Lang Resources & Evaluation (2015) 49:195–214
DOI 10.1007/s10579-014-9283-2

PROJECT NOTES

InSight Interaction: a multimodal and multifocal dialogue corpus

Geert Brône · Bert Oben

Published online: 27 September 2014

Abstract Research on the multimodal aspects of interactional language use requires high-quality multimodal resources. In contrast to the vast amount of available written language corpora and collections of transcribed spoken language, truly multimodal corpora including visual as well as auditory data are scarce. In this paper, we first discuss a few notable exceptions that do provide high-quality and multiple-angle video recordings of face-to-face conversations. We then present a new multimodal corpus design that adds two dimensions to the existing resources. First, the recording set-up was designed in such a way as to have a full view of the dialogue partners' gestural behaviour, including hand gestures, facial expressions and body posture. Second, by recording the participant perspective and behaviour during conversation, using head-mounted scene cameras and eye-trackers, we obtained a 3D landscape of the conversation, with detailed production information (scene camera and sound) and indices of cognitive processing (eye movements for gaze analysis) for both participants. In its current form, the resulting *InSight Interaction Corpus* consists of 15 recorded face-to-face interactions of 20 min each, of which five have been transcribed and annotated for a range of linguistic and gestural features, using the ELAN multimodal annotation tool.

Keywords Multimodal interaction · Video corpus · Head-mounted eye-tracking · Multifocal approach

G. Brône (✉)
Department of Linguistics, University of Leuven, Campus Antwerp, Sint-Andriesstraat 2, 2000 Antwerp, Belgium
e-mail: geert.brone@arts.kuleuven.be

B. Oben
Department of Linguistics, University of Leuven, Louvain, Belgium
e-mail: bert.oben@arts.kuleuven.be

1 Introduction

Recent years have witnessed an interesting convergence of interest in a variety of research disciplines, viz. the modelling of multimodal aspects of communicative interaction. Programs in human–computer interaction, gesture research, conversation analysis, cognitive psychology and linguistics are addressing questions pertaining to the interaction of and trade-off between different modalities. Although these different traditions take a (sometimes radically) different perspective on multimodality, they all draw on visual as well as auditory input from interactional sequences as a basis for scholarly inquiry.

Researchers focusing on multimodal aspects of interactional language use require high-quality multimodal resources to get a fine-grained view of the trade-off between verbal, visual and bodily features in communication (Massaro and Beskow 2002). Such resources are, however, notoriously hard to come by. Compiling, annotating and processing a corpus of audio–video recordings of conversational speech is a time-consuming and complex undertaking even for a relatively restricted set of conversational data. The reason is that in contrast to purely text-based corpora (of both written and spoken language use) there are significantly less (semi-) automatic annotation procedures for visual and auditory input, and for some features, there are no standardized coding schemes that are widely used (e.g. gaze, gesture,[1] head movements, etc.). As a consequence, interaction analysis for a long time resorted to transcripts and audio data alone. Despite the technical and methodological difficulties, however, a number of recent projects have started to collect and annotate video recordings of conversational speech in order to meet the growing need for multimodal dialogue corpora.

The present paper presents the design, recording set-up and annotation structure of a newly developed corpus consisting of video recorded dyadic interactions. The central goal of the project was to add a dimension to the static external camera perspective that is generally used in video corpora (with a profile or frontal shot of the interlocutors). What is generally lacking in existing corpora is a 'user perspective' of each of the conversation partners, placing the analyst in the position of one of the (or both) interlocutors, and hence tracking face, gaze, hands and body of the other from a subject-tied perspective. In this paper, we present the output of a multi-angle recording technique, which combines data from an external camera with those of state-of-the-art head-mounted cameras and eye-trackers worn by both interlocutors in a dyadic interaction. The result is a rich representation of the interactional 'landscape', which provides access to a wide variety of multimodal cues, both from the perspective of production and processing.

The paper is structured as follows. Section 2 presents an overview of some of the existing video corpora of interactional language use, with a specific focus on their potential for multimodal research. Section 3 discusses the role of gaze in interaction and the possibilities to include visual distribution of attention in multimodal

[1] Unless explicitly mentioned differently, we use *gesture* in this paper to refer to hand gestures, i.e. we do not use *gesture* as a cover term to refer to any type of body movement, such as head movement, posture or eye movements.

dialogue corpora. Section 4 builds on insights from Sects. 2 and 3 and presents the InSight Interaction Corpus, a multimodal corpus of dyadic interactions with a strong focus on language users' individual perspectives and (gaze) behaviour. Section 5 provides a general discussion and proposes avenues for the future development of multimodal interaction corpora

2 Video dialogue corpora

The growing interest in multimodal features of dialogue, as well as the technical development of multimodal corpora, has recently led to a number of projects that have started to collect and annotate video recordings of conversational speech, with the aim to provide a fine-grained data collection for theoretical, descriptive and experimental work. For the purpose of the present paper, we discuss a selection of these projects, with a specific focus on the recording set-up and design. It needs to be stressed that we do not aim to provide an exhaustive overview of existing multimodal dialogue corpora, nor a fully-fledged critical evaluation of the selected examples (see Allwood 2008; Knight 2011 for overviews).

One of the key features in developing a video dialogue corpus is the recording set-up that is used to capture multimodal speaker behaviour. Depending on the level of detail one wants to obtain, interactions can be recorded using one, two or multiple cameras. An example of a corpus using a single-camera perspective is the CID Corpus of Interactional Data for French (Bertrand et al. 2008; Blache et al. 2008) (see Fig. 1). Apart from systematic corpus projects such as CID, researchers working on different aspects of multimodality have resorted to video recordings of this type, yielding a static external perspective of the interaction, with interactants captured either in profile (Gerwing and Allison 2009; McNeill 2008; Pine et al. 2004) or facing the camera (Kimbara 2006).

Using a single external perspective on the speech situation has the disadvantage that it restricts access to gesture and gaze. In the case of recordings using a profile perspective the analyst does not have a fully unblocked view of the interactants' face and body, which significantly reduces the analytical potential of such data [as noted among others by Streeck (2009)]. The alternative with conversation partners facing the camera produces a situation that is unnatural to the conversationalists, as it deviates from prototypical *face-to-face* setting, and obviously limits the possibilities of studying the role of gaze in dialogue (cf. Sect. 3). One way to overcome this problem is to combine two or more camera perspectives to provide access to the interaction from multiple angles, and thus to allow a more fine-grained and reliable analysis of behavioural features. Several corpus projects have adopted a multi-angle approach, with either a speaker- or a scene-oriented focus. In a speaker-oriented design, the primary focus is on capturing the individual speakers in as much detail as possible. In a scene-oriented setting, the cameras are set up in such a way that the analyst obtains a 360° perspective on the interactional landscape (or an approximation thereof).

Fig. 1 Frontal shot in the CID corpus

Examples of speaker-oriented multi-angle corpora are the NMMC Nottingham Multimodal Corpus (and more specifically the data recorded as part of the HeadTalk project, Knight et al. 2008, 2009), the Spontal corpus (*Swedish Spontaneous Dialogue Corpus*, Edlund et al. 2010) and the IFADV corpus (*IFA Dialogue Video Corpus*, Van Son et al. 2008). For the dyadic interactions in the NMMC corpus, conversations in an academic setting (e.g. between students and mentors) were recorded from two perspectives, with two cameras directly facing the participants (Fig. 2a). The resulting recordings were used as input for computer vision techniques (including a 3D head tracking model), which allow for the (semi-) automatic recognition of specific gesture types (and more specifically head nods). The IFADV corpus consists of annotated video recordings of spontaneous and non-directed face-to-face conversations in Dutch. Comparable to the HeadTalk corpus, the interactions were recorded using two cameras positioned next to the speakers and facing the other (Fig. 2b). The recorded data were annotated automatically for a range of parameters (e.g. POS tagging and word alignment) and manually for the pragmatic function of utterances (e.g. reactions, grounding acts, etc.) and gaze direction.

A scene-oriented recording technique was adopted in the D64 corpus (Campbell 2009), the VACE multimodal meeting corpus (Chen et al. 2006), the NOMCO corpus (Paggio et al. 2010) and the UTEP-ICT Cross-Cultural Multiparty Multimodal Dialogue Corpus (Herrera et al. 2010). In all four examples, two- or multiparty interactions were recorded using multiple cameras (up to ten in the case of the VACE corpus), as shown in the recording layout for the D64 and VACE corpus in Fig. 3a, b. In the case of the D64 and VACE corpus, visual tracking systems were applied to include information on position and movement of torso, head and hands of the recorded speakers. In the UTEP-ICT Cross-Cultural

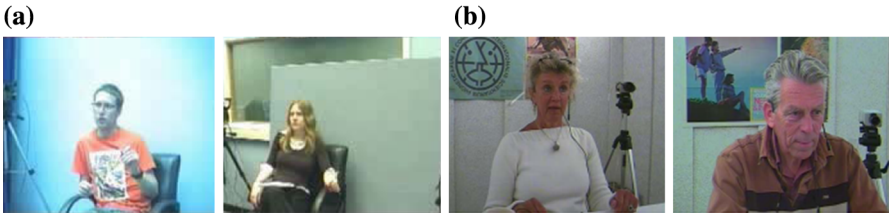

Fig. 2 **a** Double camera perspective in the NMMC Nottingham Multimodal Corpus. **b** Double camera perspective in the IFADV corpus

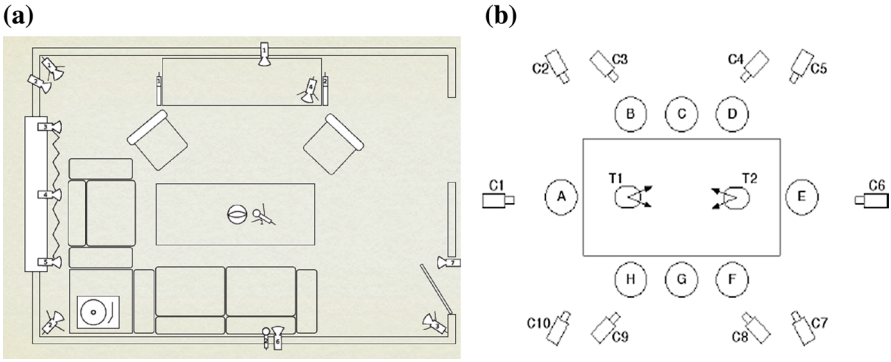

Fig. 3 **a** Recording configuration for the D64 corpus. **b** Recording configuration for the VACE corpus

Multiparty Dialogue Corpus, the recorded interactions were coded for features of turn-taking, gaze and proxemics.

One dimension of multimodal speaker behaviour that is included in several of the above-mentioned corpora, but which poses a significant challenge in the corpus annotation process, is eye gaze. Studying gaze behaviour on the basis of 'external' video data of a participant's face is problematic as it is impossible to pinpoint the exact focus of attention (Kendon 2004; Paggio et al. 2010; Streeck 2009). For instance, the VACE corpus, the IFADV corpus and the UTEP-ICT Cross-Cultural Multiparty Multimodal Dialogue Corpus all include gaze target estimations on the basis of manual coding. One way to overcome this problem of reliability and accuracy is to track participants' gaze behaviour during face-to-face conversation, using head-mounted eye-tracking equipment. Before we turn to the operationalization of gaze behaviour in our InSight Interaction Corpus, we briefly address the role of gaze in conversation.

3 Gaze in interaction

It has long been acknowledged that eye gaze is an important measure for cognitive processing and production features in language use. In the last 50 years, the

measuring of gaze points and eye movements during online behaviour, the so-called eye-tracking technique, has established itself as an instructive paradigm for studying a range of phenomena in psycholinguistics and psychology (see Rayner 1998 for an extensive review). The vast majority of studies using the paradigm explore eye gaze patterns as a measure of cognitive processing on the part of an individual cognizer who is given a physical stimulus, like e.g. a text or image projected on a computer screen. Significantly less addressed is the exact communicative function of gaze on the production side (e.g. as a referential instrument) and its relation to co-occurring utterances (Bavelas et al. 2002; Staudte et al. 2011).

With the development of unobtrusive eye-tracking equipment in the last decade, new opportunities have arisen to inquire into this hitherto underexposed potential of the paradigm. Eye gaze is not only indicative of comprehension processes, but can function as a strong (intentional) communicative instrument as well. By broadening the perspective from processing to production, the traditional scope may also be extended from a unidirectional setting (e.g. an individual reading a text or looking at a scene) to a more basic setting of communication, viz. face-to-face communication. Mobile systems, like e.g. head-mounted eye-trackers, allow for a fine-grained view on visual attention distribution in conversation, both on the part of a speaker and a hearer. This opens up a vast area of research on multimodal interaction, including the role of gaze as a directive instrument, the correlation between gaze and gesture, gaze as a disambiguation instrument, interactive alignment in various semiotic channels, etc.

To date, however, the potential of unobtrusive eye-tracking has been largely unexplored, with a few notable exceptions. Brennan and colleagues studied (a) the role of shared gaze, based on pairs of participants wearing eye-trackers and performing collaborative tasks (Brennan et al. 2008), and (b) the impact eye gaze has on the disambiguation of referring expressions in spontaneous dialogues (Hanna and Brennan 2007). Hadelich and Crocker (2006) measure eye–eye span in spontaneous interactions, defined as "the time difference between the last fixation by the speaker to a referent before the onset of the respective referring expression and the first fixation by the listener to the same referent after the respective referring expression" (ibid.: 1). A reduction of the eye–eye span may serve as an expression of convergence between speakers. Jokinen and collaborators explored the role of non-verbal signals for turn-taking and feedback in face-to-face conversations. In these studies, participants were eye-tracked in both two-party (Jokinen et al. 2009) and multi-party conversations (Jokinen 2010). The results suggest that non-verbal elements (like gesture, head movements and gaze) function primarily as "indexical signs rather than meaning carrying symbols and their interpretation is linked to the whole context in which they occur" (ibid.: 2965).

The studies focusing on gaze in interaction stress its function as a referential mechanism and the potential of eye-tracking to provide further insights into the tight coupling of processing and production in interaction on the one hand, and the integration of different sources of information over time (and in parallel) in discourse on the other. In order to address these questions more systematically, researchers may benefit from dialogue corpora that are both multimodal and multifocal in nature, providing detailed access to gaze behaviour of two (or

multiple) interlocutors during task-based or spontaneous interactions. In the next section, we present the InSight Interaction project, which was conceived as a first attempt at such a multifocal corpus that captures the participant perspectives and behaviour during dialogue, using head-mounted scene cameras and eye-trackers.

4 The InSight Interaction Corpus

This section presents the basic features of the InSight Interaction Corpus, an innovative multimodal and multifocal dialogue corpus. In three subsections we present the task design and recording set-up that led to the recorded slices of interaction (4.1), the multimodal annotation scheme that was used to code relevant aspects of speech, gesture and gaze (4.2) and the resulting corpus that is made available to the research community (4.3).

4.1 The recordings

4.1.1 Dual task design

From the onset of the project, the goal of recording the corpus has been twofold. On the one hand we wanted to compile a dataset to study specific aspects of interactional dynamics pertaining to the tight relationship between processing and production in face-to-face dialogue discussed above, and described among others in Pickering and Garrod's *interactive alignment theory* (Pickering and Garrod 2004, 2006). On the other hand, the recording and annotation design was set up in such a way as to provide a novel and relevant perspective for other topics, and researchers the like. To meet both needs, we differentiated in the type and complexity of the recorded interactions. In a first part of the recording session we asked participants to perform targeted collaborative tasks because these allow for appropriate trials and baseline conditions to emerge from the interactive dialogue: by eliciting specific types of utterances one can compare parameters and variables across subjects (Tanenhaus and Brown-Schmidt 2008). In the second part of the session, we let the participants have free-range conversations because these yield the most natural picture of multimodal interaction.

For the purpose of the present paper, we do not need to focus on how the corpus was designed to capture specific aspects of multimodal interaction particularly relevant to our own research interests (more specifically interactive alignment). It suffices here to note that the targeted collaborative tasks were designed around short animation clips which were shown to the participants and which they needed to describe to one another. Some general arguments that shaped our final task design, however, do warrant further discussion.

In a series of pre-tests, we checked to see if the task design generated sufficient dynamics between the co-participants. In a first version of the set-up, we designed a test in which one participant was shown a spatial scene on a computer screen, and was then asked to describe that scene to the co-participant. The latter was then

shown two or more possible 'solutions' and was asked to indicate the scene just described. The participants took turns in their active role as 'describer' versus 'judge'. The pre-test using this design revealed, however, that the co-participants tended to look at their screens for most of the task. Because accurately tracking the gaze behaviour of conversational partners was one of the main goals of this corpus, the substantial focus on the screen was not desirable. We solved the problem by eliciting some kind of conflict and by showing the animations simultaneously rather than by turns. In the final set-up the participants were simultaneously shown a simple animation on two screens, one for each participant. These animations were identical except for one or two features, and the participants were asked to discuss which were the relevant differences. Because the subjects no longer saw the same animations and because the animations were shown at the same time, there was a clear-cut boundary between animation and interaction. Moreover, participants interacted more naturally and looked at each other because they used non-verbal cues to express the content of the animation as well.

After a set of collaborative tasks we recorded the participants as they engaged in free conversation. To get the conversation going, however, we asked them to talk about a specific topic. The participants were entirely free how rigorously they stuck to the given topic. In the final set-up that topic was 'what should designers take into account when creating a new cell phone branded for women?' The results of the pre-tests revealed that participants relate to this topic and enjoy talking and fantasizing about it.

4.1.2 The recording set-up

4.1.2.1 Recording room A screenshot of the recording configuration is presented in Fig. 4. The pairs of participants sit opposite each other with no objects within their reach (this to avoid parts of the body being concealed on the video images and to maximize the freedom for hand gestures). Behind each of the subjects is a large screen on which the animations used for the collaborative tasks are projected. The video data that build the core of the multimodal corpus come from a fixed camera which records the ongoing conversation in profile (left image in Fig. 4), and two mobile cameras integrated in the head-mounted eye-trackers which record a full frontal image of each of the conversation partners (right images in Fig. 4). The green dots[2] on the right images are generated by the eye-trackers and indicate very precisely where the interlocutors are gazing at, i.e. where they are visually fixating on. At no point (not even during the calibration of the eye-trackers) do the participants see their own or their partner's eye gaze behaviour on the screens of the recording laptop computers. Although there is a wire connecting the eye-trackers to the computers, the subjects are free to move and gesticulate. They do not need to restrict themselves to a certain position or virtual frame.

[2] The mobile eye-trackers provide two types of data: video files from a scene-camera, and data files on the basis of the eye-movements (containing simple x and y co-ordinates that together constitute the exact location of the fixation point, at a rate of 30 Hz). The right images in Fig. 4 are an overlay of the video files from the scene-camera with the gaze co-ordinates from the data files.

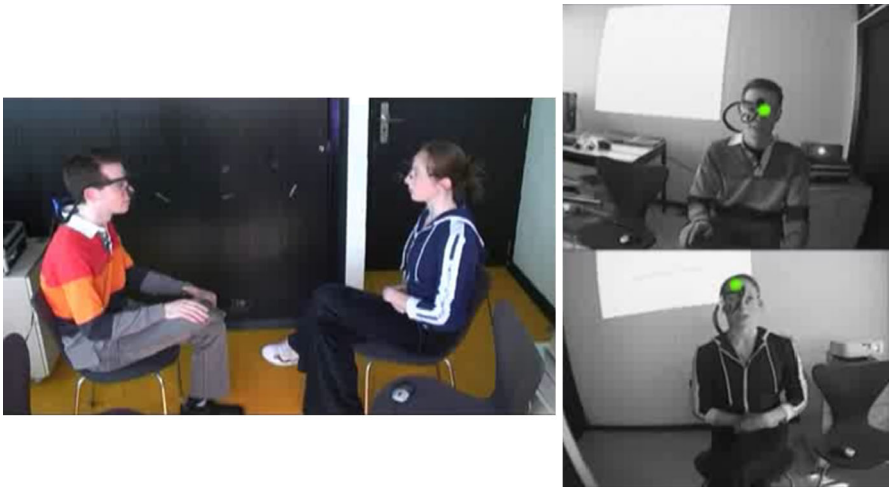

Fig. 4 Recording configuration

4.1.2.2 Recording devices Our recording set-up required us to record video from three perspectives (three cameras), eye-movements for each of the participants (two eye-trackers) as well as the audio signals (two microphones). Below is a list with some of the technical specifications of the gear used for all of the experiments.

- 1 fixed colour camera
  - Sony HDR-FX1000E
  - 25 frames per second
  - 720 × 576 pixels
- 2 head-mounted eye-trackers, with scene camera included
  - Arrington Gig-E60 eye-tracking frames with flexible cameras for different camera angles. This allowed us to adjust the cameras in such a way that the subjects saw their partners' eyes at all times
  - 30 frames/events per second
  - 320 × 240 pixels
- 2 microphones
  - Zoom H2: directed
  - Both microphones recorded in the 16bit/44.1 kHz WAV format

4.1.2.3 Participants The participants are all Belgian undergraduate students from the University of Leuven and Lessius University College and native speakers of Dutch. In the 15 successful recordings, there are 9 men and 21 women from various regions in Flanders. Each of the participating pairs is well-acquainted. This high level of acquaintance makes them share a lot of personal and conversational history, which given the unnatural lab setting, made them more relaxed and less intimidated

by the experimental circumstances. Although the data are not fully spontaneous, given the lab setting and the strictly defined tasks in the first part of the recording, there are nevertheless sufficient arguments to categorize the data as examples of naturally occurring interactions:

- The tasks create an external trigger to communicate and even impose a conversational topic, but the internal dynamics of the conversation is entirely free.
- There is no intermediate person or factor interfering with the ongoing discourse: the participants decide for themselves how long they talk, what their strategy is to perform the tasks, what they talk about apart from the ongoing task, etc.
- The final part of each recording session, which consisted of free conversations revolving around a given topic (supra), gave very satisfactory results. After ten minutes in the lab setting participants seemed relaxed, not paying attention to the recording devices and started chatting and joking cordially about topics totally unrelated to the experiment.

4.1.3 Synchronising the data

When working with time-aligned data coming from five different sources, synchronisation is crucial. Because we only needed to cross-check the waveforms, synchronising the fixed camera with the microphones was straightforward. We used the editing tool Adobe Premiere Pro to perform this first synchronisation: it simply sufficed to load the video file from the fixed perspective and the audio files from the microphones, play them at the same time and listen whether there was any echo or even larger difference between the three files. In none of the cases further manual adaptation was needed. Each of the recordings ran 100 % in sync for the entire duration of the recording session.

Synchronising the video data from the eye-trackers was more time-consuming because there were a number of dropped frames and no exact information on where those frames were dropped.[3] To avoid a post-processing frame-by-frame analysis, we adopted a workable methodology: we looked for at least two anchor points per minute at which we checked the synchronisation between the three video files.[4] The data from the fixed camera was regarded as our fixed starting point or baseline and was left unchanged. If frames were dropped in the eye-tracking data, the corresponding span between two anchor points was 'stretched' so as to synchronise with the span in the baseline. This way we made sure that the dropped frames did not accumulate into a noticeable and undesirable time lapse between the video signals.

The three separate video tracks can be accessed and used separately but for ease of use we edited a file where we combine the three recording angles into one 'trivid' video file as well (supra, Fig. 4). The advantage of that combination is that it

[3] The average length of the video files was 6.36 min. The average number of dropped frames per video file was 38, with nearly all of the dropped frames occurring in clusters of 3–7 frames.

[4] On average we had an anchor point in the video files from the eye-trackers every 21.38 s. The exact number and position of the anchor points depended on the content of the video data: the onset or offset of hand gestures were particularly frequently used as anchor points because those actions were clear signals in each of the video files.

provides a better overview of the interaction as a whole and that the processing cost can be reduced when using annotation tools. Because the 'trivid' is time-aligned with the three other videos, it is possible to switch between perspectives at any time during the annotation process. It might for example be useful to use the 'trivid' for a coarse annotation and work on the details in a next phase with one of the more close-up video perspectives or separate audio files.

4.2 The annotation

4.2.1 The annotation tools

4.2.1.1 ELAN For the transcription and annotation of the video data we used the audio and video annotation software ELAN (Brugman and Russel 2004; Lausberg and Sloetjes 2009). This versatile, user-friendly and open-source tool was developed at the Max Planck Institute for Psycholinguistics in Nijmegen (http://www.lat-mpi.eu/tools/elan/). The data files created in ELAN can easily be queried using the extensive search options within the tool. Because there are numerous export options (e.g. to tab-delimited text among many others), the output is compatible with database or query functionalities in other tools, and can be used for further statistical processing.

4.2.1.2 Praat To more accurately perform the orthographic transcriptions according to the GAT transcription norm (see 4.2.2 below), and more specifically to help the transcribers assessing the boundaries of intonation units or the intonation contours, we used Praat (Boersma and Weenink 2009). This tool is fully compatible with ELAN and provides access to the sound as well as visual representations of speech intensity and f0.

4.2.2 Annotation levels and values

The development of the annotation scheme for the InSight Interaction Corpus was inspired in part by previous work on interaction corpora, and most notably the CORINTH corpus, a 6 h audio corpus of spontaneous conversations in Dutch (Feyaerts et al. 2011). The transcription norm and part-of-speech tagging that were used for CORINTH have proven to be very flexible. Because the present data are similar, the same transcription and POS-tagging protocol were used for this corpus. The additional layers of annotation, the integration of which makes up the very core of this corpus project, will be discussed in more detail below.

4.2.2.1 Transcription For the transcription of our video data we used the GAT transcription norm (Selting et al. 1998; Selting 2000). The advantage of that norm is its modularity, i.e. there are various possibilities to personalize the norm and choose the granularity or detail at which the acoustic signal can be represented. For our corpus we use an orthographical transcription with the following additional prosodic information:

- the main accent per intonation unit
- terminal intonation contour per intonation unit
- pauses within intonation units
- manifest lengthening of vowels and consonants

4.2.2.2 POS-tagging For the part-of-speech tagging in this corpus we used the Frog tagger (Van den Bosch 2007). Because the Frog algorithm was used for the reference corpus of spoken Dutch (CGN—*Corpus Gesproken Nederlands*, Oostdijk 2000), it is well trained for interactional data in Dutch. We ran the tagger with a reliability threshold of 0.06: if the Frog tagger attributed a POS-tag with 94 % or more reliability, the tag was not checked further. Tags below that threshold were checked by a human annotator. To check for the validity and consistency of the tags attributed by the human annotators, we performed a kappa-test on 300 tags for two annotators ($\kappa = 0.892$), which turned out to be satisfactory.

4.2.2.3 Gesture Gesture is the most labour-intensive of the multimodal layers to be coded in the corpus. Much in line with the MUMIN coding scheme (Allwood et al. 2007) our annotation grid for gesture is a simplification of the system proposed in the seminal work of McNeill (McNeill 1992, 2005). As Table 1 shows, both gesture form and function were coded. "Appendix" presents an overview of all parameters for gesture and gaze, the potential values for each of these parameters, as well as the codes that were used in ELAN.

With regard to the form we first define the so-called gesture phrases and phases. The stroke phases (i.e. the phases carrying the core message of the gesture) constitute the units that are then further annotated with respect to handedness, hand form and motion. Because of the specific recording technique applied in the InSight Interaction Corpus, there is one dimension we added to the coding scheme

Table 1 Gesture coding scheme

Gesture form	Structure	Gesture unit
		Gesture phrase
		Gesture phase
	Handedness	Left hand
		Right hand
		2 hands
	Hand form	ASL alphabet
		Finger orientation
		Palm orientation
		Position in gesture space
	Motion	Complexity
		Direction and position
Gesture function	Iconic	
	Deictic	
	Symbolic	

suggested by McNeill, viz. the 'depth' of gestures. As demonstrated by Kipp et al. (2007) specific recording set-ups enable a 3D view of the communicative situation, hence allowing for an accurate coding of the depth of the gestures, i.e. how close the hands are to the body.

The annotation of multimodal data raises the issue of inter-coder agreement and the reliability of the coding scheme [as systematically indicated by Cavicchio and Poesio (2009)]. To check for the consistency of the gesture annotation we performed kappa statistics for the features gesture function, structure and hand form. We didn't check inter coder reliability for handedness and motion because those features are not *debatable*, i.e. they don't involve training or interpretation and are purely based on unambiguous, visual perception: either a gesture is performed with the left or the right hand, either the movement is from left to right or from right to left, etc. For gesture function, structure and hand form we did check for inter coder agreement: three coders annotating a random selection of three minutes (i.e. 110 cases for function, 254 cases for gesture structure and 80 cases for hand form) of interaction were cross-checked. For each of the features discussed below, this makes for 3 kappa values: each annotator is compared to each other annotator.

With regard to the gesture structure, we performed a kappa test at the most fine-grained level, i.e. that of the gesture phases (because we expected more variation there than at higher levels). The kappa values of 0.913, 0.806 and 0.769 indicate a good coding consistency at this level. For the hand form, we checked for the hand shape (ASL alphabet, cf. McNeill 1992), palm orientation and finger orientation. The kappa values for hand shape are at 0.824, 0.832 and 0.847, which given the large amount of possible values (20 pre-defined possibilities in the ASL alphabet) is a very good result. Palm and finger orientation produce somewhat different results: for palm orientation the kappa scores are satisfactory (0.742, 0.726 and 0.707), for finger orientation they are moderate (0.610, 0.552, 0.479). The latter result may be explained in part by the unit at which the annotation was done: finger orientation may change in the course of a single gesture phase. Hence, this part of the annotation needs to be treated with some caution by the users of the corpus. Finally, concerning gesture function, we found kappa values of 0.864, 0.751 and 0.770, which again indicates sufficient coding consistency.

4.2.2.4 Gaze When annotating gaze, the first important issue to be addressed is the minimum fixation duration. Gaze behaviour often seems very chaotic as people switch from one object in focus to the other extremely fast. In other words, before starting the analysis of what people look at in the course of an interaction, we need a clear definition of what is considered as 'looking at', i.e. a standard of minimum fixation duration that allows for a reliable categorization of a gaze event as a fixation. In the extensive literature on the topic, the most frequently used standard for minimal fixation duration is around 120 ms (Jacob and Karn 2003; Vertegaal et al. 2001). Although the eye-tracking device used in the present project allows for a higher frequency, we used the frame rate of the video data (25 frames per second) to define the smallest possible unit of analysis, i.e. 1/25 s or 40 ms. Hence, in our annotation of gaze, participants need to focus on an object for longer than three

frames (120 ms) before we recognized it as a gaze event. Fixations of three frames or less were disregarded as relevant units.

Although we used a mobile eye-tracking system that allowed the participants to move, the corpus only contains video data from dyads sitting opposite each other. This restricts the number of potentially relevant objects or regions that can be fixated. As a consequence, the tag set for gaze contained a limited set of items: face (of other), body (of other), gesture of other, own gesture, screen, and wall. The gaze behaviour of each of the participants is annotated for the entire duration of the video file.

As gaze behaviour may function as a strong directive instrument in interaction, with speakers driving the visual attention of interlocutors towards objects by means of their own gaze cues, we added a calculation and indication of the overlap between identical objects of focalisation. In the annotation grid, a separate tier was included that shows when people look at each other [the so-called gaze window (Bavelas et al. 2002)] or at the same object (for example a gesture performed by one of the speakers, fixated by both).

The advantage of working with mobile eye-trackers as was done for this project is (at least) twofold. First, it allowed us to let interlocutors talk to each other freely without any restrictions as to head or hand movements. When using fixed table top eye-trackers interlocutors have to keep their head within a very specific 3D region in space to guarantee accurate gaze measuring. This restricts participants in moving and gesturing as freely as they would under more 'natural' circumstances. Moreover, in most studies using fixed eye-trackers, the device is put on a table in front of the participant, which further limits the possibility of gesturing freely.

Second, the use of mobile eye-trackers not only allows researchers to monitor eye movements very accurately,[5] but also has potential for studying head movements without using e.g. motion capture techniques: if a participant wearing a mobile eye-tracker tilts his or her head only very slightly, this will immediately be apparent from the video files generated by the scene cameras (and not or less clearly from an external video perspective such as the left image in Fig. 4). The scene camera video files can thus be used to study head movements accurately, without the burden of more expensive and time-consuming methods such as motion capture.

The annotations of the eye-tracking data were checked for consistency, using the same sample as described in the section on gesture above. In the three minute sample we checked a total of 69 gaze events, with kappa values at 0.947, 0.902 and 0.911. These near perfect values show a more than sufficient coding consistency at the gaze level.

4.3 The corpus

4.3.1 Size and file formats

In its current version (v1.1.) the corpus contains 15 recorded face-to-face interactions of 20 min each, of which five have been transcribed and annotated

[5] Especially compared to studies only using video files to determine eye gaze, such as Kendon (2004), Paggio et al. (2010), and Streeck (2009).

for the range of features detailed above. The remaining ten interactions are transcribed and only annotated for gaze and POS-tags. This data set contains 18,000 words and is currently being processed further.

Since all of the annotation work was done in ELAN (cf. supra) the annotation files are available in the ELAN-format, but to allow for the use of the corpus in different editors and tools also a tab-delimited version can be obtained. The media files are all synchronised and available separately as well as edited into a single trivid file. The edited video file combining the three perspectives into one image has the advantage of offering a good overview of the speech situation and avoids the necessity to load multiple video files into annotation tools like ELAN. For researchers interested in more detailed and higher-resolution video and audio files, each of the tracks in the edited video (one static camera, two mobile eye-trackers, two microphones) is available as a separate file as well. All of the media data are synchronised, which allows for easy switching between different media files. The video data are delivered in the AVI and WMV-format, the audio data in the WAV-format.

4.3.2 Availability

As an illustration of the type of data collected in the InSight Interaction Corpus, a sample of the transcriptions, annotations, audio and video files is available online (https://www.arts.kuleuven.be/ling/midi/corpora-tools/insight-interaction-corpus). Because of the multitude of media files for one recorded interaction, it is not possible to offer the entire corpus online. However, researchers interested in working with the corpus can obtain an offline version by contacting the authors. As stated above, version 1.1 of the corpus contains a full transcription and annotation for 5 of the 15 recordings, with the rest of the annotations to be updated soon. The use of the corpus is free of charge.

4.3.3 Privacy and legal issues

All of the participants in the corpus have signed a document in which they agree that the recordings can be used for scientific research. The document explicitly mentions that excerpts of the audio or video files can be used in publications or presentations. Furthermore the corpus has been declared at the Belgian federal privacy committee Commissie voor de Bescherming van de Persoonlijke Levenssfeer (CBPL).

All of the video and audio files have been anonymized in the sense that references to existing people have been omitted in the media data and replaced with fictional references in the transcriptions.

5 Concluding remarks

Multimodal corpora are growing in quantity and quality, as is adequately shown in the overview by Knight (2011). The Insight Interaction Corpus presented in this paper is intended as a contribution to this rapid growth along three dimensions.

First, we add gaze data based on mobile eye-tracking to the realm of annotation layers in multimodal research of human communication. In doing so we include a dimension of cognitive processing to the analytic apparatus of multimodal corpus analysis, and pave the way for potentially new insights into the tight coupling of processing and production in interaction based on systematic corpus research. Second, our recording technique allows for a 3D view of the speech situation and a more accurate coding of gesture in terms of 'depth'. And third, specifically for Dutch, the corpus presented here is unique in its set-up and provides a valuable addition to the already available resources, such as the IFADV-corpus and the large reference corpus of spoken Dutch (CGN, *Corpus Gesproken Nederlands*, Oostdijk 2000). The recording set-up and multitude of annotation layers make the InSight Interaction Corpus into a useful and pioneering resource for the study of multimodal communication in Dutch.

Although the corpus still is work in progress, even at this stage we want to look at challenges ahead. To make the multimodal endeavor a truly widespread and large-scale project, it will be imperative to reduce annotation cost and improve and diversify in natural language use. This can only be done by aspiring to wide-spread annotation standards, though flexible enough to meet individual research aims, and by constantly monitoring and applying new technologies and methodologies to the benefit of the research field. Incorporating vision technology (Fanelli et al. 2010), covering multi-party dialogue (Campbell 2009) and incorporating ever more natural language use situations (Adolphs et al. 2011), to name but a few, are examples of future challenges of multimodal corpus research.

Acknowledgments This work was partially supported by Grant Number 3H090339 STIM/09/03 of the University of Leuven.

Appendix

See Table 2.

Table 2 Annotation parameters for gesture and eye gaze

Parameter	Code	Value	Code
Gesture phase	Gphase	Preparation	Prep
		Stroke	Stroke
		Hold	Hold
Gesture function	Gtype	Iconic	Icon
		Metaphoric	Metaf
		Deictic	Deict
		Beat	Beat
Finger orientation	Forient	Fingers towards up	FTU
		Fingers towards down	FTD
		Fingers towards center	FTC

Table 2 continued

Parameter	Code	Value	Code
		Fingers away from back	FAB
		Fingers towards back	FTB
		Fingers away from center	FAC
Palm orientation	Porient	Palm towards up	PTU
		Palm towards down	PTD
		Palm towards center	PTC
		Palm away from back	PAB
		Palm towards back	PTB
		Palm away from center	PAC
Motion—body	GmotB	Towards body	TB
		Away from body	AB
		Towards center	TwC
		Away from center	AwC
		Parallell to body	PAR
Motion—axis	GmotP	Parallell to front	PF
		Parallell to side	PS
Motion complexity	C/Smot	Complex motion	COMPL
		Simple motion	SIMPL
		No motion	NO
Motion 2 hands	2HANDmot	2-hand movement, same movement	SM
		2-hand movement, different movement	DM
Direction—left/right	LR/Rldir	From left to right	LR
		From right to left	RL
Direction—up/down	UD/Dudir	From up to down	UD
		From down to up	DU
Direction—front/back	FB/Bfdir	From front to back	FB
		From back to front	BF
Direction—diagnoal	DIAGdir	Diagonal	DIAG
Gesture position	Gzone	Center center	CC
		Center center	C
		Periphery, up	PU
		Periphery, up, left	PULe
		Periphery, up, right	PUR
		Periphery, left	PLe
		Periphery, right	PR
		Periphery, low	PLo
		Periphery, low, left	PLoL
		Periphery, low, right	PLoR
		Extreme periphery, up	EPU
		Extreme periphery, up, left	EPULe
		Extreme periphery, up, right	EPUR
		Extreme periphery, left	EPLe

Table 2 continued

Parameter	Code	Value	Code
		Extreme periphery, right	EPR
		Extreme periphery, low	EPLo
		Extreme periphery, low, left	EPLoL
		Extreme periphery, low, right	EPLoR
Gesture depth	Gdepth	Touch	Touch
		Close	Close
		Normal	Normal
		Far	Far
Gaze	GAZEevent	Focus on face of adversarial	Face
		Focus on gesture of adversarial	Gest
		Focus on own gesture	Own
		Focus on screen	Screen
		Focus on none of the above	Wall

References

Adolphs, S., Knight, D., & Carter, R. (2011). Capturing context for heterogeneous corpus analysis: Some first steps. *International Journal of Corpus Linguistics, 16*, 305–324.

Allwood, J. (2008). Multimodal corpora. In A. Lüdeling & M. Kytö (Eds.), *Corpus linguistics: An international handbook* (29th ed., pp. 207–225). Berlin: Mouton de Gruyter.

Allwood, J., Cerrato, L., Jokinen, K., Navarretta, C., & Paggio, P. (2007). The MUMIN coding scheme for the annotation of feedback, turn management, and sequencing phenomena. In J. Martin, P. Paggio, P. Kuenlein, R. Stiefelhagen, & F. Pianesi (Eds.), *Multimodal corpora for modelling human multimodal behaviour* (41st ed., pp. 273–287). Heidelberg: Springer.

Bavelas, J., Coates, L., & Johnson, T. (2002). Listener responses as a collaborative process: The role of gaze. *Journal of Communication, 52*, 566–580.

Bertrand, R., Blache, P., Espesser, R., Ferré, G., Meunier, C., Priego-Valverde, B., et al. (2008). Le CID—Corpus of interactional data—Annotation et exploitation multimodale de parole conversationnelle. *Traitement automatique des langues, 49*, 105–134.

Blache, P., Bertrand, R., & Ferré, G. (2008). Creating and exploiting multimodal annotated corpora. In *Proceedings of the sixth international conference on language resources and evaluation (LREC)*.

Boersma, P., & Weenink, D. (2009). *PRAAT: Doing phonetics by computer (version 5.3.05)*. http://www.praat.org/. Accessed February 27, 2012.

Brennan, S., Chen, X., Dickinson, C., Neider, M., & Zelinsky, G. (2008). Coordinating cognition: The costs and benefits of shared gaze during collaborative search. *Cognition, 106*, 1465–1477.

Brugman, H., & Russel, A. (2004). Annotating multimedia/multi-modal resources with ELAN. In *Proceedings of the fourth international conference on language resources and evaluation (LREC)*.

Campbell, N. (2009). Tools and resources for visualising conversational speech interaction. In M. Kipp, J. Martin, P. Paggio, & D. Heylen (Eds.), *Multimodal corpora: From models of natural interaction to systems and applications* (pp. 231–234). Heidelberg: Springer.

Cavicchio, F., & Poesio, M. (2009). Multimodal corpora annotation: Validation methods to assess coding scheme reliability. In M. Kipp, J. Martin, P. Paggio, & D. Heylen (Eds.), *Multimodal corpora: From models of natural interaction to systems and applications* (pp. 109–121). Heidelberg: Springer.

Chen, L., Travis-Rose, R., Parrill, F., Han, X., Tu, J., Huang, Z., et al. (2006). VACE multimodal meeting corpus. *Lecture Notes in Computer Science, 3869*, 40–51.

Edlund, J., Beskow, J., Elenius, K., Hellmer, K., Strömbergsson, S., & House, D. (2010). Spontal: A Swedish spontaneous dialogue corpus of audio, video and motion capture. In *Proceedings of the seventh international conference on language resources and evaluation (LREC)*.

Fanelli, G., Gall, J., Romsdorfer, H., Weise, T., & Van Gool, L. (2010). 3D vision technology for capturing multimodal corpora: Chances and challenges. In *Proceedings of the seventh international conference on language resources and evaluation (LREC)*.

Feyaerts, K., Oben, B., Brône, G., & Speelman, D. (2011). Corpus interactional humour. http://www.arts.kuleuven.be/ling/midi/corpora-tools.

Gerwing, J., & Allison, M. (2009). The relationship between verbal and gestural contributions in conversation: A comparison of three methods. *Gesture, 9*, 312–336.

Hadelich, K., & Crocker, M. (2006). Gaze alignment of interlocutors in conversational dialogues. In *Proceedings of the 2006 symposium on eye tracking research and applications*.

Hanna, J., & Brennan, S. (2007). Speakers' eye gaze disambiguates referring expressions early during face-to-face conversation. *Journal of Memory and Language, 57*, 596–615.

Herrera, D., Novick, D., Jan, D., & Traum, D. (2010). The UTEP-ICT cross-cultural multiparty multimodal dialog corpus. In *Proceedings of the seventh international conference on language resources and evaluation (LREC)*.

Jacob, R., & Karn, K. (2003). Eye tracking in human-computer interaction and usability research: Ready to deliver the promises. In R. Radach, J. Hyönä, & H. Deubel (Eds.), *The mind's eye: Cognitive and applied aspects of eye movement research* (pp. 573–605). Oxford: Elsevier Science.

Jokinen, K. (2010). Non-verbal signals for turn-taking and feedback. In *Proceedings of the seventh international conference on language resources and evaluation (LREC)*.

Jokinen, K., Nishida, M., & Yamamoto, S. (2009) Eye gaze experiments for conversation monitoring. In *Proceedings of the 3rd international universal communication symposium*.

Kendon, A. (2004). *Gesture: Visible action as utterance*. Cambridge: Cambridge University Press.

Kimbara, I. (2006). On gestural mimicry. *Gesture, 6*, 39–61.

Kipp, M., Neff, M., & Albrecht, I. (2007). An annotation scheme for conversational gestures: How to economically capture timing and form. *Journal on Language Resources and Evaluation, 41*, 325–339.

Knight, D. (2011). The future of multimodal corpora. *Revista Brasileira de Linguistica Aplicada, 11*, 391–415.

Knight, D., Adolphs, S., Tennent, P., & Carter, R. (2008) The Nottingham multi-modal corpus: A demonstration. In *Proceedings of the sixth international conference on language resources and evaluation (LREC)*.

Knight, D., Evans, D., Carter, R., & Adolphs, S. (2009). HeadTalk, HandTalk and the corpus: Towards a framework for multi-modal, multi-media corpus development. *Corpora, 4*, 1–32.

Lausberg, H., & Sloetjes, H. (2009). Coding gestural behavior with the NEUROGES-ELAN system. *Behavior Research Methods, Instruments, & Computers, 41*, 841–849.

Massaro, D., & Beskow, J. (2002). Multimodal speech perception: A paradigm for speech science. In B. Granstrom, D. House, & I. Karlsson (Eds.), *Multimodality in language and speech systems* (pp. 45–71). Dordrecht: Kluwer Academic.

McNeill, D. (1992). *Hand and mind: What gestures reveal about thought*. Chicago: University of Chicago Press.

McNeill, D. (2005). *Gesture and thought*. Chicago: University of Chicago Press.

McNeill, D. (2008). Unexpected metaphors. In A. Cienki & C. Müller (Eds.), *Metaphor and gesture* (pp. 155–170). Amsterdam: John Benjamins.

Oostdijk, N. (2000). The spoken Dutch corpus. Overview and first evaluation. In *Proceedings LREC 2000, Genoa, Italy*.

Paggio, P., Allwood, J., Ahlsén, E., Jokinen, K., & Navarretta, C. (2010). The NOMCO multimodal Nordic resource—Goals and characteristics. In *Proceedings LREC 2010, Valletta, Malta*.

Pickering, M., & Garrod, S. (2004). Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences, 27*, 169–226.

Pickering, M., & Garrod, S. (2006). Alignment as the basis for successful communication. *Research on Language and Computation, 4*, 203–228.

Pine, K., Lufkin, N., & Messer, D. (2004). More gestures than answers: Children learning about balance. *Developmental Psychology, 40*, 1059–1067.

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin, 85*, 618–660.

Selting, M. (2000). The construction of units in conversational talk. *Language in Society, 29*, 477–517.

Selting, M., Auer, P., Barden, B., Couper-Kuhlen, E., Günther, S., Quasthoff, U., et al. (1998). Gesprachsanalytisches transkriptionssystem (GAT). *Linguistische Berichte, 173*, 91–122.

Staudte, M., Heloir, A., Crocker, M., & Kipp, M. (2011). On the importance of gaze and speech alignment for efficient communication. In *Proceedings of the 9th international gesture workshop*.

Streeck, J. (2009). *Gesturecraft—The manufacture of meaning*. Amsterdam/Philadelphia: John Benjamins.

Tanenhaus, M., & Brown-Schmidt, S. (2008). Language processing in the natural world. In B. Moore, L. Tyler, W. Marslen-Wilson (Eds.), *The perception of speech: From sound to meaning*. *Philosophical Transactions of the Royal Society B: Biological Sciences, 363*, 1105–1122.

Van den Bosch, A., et al. (2007). An efficient memory-based morphosyntactic tagger and parser for Dutch. In *Selected papers of the 17th computational linguistics in the Netherlands meeting*.

Van Son, R., Wesseling, W., Sanders, E., & Van Der Heuvel, H. (2008). The IFADV corpus: A free dialog video corpus. In *Proceedings of the sixth international conference on language resources and evaluation (LREC)*.

Vertegaal, R., Slagter, R., Van der Veer, G., & Nijholt, A. (2001). Eye gaze patterns in conversations: There is more to conversational agents than meets the eyes. In *Proceedings of the Conference on Human Factors in Computing Systems*.