Toward a Large Spontaneous Mandarin Dialogue Corpus

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

Abstract

This paper addresses recent results on Mandarin spoken dialogues and introduces the collection of a large Mandarin conversational dialogue corpus. In the context of data processing, principles of transcription are proposed and accordingly a transcription tool is specifically developed for Mandarin spoken conversations.

Introduction

Large speech corpora have become indispensable for current linguistic research and information science applications dealing with spoken data (Gibbon et al. 1997). Concretely, they provide real phonetic data and empirical data-driven knowledge on linguistic features of spoken language. The corpus presented here is composed of conversational dialogues. Conversations contain a considerable variety of linguistic phenomena as well as phonetic-acoustic variations. Furthermore, they open up a wide range of research issues such as dialogue acts, turn-taking, lexical use of spoken language and prosodic use in conversation. From a diachronic point of view, such a large dialogue corpus archives the contemporary daily conversational use of a given language.

1 General Issues on Mandarin Dialogues

In the following, issues on Mandarin dialogues relevant to spontaneous dialogue annotation are summarized and discussed. It includes lexical distribution, discourse markers, turn-taking and prosodic characterization.

1.1 Lexical Distribution in Spoken Mandarin

Results presented by Tseng (2001) show that speakers of Mandarin adopt some 30 words for building core structures of utterances in conversation, independently of individual speakers. All subjects used these words more than three times. The occurrences of these 30 core words make up about 80% of the overall tokens in conversation. Interestingly but also expected in conversational dialogues, the distribution of token frequency across all subjects is highly symmetric (Tseng 2001). For instance, verbs “is located”, “is”, “that is”, “say”, “want” and “have” were frequently used, so were pronouns “s/he”, “you” and “I”. The negation “don’t have” was a high-frequency word, so were words “right”, “this/these” and ”that/those”. Grammatical particles as well as discourse particles were also among the core words.

1.2 Discourse Markers

It is now well known that what differentiates written texts from spontaneous speech most is the use of discourse particles. Among the core words, eleven words were discourse particles, or they were used as discourse markers. In the literature, there is still no consistent definition for discourse markers (Hirschberg and Litman 1993). Discourse markers can be defined as follows: elements whose original semantic meaning tends to decrease and their use in spoken discourse becomes more pragmatic and indicative of discourse structuring are discourse markers. In addition to several adverbs and determiners, discourse particles can also be categorized as discourse markers. They are very often observed in Mandarin spoken conversations as mentioned in Tseng (2001) and Clancy et al. (1996).

In Tseng (2001), each subject used on average 1.6 discourse particles per turn. This result leads to the consideration, if there is a need to add special categories for discourse particles or particle-like words for spoken Mandarin. Discourse particles were found to have different and specific discourse use in conversation.

Namely, there exist discourse particles appearing preferably in turn-beginning position and some other discourse particles may exclusively mark the location of repairs. Regarding the small size of data used in Tseng (2001), it is one of the reasons why the ongoing project is necessary for research of Mandarin spontaneous conversations.

1.3 Taking Turns in Dialogues

In spontaneous conversation, turn-taking usually takes place arbitrarily to the extent that every individual interacts differently with the others under different circumstances. Thus, how to annotate overlapping sequences is one of the essential tasks in developing annotation systems. In Mandarin conversation, there are words preferably used in turn-initial position (Tseng 2001, Chui 2000). They normally have their own discourse-related pragmatic function associated with their positioning in utterances. Similarly, how to mark up turn-initial positions is also directly connected with the annotation convention.

1.4 Prosody in Spoken Mandarin

Lexical tones are typically characteristic of spoken Mandarin. The interaction of lexical tones and the other prosodic means such as stress and intonation are related to a number of research issues, particularly in conversation. Falling tones may not show falling tendency anymore, when the associated words are used for specific discourse functions such as for indicating hesitation or the beginning of a turn (Tseng 2001).

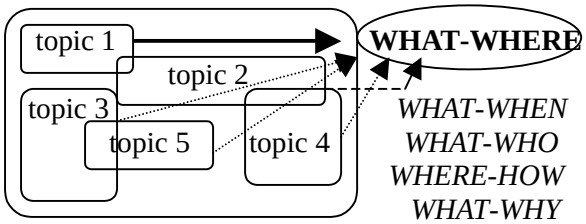
2 Mandarin Conversational Dialogue Corpus

This section deals with the design and collection of a large Mandarin dialogue corpus currently produced in Academia Sinica.

2.1 Design and Methodology

The long-term goal of this project is to collect domain-independent Mandarin spoken dialogues. They are daily conversations with specific topics given for each stage of recording. Since the design of scenario aims to collect natural and spontaneous conversations, limitations on the topics are reduced to a minimum. Different from task-oriented dialogues such as air-planning or instruction-construction tasks (Kowtko and Price 1989, Sagerer et al. 1994), subjects participating in this project were told to converse as naturally as possible. The scenario is similar to a situation where two strangers meet at the first time, try to find common topics interested by both of them and have a chat.

Figure 1: *Corpus Domain Design*



As illustrated in Figure 1, this stage of corpus collection consists of WHAT-WHERE component. The subjects have to determine on WHAT topic they'd like to talk. Usually, they do not stick to only one topic. Sometime in conversation, the participants asked each other WHERE the events mentioned in their conversation happened or WHERE they could take part in the events. There are two reasons for this design. First, we will extend the local domains to WHAT-WHEN, WHAT-WHO, WHERE-HOW as well as WHAT-WHY combinations in the next five years to cover important dialogue components used for daily talks in Taiwan. This aims to archive the use of contemporary spoken Mandarin from varied daily-life perspectives. Second, casual conversations usually do not require correctness of information. To make sure that we will obtain at least some "seriously spoken" materials, the subjects should interact with attention to fulfil the WHERE task, namely the route-asking and route-describing task.

2.2 Subjects and Instructions

60, 23 male and 37 female, Taipei residents of Taipei City who volunteered to participate in the project were recorded in pairs. Age ranges from 16 to 45. Subjects did not know each other and their task was to introduce themselves first. Then they chose topics from those given on the instruction sheet or they were also free of choosing any other not-listed topics to talk about. In addition, they asked some questions about routes within conversation. The topics given to the subjects are family, work, shopping, food, travelling, politics and economics. Both subjects

can be information-acquirer or information-giver. However, they were told that the person who asked route questions had to make sure that s/he has completely understood the described routes.

2.3 Recording

The dialogues were recorded by a SONY TCD-D10 Pro II DAT tape recorder with Audio-Technica ATM-33a handheld/stand cardioid condenser microphones at a sampling rate of 48 kHz. Each subject was recorded on a separate channel on a DAT tape. There was no time constraint given in advance. Once the subjects completed their task and wished to end the conversation, the recording was stopped. Total length of corpus is about 25 hours recording.

3 An Extensible Transcription Tool

3.1 Functional Considerations

This section discusses three principles for constructing word-based database for Mandarin dialogues from audio data to lexical database. Three functions have to be included in a transcription system, either directly or potentially: 1) connecting the tasks of transcribing, annotating and labelling of sound data, 2) being able to deal with overlapping turns and 3) making available possible tiers for later time-alignment.

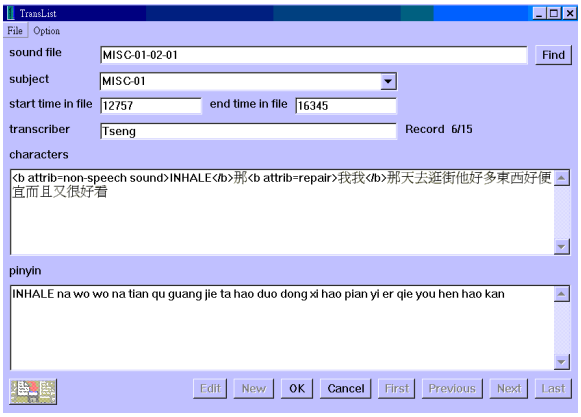
There are three working levels for processing spoken data: transcribe, annotate and label. First, transcription is the transliteration of audio data. Normally, it is verbatim transcription in plain text form. A transcription tool has exclusively been developed for broadcast spoken data, named Transcriber (Barras et al. 2001). Audio data can be nicely combined with the other information. However, it lacks flexible possibility for defining new annotation tags. It is especially difficult to use Transcriber to transcribe Mandarin conversations because of the written system of Mandarin. For the understanding of content and for the completeness of written system, Chinese characters are as representative and important as Latin transcriptions.

Secondly, to annotate spoken data is to add linguistic explanations to the plain transcript to represent linguistic structures at selected levels. And lastly, to label sound data is to temporally align transcript texts with speech signals. The tool we develop for our corpus collection aims to transcribe and annotate the speech data as well as to build potential temporal tiers for future labelling work. Traditional annotations of spoken data orient at turn-structured separation of utterances or sets of utterances (Sagerer et al. 1994). This leads to the following inconveniences. The beginning and ending boundaries between utterances are not represented, because it is presupposed that the current ending boundary is the beginning boundary of the next unit. While doing temporal alignment, pauses between utterance units and speakers may be missing. From the point of view of searching mechanism, an annotation system should also satisfy the demand on classifying sequences produced by given speakers or sequences produced in a given time period. Thus, it will be statistically effective and useful to output annotated transcription data in a database format.

3.2 TransList

Recorded dialogues are currently being transcribed by using a computer-aided program TransList, specifically developed for transcribing Mandarin conversations. The interface is illustrated in Figure 2. Input contents include location of sound file, subject, start and end times of the transcribed segment in the correspondent sound file, the person who transcribes the segment. The actual transcription is done twofold: in characters and in Pinyin.

Figure 2: TransList



Tag list can be flexibly extended to annotate different phenomena any time while doing the transcription. Each transcribed segment is

referred to its original sound file. However, a direct connection to processing audio files is not available yet. Regarding the output format of TransList, two variations are currently in use. One is conversation-typed format. In other words, all sound files split from one conversation form an independent text file. In order of time and subject, the program outputs a turn-oriented format, as illustrated in the next section. More important is the second output format. All transcribed segments belonging to one conversation will be listed in a database, having the following columns: characters, pinyin, sound file and all added tags. Words marked up by different tags will all have the values of added tags as their attribute in the database. By doing this, we plan to do word segmentation and tagging for spoken Mandarin. An automatic word segmentation and tagging system is currently available for written Mandarin (Chen et al. 1996). We intend to test this program for spontaneous Mandarin. By outputting the transcription in database format, we will make the continuing data processing and searching more effective.

3.3 An Output Example

This section gives an example produced by TransList.

<A 1 6473> cong shili yao dao shili meishuguan dehua women jiushi keyi zhiyao ni en jiushi cong women xuexiao yeshi yiyang da ersanliu zui fangbian de a </A 1 16200><B 1 16230> mhm </B 1 16380><A 2 16530> ranhou da dao yeshi dao gongguan huan danshuixian ranhou ni zhiyao zuo dao dagai yuanshanzhan[1]</A 2 22230>

In the above example, the brackets <> and </> mark up the beginning and ending boundaries of a speaker production sequence. A and B stand for speakers. Turns are not explicitly separated, but marked up in the annotation. Numbers after the speaker abbreviations indicate numbers of production sequences by the speaker. Thus, whether it is a turn-taking or it is a overlapping can be evaluated by means of the third parameter time (msec). With respect to tags added into the transcribed segments, it is optional to include or to exclude the annotation tags. As shown in Figure 2, these can be non-speech sounds, repairs or discourse markers (Heeman and Allen 1999).

[1] <A 1 6473> from city want go city gallery in the case we just can just you en just from our school too is same take 236 most convenient PRT a </A 1 16200> <B 1 16230> mhm </B 1 16380> <A 2 16530> afterwards take to too to Gongguan change Danshui-Line afterwards you just take to approximately Yuanshan-Station </A 2 22230>

Conclusion

This paper discussed general issues on Mandarin spoken dialogues and analysed components of a new developed transcription and annotation tool for spoken Mandarin.

References

Barras, C. et al. (2001) Transcriber: Development and Use of a Tool for Assisting Speech Corpora Production. Speech Communication. 33. Pp. 5-22.

Chen, K.-J. et al. (1996) *SINICA CORPUS: Design Methodology for Balanced Corpora*. *PACLIC 11*. Pp.167-176.

Chui, K.-W. (2000) *Ritualization in Evolving Pragmatic Functions: A Case Study of DUI*. In Proc. of the 7th International Symposium on Chinese Language and Linguistics. Pp. 177-192.

Clancy, P. et al. (1996) *The Conversational Use of Reactive Tokens in English, Japanese and Mandarin*. Journal of Pragmatics. Pp. 355-387.

Gibbon, D., Moore, R. and Winski, R. (1997) *Handbook of Standards and Resources for Spoken Language Systems*. Mouton de Gruyter.

Heeman, P. and Allen, J. (1999) *Speech Repairs, Intonational Phrases and Discourse Markers: Modelling Speakers' Utterances in Spoken Dialogue*. Computational Linguistics, 25/4. Pp. 527-571.

Hirschberg, J. and Litman, D. (1993) *Empirical Studies on the Disambiguation of Cue Phrases*. Computational Linguistics, 19(3), pp. 501-530.

Kowtko, J. C. and Price, P.J. (1989) *Data Collection and Analysis in the Air Planning Domain*. In Proc. of the DARPA Speech and Natural Language Workshop. Pp. 119-125.

Sagerer G. and Eikmeyer H. and Rickheit G. (1994) *"Wir bauen jetzt ein Flugzeug": Konstruieren im Dialog. Arbeitsmaterialien*, Technical Report. SFB360 "Situierte Künstliche Kommunikation. University of Bielefeld, Germany.

Tseng, S.-C. (2001) *Highlighting Utterances in Chinese Spoken Discourse*. In Language, Information and Computation. PACLIC 15. Pp. 163-174.