D64: A Corpus of Richly Recorded Conversational Interaction

C. Oertel, F. Cummins, N. Campbell, J. Edlund, P. Wagner

U. Bielefeld, University College Dublin, Trinity College Dublin, KTH, U. Bielefeld
Corresponding author: nick@tcd.ie

Abstract

Rich non-intrusive recording of a naturalistic conversation was conducted in a domestic setting. Four (sometimes five) participants engaged in lively conversation over two 4-hour sessions on two successive days. Conversation was not directed, and ranged widely over topics both trivial and technical. The entire conversation, on both days, was richly recorded using 7 video cameras, 10 audio microphones, and the registration of 3-D head, torso and arm motion using an Optitrack system. To add liveliness to the conversation, several bottles of wine were consumed during the final two hours of recording. The resulting corpus will be of immediate interest to all researchers interested in studying naturalistic, ethologically situated, conversational interaction.

1. Introduction

The D64 Multimodal Conversational Corpus has been collected to facilitate the quasi-ethological observation of conversational behavior. Conversational interaction in person is a fully embodied activity (Cassell et al., 1999). The role of posture, eye gaze, torso movement, head rotation, hand and arm gestures all contribute to the dynamic establishment, maintenance, and dissolution of domains of joint attention (Baldwin, 1995). Little is currently known about the structure of such transient collaborative domains, and how they might be indexed. However it is clear that the felicitous participation in any natural human-human conversation demands attention to a host of subtle movement cues that permit the ephemeral coupling among participants that constitutes conversational ebb-and-flow.

There is widespread agreement that the empirical investigation of conversational interaction demands multimodal data (Massaro and Beskow, 2002). This is important, both in furthering our understanding of naturally occurring human-human interaction, and in the development of systems that are required to interact in a human-like fashion with human speakers (Edlund et al., 2008). Along with audio recordings, it is now commonplace to include video recordings of at least the faces of conversation participants (van Son et al., 2008). Speech is, however, thoroughly embodied, and unfettered conversational behavior includes appropriate manual gesturing, torso positioning, head direction, gaze behavior, blinking, etc. Furthermore, conversation is often carried out in a dynamic context, with free movement of the participants, change over time in the set of conversational participants, and with an openness that is entirely lacking from most careful studio recordings.

The D64 Multimodal Conversational Corpus has been designed to collect data that transcends many of these limitations. It has been designed to be as close to an ethological observation of conversational behavior as technological demands permit (see also Douglas-Cowie et al., 2007). We first outline the recording setup, the planned model of distribution, and finally, some of our initial aspirations in the analysis of the rich data that results.

Figure 1: The apartment room in which all recording was conducted.

2. Recording

The recording setup chosen for the data collection described is built on the following premises:

[1] The setup ought to be as naturalistic as possible, whereby ”naturalistic” is taken to mean a recording situation that is radically different from a typical laboratory recording, carried out in a recording booth or anechoic chamber with speakers sitting or standing in carefully controlled positions. Instead, a naturalistic recording situation approximates a conversational situation speakers may experience in their daily lives. A scale for different degrees of naturalistic settings is sketched in Fig. 2. The motivation for this decision was to remove as many behavioural artefacts as possible resulting from placing the speakers in laboratory conditions. As laboratory settings are conventionally employed in the hope of removing as many confounding variables as possible, our decision deliberately allows for all kinds of unexpected effects that might influence our data collection.

[2] Unlike most corpus recordings (e.g. map tasks, tourist information scenarios etc.), the chosen setup was not task oriented. No agenda or set of topics was provided. The motivation behind this was to allow the speakers to focus on language use for the purpose of social interaction. In

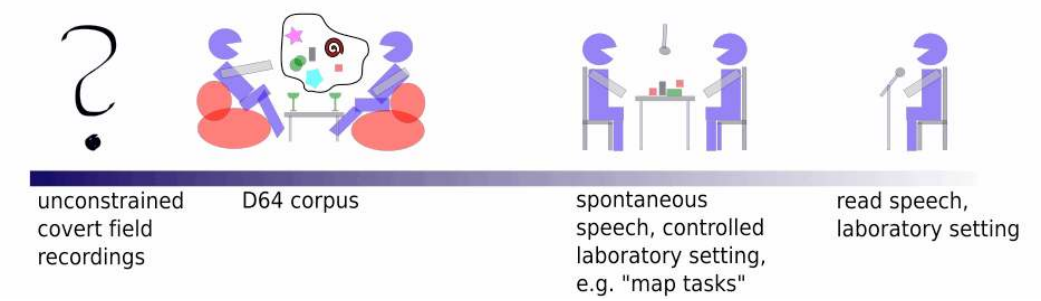

Figure 2: Spectrum of observation scenarios ranging from highly controlled to truly ethological.

task oriented dialogue, the goal of linguistic exchange is the collaborative achieve of a particular goal set by the task, e.g. to receive a particular kind of information or make an appointment. Certainly, social interaction does play an important role in task-oriented dialogue as well, but it is expected to do so to a lesser degree.

[3] Since the speakers knew that they would be recorded and filmed, our setup does not control for the observers paradox (Labov, 1997). However, it has at least the following desirable properties:

- The conversation is interpersonal, with an active and involved other (NOT just a "listener"!);
- It is both social and spontaneous;
- Participants were free to move around, or even leave;
- Speech is unprompted and unscripted;
- Recordings were made over a long period (8 hours over 2 days) thus helping to avoid stereotypical role playing;
- Subjects shared many common interests, and subjective impressions of the interaction were that it was unforced.

Figure 1 shows the domestic apartment room in which all recording was conducted. A mid-sized room with conventional furniture, with a sofa and some comfortable chairs arranged around a low coffee table was employed. Recordings were made over two days, each session being approximately 4 hours long, although the length of the corpus that will ultimately be made available has yet to be precisely determined. The first session was split into two two-hour sessions with an intervening lunch break, while the recording on the second day was continuous over 4 hours. Five participants (the first four authors and a friend) took part on Day 1, and just the 4 authors on Day 2. In order to liven up proceedings somewhat, several bottles of wine were consumed during the latter half of recording on Day 2. Participants were free to stand up, use the adjoining kitchen, change places, etc throughout. In the same spirit, no attempt was made to constrain the topic of conversation, and subject matter varied widely from technological detail (inevitable under the circumstances) to pop culture and politics.

Seven video cameras were employed. There was at least one camera trained on each participant (or one on the sofa as a whole, accommodating two participants). There were also two 360-degree cameras that captured the entire conversational field at a lower resolution. Audio was captured using both mainly wireless microphones (both head-mounted and lapel), along with a variety of strategically placed wide-field microphones. In addition, reflective markers (3 on the head, 1 on each elbow and shoulder and one on the sternum) were monitored by an array of 6 Optitrack cameras.

3. Post-processing

Figure 3: Sample flash interface. Speakers spoken contributions are color coded at the top. Several alternative displays are possible, this one being especially popular. For details, see Campbell and Tabata (2010).

Post-processing of the large amount of data is ongoing, including fragmentation into manageable chunks, cross-channel synchronization, and initial annotation. The data will ultimately be released in clearly indexed chunks of approximately 15 minutes duration, with a transparent indexing by both speaker and source. The entire corpus will be released under the Creative Commons Attribution-Noncommercial-Share Alike License. The raw data is al-

ready available for collaborative annotation, and the aspiration of the project team is that researchers availing of the data will contribute their annotations and other relevant information back to a central repository where others can make use of it. Interested parties may obtain the raw data from the URL www.speech-data.jp/nick/mmx/d64.html, and the project team request that they be informed of any annotation conducted.

As well as the video, audio, and motion-capture data files, the data will be presented in a custom-built flash interface that will allow the user to view, browse, search and export arbitrary subsets of the D64 corpus (Campbell, 2009). The graphical layout will make it particularly easy to search utterance sequences based on dialogue structure and speech overlaps. Each utterance will be accessible by mouse-based interaction. Moving the mouse over a bar will reveal its text, and clicking on the bar will play the speech recording associated with that specific utterance. Each speaker's data will be shown using a different colour to aid identification. Figure 3 shows the kind of flash interface envisaged, as applied to another set of conversational recordings.

Two modes of audio output will be offered for dialogue speech, since it is sometimes preferable to hear a stereo recording, which provides simultaneous natural access to two speakers' overlapping segments, while sometimes it is preferable to hear a monaural recoding, where overlapping speech does not intrude. Separate speech files can be employed in each case. Rapid and more detailed search facilities will ultimately be included. A Join-Play interactive-editing feature will allow the user to simply append the latest utterance segment (video and audio, or audio alone) to a list of related segments to build up a novel data sequence with the speech files and associated text files zipped in a form ready to burn to DVD for wider distribution.

4. The Ebb and Flow of Joint Interaction

Quasi-ethological conversational data of the sort provided by the D64 corpus have not been widely available. With rich capture of visual, audio, and movement data in a naturalistic setting, opportunities arise for the annotation and observation of both quantitative and qualitative aspects of the conversation, in a manner not otherwise possible.

It is our contention that domains of joint interaction arising in a naturalistic conversation are different in an important sense from any joint properties of the participants considered separately. This is graphically illustrated in the well-known experimental work of Murray and Trevarthen (1986) who had mothers and babies interact in real time over a video link. Infants (2 months old) were happy to interact with their mothers through this live link. However, if the infants were shown a prior recording of their mother in interaction with them at an earlier point in time, they objected and immediately withdrew from the exchange. The infants were exquisitely sensitive to the real-time push-and-pull of social interaction, and were not fooled for a moment by a recording that was incapable of responding to their own infantile selves. This work clearly illustrates that there is a meaningful coupling between the mother and infant that is not comparable to the sum of mother+infant.

The task of identifying empirical correlates of this kind of interactional fabric is a daunting challenge. As a first foray into the territory, we propose to attempt to annotate much of the D64 data using two novel quantitative scales that will need to be calibrated and assessed, to see if they may be of use in documenting the ebb and flow of joint interaction. Both variables we will use will initially be based on subjective assessment by trained observers. They will provide subjective ratings of the overall conversational *arousal* and the pairwise *social distance* between participants.

Arousal This variable is hypothesized to index the *joint* arousal of the entire group. Thus, when whole-hearted laughter breaks out all around, for example, we would note a relatively high degree of arousal, while boredom, or indeed silence, would be at the other end of the scale. These examples hide a deal of complexity, however. For example, nervous laughter may reflect a stagnation of the conversation, and thus receive a low arousal rating, and, conversely, a highly pregnant pause may be associated with high joint arousal. For this variable to index a coherent property of the group dynamic as a whole, it is necessary that there be a single conversation, rather than several, relatively independent, conversations. Natural conversation is very fluid, and there is no guarantee that arousal, as envisaged here, will be continuously documentable. Rather, an arousal rating will be provided for successive 5 second frames just in case all participants are mutually engaged.

Social Distance Social distance is a pairwise variable, which is expected to be at a relatively high level for most dyads, most of the time, but to decrease as two participants attend jointly, or engage in reciprocal interaction. We adopt a convention where a low distance value corresponds to relatively intense pairwise interaction, and a high value reflects the perception of greater distance by the annotator. An example of low rated distance would be where two people look at each other, smile at each other and address each other in conversation. The conversation does not need to be of a friendly kind. Two people having an argument would be rated low (close) just as two people confessing their love to each other. Another instance in which social distance would be indexed as relatively low would be when two people display the hallmarks of joint attention, in that they have the same point of focus, look at the same object for a rather long period of time and have the same body posture or move their heads in the same moments. In contrast, a scene in which two participants look in different directions and seem to be interested in different events would be rated as relatively high with respect to social distance.

Both of these proposed scales are highly speculative. It is not yet known whether a sufficiently high-degree of inter-rater reliability will be obtainable, even after considerable refinement of the criteria employed by annotators. Initially, annotators are being asked to base ratings on a combination of such observable characteristics as posture, torso-facing, eye gaze, head rotation, simultaneity of movement, etc. Importantly, annotators are required to use observable characteristics of the scene, and not linguistic interpretation, in their ratings. Ratings are on a scale from 1 to 10, and we freely acknowledge that there will be a period of calibration required in order to arrive at rating guidelines, based on ob-

servables, that lead to a relatively consistent evaluation of the character and dynamic of joint interaction. To bootstrap the process, a selection of extracts from the recordings will be made available on a website, and will be independently rated by at least 10 raters each. Feedback will be obtained about the observable features considered to be of most use, and inter-rater reliability will be assessed using Krippendorff's Alpha (Lin, 1989)

A second line of investigation we have been pursuing is harder to document in a static document, as it involves observation of simultaneous real-time movement of several participants at once. Several quite striking examples of simultaneity of movement of two participants have been observed, and can be viewed at http://tinyurl.com/yk2q34d. Much as spectators at a tennis match can be observed to display head movement locked to the to-and-fro of the tennis ball, so too listeners can be observed to be coupled to the ongoing flow of social interaction. Simultaneous onset of head nodding, whole body turning, etc are readily observable in the data, and are most clearly seen when the observer does not attend to the linguistic content of the ongoing discussion. We have found that the simple expedient of observing the data at a faster rate than normal, with the sound turned down, helps greatly in attending to the embodied participation of participants in the ongoing ebb and flow.

These two examples illustrate both the opportunity for rich observation, and the challenge in documenting conversational interaction as a rich form of human behavior extending far beyond mere linguistic content. The coordination of behavior in conversation has recently been described as *participatory sense-making* within the enactive tradition (De Jaegher and Di Paolo, 2007). In this approach, the process of interaction is seen as the basis for the creation, maintenance and transformation of domains of autonomy. The dimensions of social distance and arousal we have identified above may index the process by which interaction moves from the coordination and sense-making of distinct individuals to the joint process of participatory sense-making.

5. Discussion

Naturalistic data collection on the scale employed here has not hitherto been generally feasible. The utility of such large-scale oversampling will depend, to a great extent, on the usability of the web-based interfaces employed in the dissemination of the corpus. Conversely, with such rich data, it is not possible to anticipate with any certainty the kind of annotation, or the variables annotated, by specific research groups. While we have suggested some novel ways of potentially indexing the dynamics of conversational interaction, our plans here are highly speculative, and the variables, as yet, untested. We hope that the availability of multiple points of view, along with motion capture data, and extensive audio recording, will encourage other groups to consider new and ambitious ways of interpreting conversation in a natural setting.

Acknowledgements

This work has been supported by grants to Nick Campbell from the Visiting Professorships & Fellowships Benefaction Fund from Trinity College Dublin, and the Kaken-B Fund for Advanced Research from the Japanese Ministry of Information, Science & Technology, and also Science Foundation Ireland, Stokes Professorship Award 07/SK/I1218. Jens Edlund is supported by The Swedish Research Council KFI – Grant for large databases (VR 2006-7482). Catharine Oertel is supported by the German BMBF female professors programme (Professorinnenprogramm) awarded to Petra Wagner. Finally, thank you to Nike Stam for her generous participation.

6. References

D.A. Baldwin. 1995. Understanding the link between joint attention and language. *Joint Attention: Its Origins and Role in Development*, pages 131–158.

N. Campbell and A. Tabeta. 2010. A software toolkit for viewing annotated multimodal data interactively over the web. In *Proc. LREC*.

N. Campbell. 2009. Tools & Resources for Visualising Conversational-Speech Interaction. *Multimodal Corpora: From Models of Natural Interaction to Systems and Applications*, page 176.

J. Cassell, T. Bickmore, M. Billinghurst, L. Campbell, K. Chang, H. Vilhjálmsson, and H. Yan. 1999. Embodiment in conversational interfaces: Rea. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems: the CHI is the limit*, pages 520–527. ACM New York, NY, USA.

H. De Jaegher and E. Di Paolo. 2007. Participatory sense-making: An enactive approach to social cognition. *Phenomenology and the Cognitive Sciences*, 6(4):485–507.

E. Douglas-Cowie, R. Cowie, I. Sneddon, C. Cox, O. Lowry, M. McRorie, J. Martin, L. Devillers, S. Abrilian, A. Batliner, et al. 2007. The HUMAINE database: Addressing the collection and annotation of naturalistic and induced emotional data. *Lecture Notes in Computer Science*, 4738:488.

J. Edlund, J. Gustafson, M. Heldner, and A. Hjalmarsson. 2008. Towards human-like spoken dialogue systems. *Speech Communication*, 50(8-9):630–645.

W. Labov. 1997. Some further steps in narrative analysis. *Journal of Narrative & Life History*, 7(1-4):395–415.

LI Lin. 1989. A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, 45(1):255.

D.W. Massaro and J. Beskow. 2002. Multimodal speech perception: A paradigm for speech science. *Multimodality in Language and Speech Systems*, pages 45–71.

L. Murray and C. Trevarthen. 1986. The infant's role in mother-infant communications. *Journal of Child Language*, 13(1):15–29.

R. van Son, W. Wesseling, E. Sanders, and H. van den Heuvel. 2008. The IFADV corpus: A free dialog video corpus. In *International Conference on Language Resources and Evaluation*.