

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	d-11 22 min.	Spouses	F: 34	d-21 17 min.	Spouses	F: 36
		M: 45			M: 43			M: 36
d-02 23 min.	Friends	F: 30	d-12 21 min.	Friends	F: 23	d-22 21 min.	Siblings	M: 45
		M: 36			M: 23			F: 40
d-03 17 min.	Siblings	F: 37	d-13 26 min.	Relatives	M: 47	d-23 22 min.	Mother-daughter	F: 46
		F: 39			F: 43			F: 21
d-04 11 min.	Friends	M: 26	d-14 23 min.	Friends	F: 35	d-24 16 min.	Spouses	M: 45
		F: 22			F: 43			F: 45
d-05 19 min.	Friends	M: 29	d-15 20 min.	Friends	M: 42	d-25 22 min.	Friends	M: 22
		F: 26			M: 40			M: 23
d-06 21 min.	Siblings	F: 33	d-16 17 min.	Mother-daughter	F: 43	d-26 21 min.	Siblings	F: 17
		M: 36			F: 15			F: 14
d-07 23 min.	Spouses	F: 47	d-17 21 min.	Mother-faughter	F: 36	d-27 22 min.	Friends	M: 26
		M: 46			F: 63			M: 28
d-08 18 min.	Spouses	F: 37	d-18 21 min.	Relatives	M: 40	d-28 20 min.	Friends	M: 29
		M: 42			M: 24			M: 28
d-09 20 min.	Mother-daughter	F: 20	d-19 23 min.	Friends	F: 27	d-29 22 min.	Friends	M: 21
		F: 49			F: 27			M: 23
d-10 25 min.	Mother-daughter	F: 48	d-20 22 min.	Friends	F: 45	d-30 23 min.	Friends	F: 42
		F: 23			M: 53			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 他就馬上撤僑 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
repeat	7	2	0	0	0	1	3	4	0	0	1	1	0	0	0	0	19
rephrase	2	0	2	2	0	2	2	2	1	1	0	0	0	0	0	0	14
request	0	0	1	0	0	0	0	1	2	0	6	3	0	0	0	14	27
rhetorical_ question_ answered	8	0	0	1	0	8	0	1	2	2	0	3	0	3	0	3	31
rhetorical_ question	2	0	0	0	0	0	0	0	0	0	0	0	2	1	0	0	5
suggest_topic	1	1	0	1	1	0	1	1	0	2	1	0	0	1	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.

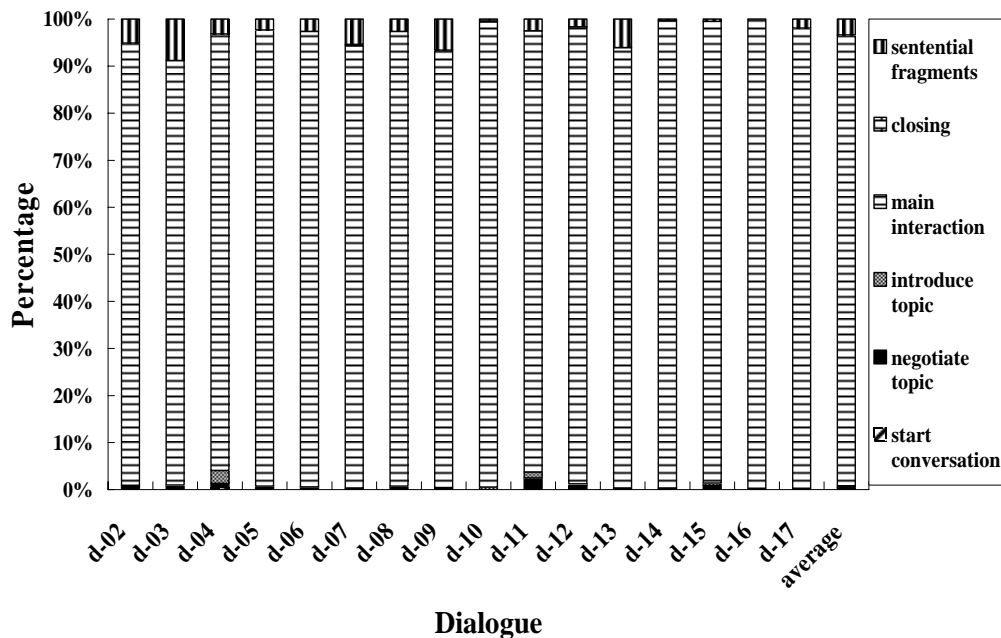


Figure 1. Dialogue structure

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].

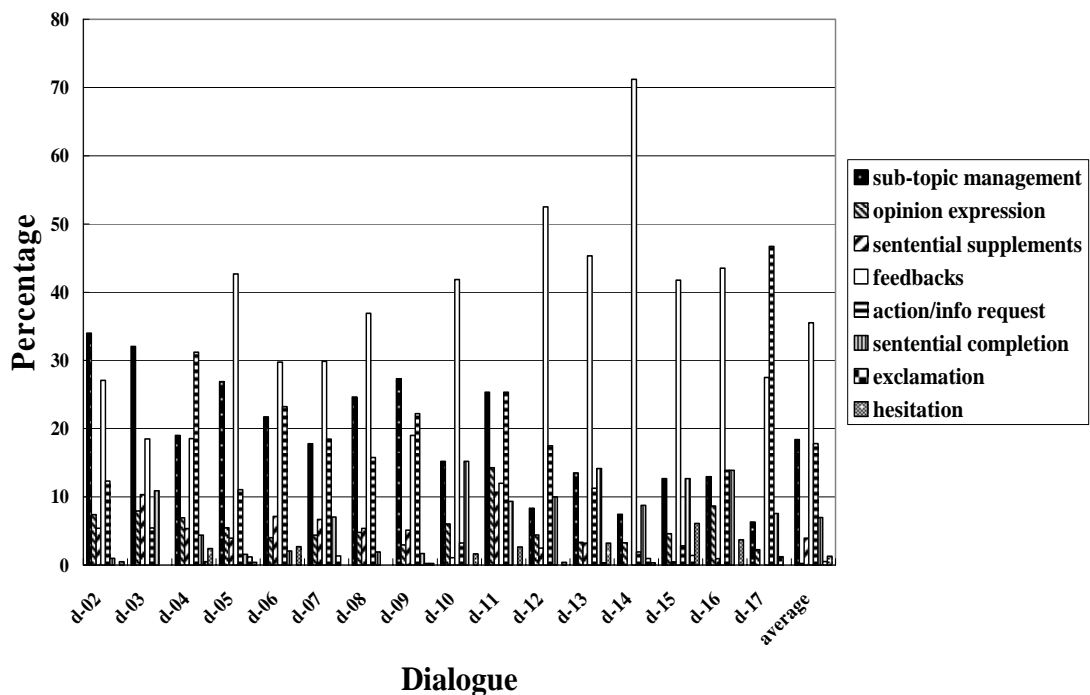


Figure 2. Major interaction of topic-oriented conversations

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 - 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	標記 <input type="text" value="agree_part"/> <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