Toward a Mandarin-French Corpus of Interactional Data

Helen K.Y. Chen[1,2], Laurent Prévot[1], Roxane Bertrand[1], Béatrice Priego-Valverde[1], Philippe Blache[1]

(1) LPL, CNRS & Aix-Marseille Université
5 Ave Pasteur
13604 Aix en Provence-France

(2) Dept. of Chinese and Bilingual Studies
The HK Polytechnic University
Hung Hom, KLM, Hong Kong

FirstName.LastName@lpl-aix.fr

1. Introduction

This paper provides a preliminary description of the construction of an audio-video speech corpus of face-to-face Mandarin interaction. The corpus consists of 5 hours of two-party, face-to-face, spontaneous Mandarin interaction. Following the construction of the *Corpus of Interactional Data* (CID), an audio-video corpus in French recorded and processed at the Laboratoire Parole & Langage (LPL), Aix-Marseille Universités (Bertrand et al., 2008), the current project proposes the construction of a Mandarin speech corpus that will be annotated, based on a *multimodal* perspective, at various linguistic levels including prosody, morphology, syntax, as well as discourse and non-verbal representations such as gestures. The objective of building such a corpus is to provide a speech resource annotated with wealthy and detailed information at each linguistic level. The eventual goal is to facilitate analyses of conversational interaction from a multimodality perspective.

2. Design of the corpus – the protocol

It is preliminarily proposed that the corpus will consist of 5 segments of two-party, face-to-face Mandarin conversational recordings. Each recording lasts about 1 hour. Thus at the initial stage of the creation of the corpus, the total recording time will add up to 5 hours. The eventual goal is to create a corpus of Mandarin spontaneous conversations of at least 10 hours.

2.1 The subjects

5 males and 5 females are involved in the recordings of face-to-face interaction. All speakers are native speakers of Taiwanese Mandarin. Some of the speakers have grown up bilingually speaking also Taiwanese. Also some speakers may have worked in the same lab where the recording took place and are familiar with each other and the recording environment.

2.2 The task

The participants are invited to join the experiment, in which they are instructed to "have a chat" with another speaker. There are 3 sets of guidelines provided randomly to the participants prior to each recording session: 1. the participants may be instructed to talk about the most difficult things they've encountered while living in France; 2. the participants may be instructed to talk about one of the unusual things s/he has encountered during a recent trip; 3. the participants may not be given any instruction about what to talk about and simply started the conversation on their own. The reason to provide a guideline (especially for guideline 1 and 2) is to enable speakers to be engaged in the interaction rather quickly as soon as the recording starts. Although the speakers may be provided with one of the 3 guidelines as the initial topic for interaction, there was no further instruction about how long the topic should last. The speakers are free to switch to other topics as the interaction carries on.

2.3 The setting of the recordings

Following the original French CID setting (Bertrand et al., 2008), all recording session took place in a soundproof chamber at LPL. The two speakers in each session sat side-by-side and slightly tilted towards each other. During the recording, each subject wore a headset and the voice from each speaker was recorded onto a separated track. As result, the optimal quality of sound files of the spoken data can be obtained for the purpose of detailed annotations on the phonetic and prosodic levels. Moreover, the recordings with two separated sound tracks have the advantage of allowing a more detailed analysis on the content in the overlapped sequences (Bertrand et al., 2008). In addition to the specific setting for the sound recording, subjects were also filmed in long and fixed shot. The video recordings otherwise provide data for non-verbal cues such as gestures.

2.4 The characteristics of the corpus

As result of the aforementioned experimental design and settings, the conversations recorded for the CID corpus is presented as closely simulating the data of naturally occurred, face-to-face interaction. The dialogues of the current corpus resemble daily Mandarin interactions and can serve as a speech resource with rich information on turn-taking and sequential organizations of conversation (Sacks et al., 1974). Such characteristic of the corpus can contribute to the further analysis of the interactions between Mandarin speakers from an

interesting range of perspectives: the correlation between conversational interaction and sound realization (such as at the phonetic or prosodic levels, see Chen 2011), the examination of interaction between the speakers in terms of syntactic constructions or semantic/pragmatic implicatures, and finally, but not the least, the non-verbal cues such as gestures.

3. Levels of transcriptions and annotations

The data will be transcribed and then annotated at various linguistic levels. The following describes the process of transcription and various levels of annotations proposed.

Following Blache et al. (2009), prior to the transcriptions the data will be pre-processed by an automatic segmentation into blocks of sound stream by silent pauses of 200ms. The purpose of the segmentation is to facilitate further orthographic and phonetic/phonological transcription, as well as the alignment of the signals and annotations. The segmented data will then be transcribed orthographically (by using standard Hanyu Pinyin). Additional phonetic and/or phonological transcriptions may be added later on.

At the prosodic level, the annotation scheme distinguishes two levels: a higher level of intonational phrases; and a lower one of interactionally related prosodic cues: including duration, cut-off, lengthening, special voice quality (such as laryngealized voice). For the lower level of prosodic annotations the unit will be the tokens.

Concerning discourse and interaction, we will start from the Turn Constructional Unit (TCU). Following Schegloff (2007), a TCU corresponds to an action in interaction, i.e. a question, an answer, a request, etc. It should be noted that a TCU does not necessarily correspond to a complete sentence; it can correspond to either a lexical item (as a reactive token), or a TCU may consist of more than one sentence. The discourse/interaction information will be annotated by TCU, with notations about the specific action that consists of the TCU: i.e., a question/answer pair, a request or refusal to the request, etc. Moreover a parallel annotation of disfluencies will identify the *reparandum, reparans* and *break interval* in the spirit of Shriberg (1994) (see also Pallaud, 2006, Chen, 2011).

Finally, with regard to the non-verbal gestures, 5 types of gestures will be identified, following McNeill (1992), as well as Chui (2003): *iconic, metaphoric, deictic, spatial* gestures and *beats*. Each type of gesture will be further marked by its *preparation*, *stroke*, and *retraction* (McNeill 1992).

4. Contribution of the corpus

In addition to obvious benefits of such a multimodal richly annotated corpus, the perfect replication of the CID experimental setting will allow for systematic cross-linguistic studies. In addition, the Mandarin CID corpus will be closely related to the Mandarin Conversational Dialogue Corpus (MCDC) (Tseng, 2004). However the CID corpora will be perfectly comparable while the MCDC is different in terms of the experimental settings (e.g the speakers did not know each other when doing the recoding.) Comparisons between these corpora would be nevertheless worthy further exploration.

References

Bertrand et al. (2008). "Le CID - Corpus of Interactional Data - Annotation et Exploitation Multimodale de Parole Conversationnelle." *Traitement automatique des langues* (TAL), vol. 49, no. 3. 105-134.

Blache et al. (2009). "Creating and Exploiting Multimodal Annotated Corpora: The ToMA Project." *Multimodal Corpora*: *From Models of Natural Interaction to Systems and Applications*, Springer.

Chen, K. (2011). *Sound Patterns in Mandarin Recycling Repair*. Ph.D. dissertation. University of Colorado at Boulder.

Chui, K. (2003). Categorization of Gestures in Communication. *Form and Function*: *Linguistic Studies in Honor of Shuanfan Huang*. 105-129. Taipei: Crane.

McNeill, D. (1992). *Hand and Mind: What Gestures Reveal about Thought*. University of Chicago Press.

Shriberg, E. (1994). *Preliminaries to a theory of speech disfluencies*. Berkeley, CA: University of California at Berkeley.

Sacks et al. (1974). "A simplest systematics for the organization of turn-taking for conversation." *Language* 50. 696–735.

Pallaud, B. (2006). Une base de données sur les troncations involontaires de mots en français parlé. *Travaux Interdisciplinaires de Parole et Langage* (TIPA), no. 25. 173-184.

Schegloff, E. (2007). *Sequence organization in interaction: A primer in conversation analysis I.* Cambridge & New York: Cambridge University Press.

Tseng, S-C. (2004). Processing Spoken Mandarin Corpora. *Traitement automatique des langues*. Special issue: Spoken corpus processing 45.89-108.