Transcribing and Annotating Mandarin Conversational Dialogues

Shu-Chuan Tseng

Institute of Linguistics, Academia Sinica, Nankang, 115 Taipei, Taiwan
tsengsc@gate.sinica.edu.tw
http://www.ling.sinica.edu.tw

Abstract. Presentation of spoken data content is based on the annotation system designed to fulfil the attempted research goals. It should include the orthographic transcription, the linguistic annotation such as markings of extragrammatical spontaneous speech sequences, and the documentation of data collection and processing. Especially, spontaneous conversation contains a great variety of sentential construction, pronunciation variation, and conversational interaction. This paper gives an overview of the linguistic annotation system and the tool we have developed for transcribing Mandarin spontaneous conversational dialogues. By means of the annotated data, a character-based linguistic database is construed to provide quantified materials for queries.

1 Introduction

Spoken corpora have recently become a powerful tool for studying spontaneous speech phenomena and human-human (human-machine) interactions. This paper presents the content of the project "Mandarin Conversational Dialogue Corpus" (MCDC) which is currently being executed at the Institute of Linguistics, Academia Sinica, Taipei. By collecting and annotating spontaneous speech data, we obtain invaluable materials on special syntactic constructions and sequencing, which are only produced in verbal communication [6]. In general, syntactic structures in spoken language can be clearly distinguished from those in written language. Spoken sentences hardly satisfy the grammatical rules written for texts. But they are in no way ungrammatical. The linguistic annotation system developed for MCDC data aims to mark as many *extragrammatical sequences* as possible, because it is also one of our main goals to systematize syntactic phenomena in spoken Mandarin working towards a spoken grammar of Mandarin [2]. It is noted that throughout this paper all Mandarin examples are written in Pinyin with a number representing lexical tone. Four lexical tones: high flat, rising, contour, and falling tones plus one neutral tone are used in Mandarin. We use 1, 2, 3, 4, and 5 to represent them in this paper, for instance the bi-syllabic word *tomorrow* is written as "ming2tian1".

2 Mandarin Conversational Dialogue Corpus

With the help of the Office of Survey Research, Academia Sinica, 1080 residents of Taipei City were randomly chosen from three age ranges, 16-25, 26-35 and 36-45. An invitation letter was sent to all 1080 candidates to ask if they are interested in participating the project. 37 female and 23 male subjects were recorded in pairs, arranged in the order of the time they responded to the invitation letter. Among the recorded subjects, 20 of them age between 16 and 25, 19 between 26 and 35, and 21 between 36 and 45, covering a wide range of occupations: pupils, students, managers, teachers, workers, and bankers etc. Subjects did not know each other before the time of recording. Their task was to introduce themselves, to choose topics from those given on the instruction sheet or any other topics they can think of, then to talk on the chosen topics. By means of the self-introduction, we attempt to collect data on how strangers introduce themselves in formal situations. 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.

3 Linguistic Annotation

Orthographic transcription is done in both traditional Chinese characters and in Pinyin. The writing of characters and the spelling of Pinyin all refer to two standard dictionaries: Guoyu Dictionary [3] and Contemporary Mandarin Dictionary [8].

3.1 Annotation Tags

The annotation system of the MCDC data mark linguistic and paralinguistic phenomena which are dominantly produced in spontaneous speech [4], [7]. Linguistic sequences such as repairs and disfluent repetitions are hardly found in written texts, so are turn taking and interruption in conversation. Phenomena such as noise and other verbal non-speech sounds are annotated, too. In summary, the MCDC annotation system consists of four groups: 1) disfluency, 2) socio-linguistic phenomena, 3) pronunciation variation, and 4) unintelligible and non-speech sounds. Detailed annotation tags are given in Table 3.1.

Among them, *disfluency* is the most complicated and at the same time the most frequently produced phenomena differentiating spontaneous speech from the other forms of language use. Disfluent speech sequences lead to discontinuity of speech flow and often pitch contour, too. Discourse markers and discourse particles are included in the group of disfluency because their use in spontaneous conversation normally deviates from their original semantic content. *Socio-linguistic phenomena* indicate socio-linguistic characteristics such as code switching, deviated pronunciation due to influences by other languages or dialects, and newly created fashion words which are mostly invented and used by young people. In addition, languages other than English unavoidably

disfluency	**socio-linguistic**	**pronunciation**	**unintelligible &**
	phenomena	**variation**	**non-speech sounds**
pause	code switching	syll. contraction	unrecog. speech sound
short break	new word	assimilation	unre. non-sp. sound
stutter		lengthening	uncertain
restart		nasalized	laugh
repetition		inappr. pronun.	cough
overt repair			breathe
editing term			smack
error			click
word fragment			clear throat
inappropriate usage			sigh
abridged utterance			sneeze
interrupted utterance			swallow
discourse marker			.
discourse particle			.

Table 1. Linguistic Annotation of Spontaneous Speech

have to take into account the use of common English words in everyday life or technical terminologies in different fields such as "bye-bye" and "CPU", respectively. A wide range of *pronunciation variation* occur in spoken conversation such as lengthening, nasalized sounds, syllable contraction, assimilation etc. *Non-speech sounds* include all recognizable verbal but non-speech sounds for instance laughing, clearing throat etc. as well as non-verbal sounds such as noises. *Unintelligible* sounds may result from inappropriate pronunciation by the speakers themselves or defective recording quality. Noises, slightly deviated pronunciation, non-speech sounds result in relatively minor problems for speech recognition and parsing systems, because they can be treated as sequences containing no substantial linguistic contents and can be disregarded or be compensated if their context is identifiable. Foreign words result in a different sort of problem in speech recognition, which should be discussed elsewhere. What constitutes the most serious problem is the group of disfluency. Thus, we focus on the disfluent tags in this section; detailed definitions of the annotation systems see [7].

3.2 Prosodic Disfluency

Silence is used to annotate a silent pause, where none of the conversation partners utters anything within the course of conversation. Sometimes, it is so short, that it hardly bothers the conversation participants. But sometimes it can last for a few seconds and often embarrasses all participants. Different from silence, *pause* only occurs within a certain speaker's utterances. In spontaneous speech, depending on how accenting the speaker wants the listeners to perceive his/her speech, the pausing can be located anywhere within utterances, irrespective of the syntax. A *short break* annotates pausing on a smaller scale than a *pause*. In human-human conversation, speakers may hesitate or *stutter*, when they have problems in finding the correct words. Stuttering not only makes the speech flow sound discontinuous, each uttered Chinese syllable may possibly be an "independent" word, because in Mandarin each syllable forms a separate character and carries to a certain degree semantic content.

3.3 Repairs

Annotation of repairs is divided into six categories: *repetition*, *restart*, *overt repair*, *editing term*, *error*, and *word fragment*. Repetitions in Mandarin Chinese are in a number of cases perfectly legal syntactic constructions to put emphasis on particular components or to express subtle semantic nuance. Repetitions can be used for different grammatical categories including verbs, adverbs, and adjectives. To take "da4 da4 de5" as an example: "da4" is the adjective for *big* and "de5" is a structural particle, but both "da4 de5" and "da4 da4 de5" mean *big*. We need to exactly distinguish what is a syntactically well-formed repetition and what is a disfluent repetition in spoken Mandarin. By definition, disfluent repetitions are *direct repetitions*, which cannot be explained or justified by Mandarin grammatical rules. Direct repetitions immediately imply fully repeated word sequences. Partial repetitions are grouped into the category *restart*, for instance the disfluent sequence "gong1cheng2 gong1cheng2shi1", where "gong1cheng2" means construction and "gong1cheng2shi1" means engineer. An *overt repair* contains both the reparandum and the reparans [5]. Only disfluent sequences in which we can clearly identify what is to be corrected and what is the correction are regarded as repairs. In the sequence "shi4 jin4kou3 EN chu1kou3 ma1" ("shi4" means the verb BE, "jin4kou3" import, "chu1kou3" export, "EN" is a discourse particle and "ma1" is a grammatical particle for interrogative sentences), "jin4kou3 is the reparandum and "chu1kou3" is the reparans. Also found in this example, the particle "EN" is an *editing term* used to bridge the gap between the reparandum and the reparans. Editing terms can be particles, fillers or lexicalised items. Items used between the to-be-repeated and repeated elements in direct repetitions are also counted as editing terms. Words, which are not completely uttered, are annotated with the tag *word fragment*. In Mandarin, phonetic errors are hardly identifiable without knowing the subsequent syllables. Thus, the category *error* only implies for lexical or syntactic errors such as false combination in compound words, idioms, and false classifiers for nouns.

3.4 Syntactic Disfluency

When an utterance sounds inappropriate and there is more than one way (to add, to delete or to substitute something) to make the utterance sound correct, the utterance is labelled *inappropriate usage*. If an utterance is incomplete from the syntactic point of view, it is defined as *an abridged utterance*. Utterance fragments resulting from interruption by other speakers in human-human interaction, are defined as *interrupted* utterances. Abridged utterances may have a stronger relationship to the next utterance than an interrupted utterance that simply stops at the point of interruption.

3.5 Discourse Particles and Markers

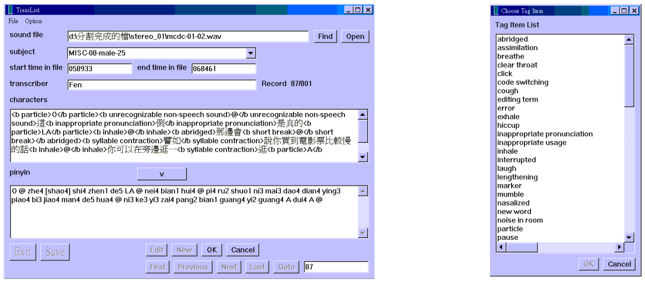
Discourse particles and *markers*, which are exclusively used in spoken language or in written language of colloquial style, also identify a kind of disfluency. In our notation, we consistently use capital letters to transcribe discourse markers, particles and other non-lexicalised items in both character and Pinyin transcription to distinguish them from words having the same pronunciation or similar meaning. Some of the discourse particles used in Chinese can be exactly mapped to characters under the condition that an agreement on a unified semantic use is shared among the native speakers. However, for some discourse particles, it is difficult to choose the exactly matched character to transcribe. For instance, discourse particles originated from Taiwanese cannot be transcribed with traditional Chinese characters. NA and NAGE are two frequent *discourse markers* found in Mandarin human-human conversation. ”Na4” originally means *that* as a determiner or *that* as a pronoun. ”Na4ge5”, *that* + classifier, is used for proper nouns such as ”na4ge5 ren2” *that man*. When used in discourse, NA often appears in utterance-initial position to signal the speaker’s intention to begin a new turn. Clearly different from NA, NAGE is usually found in mid-utterance position before a piece of new information. However, deviated from its function as a determiner, this piece of new information does not necessarily have to be a noun.

4 Transcription of Mandarin Dialogues

Traditionally, orthographic transcription is regarded as a necessary step to obtain materials for discourse analysis. But seldom has it received attention as to how to efficiently integrate the process into the analysis procedure. Transcription should be more than writing down what transcribers hear in the format of a simple text file. Functional links to audio files, documentation notes such as who has transcribed this file, where is this audio file located, important information about the subjects such as gender and age are just like the orthographic transcription part of the transcription content. For labelling spoken data, it is difficult to integrate the IPA fonts, not to mention thousands of characters, as most of the labelling software has problems in encoding and decoding special fonts. Therefore, Pinyin, a nation-wide used Latin transcription system in China, is adopted for transcribing and labelling Mandarin speech. Because of the large number of homophones in Mandarin, characters have to be used to identify which word is actually meant and to differentiate subtle semantic nuance. Thus, the tool developed for transcribing broadcast speech data, the ”Transcriber” [1], is not the best choice in our case, if we want to have these two kinds of possibilities to input transcription contents. Moreover, another concern is that in addition to characters and Pinyin in their standard dictionary forms we also need flexible tags to note down the real pronunciation.

4.1 TransList

For the above reasons, we developed the interface ”TransList” to help the transcription work and the construction of linguistic database. Each segment stored in TransList is a speaker turn. In addition to the character and Pinyin transcription, TransList embeds speaker and transcriber information, as illustrated in Figure 1. As a preparation to a direct link to simultaneously process the audio files, the location of audio files is indicated, too. The start and end time of transcribed segments also aims to provide temporal information (time code) to cooperate with future labelling work. Within our framework, these temporal records also help to arrange items in the database in the order of time. What differentiates other transcribing tools from the interface we are developing most is the insertion of annotation tags marking special linguistic or paralinguistic features in conversation.



sound file: d:\分割完成的檔\stereo_01\mcdc-01-02.wav

subject: MISC-08-male-25

start time in file: 058933

end time in file: 068461

transcriber: Fen

Record 87/801

characters:

<b particle>0</b particle><b unrecognizable non-speech sound>@</b unrecognizable non-speech sound>這<b inappropriate pronunciation>倒</b inappropriate pronunciation>是真的<b particle>LA</b particle><b inhale>@</b inhale><b abridged>那邊會<b short break>@</b short break></b abridged><b syllable contraction>譬如</b syllable contraction>說你買到電影票比較慢的話<b inhale>@</b inhale>你可以在旁邊逛一<b syllable contraction>逛<b particle>A</b

pinyin:

O @ zhe4 [shao4] shi4 zhen1 de5 LA @ nei4 bian1 hui4 @ pi4 ru2 shuo1 ni3 mai3 dao4 dian4 ying3 piao4 bi3 jiao4 man4 de5 hua4 @ ni3 ke3 yi3 zai4 pang2 bian1 guang4 yi2 guang4 A dui4 A @

Goto: 87

Tag Item List

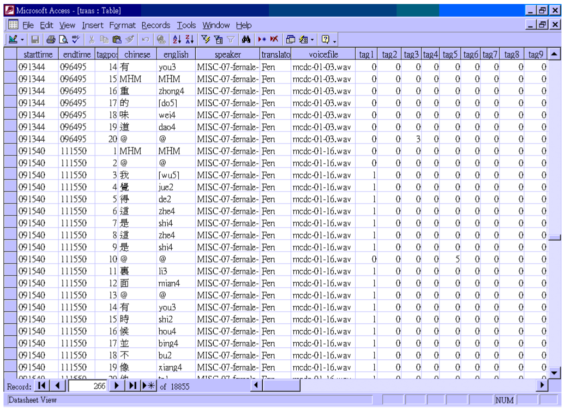
abridged
assimilation
breathe
clear throat
click
code switching
cough
editing term
error
exhale
hiccup
inappropriate pronunciation
inappropriate usage
inhale
interrupted
laugh
lengthening
marker
mumble
nasalized
new word
noise in room
particle
pause

(a) Interface

(b) Tags

Fig. 1. A Tool for the Transcription and Annotation of Mandarin Conversational Dialogue Data

All tags inserted into transcript files are put into the resulted database having attributes as the named tags. For this stage of development, the database is in Microsoft Access format, as illustrated in Figure 2. In Chinese, the definition for a ”word” is ambiguous and usually principles for word segmentation are obligatory for linguistic analyses on Chinese. For this reason, we chose ”character” as our unit in the database. We list Chinese characters and their Pinyin spelling and lexical tone in columns with all related annotated attributes in rows. Queries can be created on the basis of the database. Word frequency and distribution of annotated sequences in terms of subjects, conversation index, sound file etc. can all be automatically obtained.


Microsoft Access - [trans : Table]
File Edit View Insert Format Records Tools Window Help
Record: 266 of 18855
Datasheet View
NUM

starttime	endtime	tagpo:	chinese	english	speaker	translato	voicefile	tag1	tag2	tag3	tag4	tag5	tag6	tag7	tag8	tag9
091344	096495	14	有	you3	MISC-07-female-	Fen	mcdc-01-03.wav	0	0	0	0	0	0	0	0	0
091344	096495	15	MHM	MHM	MISC-07-female-	Fen	mcdc-01-03.wav	0	0	0	0	0	0	0	0	0
091344	096495	16	重	zhong4	MISC-07-female-	Fen	mcdc-01-03.wav	0	0	0	0	0	0	0	0	0
091344	096495	17	的	[do5]	MISC-07-female-	Fen	mcdc-01-03.wav	0	0	0	0	0	0	0	0	0
091344	096495	18	味	wei4	MISC-07-female-	Fen	mcdc-01-03.wav	0	0	0	0	0	0	0	0	0
091344	096495	19	道	dao4	MISC-07-female-	Fen	mcdc-01-03.wav	0	0	0	0	0	0	0	0	0
091344	096495	20	@	@	MISC-07-female-	Fen	mcdc-01-03.wav	0	0	3	0	0	0	0	0	0
091540	111550	1	MHM	MHM	MISC-07-female-	Fen	mcdc-01-16.wav	0	0	0	0	0	0	0	0	0
091540	111550	2	@	@	MISC-07-female-	Fen	mcdc-01-16.wav	0	0	0	0	0	0	0	0	0
091540	111550	3	我	[wu5]	MISC-07-female-	Fen	mcdc-01-16.wav	1	0	0	0	0	0	0	0	0
091540	111550	4	覺	jue2	MISC-07-female-	Fen	mcdc-01-16.wav	1	0	0	0	0	0	0	0	0
091540	111550	5	得	de2	MISC-07-female-	Fen	mcdc-01-16.wav	1	0	0	0	0	0	0	0	0
091540	111550	6	這	zhe4	MISC-07-female-	Fen	mcdc-01-16.wav	1	0	0	0	0	0	0	0	0
091540	111550	7	是	shi4	MISC-07-female-	Fen	mcdc-01-16.wav	1	0	0	0	0	0	0	0	0
091540	111550	8	這	zhe4	MISC-07-female-	Fen	mcdc-01-16.wav	1	0	0	0	0	0	0	0	0
091540	111550	9	是	shi4	MISC-07-female-	Fen	mcdc-01-16.wav	1	0	0	0	0	0	0	0	0
091540	111550	10	@	@	MISC-07-female-	Fen	mcdc-01-16.wav	0	0	0	0	5	0	0	0	0
091540	111550	11	裏	li3	MISC-07-female-	Fen	mcdc-01-16.wav	1	0	0	0	0	0	0	0	0
091540	111550	12	面	mian4	MISC-07-female-	Fen	mcdc-01-16.wav	1	0	0	0	0	0	0	0	0
091540	111550	13	@	@	MISC-07-female-	Fen	mcdc-01-16.wav	1	0	0	0	0	0	0	0	0
091540	111550	14	有	you3	MISC-07-female-	Fen	mcdc-01-16.wav	1	0	0	0	0	0	0	0	0
091540	111550	15	時	shi2	MISC-07-female-	Fen	mcdc-01-16.wav	1	0	0	0	0	0	0	0	0
091540	111550	16	候	hou4	MISC-07-female-	Fen	mcdc-01-16.wav	1	0	0	0	0	0	0	0	0
091540	111550	17	並	bing4	MISC-07-female-	Fen	mcdc-01-16.wav	1	0	0	0	0	0	0	0	0
091540	111550	18	不	bu2	MISC-07-female-	Fen	mcdc-01-16.wav	1	0	0	0	0	0	0	0	0
091540	111550	19	像	xiang4	MISC-07-female-	Fen	mcdc-01-16.wav	1	0	0	0	0	0	0	0	0

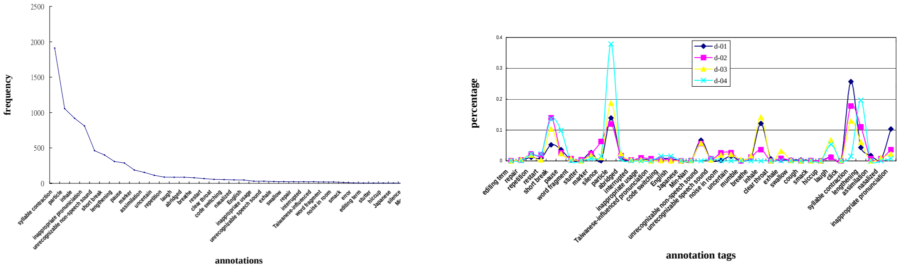
Fig. 2. Linguistic Database

4.2 Annotated Results

Applying the system of linguistic annotation presented above, we first annotated one dialogue data (1 female and 1 male), about one hour recording. In total, 38 different tags were used to annotate 7842 spontaneous speech sequences. The female speaker produced 583 turns and the male speaker 289 turns. Results of the annotated data are illustrated in Figure 3 in the order of the frequency of labelled tags. The most frequently annotated pronunciation variations are syllable contraction, inappropriate pronunciation, lengthening and assimilation, whereas simple disfluency such as particle, marker, short break, pause and complex disfluency such as repetition and abridged utterance occurred most often. Comparing the data (abbreviated as d-01) with the other annotated dialogues (abbreviated as d-02, d-03, and d-04), we found a clear symmetric pattern of distribution across these four different dialogues transcribed by four different transcribers. Although we still need to double-check the inter- and intra-transcriber consistency, nevertheless, this result supports the adequacy of the linguistic annotation system we use at this preliminary stage. Transcription and annotation of the data as well as a number of elaborated analyses (pronunciation variations, Mandarin repair types, and syntactic inappropriateness in spontaneous speech) on the basis of the linguistic database are currently in progress.

5 Conclusion

This paper gives an overview of the linguistic annotation system and the transcription tool used for Mandarin conversation data. Not only do we want to provide well-annotated spoken data for linguistic and engineering studies. By developing a system of linguistic annotation, we also hope to be able to more systematically and consistently identify spontaneous speech phenomena than



(a) 1 Dialogue

(b) 4 Dialogues

Fig. 3. Annotated Results

till now. Prosodic annotation has been neglected at this stage of research for the reason of time and budget. In near future, this area will certainly form an independent project.

References

1. Barras, C. et al.: Transcriber: Development and Use of a Tool for Assisting Speech Corpora Production. Speech Communication. **33** (2001) 5–22
2. Chao, Y.-R.: A Grammar of Spoken Chinese. University of California Press. (1968)
3. Guoyucidian: Jiaoyubu Chongbian Guoyucidian. Taiwan Shangwu Yinshuguan (1995)
4. Labov, W.: On the Grammaticality of Everyday Speech. In: Annual Meeting of the Linguistic Society of America. **41** (1966)
5. Levelt, W.J.: Monitoring and Self-Repairs in Speech. Cognition. **14** (1983) 41–104
6. Nakatani, C., Hirschberg, J.: A Corpus-Based Study of Repair Cues in Spontaneous Speech. Journal of the Acoustical Society of America. **95** (1994) 1603–1616
7. Tseng, S.-C., Liu, Y.-F.: Mandarin Conversational Dialogue Corpus. MCDC Technical Note 2001-01. Institute of Linguistics. Preparatory Office. Academia Sinica. (2001)
8. Xiandaihanyu Cidian. Shangwu Yinshuguan (2001)