by prosodic means. We segmented a spoken discourse into smaller prosodic units to allow for a more precise study of discourse items, prosodic features, and disfluency. These issues are correlated with each other, especially through prosodic markings.

*Shu-Chuan Tseng*

## 1. CORPUS CREATION/LANGUAGE ARCHIVES OF SPOKEN TAIWAN MANDARIN

The importance of spoken language resources has been increasingly noticed in recent research works on speech technology and linguistic studies. For some languages, spoken corpora of a considerably large data size have been made available and subsequent research applications have already shown how useful spoken corpora can be in spite of the time- and labor-consuming collection and annotation processes, e.g., the Map Task Corpus, the SWITCHBOARD Corpus, and the Corpus of Spontaneous Japanese (Anderson *et al.* 1991, Godfrey *et al.* 1992, Maekawa 2004). In this paper, three corpora of Mandarin conversations which have been collected and processed at the Institute of Linguistics in Academia Sinica will be introduced. These corpora vary in the recruitment of subjects, the setting of scenario, and the length of the conversation. Some of them were collected for language-archive purposes. But they were all processed in the same way, so we can use them as structured language resources for linguistic analysis. A brief summary is given in **Table 1**. Details of the corpus description and annotation principles can be found at the website http://mmc.sinica.edu.tw; or refer to Tseng (2004).

**Table 1**: Overview of spontaneous Taiwan Mandarin speech corpora[1]

| Corpus | Mandarin Conversational Dialogue Corpus (MCDC) | Mandarin Topic-Oriented Conversation Corpus (MTCC) | Mandarin Map Task Corpus (MMTC) |
|---|---|---|---|
| Scenario | Free conversation between strangers | Subjects knew each other well | Subjects knew each other well |
| Period | 2001.03 – 2001.07 | 2002.01 - 2002.03 | 2002.01 - 2002.03 |
| Data | 8 conversations, 8 hrs. 4.8 GB (90K words) Orthographically transcribed/annotated (disfluency) | 29 conversations, 9.7 hrs. 6 GB (123K words) Orthographically transcribed/annotated (dialogue acts) | 26 dialogues, 4.3 hrs. 2.8 GB (44K words) Orthographically transcribed/annotated (phonetic variations) |
| Funded by | Academia Sinica | National Digital Archives Project | National Science Council |

**1.1 Free Conversation between Strangers (2001)**

The Mandarin Conversational Dialogue Corpus (MCDC) was recorded in 2001. Sixty subjects were recruited from a random sampling of Taipei City citizens. The recorded data are natural conversations between two strangers. The conversation partners had to introduce themselves at the beginning of the conversation and were then free to talk about anything they wanted. In total, 25.6 hours of data were recorded; each conversation is about 50 minutes long. Eight conversations are orthographically transcribed (136K characters, equivalent to 90K words) and a number of spontaneous speech phenomena have been annotated, e.g., disfluency and pronunciation variations. The MCDC data is publicly distributed via the Association for Computational Linguistics and Chinese Language Processing (ACLCLP, website http://www.aclclp.org.tw).

---

[1] Only eight conversations of the MCDC have been processed due to a lack of funding. Each conversation is about one hour long and, thus a huge of amount of manpower would be required to process the remaining 22 conversations. Moreover, the reason for the one missing conversation in the MTCC and four missing conversations in the MMTC is that Taiwan Southern Min was spoken most of the time in these conversations and we did not process these data because our aim was to collect Taiwan Mandarin spoken data.

**1.2 Topic-oriented Conversation between Familiar Persons (2002)**

The Mandarin Topic-Oriented Conversation Corpus (MTCC) was recorded in 2002, as a follow-up project to the MCDC. Thirty subjects from the MCDC were recruited again, but this time they were asked to bring a person with whom they were familiar with them. The subjects were requested to talk about one specific topic or event which had happened in the year of 2001. They were asked to talk only about that particular topic or event. Eleven hours of data were recorded in total; each conversation is about 20 minutes long. The transcription convention of the MTCC follows that of the MCDC with regard to the transcription of the discourse items, paralinguistic phenomena, and word fragments. An annotation scheme was designed to mark up the associated dialogue acts in conversation. Twenty-nine conversations are orthographically transcribed (185K characters, equivalent to 123K words).

**1.3 Task-oriented Conversation between Familiar Persons (2002)**

The Map Task project recruited the same subject group as that of the MTCC. The subjects had to complete a task in a Map Task scenario (Anderson *et al.* 1991). In each conversation, the MCDC subject had a detailed map with three destinations marked on the map, whereas his or her partner had a simplified map. The person with the detailed map had to give oral instructions to his or her partner, who had the simplified map, in order to find the pre-given locations on the map. The one who gives instructions has to talk more than the one follows the instructions. The task was arranged in this way because we wanted to collect more data from the same subjects who already participated in the previous project. In total, five hours of data were recorded in the Mandarin Map Task Corpus (MMTC); each conversation is about ten minutes long. Twenty-six conversations are orthographically transcribed (66K characters, equivalent to 44K words). Particular pronunciation such as lengthening, assimilation, contraction, and nasalization are annotated. Part of the data is phonetically labeled (approximately 12,000 segments).

**1.4 Read Speech (2002)**

The same 60 subjects from the MTCC and the MMTC were asked to read aloud newspaper articles which had been shortened to about

one-page in length. The subjects were told that they could practice reading the text as many times as they wanted before the recording.

**1.5 Questionnaire-based Street Interviews (2007-2012)**

In 2007, we started running a six-year Language Archives Project. Questionnaire-based street interviews will be collected, especially in the counties of middle and southern Taiwan within this project. Information on personal, educational, and language background of the interviewees will also be collected in conjunction with the linguistic data collection. Some of the questions are asked to collect targeted items of particular phonetic structures. All interviews will be transcribed and segmented into prosodic units.

**2. LANGUAGE VARIATION AND CHANGE**

In principle, diachronic changes in a language are results of synchronic variations, especially in spoken language use. After the changes are completed[2], traces may be left which mark the differences existing before and after the changes. For instance, sociophonetics puts together studies of sociolinguistics, language change and variation, and phonetics to identify the phonetic traces left in the contemporary language use (Labov 2006). More specifically, we know that different kinds of phonetic reduction are found in spontaneous speech. They may be the results of language-intrinsic conditions due to phonological environment and constraints in the sound system of a particular language. To take *wo3+men5* (我們, we) as an example, the syllabic nasal m̩ is an extreme reduced form. One reason for this phonetic reduction is that there are no homophones for *wo3* in Chinese, so an extreme reduction to the syllabic nasal will not hinder the understanding of the listener. On the other hand, phonetic reduction can also be language-independent due to physiological reasons, for instance assimilations such as palatalization.

In Mandarin Chinese, the extreme form of phonetic reduction, namely syllable merger, may be one of the most essential reasons for coalescence compounds which consequently influence the writing

---

[2] "The change is completed" here means that there is a change in the sound system or in the writing system.

system (Lung 1979), for instance *zhi1+hu1 -> zhu1* (之＋乎->諸)[3]. Not-so-extreme phonetic reductions such as the reduced form of *wo3+yao4 ->* wɔ♯a (我要, I want) can also be investigated, if we have appropriate spoken language resources (Tseng 2005a). However, this type of study of phonetic reduction is not found in the literature, because natural speech data was not easy to analyze until the power and the compatibility of the related software and hardware was substantially improved. Our initial work on phonetic reduction based on spontaneous speech data shows that vowel shifts may indeed occur in natural speech of Taiwan Mandarin (Section 2.2). The issue of grammaticalization is also related to issues of language change (Section 2.3). We took directional complements as our first object of study and investigated how different types of directional complements are related to the tendency of syllable merger. We will discuss these works in this section; and it is hoped that the results will demonstrate how important and useful the spoken language resources are for linguistic studies.

For notations, Pinyin is used to transcribe Chinese characters and the lexical tones used in Taiwan Mandarin are represented by 1, 2, 3, 4, and 5 for the high level tone, the rising tone, the contour tone, the falling tone, and the neutral tone, respectively.

**2.1 Syllable Contraction**

When two syllables are merged, maybe into one or maybe partly into one, it is called syllable contraction. For the discussion of syllable contraction in Chinese, there are two different implications. In the history of Chinese phonology, syllable contraction is associated with coalescence compounds (which may or may not be fixed in the writing system), and the target form is mostly phonologically predictable (Cheng 1985, Chung 1997). Another type of syllable contraction refers to a kind of extreme phonetic reduction in natural speech. Phonetic reduction across syllable boundaries can occur in various forms, in other words, reduction to different degrees. In our study, we are mainly concerned with phonetic reductions. Examples of phonetic reduction by syllable contraction are by definition, when the syllable boundary between two syllables is omitted, or the two nuclei are merged into one, or the syllable

---

[3] 之,乎, and 諸 are particles in Mandarin.

structure of the original two syllables dramatically changes. A typical process a bi-syllabic contraction may undergo is illustrated in (1).

(1) $s_1$        +      $s_2$          =>        s

    $\mathbf{C_1}V_1X_1$        $C_2V_2\mathbf{X_2}$             $\mathbf{CVX}$

C represents the onset consonants, V the nuclei including glides, and X the coda consonants. When a syllable $C_1V_1X_1$ is contracted with another syllable $C_2V_2X_2$, the resulting syllable CVX will normally undergo certain of the following processes:

(2) a. deletion of $C_2$,
     b. retention of $C_1$ and $X_2$ (i.e., $C_1 = C$, $X_2 = X$),
     c. spreading of $X_1$ (nasals [n] and [N]) in some cases,
     d. $V_1/V_2$ reduction, and
     e. reduction of $C_1$.

The above process of syllable contraction is also called the Edge-in Theory, where the first onset and the second coda are retained and the original nuclei changes to a new nucleus (Chung 1997, Hsu 2003). The Edge-in Theory may explain the phonological process which may be undergone in the merger of a syllable, but this theory needs certain modifications, as the forms produced in spontaneous speech do not always take those predicted by the Edge-in Theory. For example, there are cases where the speaker does not prefer a deletion of the second onset consonant, but keeps it instead and deletes the second nucleus, e.g. ni#m in **Table 2**, where /m/ can be pronounced as a syllabic nasal as in Southern Min, or alternatively, only a single consonant may be left, the target consonant may be in its original form or in a derived form. The function of the coda is to signify the existence of the second syllable, e.g., ni#z and ni#s. The crucial issue is that the listener should be able to recognize both syllables from the target form, no matter whether through the nucleus or the consonant.

**Table 2**: Merged pronouns (a revised version of Tseng 2005a: 247)

| | *ye3* | *yao4* | *jiu4* | *you3* | *yi1* | *hai2* | *hui4* | *men5* | *de5* | *shi4* | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|
| # of *ni3* | 12 ni#ɛ nɪ#ɤ | 20 ni#ao ni#a | 4 ni#o ni#z | 11 ni#u nɪ#ɔ nɪ#ɤ | 10 ni: ni | 6 ni#ʔai ni#ai | 23 ni#wei ni#h | 43 ni#m̩ ni#m n̩#m | 12 n#də nɪ#ɔ | 9 ni#z ni#s nɪ#ɤ | 150 (71% of *ni3*-pairs) |
| # of *wo3* | 65 w#eɤ w#e w#ɤ | 16 wɔ#a wɔ#ɤ w#ao | 24 wɔ#ʒɤ, o#ʒɤ wɔ#ɤ, ɤ#ʒ wɔ#z, o#ɤ | 17 wɔ#ɪu wɔ#ɤ w#ɤ | 14 wɔ#ɪ w#e ɔ#ɪ | 8 wɔ#aɪ w#ai | 20 wɔ#e wɔ#ɤ w#ɤ | 297 wɔ#m̩ wɔ#m o#m m | 15 wɔ#ɤ w#ɤ | 15 wɤ#sɤ wɔ#z wɤ#z | 491 (83% of *wo3*-pairs) |
| # of *ta1* | 23 tʰa#ɪe tʰa#ɪɤ tʰa#i | 4 tʰɤ#ɪao tʰ#iao | 26 tʰa#ʒɤ, tʰa#ɪu, xa#ʒɤ, ta#o tʰa#ʒ, tʰa#ɤ | 16 tʰa#ɪu tʰa#ɪɔ tʰa#ɪɤ tʰa#i | 6 tʰa#i xa#i ʔa#i | 10 tʰa:#i tʰa#i tʰa#o | 24 tʰa#ue tʰa#uɪ tʰa#u | 144 tʰa#m̩ tʰa#m tʰ#m̩ | 26 tʰa#ʔe tʰa#ʔɤ tʰa#ɤ | 10 tʰa#lu tʰa#z tʰa#s | 289 (82% of *ta1*-pairs) |
| Σ | 100 | 40 | 54 | 44 | 30 | 24 | 67 | 484 | 53 | 34 | 930(81%) |

**Table 2** lists some of the possibilities in the syllable contraction of word pairs of personal pronouns and also of a number of high-frequency words consisting of only one syllable, extracted from the MCDC (**Appendix A**). The pronouns are *ni3* (you), *wo3* (I), and ta1 (she/he/it). Monosyllabic functions are *ye3* (too), *yao4* (want), *jiu4* (then), *you3* (have), *yi1* (one), *hai2* (still), *hui4* (will), *men5* (plural suffix), *de5* (structural particle), and *shi4* (to be). If the target contraction form is a legitimate syllable satisfying Mandarin phonotactics, it may be possible that it may lead to a change in the writing system. The target form can be a direct derivation from the Edge-in Theory, such as ***zhi1***+***hu1*** -> ***zhu1*** 之+乎->諸, or it can be an adjusted form (to conform to the phonotactics) from the predicted form such as ***zhe4***+***yang4*** -> *jia****ng****4* 這+樣->醬 or 降. The predicted form *zhiang4* does not exist in Mandarin, so a "near-homophone" *jiang4* (sauce or descend) is adopted[4]. This spoken form is widely used in spoken language as well as in web-related text

---

[4] The choice of *jiang4* to represent *zhiang4* is, in the opinion of the author, due to their similarity in pronunciation. *Zhiang4* does not need to undergo any phonological process such as palatalization.

communication. The new contraction syllable representing the semantic meaning of the original syllables will be orthographically written either by borrowing an already existing character (*zhe4+yang4 -> jiang4* 這＋樣->醬or 降) or by a newly invented character (*bu2+yong4->beng2* 不＋用->甭)[5]. In Taiwan Mandarin, the young generation likes to use the contraction form *biang4* for the word *different*, originating from *bu4+yi2+yang4* (不一樣). Because *biang* is not a legitimate syllable in Mandarin Chinese, it is unlikely that any character, either borrowed or newly invented, will be used to represent the contracted form in the current standard form of Mandarin Chinese. In another study we also found that geminate vowels and vowel pairs with a [±back] contrast are more likely to merge (Tseng 2005b).

**2.2 Phonetic Reduction**

There are different kinds of reductions in spontaneous speech. Syllable contraction is the most radical form. We took data produced by one female speaker[6] from the MCDC and MMTC to analyze the phonetic reduction. We asked the same female speaker to read a list of isolated words and sentences in order to obtain her read speech data to compare with the natural speech data. The formants F1 and F2 were measured by PRAAT (Boersma and Weenink 2006). The vowel chart illustrated in **Figure 1** is drawn by using the isolated speech data.[7] From the vowel chart, we noticed that the vowel /a/ is realized more like a middle low vowel. We also noticed that the allophonic variations of /a/; /a/ followed by a dental nasal (a-(n) in **Figure 1**) are spatially clearly distinct from /a/ followed by a velar nasal (a-(N) in **Figure 1**). This variation is not observed in the case of the front vowel /i/. /i/ followed by the dental and the velar nasal coda are not as clearly distinguished as in the case of /a/. In natural speech of Taiwan Mandarin, we often observe that /in/ and /iN/ are not clearly distinguished. However, whether such phenomenon indicates the occurrence of a language-specific development which has produced vagueness in the distinction between

---

[5] 這樣 means "this way", whereas 醬 means "sauce" and 降 means "to lower". When borrowing an already existing character to represent the meaning of the original syllables, the choice of the character usually avoids any semantic or syntactic ambiguities between its original use and the new use for contraction. 不用 and 甭 both mean "need not to".
[6] The speaker is S-01 in **Appendix A**.
[7] For the symbol notations please refer to our website http://mmc.sinica.edu.tw/sampa.htm.

/in/ and /iN/ or it only reflects a general fact that the variability range of formants of /i/ is intrinsically smaller than that of /a/, needs closer investigation. **Figure 1** also illustrates that the diphthong /ou/ and the vowel /o/ are pronounced similarly, too. It may be related to the influence of Southern Min, because Southern Min does not contain /ou/. We are currently processing data of more speakers, and will study the dialectal influence in more detail.
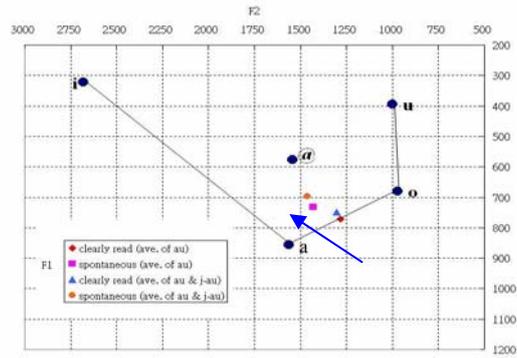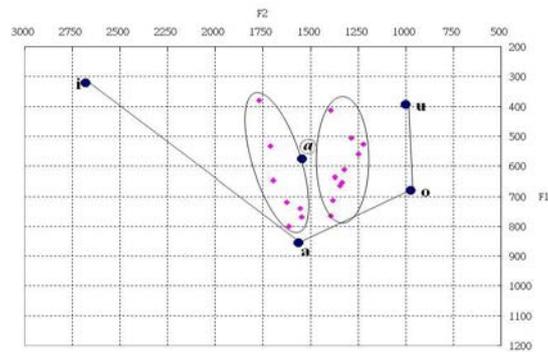
**Figure 1**: Vowel chart – reading words in isolation
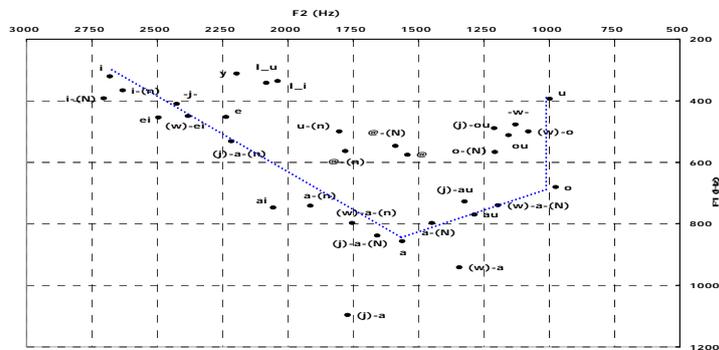


**Figure 2** is the vowel chart drawn by extracting all of the vowels produced in natural speech by the same female speaker as in **Figure 1**. When we compare results for **Figures 1** and **2,** two main characteristics of the speech reduction of vowels, namely vowel space shrinkage and centralization of the vowels are clearly shown (Harrington and Cassidy 1999). Vowel reduction may be due to rate of speech, context assimilation or gesture overlap while articulating the sounds. Using spoken corpora containing different speaking styles produced by different speakers, we can extract reliable empirical evidence to study the phenomenon of speech reduction. Moreover, in **Figure 2**, it is shown that the distance between /in/-/iN/ and /o/-/ou/ is even closer in natural speech than in reading isolated words. It would be interesting to compare spoken data of Taiwan Mandarin and Beijing Mandarin, where the dialectal influences are different, to see if it is a general reduction phenomenon or if it is related to a specific language environment.

**Figure 2:** Vowel chart – natural speech



To take the diphthong /au/ as an example, **Figure 3a** clearly shows a kind of centralization[8]. As we compare the clearly read items and the spontaneous items found in the MMTC, /au/ moves towards the centre of the vowel chart no matter whether it is with or without a front on-glide. Furthermore, on examining the contraction occurrences containing /au/, we found that the nuclei preceding /au/ can be grouped into two main categories by the [±back] feature, as shown in **Figure 3b.** This supports the point made in Section 2.1 that nuclei with the [±back] contrast are more likely to merge. This result also suggests that the sonority hierarchy may not be the only principle determining the final form of the nucleus of a syllable merger. The phonological environment of the original syllables may also play a role.

---

[8] The author would like to thank Tzulun Lee for drawing the charts in **Figures 3a** and **3b**.

**Figure 3a**: Reduction of /au/



**Figure 3b**: Preceding nuclei merged with /au/



**2.3 Grammaticalization of Directional Complements**

The grammatical structure, the prosodic phrasing and the stress placement of directional complement constructions have been thoroughly investigated in the literature of Chinese linguistics, for instance Liu (1998), Lamarre (2004), and Fan (1963). The discussion has reached a consensus that directional complements stem from the verbs *lai2* (come) and *qu4* (go). As grammaticalization continues, three different types of directional complements are differentiated: (I) verb + deictic directional, (II) verb + path directional, and (III) verb + path

directional + deictic directional (**Table 3**)[9]. Type **a** is the base form. Type **b** uses both locative and object NPs. And Type **c** is the metaphorical use, which is not considered in the present study.

**Table 3**: Type of Directional Complements in Taiwan Mandarin

| | Type **I**: Verb + Deictic Directional | | | |
|---|---|---|---|---|
| | Path verb | Co-event verb | Other verb | Deictic directional |
| **Ia** | 上 *shang4* (move upward) | | | 來 *lai2* |
| **Ib** | | 跑 *pao3* (run) | | 去 *qu4* |
| **Ic** | | | 醒 *xing3* (awake) | 來 *lai2* |
| | Type **II:** Verb + Path Directional | | | |
| | Verb | Path directional | Locative NP | |
| **IIa** | 衝 *chong1* (rush) | 進 *jin4* (inside) | | |
| **IIb** | 衝 *chong1* | 進 *jin4* | 學校 *xue2xiao4* (school) | |
| **IIc** | 裝 *zhuang1* (fill) | 出 *chu1* (outside) | (pretend) | |
| | Type **III:** Verb + Path Directional + Deictic Directional | | | |
| | Verb | Path directional | Locative/Object NP | Deictic directional |
| **IIIa** | 衝 *chong1* | 進 *jin4* | | 去 *qu4* |
| **IIIb** | 衝 *chong1* | 進 *jin4* | 學校 *xue2xiao4* | 去 *qu4* |
| **IIIc** | 問 *wen4* (ask) | 回 *hui2* (back) | | 去 *qu4* |

As a kind of phonetic absence can sometimes signal the prominence of its syntactic function (Ansaldo and Lim 2004), we investigated directional complements using our contraction data in the MCDC. Details of the MCDC are listed in **Appendix A.** Contractions in our data show three main characteristics: (1) the second original syllable in a contraction is usually semantically less essential, (2) the second original syllable in a contraction is seldom a proper noun or a full verb, and (3)

---

[9] **Table 3** lists only a few examples from the data. **Type I** takes 來 *lai* or 去 *qu* as the directional complement. **Type II** takes 上 *shang4*, 下 *xia4*, 進 *jin4*, 出 *chu1*, 起 *qi3*, 回 *hui2*, 過 *guo4*, 開 *kai1* or 到 *dao4* as the directional complement. **Type III** combines 來 *lai2* or 去 *qu4* with 上 *shang4*, 下 *xia4*, 進 *jin4*, 出 *chu1*, 起 *qi3*, 回 *hui2*, 過 *guo4*, 開 *kai1* or 到 *dao4*.

the second original syllable in a contraction is usually phonetically more reduced. The contraction pairs usually have a strong-weak stress pattern. We then used the contraction data to observe the extent to which the deictic directional appears phonetically reduced in contractions. The result shows that in Type **Ia**, the deictic directional occurrences in the contraction-strong position make up about 51% of the overall **Type Ia** occurrences, whereas in Type **Ib** it is 65%. This indicates that there is a prosodic difference in the production of the deictic directional according to whether it is used with a path verb or with a co-event verb. In Type **III**, only 38% of the deictic directional occurrences are realized as contraction-strong syllables. If we consider the deictic directional in Type **III** to be more grammaticalized than that in Type **I**, this result supports the notion that the clearer the syntactic function is, the more phonetically reduced the item is. With regard to the semantic content and the syntactic roles, the deictic directional in Type **III** loses part of its original meaning by sharing it with the path directional. But in Type **I** the deictic directional still preserves its function conveying the signified directions to a large degree. So, it is not surprising that the deictic directional appears more reduced phonetically in Type **III.** However, these are only indirect and primitive observations obtained by making use of our contraction data. In order to study the phonetic reduction and prosodic boundaries in directional complements more closely, we are currently processing a database of directionals by labeling the boundaries of prosodic units and syllables. With the use of these concrete acoustic measurements, the issue of grammaticalization can then be examined in more detail.

### 3. SPOKEN DISCOURSE ANALYSIS

Spontaneous speech not only provides empirical data for analysis of segments, syllables or words as mentioned in the previous section, but also provides pieces of data for spoken discourse analysis. In spoken discourse, linguistic representations at all levels interact with each other. Phonetic reduction may be the result of an unstressed syllable due to its less essential semantic content, or the result of the speaker's intention to fit some words into a phonologically-conditioned duration pattern, or simply because the speaker wants to use an utterance-final particle, so the previous syllable has to be unstressed. Thus, we processed our data

first by means of prosodic segmentation, partly because prosody is important for spoken language in general, and partly because we needed an intermediate unit to obtain a concrete access to the content of speech. The results suggest that there exists a temporal pattern within prosodic units. Chinese discourse particles have no lexical tones, but instead use different intonation contours to express different attitudes of a speaker (Chao 1968). By extracting the occurrences of the discourse particle A from the MCDC data, four typical pitch contours may be identified. Disfluency is important for research on spontaneous speech and has been studied carefully already. However, we want to know whether non-native speakers can also recognize prosodic boundaries and disfluency in a foreign language. This was investigated in a pioneer judgment experiment. We used data extracted from the aforementioned speech corpora to study the following issues. It is hoped that the different aspects of spoken discourse may be examined at a later date for correlations.

### 3.1 Prosodic Units and Duration

When analyzing phenomena in spontaneous speech, we need a kind of unit, similar to punctuation marks in written texts, to provide a structural framework. Intonation units have been widely used for discourse and conversation analysis. Tao (1996) has applied the notion of intonation units (IU) to analyze Mandarin conversation. An IU usually has one or more of the following features: (1) it often ends with a pause, (2) the initial pitch value of an IU is higher than the ending pitch value of the immediately preceding IU, and (3) the final lexical items of an IU are likely to be lengthened. So, for our data, we mainly adopted the notion of IU, but with some modifications.

We would like to use the term prosodic units (PU) to make it clear that all prosodic features, not only intonation, are in fact considered in the definition. A prosodic unit is defined as a perceptually coherent prosodic constituent. A number of the prosodic cues characterized in ToBI and IU are also adopted in our work including coherent contour type, pitch reset, final syllable lengthening, and disjunction of adjacent words (such as break, pause, and laughter). In our labeling guidelines, we add one more feature to identify prosodic units: tempo alternation. Supported from the studies on spoken Czech and Mandarin in which speech tends to start fast at the beginning of an IU and to end slowly

(Dankovičová 1997, Tseng 2006b), temporal variability proves to be a useful cue for unit boundary identification.

The operational principles defined for the labelers are (1) pitch reset: a shift upward in overall pitch level. In other words, a new prosodic unit may begin with a pitch value higher than the ending pitch value of the previous prosodic unit, (2) lengthening: lengthening of syllables, changes in duration, (3) occurrences of paralinguistic sounds: disjunction or disruption of utterances by pauses, laughter etc., and (4) alternation of speech rate: changes in rhythm within the same speaker turn.

We have segmented the data of the same female speaker from the MCDC as in the previous section in terms of the syllabic boundaries.[10] **Figure 4** shows the distribution of the occurrences of different syllable durations. The distribution of syllable duration over its frequency is similar to the result presented in Greenberg (1999:171). In spite of the fact that English is often regarded as a stress-timed, and Mandarin a syllable-timed language, the deviation in Mandarin seems to be more obvious than in English. Notice that the discourse items used in Mandarin are often monosyllabic and that they also often form an independent prosodic unit. Syllables in monosyllabic prosodic units are often much longer than in non-monosyllabic prosodic units. This factor may be one of the reasons accounting for the greater deviation in Mandarin.

---

[10] We would like to thank Professor Yih-Ru Wang for using the MCDC data in training their automatic speech recognizer and for automatically segmenting the syllabic boundaries of our data. It was a great help although some post-editing by hand was still needed. The prosodic units were labeled by Ya-Fang He, Yun-Ju Huang, Tzulun Lee, and Yi-Fen Liu.

**Figure 4:** Syllable duration
(N=4,964, mean=204.5msec, standard deviation=150.9msec)
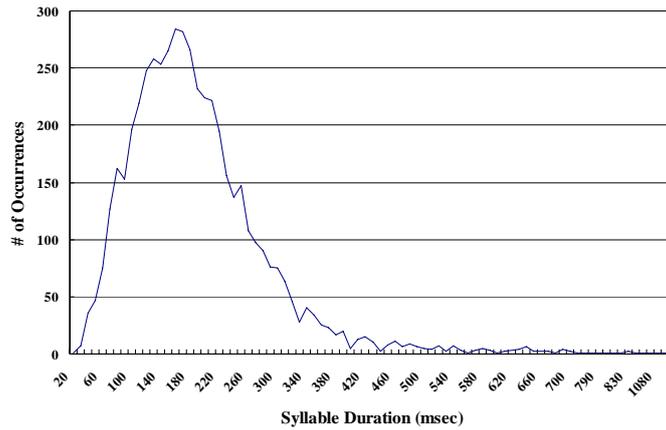


Since we had data for syllable duration, we subsequently studied the duration of the PU boundary items. As shown in **Figure 5**, PU-initial syllables (the yellow curve) are shorter than the means for the average syllable of the corresponding PU, and PU-final syllables (the red curve) are longer than the means of the average syllable (Tseng 2006b). This suggests that prosodic units are concrete units which are marked, at least, within the duration pattern.

**Figure 5**: Syllables at PU boundaries and within PUs



17
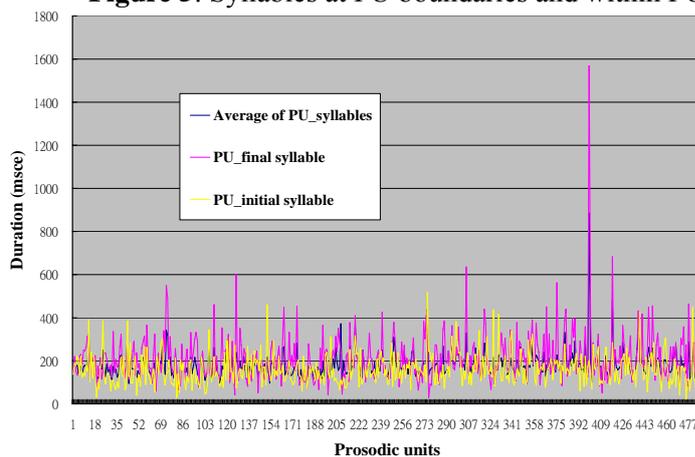
The same dataset was also labeled at the phonemic level. The preliminary results show that there are different levels of tempo phrasings: word-level and PU-level. For the word level, we found that the onset of the first syllable of a multi-syllabic word is lengthened and also that the rhyme of the final syllable of a multi-syllabic word is lengthened, too. For the PU level, the rhyme of the final syllable of a prosodic unit is lengthened and the onset and rhyme of the first syllable of a prosodic unit are shortened.

**3.2 Discourse Particles**

From the MCDC data (**Appendix A**), we extracted all 2,008 occurrences of the discourse particles *A*. *A* is the most frequently used discourse particle, making up approximately 25% of all discourse particles found in our data. After normalizing the pitch values (from 0 to 1) and the time (from 1 to 10) and undertaking cluster analysis, we obtained four different pitch contours: level-falling, high-level, mid-falling, and mid-rising[11]. In discourse analysis, only two major pitch patterns are distinguished, namely high-pitched and low-pitched contours (Chu 2002). High-pitched *A* may indicate an active intention associated with the addressee, whereas the low-pitched *A* expresses a speaker-related attitude. But our result shows that there are more subtle differences among the high-low contrast, as shown in **Figure 6**. Each pitch contour can be associated with more than one pragmatic function. To take *Dui A* (which can mean 'right' or 'well' in English, and is often used as a discourse marker) as an example, a level-falling contour can indicate a kind of assertion-endorsement; a high-level contour an indication for an intra-turn exchange; a mid-falling contour an answer to a question; and a mid-rising contour can indicate a kind of summary after a long statement.

---

[11] The occurrences of *A* were labeled by Vincent Liu. The normalization and clustering were carried by Che-Kuang Lin at National Taiwan University.

**Figure 6**: Different pitch contours of *A,* Liu (2005: 47)
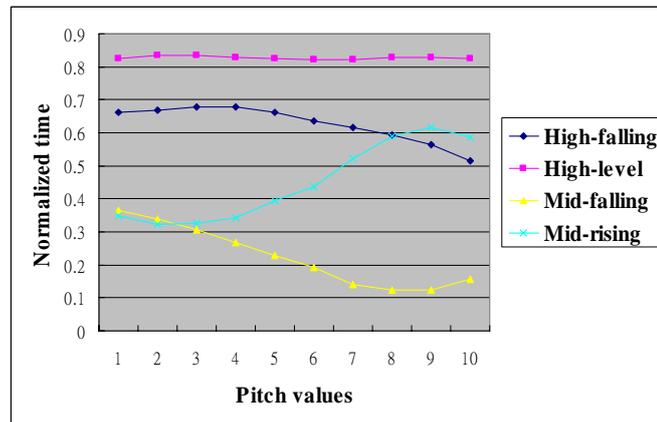


### 3.3 Disfluency

Disfluency is one of the most important features of spontaneous speech. A number of works have been done for disfluency in Mandarin (Tseng 2006a, Lin *et al.* 2005). To test the prosodic boundaries and the location of the disfluency, a perceptual recognition experiment was run in the summer of 2007 to evaluate how good the criteria are and how they may work the same or differently for native and non-native speakers. Thirty-three utterances from the MCDC data were used as stimuli to test the number of prosodic units that the subjects think there are. All 33 utterances are similarly long and contain one, two, or three prosodic units, according to our final labeled version. Twelve utterances were used to test whether the listener perceived them as fluent or disfluent. Among them, four are fluent speech. In the training phase, the criteria for prosodic boundaries and disfluency were first explained to the subjects. Then, examples were shown to the subjects, including utterances containing different numbers of prosodic units as well as fluent and disfluent utterances. The subjects were allowed to repeatedly hear the examples until they thought that they understood the principles.

Twenty-four subjects between the ages of 20 and 40 participated in this initial experiment in each native and non-native group, respectively. The native-group experiment was run at Academia Sinica. The

non-native-group experiment was run in Edinburgh[12]. The majority of both groups of subjects reported a relatively high degree of self-confidence in their judgments.

**Figure 7**: Judgment result of prosodic units and disfluency
(M: native group, E: non-native group)
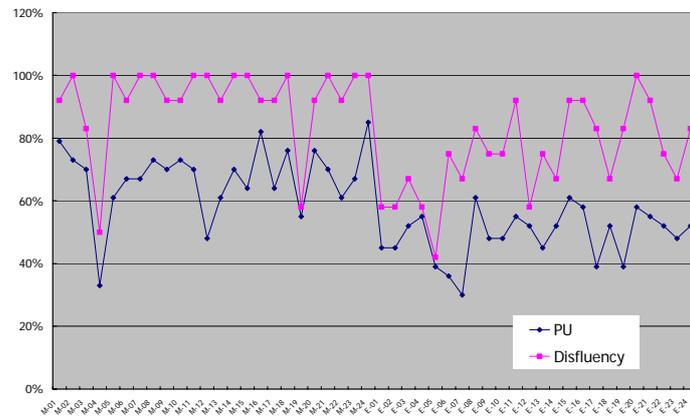


**Figure 7** shows that the native speakers are able to tell the number of prosodic boundaries and the existence of disfluency more correctly than the non-native speakers. Although this experiment was not designed to collect data of the positions where the subjects think the prosodic boundaries are located, the difference of the correction rate of the number of prosodic units between the native and the non-native speakers is clear. The detection of disfluency (93% vs. 74%, native vs. non-native) in both groups is better than the detection of the prosodic boundaries (67% vs. 49%, native vs. non-native). This may be due to the fact that prosodic means are used to mark the occurrence of disfluency such as filled pauses, silent pauses, and repetitions.

---

[12] The author would like to thank Yi-Fen Liu for running the experiment with the native group and Eleanor Drake for running the experiment with the non-native group.

**Figure 8:** Prosodic units and their boundary features
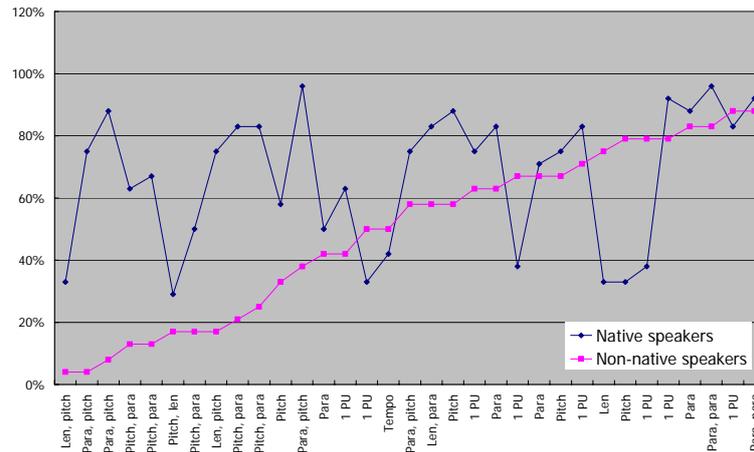


In the x-axis of **Figure 8**, prosodic features of the relevant prosodic boundaries are given. These include *len* (lengthening), *pitch* (pitch reset), *para* (paralinguistic sounds), and *tempo* (change in the speaking rate); 1PU means that the whole of the speech stretch is one prosodic unit. **Figure 8** shows that pitch reset seems to be a difficult prosodic feature for the non-native speakers. This is an interesting result, because originally we expected that native speakers would be biased by lexical meaning when they are judging prosodic chunks. But we did not have any prior hypothesis for non-native speakers. However, this experiment was a very primitive try. We are currently planning a revised experiment where we will collect the positions of the prosodic boundaries for analysis. We hope to better understand how prosodic boundaries are perceived by using the identified positions and the prosodic features associated with the positions and to what extent acoustic properties are associated with perception.

## 4. CONCLUSION

The results presented in this paper demonstrate that an analysis of spoken corpus contributes to the understanding of linguistic phenomena. We used spoken corpora to investigate syllable contraction, phonetic

reduction and grammaticalization observed in spoken data. The study clearly supports the notion that spoken language reflects the processes of language change through observations of different variations. The phenomena of prosodic segmentation, discourse particles and disfluency are important issues in relation to spontaneous speech, and their phenomena are much more complicated than other types of speech. Thus, linguistic knowledge is necessary for constructing useful corpora in order to properly process spoken corpus (including the transcription and annotation). Although some of the studies we presented here were initial work, and certainly more work needs to be carried out, we hope this paper has shown that spoken corpora can be used to empirically confirm linguistic theories and also to explore unknown linguistic phenomena.

**REFERENCES**

Anderson, Anne H., Miles Bader, Ellen G. Bard, and Elizabeth Boyle. 1991. The HCRC Map Task Corpus. *Language and Speech* 34: 351-366.

Ansaldo, Umberto, and Lisa Lim. 2004. Phonetic Absence as syntactic prominence. Grammaticalization in isolating tonal languages. *Up and Down the Cline - The Nature of Grammaticalization*, ed. by Fischer, Norde, and Perriden. John Benjamins. 345-362.

Boersma, Paul, and David Weenink. 2006. Praat: Doing phonetics by computer. http://www.fon.hum.uva.nl/praat 6.7.2006.

Chao, Yuen-Ren. 1968. *A Grammar of Spoken Chinese*. Berkeley, California: University of California Press.

Cheng, Robert L. 1985. Sub-syllabic morphemes in Taiwanese. *Journal of Chinese Linguistics* 13(1): 12-43.

Chu, Chauncey C. 2002. Relevance theory, discourse markers and the Mandarin utterance-final particles A/YA. *Journal of the Chinese Teachers Association* 37(1): 1-42.

Chung, Raung-Fu. 1997. Syllable contraction in Chinese. *Chinese Languages and Linguistics III. Morphology and Lexicon*, ed. by Tsao and Wang. Institute of History and Philology, Academia Sinica. 199-235.

Dankovičová, Jana. 1997. The domain of articulation rate variation in Czech. *Journal of Phonetics* 25: 287-312.

Fan, Jiyan. 1963. Structural Analysis of Verbs and Directional Complements (in Chinese). *Zhongguo Yuwen* 2: 136-160.

Godfrey, John J., Edward C. Holliman, and Jane McDaniel. 1992. SWITCHBOARD: Telephone speech corpus for research and development. *Proceedings of ICASSP. San Francisco*, CA. 517-520.

Greenberg, Steven. 1999. Speaking in shorthand - A syllable-centric perspective for understanding pronunciation variation. *Speech Communication* 29: 159-176.

Harrington, Jonathan and Steve Cassidy. 1999. *Techniques in Speech Acoustics*. Kluwer Academic Press.

Hsu, Hui-Chuan. 2003. A sonority model of syllable contraction in Taiwanese Southern Min. *Journal of East Asian Linguistics* 12(4): 349-377.

Labov, William. 2006. A sociolinguistic perspective on sociophonetic research. *Journal of Phonetics*, 34: 500-515.

Lamarre, Christine. 2004. Verb Complement Constructions in Chinese Dialects: Types and Markers. *Chinese Grammar. Synchronic and diachronic perspectives*, ed. by Chappell. Oxford University Press. 85-120.

Lin, Che-Kuang, Shu-Chuan Tseng, and Lin-Shan Lee. 2005. Important and new features with analysis for disfluency interruption point (IP) detection in spontaneous Mandarin speech. *Proceedings of DISS 05*. 117-121.

Liu, Vincent. 2005. *Utterance-final Particles in Mandarin Conversations: Function and Representation of A*. MA thesis, Fu-Jen University.

Liu, Yuehua (Ed.) 1998. *On Directional Complements*. Beijing: Beijing Language and Culture University.

Lung, Yu-Chun. 1979. A Discussion of the Theory that Yin-sheng Words End with Final Consonants. *Bulletin of the Institute of History and Philology* 50(4):679-716. (in Chinese)

Maekawa, Kikuo. 2004. Design, compilation, and some preliminary analyses of the Corpus of Spontaneous Japanese. *Spontaneous Speech: Data and Analysis*, ed. by Yoneyama, K., and K. Maekawa. Tokyo: The National Institute for Japanese Language. 87-108.

Tao, Hongyin. 1996. *Units in Mandarin Conversation: Prosody, Discourse and Grammar*. Amsterdam: John Benjamins.

Tseng, Shu-Chuan. 2004. Processing spoken Mandarin corpora. *Traitement automatique des langues*. Special Issue: Spoken Corpus Processing 45(2): 89-108.

Tseng, Shu-Chuan. 2005a. Monosyllabic word merger in Mandarin. *Language Variation and Change* 17(3):231-256.

Tseng, Shu-Chuan. 2005b. Syllable contractions in a Mandarin conversational dialogue corpus. *International Journal of Corpus Linguistics* 10(1): 63-83.

Tseng, Shu-Chuan. 2006a. Repairs in Mandarin conversation. *Journal of Chinese Linguistics* 34(1): 80-120.

Tseng, Shu-Chuan. 2006b. Linguistic markings of units in spontaneous Mandarin. *ISCSLP 2006, Lecture Notes in Artificial Intelligence 4272*, ed. by Huo, Q., B. Ma, E.-S. Chng, and H. Li. Berlin-Heidelberg: Springer Verlag. 43-54.

*Shu-Chuan Tseng*

*Shu-Chuan Tseng*
*Institute of Linguistics*
*Academia Sinica*
*Nankang, 115 Taipei, Taiwan*
*tsengsc@gate.sinica.edu.tw*

**APPENDIX A CORPUS STATISTICS (MCDC)**

| Speaker | Sex | Age | Syllables | Word types | Word tokens | Types of monosyll. words (%) | Tokens of monosyll. words (%) | Types of disyllabic words (%) | Tokens of disyllabic words (%) |
|---|---|---|---|---|---|---|---|---|---|
| S-01 | Fe | 29 | 4,789 | 921 | 3,334 | 309 (34%) | 1,846(55%) | 509(55%) | 1,321(39%) |
| S-02 | M | 25 | 9,262 | 1,445 | 6,913 | 416(29%) | 3,531(51%) | 783(54%) | 2,586(37%) |
| S-03 | F | 37 | 8,522 | 1,140 | 5,853 | 341(30%) | 3,398(58%) | 659(58%) | 2,155(37%) |
| S-04 | M | 35 | 6,202 | 965 | 4,234 | 292(30%) | 2,476(59%) | 552(57%) | 1,523(36%) |
| S-05 | F | 16 | 9,273 | 1,093 | 6,339 | 413(38%) | 3,585(57%) | 598(55%) | 2,541(40%) |
| S-06 | F | 17 | 6,659 | 874 | 4,497 | 307(35%) | 2,450(55%) | 510(58%) | 1,936(43%) |
| S-07 | M | 40 | 8,887 | 1,283 | 6,946 | 372(30%) | 4,147(60%) | 759(59%) | 2,434(35%) |
| S-08 | F | 46 | 7,360 | 1,140 | 5,497 | 348(31%) | 3,171(58%) | 665(58%) | 2,047(37%) |
| S-09 | F | 30 | 2,687 | 572 | 1,967 | 210(37%) | 1,152(59%) | 316(55%) | 728(37%) |
| S-10 | F | 35 | 13,534 | 1,577 | 9,103 | 461(29%) | 5,288(58%) | 953(60%) | 3,500(39%) |
| S-11 | M | 35 | 7,140 | 1,104 | 4,399 | 312(28%) | 2,323(53%) | 665(60%) | 1,859(42%) |
| S-12 | M | 23 | 6,057 | 882 | 3,723 | 257(29%) | 2,004(54%) | 540(61%) | 1,528(41%) |
| S-13 | M | 43 | 7,847 | 1,066 | 5,301 | 308(29%) | 3,063(58%) | 640(60%) | 2,049(39%) |
| S-14 | F | 45 | 7,808 | 864 | 4,859 | 265(31%) | 2,757(57%) | 495(57%) | 1,839(38%) |
| S-15 | F | 37 | 4,437 | 858 | 3,255 | 268(31%) | 1,878(58%) | 501(58%) | 1,199(37%) |
| S-16 | M | 24 | 6,751 | 1,150 | 4,833 | 372(32%) | 2,703(56%) | 646(56%) | 1,881(40%) |

*Spoken Corpora and Analysis of Natural Speech*

口語語料庫與自然語音分析

曾淑娟
中央研究院

本文介紹中央研究院語言學研究所所收集的口語語料庫,並且討論如何使用這些語料於研究不同的語言學議題。除了語言學研究,我們處理的口語語料庫也具有語言典藏的價值。在研究議題上,分為兩個主軸:1) 語言變體與變化與 2)口語言談特性。語音弱化是語言演變的重要因素之一,因此考慮自然語音在不同語言學層次的變體也相對必要。在這部分,我們分析了音節縮讀,元音弱化與趨向補語的語音弱化等幾個現象。在言談分析層面,我們量化了言談詞語所搭配的不同語調形式。利用聲韻單位,我們也進行了聲韻與不流暢語流在自然語音的表現與母語一非母語者如何辨識的實驗。這些議題其實是相互纏繞的。本文除了介紹口語語料庫在研究方法上的價值,也希望點出聲韻的表現是不同層次相互連結的重要指標。