NeuroImage 264 (2022) 119734

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/neuroimage

The CABB dataset: A multimodal corpus of communicative interactions for behavioural and neural analyses

Lotte Eijk[a,1], Marlou Rasenberg[a,b,1], Flavia Arnese[c], Mark Blokpoel[c], Mark Dingemanse[a], Christian F. Doeller[d,e,f], Mirjam Ernestus[a], Judith Holler[c,b], Branka Milivojevic[c], Asli Özyürek[a,b,c], Wim Pouw[c,b], Iris van Rooij[c,g], Herbert Schriefers[c], Ivan Toni[c], James Trujillo[c,b], Sara Bögels[c,h,*]

[a] Centre for Language Studies, Radboud University, Nijmegen, the Netherlands
[b] Max Planck Institute for Psycholinguistics, Nijmegen, the Netherlands
[c] Donders Institute for Brain, Cognition, and Behaviour, Centre for Cognitive Neuroimaging, Radboud University, P.O.Box 9010, Nijmegen, Gelderland 6500, the Netherlands
[d] Max Planck Institute for Human Cognitive and Brain Sciences, Leipzig, Germany
[e] Kavli Institute for Systems Neuroscience, Centre for Neural Computation, The Egil and Pauline Braathen and Fred Kavli Centre for Cortical Microcircuits, Jebsen Centre for Alzheimer's Disease, Norwegian University of Science and Technology, Trondheim, Norway
[f] Wilhelm Wundt Institute of Psychology, Leipzig University, Leipzig, Germany
[g] Department of Linguistics, Cognitive Science, and Semiotics, and the Interacting Minds Centre at Aarhus University, Denmark
[h] Department of Cognition and Communication, Tilburg University, the Netherlands

ARTICLE INFO

Keywords:
Multimodal data
Face-to-face interaction
Referential communication
Motion tracking
fMRI
Conceptual alignment

ABSTRACT

We present a dataset of behavioural and fMRI observations acquired in the context of humans involved in multimodal referential communication. The dataset contains audio/video and motion-tracking recordings of face-to-face, task-based communicative interactions in Dutch, as well as behavioural and neural correlates of participants' representations of dialogue referents. Seventy-one pairs of unacquainted participants performed two interleaved interactional tasks in which they described and located 16 novel geometrical objects (i.e., Fribbles) yielding spontaneous interactions of about one hour. We share high-quality video (from three cameras), audio (from head-mounted microphones), and motion-tracking (Kinect) data, as well as speech transcripts of the interactions. Before and after engaging in the face-to-face communicative interactions, participants' individual representations of the 16 Fribbles were estimated. Behaviourally, participants provided a written description (one to three words) for each Fribble and positioned them along 29 independent conceptual dimensions (e.g., rounded, human, audible). Neurally, fMRI signal evoked by each Fribble was measured during a one-back working-memory task. To enable functional hyperalignment across participants, the dataset also includes fMRI measurements obtained during visual presentation of eight animated movies (35 min total). We present analyses for the various types of data demonstrating their quality and consistency with earlier research. Besides high-resolution multimodal interactional data, this dataset includes different correlates of communicative referents, obtained before and after face-to-face dialogue, allowing for novel investigations into the relation between communicative behaviours and the representational space shared by communicators. This unique combination of data can be used for research in neuroscience, psychology, linguistics, and beyond.

1. Introduction

Language is a key socio-cognitive human function predominantly used in interaction. Yet, much work in linguistics and cognitive neuroscience has focused on individuals' coding-decoding of signals according to their structural dependencies. Understanding the communicative use of language requires shifting the focus of investigation from individual competencies to the mechanisms used by interlocutors to understand each other during live interactions. Here, we provide a dataset that can be used to study face-to-face, multi-turn referential communication be-

* Corresponding author at: Donders Institute for Brain, Cognition, and Behaviour, Centre for Cognitive Neuroimaging, Radboud University, P.O.Box 9010, Nijmegen, Gelderland 6500, the Netherlands.
E-mail address: sara.boegels@donders.ru.nl (S. Bögels).
[1] These authors contributed equally to this work.

https://doi.org/10.1016/j.neuroimage.2022.119734.
Received 19 May 2022; Received in revised form 7 October 2022; Accepted 3 November 2022
Available online 4 November 2022.

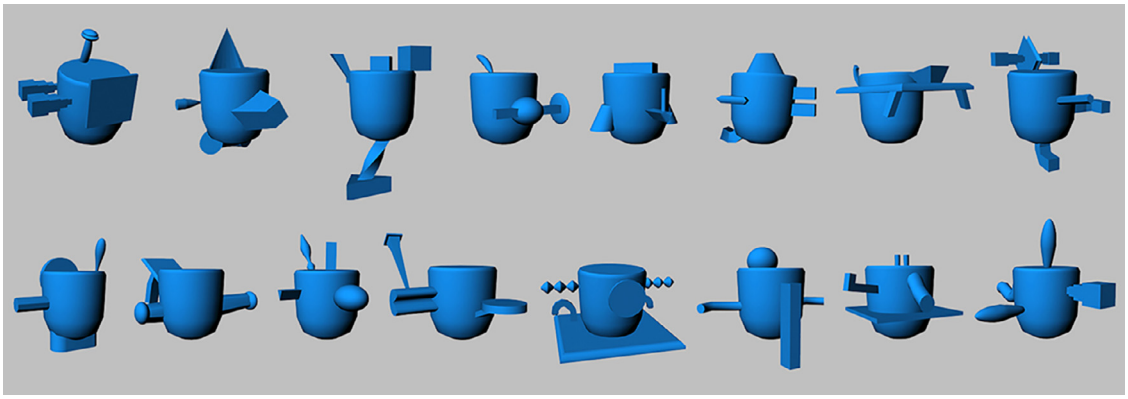

Fig. 1. The 16 stimuli (Fribbles, based on Barry et al., 2014) used in the different tasks of the study, designed to evoke various conceptualisations.

tween pairs of interlocutors (through audio/video and motion-tracking recordings), as well as individuals' representations of the dialogue referents (as estimated from behavioural and fMRI data collected before and after the dialogue).

The dataset presented here emerges from CABB (Communicative Alignment of Brain and Behaviour), a research program focused on studying interactive language use. This program builds on the notion that interlocutors can disambiguate referentially flexible signals by building a shared cognitive space (e.g., Clark, 1997; Clark and Brennan, 1991; Hutchins and Hazlehurst, 1995; Stolk et al., 2016; 2022). A shared cognitive space involves not only presumed common ground, the propositions jointly taken for granted or communicated, but also mutual awareness of the circumstances of communication, and thus the likely joint goals, norms, and affordances of the event, embedded in the recent interactional history. Besides the traditional focus on transfer of propositional content, this research initiative considers how language use is organised to achieve interactional goals and to monitor mutual understanding, and how interlocutors create and control a shared cognitive space during live communicative interactions. CABB considers the contribution of multimodal communicative resources (speech, gestures) at different levels of linguistic structure (from phonology to pragmatics) during interactive task-based dialogue.

The interactional part of the dataset consists of audio, video, and body-movement recordings of face-to-face communicative interactions in Dutch between 71 pairs of participants, without restrictions on communicative means (e.g., speech, gestures), timing, turn-taking, or feedback. Participants communicate about 16 novel visual objects which lack conventional labels - called "Fribbles" (Barry et al., 2014). We designed and pre-tested the Fribbles (see Fig. 1) for their ability to evoke different conceptualisations across individuals. As such, different pairs would need to work together to create their own pair-specific conceptualisations and labels for them, enabling us to see effects of the interaction rather than mere exposure to the stimuli. The participants were instructed to communicate in order to identify (Referential task) and localise (Localisation task) the Fribbles on a screen. These tasks are designed to capture a core element of everyday human communication: each pair needs to create mutually understood utterances, dependent on the situated context of the ongoing interaction (Clark, 1996; Stolk et al., 2022). Participants were not familiar with the Fribbles at the onset of the study, a task feature designed to amplify this process of negotiating a common referent that arguably occurs in many communicative interactions. More precisely, in the Referential task, participants need to negotiate referential expressions for the Fribbles to be able to identify them amidst the total set, similar to referential tasks with tangrams (see e.g., Clark and Wilkes-Gibbs, 1986; Holler and Wilkin, 2011). In the Localisation task, each pair needs to work out collaboratively whether a particular Fribble is located at the same position on their respective screens. Participants had equal opportunities to speak in the interaction, since they switched roles throughout the task.

The dataset contains high-quality audio recordings using head-mounted microphones, along with time-aligned orthographic transcriptions of the speech for 47 out of the 71 interactions (see *Methods*). These enable different types of linguistic analyses of individual participants' speech (e.g., lexical, semantic, phonetic), as well as investigations of alignment between participants on these levels (e.g., Pickering and Garrod, 2004). Moreover, high-quality video recordings from three different angles, as well as 3D body motion-tracking data from two Microsoft Kinects (V2), allows researchers to analyse participants' movements, postures, and gestures, as well as their alignment between participants. The face-to-face set-up, where participants stood opposite each other and had full vision of each other's torso, facilitated the use of gestures (although this was in no way explicitly encouraged).

The dataset also provides estimates of participants' individual representations of the Fribbles using two behavioural measures and one neuroimaging (fMRI) measure. These measures are taken both before (pre) and after (post) the face-to-face interaction. Behaviourally, participants named each Fribble using one to three words (Naming task), and rated each Fribble on 29 different visual and semantic features (Features task; based on Binder et al., 2016). Neurally, participants' brain responses to the Fribble images were measured using fMRI while they performed a one-back working memory task to monitor their attention to the stimuli, following earlier studies using neural representational approaches (e.g., Bracci et al., 2015; Dobs et al., 2019). By containing both the pre and post measures, this dataset is well suited for measuring changes in estimated individual representations of each referent Fribble, as well as the extent of convergence of such estimated representations within each pair, brought about by the interaction. Comparison of across-voxel activity patterns of fMRI responses to the Fribbles across participants is enhanced by the possibility of implementing so-called "hyperalignment" (Haxby et al., 2011). Namely, the dataset includes fMRI data of participants watching the same eight animated movies (about 35 min in total). This enables the fMRI pre-processing step of aligning individual brains to a common information space across the sample, based on functional (instead of anatomical) similarities between the brains. That is, voxels from different brains that are similarly activated in response to the same stimuli while watching the movies are aligned to each other. Hyperalignment is especially relevant for the present dataset because it allows for more direct comparisons between activation patterns caused by the same Fribble in different brains.

To date, this dataset is unique in that it combines multimodal interactional data with behavioural and neural characterisation of the representational consequences of a face-to-face communicative interaction. The interactional, behavioural, and neuroimaging data can be used for

addressing a wide range of research questions within and across various disciplines such as linguistics, neuroscience, and psychology. Furthermore, the dataset offers the possibility to combine those measures and investigate how face-to-face multimodal naturalistic communication changes the estimated representations of the referents within and across interlocutors.

In recent years, open access brain-imaging datasets have increasingly become available, providing different types of data (e.g., resting state, task-related) from multiple brain-imaging methods (i.e., EEG, MEG, fMRI), such as the Human Connectome Project (Van Essen et al., 2013), the CamCan dataset (Taylor et al., 2017), and the MOUS dataset (Schoffelen et al., 2019). However, none of these quantify the consequences of communicative interactions with (behavioural and) fMRI observations. The unique characteristics of this dataset can also be appreciated by comparing it to existing corpora with recordings of social interaction. Interactional corpora consisting of audio data are rather numerous (for an overview, see Ernestus and Baayen, 2011), containing for example spontaneous face-to-face and telephone conversations in Dutch as in the Corpus Gesproken Nederlands (CGN; Oostdijk, 2000), or task-based interactions in Scottish English as in the HCRC Map Task corpus (Anderson et al., 1991). Examples of multimodal corpora, consisting of both video and audio data, are the InSight Interaction Corpus (Dutch; Brône and Oben, 2015), the IFADV corpus (Dutch; Van Son et al., 2008), the Spontal corpus (Swedish; Edlund et al., 2010), and the Nijmegen Corpus of Casual French (Torreira et al., 2010). These corpora include many aspects of multimodal communication, but do not provide the combination of high quality audio, video, and motion tracking necessary to implement fine grained integrative analyses of both gestures and speech. At least one other dataset (Rauchbauer et al., 2019) also combines multi-modal interactive data (speech, eye-movements, and face-recordings) with fMRI measurements. Differently from our dataset, the fMRI data were acquired in individual participants while they were interacting with a human or a robot. With 71 interactions (47 fully transcribed), the present corpus provides ample possibilities for rich qualitative and quantitative studies of communicative interactions. This dataset also opens up new research avenues, as observations from the interaction can be related to correlates of individuals' representations of the dialogue referents as estimated from the behavioural and neuroimaging measures.

A precursory dataset of the CABB team (with a similar paradigm, but without fMRI data) has been used in earlier reports (Pouw et al., 2021a; Rasenberg et al., 2022; Rasenberg et al., in press), and further reports on the present dataset are in preparation. This contribution is intended to describe the dataset with respect to the procedures used in the acquisition as well as some example analyses, and make it available for use by other researchers. From here onwards we refer to this as the Dataset (along with a folder name). See Section 2.7.2 for information on how to access the Dataset.

2. Methods

2.1. Participants

In total, 142 right-handed, native Dutch speakers (71 pairs; 30 all-female, 7 all-male, and 34 mixed gender pairs, according to self-reported data) participated in the study, with an average age of 22.86 years (*SD* = 3.63, *range* = 18–33 with one outlier of 45). All participants reported no neurological or language-related disorders, no metal implants (except for dental) in their body, no history of brain surgery, no hearing impairments, and normal or corrected-to-normal vision. The participants were recruited via the Radboud SONA participant pool system. Data and transcriptions of 37 pairs (74 participants) from all tasks are fully complete and shared (see Section 2.7.1 for details on the availability and quality of various parts of the Dataset).

2.2. Ethical approval and participant consent

This study met the criteria of the blanket ethical approval for standard studies of the Commission for Human Research Region Arnhem-Nijmegen (DCCN CMO 2014/288). Participants were emailed information about the study in advance and verbally informed on the testing day itself. Written informed consent was obtained before data collection started. Participants agreed to the sharing of the fully anonymised data,[2] and could optionally agree to the sharing of potentially identifiable audio/video data with researchers for scientific purposes and/or for educational and/or promotional purposes, through (a) presentations/lectures (not publicly available), (b) newspapers, magazines/journals or other (online) news outlets, (c) social media, and (d) television. See the Participants folder in the Dataset for the full overview of data sharing consent.

2.3. Materials

The experimental stimuli consisted of 16 pictures of blue 3D objects made up of geometrical figures attached to each other, on a grey background, which we refer to as Fribbles (see Fig. 1; note that the term "Fribbles" was never mentioned to participants). We adapted these stimuli from objects also called Fribbles (Barry et al., 2014). The adaptation was based on pilot tests, in which participants individually named each Fribble using one to three words (see Naming task explained in Section 2.5.1 below) and/or played the Referential communication game in pairs (see Section 2.5.5 below). These pilots resulted in a final set of Fribbles (Fig. 1) which evoked variable conceptualisations (names) across both individuals and pairs. This was important to be able to control for general aspects of the interaction by comparing convergence (e.g., in labels) between real interacting pairs and pseudo-pairs, i.e. pairs who did not interact with each other (see e.g., Section 3.5 below).

2.4. Set-up and apparatus

2.4.1. MRI apparatus & (f)MRI image acquisition

Magnetic resonance images were acquired using two 3T MAGNETROM MR scanners: Prisma and PrismaFit (Siemens AG, Healthcare Sector, Erlangen, Germany). For the functional acquisition a multi-band 2D-EPI sequence released as part of the Human Connectome Project (Uğurbil et al., 2013) was used. Functional images were acquired using a multi-band six sequence. The parameters of the acquisition were: TR/TE = 1,000/34 ms, flip angle = 60°; 2mm^3 isotropic resolution over a FOV = 208 × 208 × 132 mm; Multi-band acceleration of six was used in the slice direction and no parallel imaging was applied in-plane. Phase encoding was applied on the AP direction with a partial Fourier coverage of 7/8, including five volumes with reversed phase encoding (A >> P), which can be used to correct image distortions. Approximately 750 vol were acquired in each of the four one-back runs (two in the pre session and two in the post session) and 2,074 vol in the movies run (session three).

A T1-weighted scan was acquired at the end of the second session in the sagittal orientation using a 3D MPRAGE sequence with the following parameters: TR/TI/TE = 2,400/1,000/2.22 ms, 8° flip angle. Following, a T2-weighted scan was also acquired in the sagittal orientation using a variable flip angle TSE with the following parameters: TR/TE = 3,200/563 ms, echo spacing = 3.52 ms, Turbo Factor = 314. Both the T1 and T2 used a FOV of 256 × 240 × 167 mm, a 0.8 mm^3 isotropic resolution, and parallel imaging (iPAT = 2) to accelerate the acquisition resulting in an acquisition time of 6 min 38 s for T1 and

[2] Note that defaced structural MRI data are technically only pseudonymised and not fully anonymised. However, the consent form for standard studies that we used dated from 2018, which was before this issue was recognised in the scientific and legal community.

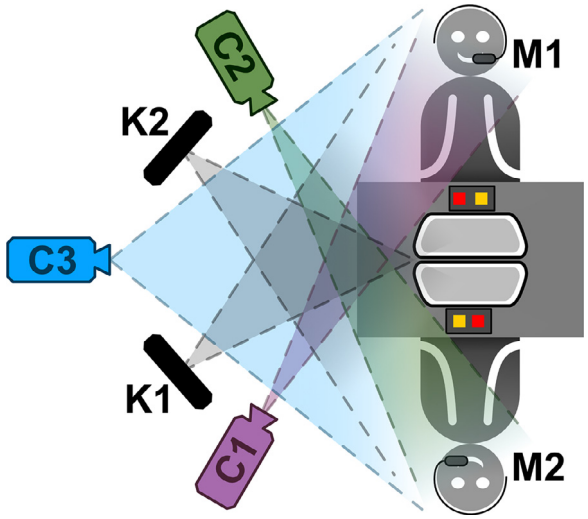

Fig. 2. Recording set-up for the interaction (C1-C3: cameras, K1-K2: Kinect, M1-M2: microphones).

5 min 57 s for T2. At the start of each of the three sessions, an additional fast T1 weighted scan was obtained using a spoiled gradient echo sequence with the following contrast parameters: TR/TE = 6.31/3.2 ms; flip angle = 11°. Acquisition was performed in the sagittal orientation with a FOV of 176 × 256 × 256 mm and 1 mm^3 isotropic. A five-fold controlled aliasing acceleration was used resulting in a total acquisition time of 1 min 17 s.

Stimuli were presented using an EIKI LC-XL100 beamer with a resolution of 1,024 × 768 and a refresh rate of 60 Hz, and were projected onto a screen behind the scanner bore. Participants were able to see the screen via a mirror. Given different characteristics of the two scanners used, the image sizes for the Fribble stimuli were adjusted such that all participants experienced all Fribbles at the same visual angle in both scanners.

During fMRI acquisition, participants' attention levels were monitored by single-eye recording, using an infrared source eye-tracker. Also, respiration and heartbeat were recorded using a respiration belt and a pulse wave sensor, respectively; both required the same MRI-compatible amplifier from BrainAmp ExG MR.

2.4.2. Set-up and apparatus of the interaction

The interaction took place in a sound-attenuated booth. Participants of a pair faced each other about two meters apart while standing in front of a table (see Fig. 3, middle panel). Each participant faced a 24″ screen (BenQ XL2430T), slightly tilted for an optimal viewing angle, and positioned at hip height. This ensured that participants could see each other, and prevented interference with the participants' gesture space (McNeill, 1992). All 16 Fribbles were simultaneously presented on each participant's screen, in a random arrangement over a grey background, each Fribble covering 4 × 4 cm on the screen (see Fig. 4). The Fribbles were labelled with numbers for one participant and letters for the other. Button boxes (with a red and a yellow button) were positioned below the screen and were used by the participants to provide answers (for the Localisation task, but not the Referential task) and/or to move to the next trial (see Section 2.5).

Video recordings were made with a frame rate of 29.97 frames per second (fps) at 1,920 × 1,080 resolution using three HD cameras (JVC GY-HM100/150); cameras 1 and 2 were positioned to the side to yield (semi-)frontal views of each participant, while camera 3 was positioned in the middle to yield an overview of both participants (Fig. 2). Two head-mounted microphones (Samson QV) were used to record speech for each participant separately. These microphones were connected to an AudiTon pre-amplifier and then to a Roland R-05 recorder, which were both situated in the control room, where the experimenter could listen to and adjust the volume of the incoming audio. The output of the pre-amplifier (which consisted of two separate audio channels, one for each participant) was transmitted to the recorder (where the audio was digitised at a sampling rate of 44.1 kHz and a 16-bit quantisation), the output of which was transmitted to cameras 1 and 2, respectively (digitised at 48 kHz and 16-bit). Two Microsoft Kinects (V2), positioned next to cameras 1 and 2 (Fig. 2) were used to collect 3D positional joint tracking data (for 25 joints) at 30 fps. During data collection, the experimenters monitored the Kinect pose skeleton tracking which served as an online quality check of the Kinect tracking.

Since recordings were started manually on the various devices, all audio, video, and motion-tracking data was synchronised off-line (see Section 2.6). To facilitate this process, a dedicated "synchronisation signal" device was used: every 60 s the device sent a digital code to the laptops controlling the Kinect (stored in log files), and a beep as audio input to the cameras (recorded on a secondary audio channel, separately from the speech). See Fig. S1 in Supplementary Materials for a schematic overview of all materials in the interaction setup and their connections.

2.5. Procedure

Participants came to the lab in pairs (but did not know each other beforehand) performing several individual tasks (i.e., Naming task, Features task, one-back task in the fMRI) before and after a joint interactional task (i.e., Referential and Localisation tasks) followed by another fMRI session (movie watching) and a questionnaire. All tasks were programmed using the Presentation software (Version 20.2, Neurobehavioral Systems, Inc., Berkeley, CA, www.neurobs.com). See Fig. 3 for an overview of all tasks and the next sections (Section 2.5.1 – Section 2.5.6) for a detailed description (for full Dutch Instructions for all tasks, see the Presentation NBS scripts in the Presentation_scripts folder in the Dataset). Before starting, participants provided informed consent (Section 2.2) and were asked not to talk to each other before the interaction part of the study and not to talk about the tasks in the break(s) between tasks during the entire session. The session lasted for six to eight hours in total and included a lunch break of at least 30 min immediately after the interaction. Whenever possible, two pairs were tested on the same day in an interleaved fashion, which meant participants had another break of maximally 45 min before the last scanner session and were asked to fill out most of the questionnaire in this break.

2.5.1. Naming and features tasks

Participants performed two behavioural tasks, the Naming and the Features task, individually in sound-proofed cubicles, while sitting in front of a 24 inch, full HD screen and responding using a keyboard and a mouse. The two tasks were presented in an interleaved fashion so that for each Fribble, participants first performed the Naming task and then the Features task before moving on to the next Fribble. The order of presentation of the Fribbles was randomised per participant. Participants received written instructions for both tasks on the screen, were given the opportunity to ask questions, and then received an oral summary of the instructions from the experimenter.

For the Naming task, all Fribbles were presented on the screen simultaneously. The position of the 16 Fribbles was randomised separately for each participant, but was the same for all trials within participants. On each trial, one Fribble was marked with a red square. Participants were instructed to name or describe that Fribble using one to three words in such a way that the other participant would be able to find it amongst all other Fribbles on the screen (see Fig. S2 in Supplementary Materials for an example screenshot of a Naming task trial).

For the Features task, the same Fribble they had just named was presented in the left top corner of the screen with a lead-in sentence next to it ("To what extent do you view this picture as…"). Underneath, 29 different features were shown in the form of linguistic labels (to be read as completing the lead-in sentence). The features were based on a

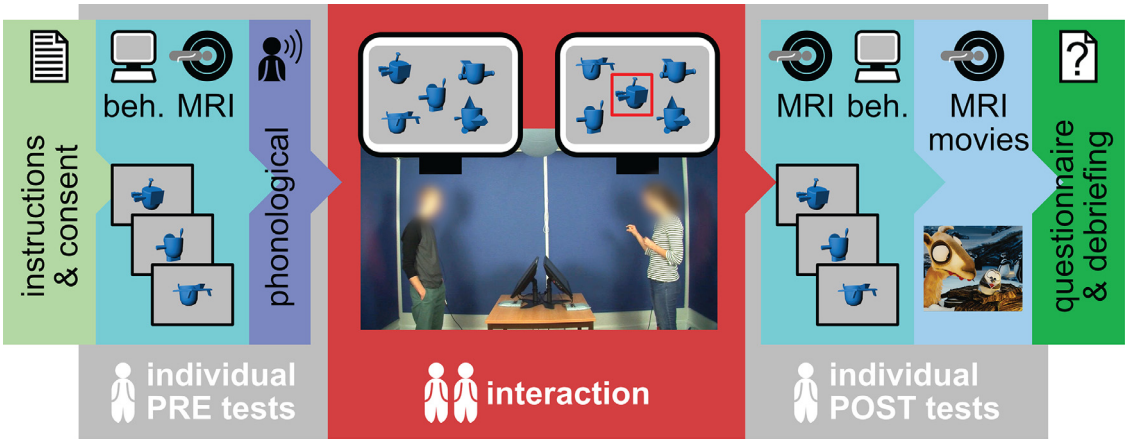

Fig. 3. Overview of participants' tasks during the testing day. *Beh.* = behavioural tasks (Naming and Features). *MRI* = magnetic resonance imaging (task: one-back task in sessions one and two; movies in session three). *Phonological* = phonological pre-test. *PRE* = before the interaction, *POST* = after the interaction.

study by Binder et al., (2016) and our own selection given the range of results in the pilot Naming task (categories of objects). Some examples of features are "Rounded", "Symmetrical", "Human", and "Positive" (see Table S1 in Supplementary Materials or Dataset: Data for the full list). Participants were instructed to judge to what extent each feature was compatible with their view of the Fribble by moving a slider underneath the feature label (left: "not at all", right: "very strongly"). They were instructed to decide within a few seconds for each feature and to choose the leftmost position on the bar if a feature was not applicable or neutral. Only after participants had moved the slider for each of the 29 features, they could press enter to continue to the next Fribble (see Fig. S3 in Supplementary Materials for an example screenshot of a Features task trial).

The tasks were the same when participants performed them for the second time (after the interaction), but with a different random order of Fribble presentations per participant (both on the screen and over trials). Participants were told that they were allowed to give the same name as the first time, but that they did not have to. They were again instructed to describe each Fribble such that their partner would be able to find it amongst the others. Both before and after the interaction 45 min were planned for the Naming and Features task.

2.5.2. *fMRI one-back task*

While still in the cubicles, participants received written instructions for the scanner-based one-back task. They were instructed to press one button when the picture they saw was the same as the previous picture and another button in all other cases (also for the first picture). They were then given an oral summary and performed a short test block with the one-back task using different pictures than the ones used in the actual task (seven trials). This test block was repeated until all responses were correct. In the MRI scanner, participants read the instructions on the screen again while localiser scans were acquired and then again performed the same practice block (to practice with the response buttons in the scanner) until all responses were correct. When necessary, additional instructions were given over the intercom. In the scanner, participants gave responses on a button-box using the index and middle fingers of their right (dominant) hand. The allocation of the fingers to "same" and "different" responses was counterbalanced over participants, but was the same for all sessions per participant.

After the onset of the fMRI sequence, Fribbles were presented one by one, slightly below the centre of the screen (to avoid vertical head movement at the onset of a Fribble), on a grey background, for two seconds, with a visual angle of about five degrees. This visual angle ensured visibility of the Fribble, while discouraging large saccades. In between Fribble presentations, a fixation cross appeared centred at the same position. In each scanning session, participants saw 12 presentations of each Fribble, as well as 32 catch trials (two per Fribble) in which the Fribble was repeated. These catch trials were used to monitor participants' attention to the stimuli. The Fribbles were presented in a jittered design, with inter-stimulus-intervals (ISI) of three, four, or five seconds. Each Fribble was preceded by all ISIs four times. The order of presentation was different per participant but the same for both one-back fMRI sessions (i.e., pre and post interaction). Each session was divided into two runs, to give participants a break, stopping scanning in between but leaving participants in the scanner for a few minutes. Each run consisted of three blocks, each containing two presentations of all Fribbles in random order plus five to seven pseudo-randomly interleaved catch trials. In between blocks, there was a 20 s break while a summary of the instructions was presented on the screen. The last five seconds of the pause were counted down on the screen. Each session lasted about 30 min in total.

2.5.3. *Phonological pre-test*

After the fMRI session, but before the interaction, we implemented a pre-test to provide a baseline for potential analyses of participants' speech production in the interaction. Participants were tested one-by-one in the same sound-attenuated booth and using the same audio equipment as for the recording of the interaction. When one of the participants was doing the pre-test, the other participant waited in a separate room, wearing Bose Quietcomfort 35 ii noise-cancelling headphones. The pre-test consisted of two parts. The first part provided a baseline for vowel and diphthong productions. Participants were instructed to read aloud 16 Dutch (non-)words that were presented on the screen. These words consisted of Dutch vowels and diphthongs ([ʏ], [ɛ], [ɪ], [ɔ], [ɑ], [a], [e], [ə], [o], [y], [i], [ø], [u], [ɛɪ], [œy], [ɑʊ]) preceded by an <h> and followed by a <t>, which served as a neutral and constant phonetic context across vowels and diphthongs. The second part of the pre-test served as a baseline for other acoustic characteristics (such as articulation rate, pitch, and /x/ in particular, since this consonant shows clear variation in Dutch) and elicited semi-spontaneous speech. Participants were instructed to read aloud five beginning parts of sentences (ranging from seven to ten words) and to complete them with the first completion that came to their mind. In both parts of the pre-test, participants could click any button on the button box to go to the next (non-)word or sentence. The (non-)words and sentence beginnings were presented centred on the screen. The pre-test lasted about three minutes in total.

2.5.4. *Interaction*

Participants received instructions about the interaction prior to the phonological pre-test. After participants indicated they had finished reading the instructions on their screen, the most important points of

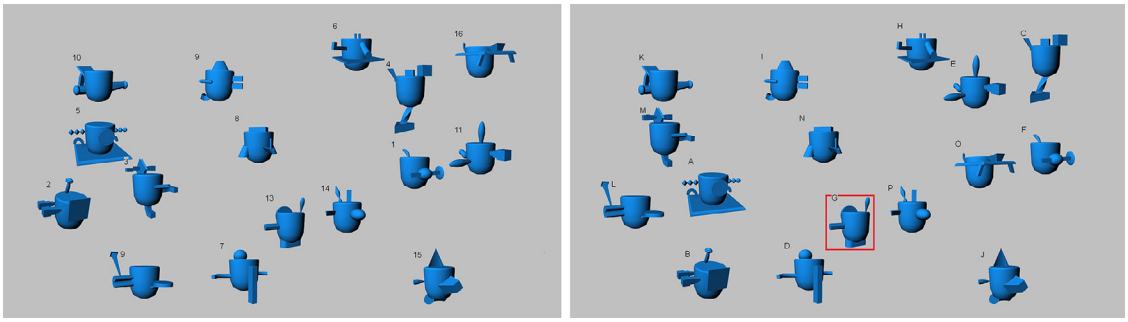

Fig. 4. Example of a trial in the interactional task as shown to the participants (left: the Matcher in this trial and right: the Director in this trial). The Fribble with the red rectangle was the target Fribble for this trial.

the instructions were verbally repeated by the experimenter and participants could ask questions. They then jointly received instructions for the phonological pre-test, which they individually completed in the sound-attenuated booth. When the participants were ready to start with the interaction, they entered the same sound-attenuated booth together and positioned themselves (standing up) behind their respective screens (see Section 2.4.2 and Fig. 3, middle panel). A short summary of the interaction task instructions was presented on the screen, which participants read in silence.

During the interaction, participants saw the 16 Fribbles on their screen in a random arrangement with corresponding numbers or letters (see Fig. 4). Both participants saw the same 16 Fribbles in the same general spatial layout, but 50% of the Fribbles were not positioned in the same locations within this layout (see Fig. 4). On each trial, one of the 16 Fribbles was marked by a red square (the target for that trial) for one of the two participants (the "Director").

The participants completed two different Fribble-related tasks in each trial: the Referential task and the Localisation task. In the Referential task, participants were instructed to communicate with one another so that the Matcher would understand which Fribble was the target on any given trial (i.e., the one marked by the red rectangle in the Director's display). They were informed that they could communicate in any way they wanted (without explicitly mentioning speech and gesture). Once the participant without the red square (the "Matcher") was certain what the target Fribble was, the Matcher said the number or letter of that Fribble out loud and clicked on the yellow button of the button box to move to the Localisation task.

In the Localisation task, participants were instructed to communicate the location of the target Fribble on their screens. They then had to decide whether it was located in the same position on the screen for both participants or not. After reaching agreement, the Matcher pressed the yellow button to indicate "same position" or the red button for "different position". After completing the Referential and Localisation tasks for one Fribble, participants switched roles for the next trial. Participants completed six rounds, each with different spatial layouts for all 16 Fribbles, resulting in 96 trials. The trial order and spatial layout was the same for all pairs. The total interaction phase took about one hour on average (M = 52.24 min, SD = 10.75 min, *range* = 35.20 - 77.48 min).

2.5.5. fMRI movies

A third fMRI session served to enable later hyperalignment of all participants' brains based on functionally similar responses to complex stimuli (Haxby et al., 2011; see *Introduction*). Participants viewed eight animated movies (see Table S2 in Supplementary Materials for details), presented on a part of the screen slightly below the centre, on a black background with a height of 360 pixels and a width of 640 pixels, and a visual angle of about 9–11° vertically and 16–20° horizontally (see Table S2 for details). The movies were in .avi format and were played at 30 frames per second. Each movie was played in its entirety, except for the start and end of the movie (i.e., titles and credits). These were cut off, so that no text was shown to participants. The duration of the movies was 4.1 min per movie on average (*range* = 2.2 - 6.1 min) and 35 min in total, including breaks. The movies were selected to contain categories of objects that were mentioned often in the pilot Naming tests of the Fribbles (see Section 2.3), such as humans, plants, tools, toys, and food. The movies were preceded by a filler video clip that lasted a few seconds. There was a 12 s break between movies. Participants were instructed to simply attend to the movies and to lie still. The movies did not contain (spoken) language, but participants were still provided with the sound of the movies via earphones within the ear, in order to make it easier to stay focused. Before scanning started, the sound of the movies was adjusted to the participants' individual preferences. Note that seven out of the eight movies could not be shared as part of the Dataset, due to copyright issues. The last movie is open source and can be found on the internet (see Table S2 in Supplementary Materials).

2.5.6. Questionnaire

The questionnaire consisted of 30 questions in total and was administered via a computer in the cubicles either after the movie fMRI session or before. In the latter case, the last two questions (about the movies) were administered on paper. The questionnaire consisted of 15 questions relating to the different aspects of the study, (e.g., the goal, strategies in the different tasks, difficulty, level of attention, etc.); nine questions about the other participant (e.g., about their personality, voice, whether they were a real participant, etc.): one open question and eight to be indicated on a 7-point Likert scale (1="not at all", 7="very much"); and six questions about the participant: four demographic (age, sex, occupation, and studies) and two judgement questions, to be indicated on a 7-point Likert scale (intro-/extraversion and whether they were proud of their own accent). For the English translation of the questions asked to participants, see Table S3 in Supplementary Materials or the Data folder in the Dataset.

2.6. Preprocessing

2.6.1. Transforming fMRI data into BIDS structure

All raw MRI data were converted to BIDS (Brain Imaging Data Structure; Gorgolewski et al., 2016) using BIDScoin (version 1.5; Zwiers et al., 2021), including conversion to NiFTi format, and supplemented with standardised metadata. All anatomical MRI scans were defaced to remove identifiable features using a wrapper tool around pydeface (DOI: https://zenodo.org/badge/latestdoi/47563497).

2.6.2. Processing and synchronising audio, video, and Kinect data

The cameras saved multiple consecutive .mp4 files for each interaction (of about 14 min / 3.45 GB each), which were first concatenated and saved as a single .mp4 file for each camera per interaction. This was done for all recordings. For the majority of pairs (see Section 2.7 and Table 1), we also manually synchronised the three videos with Adobe

Table 1
Numbers of participants and pairs for which different types of data are available and/or shared in the Dataset.

Data	Participants (*n*)	Pairs (*n*)
Total collected	142	71
Usable behavioural data (Naming & Features)	140	69
Usable (f)MRI data all sessions	124	56
Usable (f)MRI data sessions one and two	127	59
Usable interactional data	138	69
of which audio & video shared (fully preprocessed)*	*126 (98)*	*63 (49)*
Transcribed interactional data	94	47
of which audio & video shared (fully preprocessed)*	*84 (84)*	*42 (42)*
All usable data (interaction/transcription, MRI, behavioural)	84	42
of which audio & video shared	*74*	*37*

*fully preprocessed = video is synchronised and processed as described in Section 2.6.2.

Premiere Pro CC (version 2018) with the help of the auditory synchronisation signals (see Section 2.3.2).[3] The videos were trimmed and the audio (from the head-mounted microphones as recorded on the cameras) from participants A and B were set to the left (−100) and right (100) audio channel respectively. We then exported six media files which were used for the transcriptions: three video files (as .mp4 files with H.264 codec) and three audio files (as .wav files). The video files were exported with the recorded audio from both participants as stereo channel (where one participant is audible on the left, and the other on the right channel). The audio files were exported at 16-bit sample size; the audio of the individual participants was exported as mono channel in two separate files (one for each participant; in which the other may still be slightly audible), and the combined audio of both participants with stereo channel (similarly to the video exports). All video and audio files were exported with a sample rate of 44,100 Hz. Finally, to enable synchronisation of the video and Kinect data, one additional audio file was exported which included the auditory synchronisation signal. To this end we used custom-made Matlab scripts, where the principle for synchronisation was the same as described above for the videos; the script adjusted the timestamps of the Kinect data to match the videos by time-aligning the digital and auditory synchronisation signals (which were transmitted every 60 s; see Dataset: Preprocessing for the script).

2.6.3. Orthographic transcription procedure for the interaction data

Orthographic transcription of the speech in the interaction phase was done in ELAN (Wittenburg et al., 2006), see Fig. 5. The ELAN files included all synchronised media files for each pair (see Section 2.5.2): three videos (from cameras 1, 2, and 3) and three audio files (from the head-mounted microphones as recorded on the cameras; one file for each participant and another file containing both channels). This allowed the transcribers to inspect the audio waveforms and to listen to each participant separately or both participants simultaneously (which is particularly useful in case of overlapping speech). Three tiers were used for the transcription: two for the transcribed speech, and one on which the transcriber added comments.

Speech was first segmented into Turn-Constructional Units (TCUs; i.e., potentially complete, meaningful utterances, Clayman, 2013; Couper-Kuhlen and Selting, 2017; Schegloff, 2007). If TCUs exceeded 10 s, they were divided into multiple segments of under 10 s length. This was done to allow for optimal automatic forced alignment of the speech into phones for future phonetic analysis.

[3] The off-line synchronisation of the audio/video data from the different cameras did not allow for time-alignment with millisecond precision, since the sampling rate was 29.97 fps (i.e., one frame every 34 ms). We checked the lag between the two audio channels from cameras 1 and 2 after synchronisation, and found this to be 11 ms on average (range: 0-33 ms).

Speech was orthographically transcribed based on the standard spelling conventions of Dutch. All words, discourse particles (e.g., “oh”, “ah”, etc.) and filled pauses (transcribed as “uh” or “um”) were transcribed. Unfinished words were also transcribed but marked. When the transcriber was not certain about their transcription, the respective element was placed between parentheses. When the transcriber could not determine what was said at all, this part was transcribed as a question mark enclosed by parentheses. Non-lexical vocalisations and other sounds were transcribed between asterisks (e.g., *laugh*, *cough*, *lip smack*, etc.). In addition to being transcribed between asterisks, long stretches of laughter during speech were also commented on in the comments-tier.

2.6.4. Linking transcribed speech to task structure

The task structure of the interaction is indicated on the “trial” tier in ELAN (“1.1_ref” etc.; indicating round number (1–6), trial number (1–16), and task (“ref” or “loc”), see Fig. 5). The onsets and offsets of these annotations were manually adjusted; by default, a task ended when the Matcher pressed a button to move to the next task. However, sometimes there was a mismatch between the moment at which participants pressed the button and their speech productions relating to either of the two tasks (e.g., participants would start talking about the location of the Fribble before pressing the button to end the Referential task). In these cases, the onset/offset of the task was placed earlier or later such that the speech about the respective tasks would fall under the right trial annotation.

Once these annotations were finalised in ELAN, we derived the answers for the Referential task (i.e., the letters and numbers that participants said out loud) from the transcripts (note that participants indicated answers for the Localisation task with the button box). We then finalised the ELAN files by adding the following tiers about the task structure and performance: target (Fribble number), director, correct_answer, given_answer, and accuracy.

These ELAN files were then exported to text and Praat TextGrid files for further analyses. The text files were transformed into two datafiles for each pair: one containing all speech annotations (which are linked to trials based on the annotation onsets) and one containing all trial information (i.e., trial onset and offsets, and information about task, target, director, answers, and accuracy). The annotations in the Praat TextGrid files were readjusted to match the original audio files recorded with the recorder (by moving the boundaries of all annotations using a script that can be found in the Preprocessing folder in the Dataset). This was necessary since the original audio files also included the phonological pre-test, and because the audio files from the camera and the recorder were not exactly aligned due to internal clock drift.

2.7. Dataset

2.7.1. Data availability and quality

Incomplete or non-usable data is not shared (see Table 1 for an overview of the number of participants and pairs for which data is shared). First, two participants from different pairs exceeded our age criteria and misunderstood instructions, leaving 140 individual participants, and 69 complete pairs.

MRI data from a total of 16 participants were excluded due to MRI scanner/software malfunction ($n = 6$); missing data due to participant claustrophobia ($n = 4$); excessive motion within or between sessions ($n = 3$); bad performance on the one-back task ($n = 2$); and experimenter error ($n = 1$); see the Participants folder in the Dataset for more details. Note that exclusion of MRI data based on motion or one-back performance was only done for extreme cases. Researchers may still want to deal with motion artifacts and/or errors in the one-back task while preprocessing and analysing the shared data.

There are 124 individual participants and 56 complete pairs with complete and shared (f)MRI data. Furthermore, (f)MRI data of three

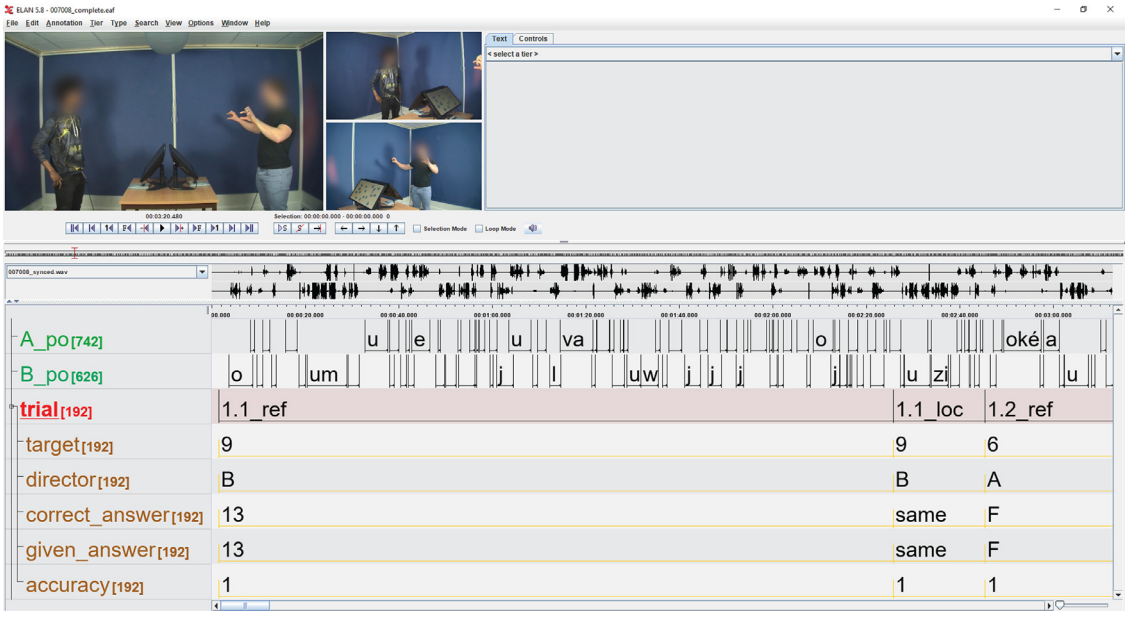

Fig. 5. Screenshot from ELAN file. It includes the synchronised videos from the three cameras (top left; faces are blurred in this figure but not in the videos in the shared dataset), the synchronised audio from both head-mounted microphones (waveforms in the middle), as well as annotations and transcriptions on various "tiers" (rows at the bottom of the window). The first two tiers include the transcribed speech for the participant on the left ("A_po") and right ("B_po"; where "po" stands for practical orthography). The remaining tiers provide information about the task structure and accuracy (see Section 2.6.4).

extra participants is still shared because only session three is incomplete and this may be irrelevant for some users of the data.

Audio/video data from the interactive sessions are shared in the Dataset, except for seven cases where one or both participants did not provide consent. We focused our transcription resources on a selection of 47 interactions, based on their (audio) quality and completeness of the MRI data (see above). However, audio/video data is shared for 42 of these pairs, four pairs contain individuals with non-usable MRI data, and questionnaire data are missing for one pair (these issues were encountered only after transcription had finished). Thus, all data (i.e., fMRI, behavioural, and transcription/interaction data) is usable for 42 pairs (84 participants) in total. For 37 of these 42 pairs, thanks to participant consent, the video and audio data can be shared as well.

2.7.2. Data accessibility

See Data and code availability statement for instructions on how to access the data.

2.7.3. Data structure and format

The Dataset contains the stimuli and the raw and minimally processed data, as well as data and scripts needed to reproduce the results reported in Section 3. It also contains scripts used for running the tasks and scripts or files used for preprocessing the data. Table S4 shows an overview of the different data types and formats in the Dataset.

3. Results

In this section, we present analyses implemented on different parts of the Dataset, chosen to offer an intuition of its characteristics and potential for analysis.

First, we present a short overview of participants' answers to the questionnaire (Section 3.1). Following, we present several measures obtained within the interaction, in which participants communicated to match and localise novel objects (Fribbles). To give a sense of the interactional data, we present data for time on task and accuracy (Section 3.2), number and type of words used (Section 3.3), and gesture characteristics based on an automated analysis pipeline for motion tracking (Kinect, Section 3.4). In these three sections, we report the results for the participant pairs for which the audio/video data has been processed, trial annotations adjusted and the interactions transcribed ($n = 47$).

We also probe correlates of the lexical and conceptual representations of each Fribble in each participant before and after the interaction, using two behavioural measures (Naming and Features) and a brain measure (fMRI). Below, we show how pair members converged on a representation after the interaction (Section 3.5) using one of the behavioural measures (Naming). For the fMRI measurements, we show an indication of the data quality based on the fMRI data before the interaction only (Section 3.6).

Note that all (pre)processing of the data specific for the analyses reported here is also included in this Results section, so that the Methods section could be reserved for a description of the dataset itself. Scripts used for preprocessing, analysis, and figure generation are shared in the Results folder in the Dataset.

3.1. Questionnaire

This section describes participants' experiences and task strategies, as reported in the questionnaire (see Table S3 in Supplementary Materials for all questions asked).

Most participants did not correctly guess the goal of the study. Some participants thought that the study was about object names/perceptions changing (or occasionally: becoming more similar) as a consequence of the interaction.

Main strategies reported for the Naming task were: describing a unique part of the Fribble, describing the objects holistically as existing concepts, or a mix of both strategies. Participants generally did not report using the list of features in their names, only "compact" and "human" were mentioned occasionally. Most participants reported using descriptions from the interaction for all or some of the objects in the post-interaction Naming session.

For the Features task, no clear strategies were mentioned. Participants did report to take into account their name from the Naming task when evaluating the features, especially in the post-interaction session.

During the one-back task in the MRI scanner, most participants used the names they used in the Naming task or in the interaction to remem-

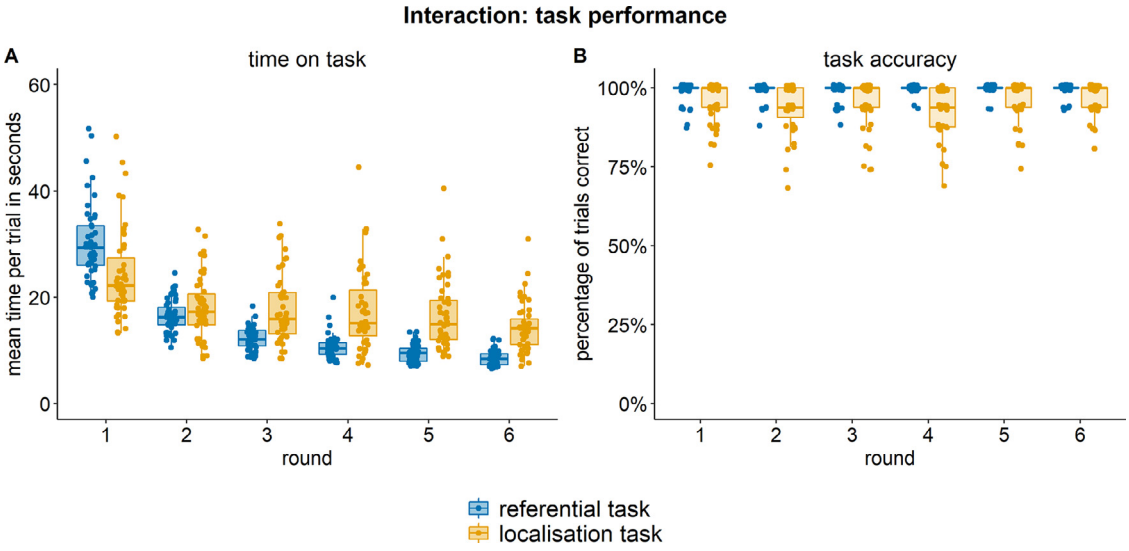

Fig. 6. Distribution of time spent per trial (panel A) and task accuracy (panel B) in the six rounds of the interaction, for the Referential and Localisation task separately. Dots represent pairs (n = 47).

ber the relevant Fribble during the inter-stimulus interval (e.g., through phonological rehearsal), even more so in the post-session.

Strategies for the Referential task in the interaction entailed describing the Fribbles' prominent visual features at first and associating them with known concepts in later rounds. Other mentioned strategies were repeating successful names from earlier rounds and using one's own names from the Naming task. Participants reported using either their own or their interlocutor's initial label or coming up with labels collaboratively. Some participants reported to have retained names that were successful in earlier rounds.

Main strategies for the Localisation task included dividing the screen into four quadrants; dividing the screen into rows and columns, proportions, or percentages; finding patterns or clusters of Fribbles (e.g., triangles, squares); or describing Fribble positions relative to features of the monitor (e.g., centre, logo).

3.2. Interaction: time on task and accuracy

On average, pairs spent almost one hour on the interaction tasks (M = 52.24 min, SD = 10.75 min, *range* = 35.20 - 77.48 min). Pairs spent more time on the Localisation task (M = 28.90 min, SD = 9.00 min, *range* = 15.21 - 54.20 min) than on the Referential task (M = 23.34, SD = 3.44, *range* = 16.54 - 31.85). Fig. 6, panel A shows that the time spent per trial decreased as the interaction progressed, and that this pattern was more pronounced for the Referential task (in blue) compared to the Localisation task (in orange). As for accuracy, pairs performed well (near ceiling) for both the Referential task (M = 99.24% of trials correct, SD = 1.06%, *range* = 94.79 - 100%) and the Localisation task (M = 95.28% of trials correct, SD = 4.60%, *range* = 79.79 - 100%), see Fig. 6, panel B.

3.3. Interaction: number of words and word types

Word counts from the interactional tasks (47 pairs) were analysed to give an insight into the content of the corpus. The transcriptions were first cleaned: we removed unfinished words, non-speech noises, punctuation (indicated here between <>: <#>, <()>, <'>, <,>, <.>, <‘>, <?>, <->), and converted back- and forward slashes (<\> and </>) into spaces. All words that were not completely clear to the transcribers are included in the analyses.

The overall average word count per pair was 8,552 (SD = 2,125, *range* = 5,007 – 15,233, Fig. 7; see Fig. 8 for word types). We distinguished between function words, content words, and interactional markers. The function word list was created from the Dutch Molex lexicon (Gigant-Molex, 2019) and can be found in the script in the Results folder in the Dataset. Interactional markers (also known as procedural conventions; Mills, 2011; Knutsen et al., 2019) are linguistic resources used to manage the interaction. To create the interactional marker list, we took all words in the corpus that did not appear in the CELEX lexical database (Baayen et al., 1996). From this list, we manually removed task responses, content words not present in CELEX (such as names), English words, typos, and spelling variations of words present in CELEX. All words that did not fit the function word list nor the interactional marker list were automatically marked as content words (see Dataset: Results).

Visual inspection of the figures shows that both word token counts and word type counts decrease over the rounds for the Referential task (in blue). For the Localisation task (in orange), these counts are more stable and seem to mostly decrease between rounds 1 and 2, but remain rather stable over the rest of the rounds. This is consistent with the observation from Section 3.1 above that the time used overall in the Localisation task is more stable over rounds as well.

3.4. Motion tracking data and automatic coding of gestures

To inspect the characteristics of manual gestures produced during the interaction, the motion tracking (Kinect) data was analysed for 94 participants (i.e. 47 pairs with complete trial annotations). Kinect sampling rate was regularised at 30 Hz (linear interpolation), timeseries of the right hand were high-pass filtered (2nd order Kolmogorov–Zurbenko filter with a span of 3), followed by filtering of the timeseries derivative. The right hand was chosen for this report to illustrate communicative movements of the dominant hand (all participants were right-handed). The hand tip (middle finger) was chosen as this was a point that captures movement of both the upper and lower arm, wrist, and finger. The resulting time series data can be inspected alongside the videos in ELAN (see Dataset: Results for an example).

3.4.1. Kinematic measures

Our key descriptive measures for communicative manual movements are (autocoded) gesture counts, submovements, gesture duration, and average vertical height.

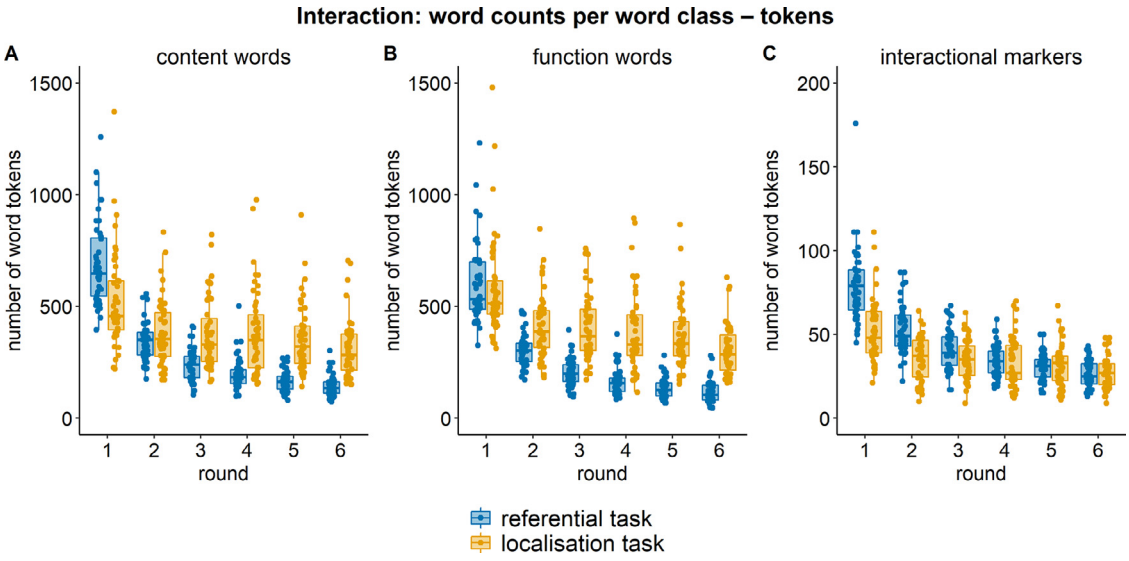

Fig. 7. Word token counts visualised per round, per task, and per pair. Plots for content words (Panel A), function words (Panel B), and interactional markers (Panel C). Note that the scales of the y-axes differ. Dots represent pairs (n = 47).

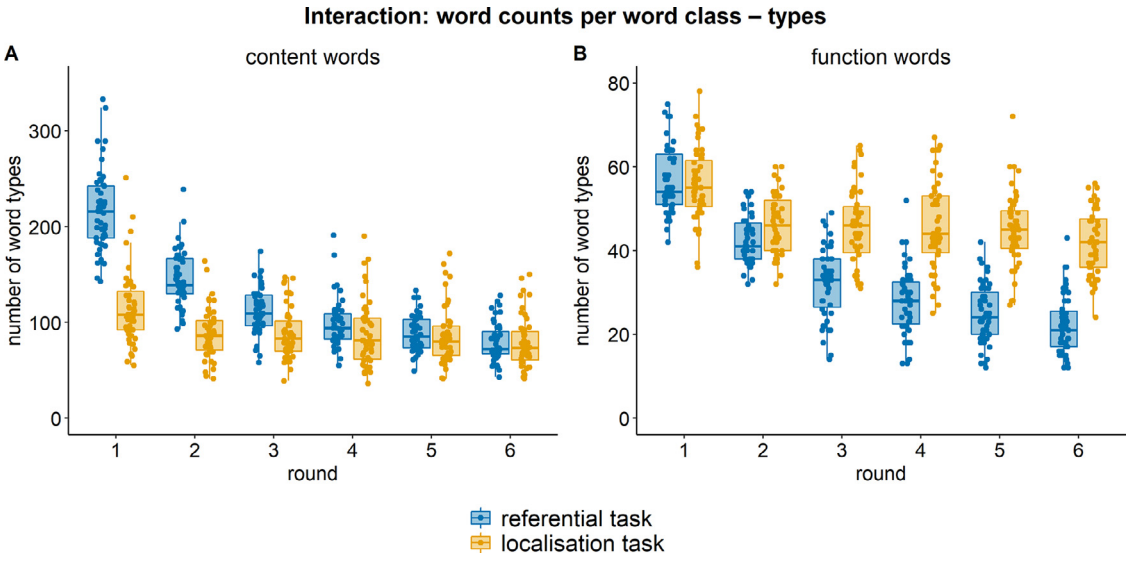

Fig. 8. Counts for different word types per pair visualised per round and per task. Plots for content word types (Panel A) and function word types (Panel B). Note that the scales of the y-axes differ. Dots represent pairs (n = 47).

Autocoder of communicative gestures. We employed a rule-based automatic movement coder (Pouw et al., 2021a) to approximate communicative gestures that were made during the interactions. The autocoder takes as input the speed and position of the hands to approximate gesture events (see Section S1 in Supplementary Materials for details). Previously we have applied our autocoder on a dataset with a similar design, where we tested the performance of the autocoder relative to human annotations of iconic gestures (Pouw et al., 2021a). In that study we found that (p. 11) the human coded iconic gestures were positively related to the auto-coded gestures, r == 0.60, $p < 0.001$, with a 65.2% accuracy in time overlap between these codings (*true positive* = 70%, *false positive* = 86%, *true negative* = 93%, *false negative* = 1%).

Submovements. Kinematic submovements are computed on the right hand tip speed by counting the number of positive local peaks that exceed 15 cm/s during an autocoded gesture event. Following earlier work (Trujillo et al., 2018, 2019), we assume that gestures designed to communicate tend to have more submovements. Measures akin to submovements have been found to strongly correlate with the number of information units human annotators perceive in the gesture (Pouw et al., 2021b). Thus the number of submovements is a kinematic measure that approximates the number of semantic units of the gesture.

Gesture duration. Duration in seconds of the autocoded gesture event.

Average vertical height. Average vertical position within each gesture event. Following earlier work (Trujillo et al., 2018, 2019), we assume that the degree to which a gesture is forefronted in a more prominent gesture space is an informative kinematic quality about the degree of saliency of the gesture.

3.4.2. Descriptive results kinematics

Fig. 9 shows the main results of the kinematic measures as they develop over the rounds. It can be seen that autocoded gesture count drastically decreases over the rounds, and the number of submovements of these gestures also decrease over time. Gesture duration and vertical height also follow this pattern but in a less pronounced way. These

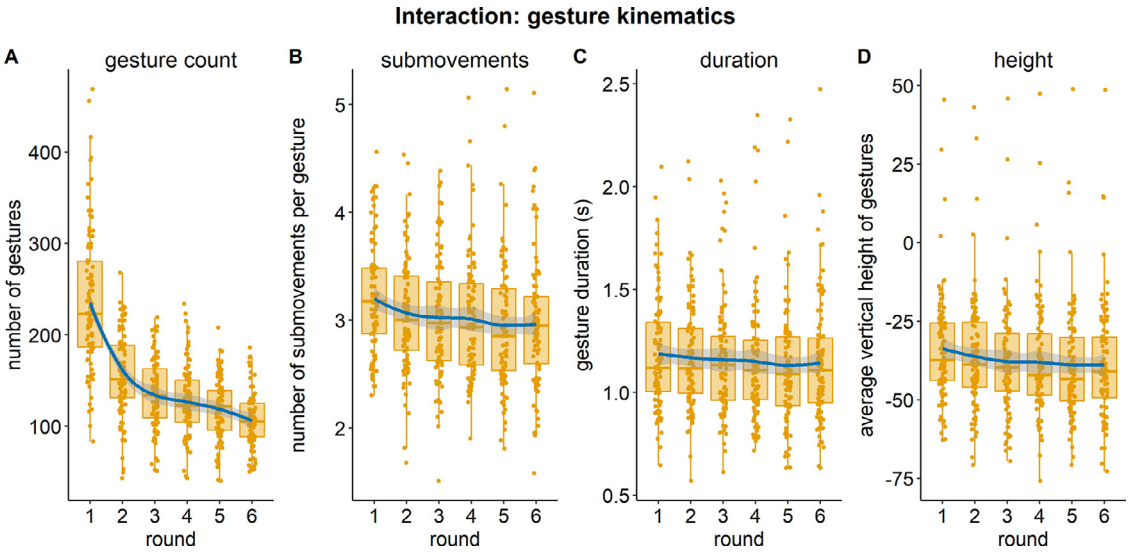

Fig. 9. Descriptives of the motion tracking data; each jitter point represents kinematic descriptives (of autocoded right-handed gestures) for one round for one participant (n = 94 participants). The trend line reflects the smoothed means using a Local Polynomial Regression (loess) Fit.

quantitative patterns in the kinematic data likely relate to the common ground that is built over the rounds (e.g., Holler and Bavelas, 2017; Holler et al., 2022; but see Hoetjes et al., 2015 for slightly different results in a referential task with stimuli similar to the Fribbles), and to kinematic optimisation of gestural signalling (e.g., Pouw et al., 2021b).

3.5. Lexical similarity in the Naming task

Since the interactional tasks were expected to lead to conceptual alignment between participants of a pair, here we report their degree of lexical alignment, as one of the proxies for conceptual alignment. In order to do so, we calculated the similarity of the names for the same Fribble between two pair members (that formed an actual pair in the interaction), both before and after the interaction, for the 69 pairs that had usable behavioural data (see Table 1, second row). These are referred to as "real pairs" from here on. To ensure that such lexical alignment was specific to the individual interactions, we also calculated the increase in lexical alignment from before to after the interaction in all possible "pseudo-pairs": pairs of participants who engaged in the tasks in different roles but who did not interact with each other (n = 4,692). The increase of the real pairs was compared to a permutation distribution (n = 10,000) of lexical alignment difference scores from all possible (real and pseudo-)pairs. Pseudo-pairs provide a rigorous control for systematic but communicatively un-specific effects of task performance and task structure.

The text written by the participants in the Naming task (one to three words per Fribble) was regularised by removing special characters (indicated here between <>: <'>, <">, <()>, <&>, <+>, <.>, <;>) except if they were part of a word, converting the character </> into a space, <=> into the word <is>, correcting spelling errors for a small number of frequently occurring words/names (i.e., *Pippi Langkous* ("Pippi Longstocking"), *Pinokkio* ("Pinocchio"), *plateau* ("plateau"), and *trofee* ("trophy"), converting uppercase characters to lowercase, and converting numbers into the corresponding words, except if the number was part of a word (e.g., 3d). The words were then checked against the NLPL Dutch CoNLL17 corpus (Zeman et al., 2017). Only words missing from the corpus were changed by correcting their spelling or dividing compounds (or words split with a <-> character) into two (or more) words. Note that this procedure may have led to an underestimation of name similarity if two differently spelled versions of the same word were both present in the corpus. For two unidentified words in the corpus it was not clear how they should be corrected, so these were left as such.

Naming similarity between names for the same Fribble was operationalised here as lexical similarity, that is, how many words were (exactly) the same between the two names, normalised by the number of words. To compute this, the cosine similarity of the names was taken, resulting in a score between 0 (no words are the same) and 1 (all words are the same; see also Duran et al., 2019; Rasenberg et al., 2022). As an example, the names "trophy triangle plateau" and "trophy with blocks" led to a similarity of 0.33 because one out of three words was the same.

Real pairs numerically showed an increase in lexical similarity from before the interaction (M = 0.06, $Median$ = 0) to after the interaction (M = 0.38, $Median$ = 0.33), see Fig. 10, left panel. The average difference score (M = 0.32) was tested against a permutation distribution of 10,000 average difference scores each calculated from 69 pairs that were randomly drawn from a pool of all possible pseudo-pairs (n = 4,692; see Fig. 10, right panel) and all real pairs (n = 69). The average difference score for real pairs clearly lies above this distribution (p = 0), showing that the increase in lexical alignment for real pairs cannot be due to mere experience with the task.

3.6. Relation of fMRI data to visual similarities of the Fribbles

As a general check of the fMRI data quality, we performed a correlation analysis between pairwise visual similarities of the Fribble-images and pairwise brain-pattern similarities related to viewing the Fribbles before the interaction. We entered 112 participants in this analysis, which constitute the 56 pairs for which both participants had usable fMRI data in all sessions (see Table 1, third row).

Each fMRI data run was spatially aligned, coregistered to the corresponding anatomical T1 scan, and spatially normalised (MNI space). Then, a general linear model was fitted to the data, considering a regressor for each of the Fribbles per session (as well as 9 nuisance regressors capturing signal variance related to head motion, and signals from the cerebro-spinal fluid, white matter, and out-of-brain compartments). The resulting stimulus specific beta weights were then used to create 16-by-16 dissimilarity matrices containing all pairwise dissimilarities between brain patterns of the 16 Fribbles for searchlights (radius = 9 mm/4.5 voxels, within a grey matter mask) through the brain. The neural dissimilarity matrices were correlated with a 16-by-16 Fribble dissimilarity matrix calculated as one minus the Structural Similarity Index (Wang et al., 2004), a metric of visual similarity between Fribbles. The resulting Fisher-Z transformed correlation values per participant and searchlight were subjected to a second level permutation analysis

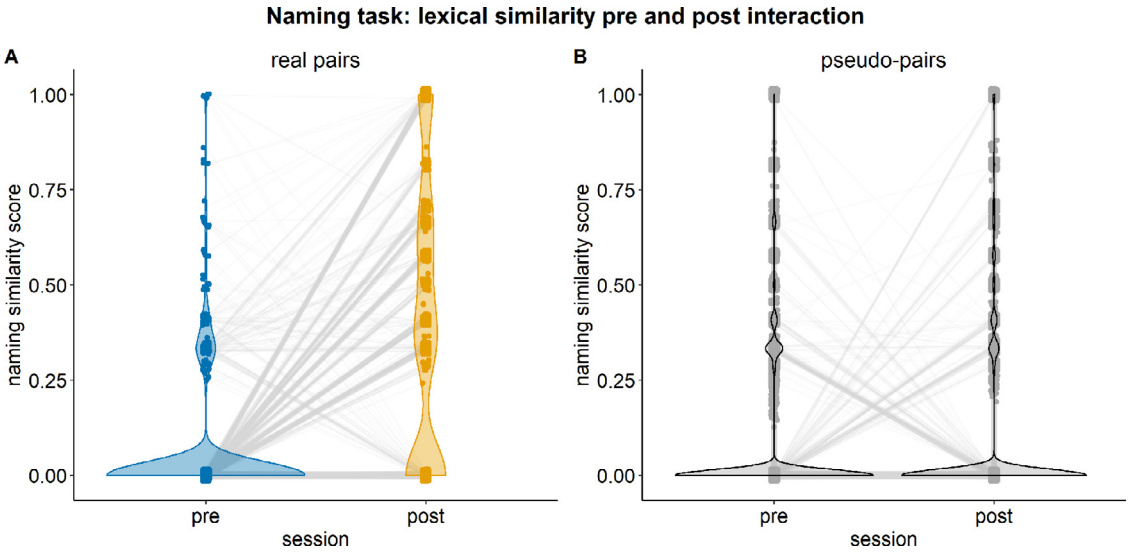

Fig. 10. Distribution of Naming similarity scores (i.e., cosine similarity between the names provided by two participants of a pair for a particular Fribble), before (pre) and after (post) the interaction. Panel A shows results from real pairs, and panel B from pseudo-pairs. Dots represent individual data points (n = 1,004 (69 pairs by 16 Fribbles) for real pairs and n = 75,072 (4,692 pairs by 16 Fribbles) for pseudo-pairs).

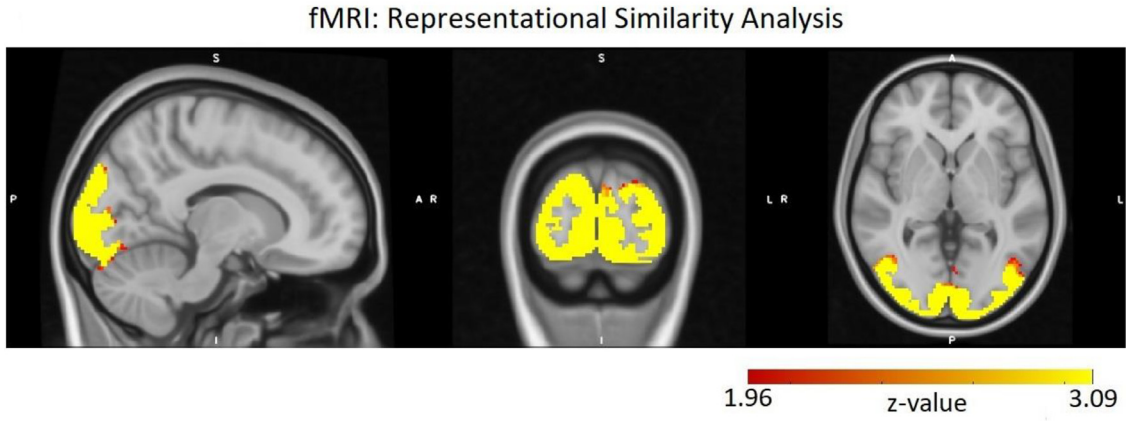

Fig. 11. Visualisation of significant correlations between visual similarities and similarities of brain patterns between the 16 Fribbles throughout the brain, shown in a sagittal (left), coronal (middle), and axial (right) slice. L = left, R = right, A = anterior, P = posterior. Colours indicate TFCE-corrected Z-values from the second level analysis across participants, above the two-sided significance threshold of 1.96.

(n = 1,000 iterations) over participants with TFCE correction for multiple comparisons (Smith and Nichols, 2009). The resulting Z-values are shown in Fig. 11 with a significance threshold of 1.96 (corresponding to two-sided $p < 0.05$). As expected, the areas with significant positive correlations are mainly located around the visual cortex.

4. Discussion

This paper describes a large dataset consisting of (transcribed) speech, audio, video, and motion-tracking data during face-to-face task-based interaction about novel objects (Fribbles), as well as pre- and post-interactional behavioural and fMRI measures, estimating representations of the Fribbles from 71 pairs of participants.

We discuss aspects of this dataset on the basis of the reported results to demonstrate its quality and to provide suggestions for potential uses of the data. We deliberately refrain from embedding the dataset in strong theoretical assumptions to avoid biasing potential uses and to allow for different hypotheses to be tested by a wide range of researchers.

On average, each pair spent about an hour performing the two interactional tasks (i.e., the Referential and Localisation tasks), spending slightly more time on the Localisation task. Participants performed near ceiling on both tasks. The time pairs spent per trial, as well as the number of function and content word types and tokens they used, descriptively decreased over the six rounds for the Referential task, whereas these measures appeared more stable for the Localisation task. An exception to this pattern is that participants appeared to use a similar number of interactional markers within the two tasks in later rounds. Over the rounds, decrease in time on task and word counts in the Referential task and, to a reduced extent, in the Localisation task, is likely to be related to the building of common ground. This pattern is in line with earlier work using repeated reference games, which are known to elicit increasingly shorter referential expressions from participants over rounds (e.g., Hawkins et al., 2020; Kraus and Weinheimer, 1964; Clark and Wilkes-Gibbs, 1986). Given that the Fribbles themselves remained the same in each round, whereas their locations changed, it is to be expected that more common ground can be built up in the Referential than in the Localisation task.

Furthermore, the amount, duration, and average vertical height of gestures, as well as the amount of gesture sub-movements decreased over interactional rounds. This general decrease in gesture count, size, and complexity over the interaction was expected on the basis of previous research showing that such modulations follow from the building

of common ground (e.g., Holler and Bavelas, 2017; Holler et al., 2022) and the kinematic optimisation of gestural signalling (e.g., Pouw et al., 2021b). Taken together with the previously described results regarding speech, it appears that, descriptively, attenuations in speech and gesture over time go hand in hand, which is also in line with earlier work (Holler and Bavelas, 2017).

In the Naming task, participants named all Fribbles using one to three words. Pairs generally showed a larger lexical similarity between their names or descriptions of the Fribbles after the interaction than before, an increase that proved highly reliable when compared to permutations including pseudo-pairs (formed post-hoc by pairing up participants who did not interact with each other). This result confirms that the interaction led to convergence of naming conventions for the Fribbles.

Regarding the fMRI data, correlations between the similarity of brain-activation patterns in response to the Fribbles and objective visual similarity of the Fribbles were highest around the visual cortex. This expected result shows the quality of the (f)MRI data.

These results show that the interactional data (linguistic and kinematic), the computer-based behavioural measures, as well as the brain imaging data are of high quality. This means that this dataset can be used for a large range of analyses on interactions (e.g., phonetic, lexical, syntactic, semantic, pragmatic, and gestural analyses). One key objective of the CABB team for collecting the data was to investigate alignment in the interactions, and therefore the dataset is well suited to quantify the degree to which participants align their linguistic and/or bodily behaviour at different levels (Pickering and Garrod, 2004; see Rasenberg et al., 2020 for an overview of different definitions and measures of alignment). A wide range of analyses is possible, which may be qualitative or quantitative in nature, recruiting manual or (semi-)automatic procedures to analyse the audio-video data (e.g., with OpenPose, Cao et al., 2017) or transcripts (e.g., with the Python package ALIGN, Duran et al., 2019).

For example, one could measure the degree of similarity between participants' realisations of different phonemes over the course of the interaction (e.g., Pardo, 2006). At the prosodic level, one may compare pitch or articulation rate (e.g., Eijk et al., 2019). The phonological pre-test (see Section 2.5.3) is useful for such analyses, since it provides a baseline of participants' speech before they start interacting with each other. At the syntactic level one could compare N-grams (e.g., Fusaroli et al., 2017; Reitter and Moore, 2014) or look at specific syntactic constructions (e.g., Hartsuiker et al., 2008). At the lexical level, one could quantify how often participants use the same words (e.g., Bangerter et al., 2020; Brennan and Clark, 1996). At the semantic level, word2vec or similar distributional models (e.g., Mandera et al., 2017) could be used to measure semantic similarity between participants' speech turns (e.g., Dideriksen et al., 2019). In terms of bodily behaviour, one could for example look at how people align their posture (e.g., Shockley et al., 2003) or at the type and form of their gestures (see e.g., Bergman and Kopp, 2012; Chui, 2014; Holler and Wilkin, 2011; Louwerse et al., 2012). The present dataset is especially suited to perform such gestural analyses, given the rich set of (mostly iconic) gestures elicited by the task (see Section 3.4.2) and the availability of Kinect measurements for quantitative analyses. It is also possible to test hypotheses about how alignment at different levels and/or modalities is related to each other (e.g., Cappellini et al., 2022; Mahowald et al., 2016; Oben and Brône, 2016; Pickering and Garrod, 2004; Rasenberg et al., 2022).

Furthermore, the task-based interactions – in which the establishment of mutual understanding is a challenge – allow researchers to investigate the interactional mechanisms that people use to solve coordination problems, such as other-initiated repair (Schegloff et al., 1977; Schegloff, 2000). Given the relatively free-form of the interactions (in which people were free to communicate in any way they wanted, without time constraints), the data can be used to analyse various (multimodal) interactional phenomena, such as turn-taking (Sacks et al., 1974) or the use of backchannels or acknowledgements (Allwood et al., 1992; Jefferson, 1984; Yngve, 1970).

In addition, the dataset allows researchers to examine whether the interactions result in changes in the estimated representations of the Fribbles, and whether the representations of pair members tend to converge. Such hypotheses could be tested in several ways using the present dataset, given the availability of both brain data and two types of behavioural data. The results provided here (see Section 3.5) show that interacting participants converge in the sense that they more often use the same words to refer to the Fribbles after the interaction than before the interaction. In a similar vein, one could investigate such convergence in terms of semantic similarity of the names, similar scores given to the features, and similar brain activation patterns in fMRI measurements between participants. The latter analysis is further facilitated by the possibility to implement functional hyperalignment of participants (Haxby et al., 2011; see *Introduction*).

Moreover, the unique feature of this dataset is the combination of linguistic, behavioural, and neural data within the same paradigm and for the same stimuli and participants, opening up the possibility for a systematic investigation of the relation between them. This in turn, may make it possible to find support for or against specific hypotheses regarding the relationship between certain characteristics of the interactions, which may support mutual understanding and convergence between participants in estimated representations. As clearly shown by Fig. 10, panel A, in Section 3.5, pairs display quite some variability with regards to lexical alignment in the Naming task after the interaction. It may be possible to find characteristics of the interaction that can explain such variance to some extent. In conclusion, the present dataset ultimately allows researchers to provide a comprehensive picture of both the behavioural aspects of multimodal interaction and associated changes in representations of the interactional referents, estimated using behavioural as well as neural measures.

Data and code availability statement

The Dataset is stored as a Research Documentation Collection in the Donders Repository (https://data.donders.ru.nl/). Note that the Dataset is not publicly available, since participants specifically consented to their sensitive (audio and video) data being used by researchers for scientific purposes only. To ensure this and to warrant secure data storage and sharing of these sensitive data, a request for access must be submitted to the Dataset managers by signing a Data Use Agreement (provided as a separate pdf file in Supplementary Materials), which specifies the conditions and restrictions under which the data is shared. Specifically, conditions are specified regarding the secure data storage of the data (see the Appendix of the Data Use Agreement for details) and the restriction that the data is used for scientific purposes only. Furthermore, it is specified that users should acknowledge the origin of the data as follows: "Data were provided (in part) by the Radboud University, Nijmegen, The Netherlands" and that they should cite the present paper in papers or other presentations using the data. Importantly, it is specified that "neither the Radboud University, nor the researchers that provide this data should be included as an author of publications or presentations if this authorship would be based solely on the use of this data."

In short, to be able to access and download the data, two steps are required. First, you need to create a user profile in the Donders Repository by logging in with your SURFconext or ORCID account (https://data.donders.ru.nl/login). For more information about the ORCID option and alternative ways to login, see: https://data.donders.ru.nl/doc/help/helppages/user-manual/login-profile.html?8. Second, the Data Use Agreement needs to be completed and sent to ivan.toni@donders.ru.nl. Upon completion of these steps, users will be granted access to the collection and can view and download files through the Donders Repository website (for more information, see: https://data.donders.ru.nl/doc/help/user-manual/transfer-data.html).

Declaration of Competing Interest

The authors declare no competing interests.

Credit authorship contribution statement

Lotte Eijk: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Resources, Software, Supervision, Visualization, Writing – original draft, Writing – review & editing. **Marlou Rasenberg:** Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Resources, Software, Supervision, Visualization, Writing – original draft, Writing – review & editing. **Flavia Arnese:** Data curation, Formal analysis, Investigation, Software, Writing – review & editing. **Mark Blokpoel:** Conceptualization, Formal analysis, Funding acquisition, Project administration, Visualization, Writing – review & editing. **Mark Dingemanse:** Conceptualization, Methodology, Supervision, Validation, Writing – review & editing. **Christian F. Doeller:** Conceptualization. **Mirjam Ernestus:** Conceptualization, Methodology, Supervision, Writing – review & editing. **Judith Holler:** Conceptualization, Methodology, Writing – review & editing. **Branka Milivojevic:** Conceptualization, Data curation, Methodology, Project administration, Resources, Software, Supervision, Writing – review & editing. **Asli Özyürek:** Conceptualization, Funding acquisition, Methodology, Supervision, Writing – review & editing. **Wim Pouw:** Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. **Iris van Rooij:** Conceptualization, Funding acquisition, Methodology, Writing – review & editing. **Herbert Schriefers:** Conceptualization, Methodology, Supervision, Writing – review & editing. **Ivan Toni:** Conceptualization, Formal analysis, Funding acquisition, Methodology, Supervision, Writing – original draft, Writing – review & editing. **James Trujillo:** Data curation, Methodology, Software, Writing – review & editing. **Sara Bögels:** Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Resources, Software, Supervision, Visualization, Writing – original draft, Writing – review & editing.

Data availability

The data will be shared under conditions specified within the Data and Code Availability Statement.

Acknowledgements

We would like to thank everyone who helped us realise the creation of this dataset. Specifically, we thank all student and research assistants for help with data collection: Yvonne van der Hoeven, Sebastian Idesis, and Inez Wijnands, and with data processing: Maarten van den Heuvel, Anouck Hollander, Daniëlle Reuvekamp, and Joost Vossers. We thank Margret van Beuningen, Paul Gaalman, Jeroen Geerts, Uriel Plönes, and Bob Rosbag for their technical support. We thank Lisette Albers (YourType) and Yvonne van der Hoeven for transcribing the speech from the interactions.

We furthermore thank these researchers for their input during various stages of this project: Jana Basnakova, Laura van de Braak, Naomi de Haas, Stephen Levinson, Rui Liu, David Neville, Jan-Matthijs Schoffelen, Arjen Stolk, and Marieke Woensdregt.

This work was funded by the Netherlands Organisation for Scientific research, through a gravitation grant (024.001.006) to the Language in Interaction Consortium.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.neuroimage.2022.119734.

References

Allwood, J., Nivre, J., Ahlsén, E., 1992. On the semantics and pragmatics of linguistic feedback. J. Semant. 9 (1), 1–26. doi:10.1093/jos/9.1.1.

Anderson, A.H., Bader, M., Bard, E.G., Boyle, E., Doherty, G., Garrod, S., Isard, S., Kowtko, J., McAllister, J., Miller, J., Sotillo, C., Thompson, H.S., Weinert, R., 1991. The HCRC map task corpus. Lang. Speech 34 (4), 351–366. doi:10.1177/002383099103400404.

Baayen, R.H., Piepenbrock, R., Gulikers, L., 1996. The CELEX Lexical Database (CD-ROM).

Bangerter, A., Mayor, E., Knutsen, D., 2020. Lexical entrainment without conceptual pacts? Revisiting the matching task. J. Mem. Lang. 114, 104129. doi:10.1016/j.jml.2020.104129.

Barry, T.J., Griffith, J.W., De Rossi, S., Hermans, D., 2014. Meet the Fribbles: novel stimuli for use within behavioural research. Front. Psychol. 5. doi:10.3389/fpsyg.2014.00103.

Bergmann, K., Kopp, S., 2012. Gestural Alignment in Natural Dialogue. In: Miyake, N., Peebles, D., Cooper, R.P. (Eds.), Proceedings of the 34th Annual Conference of the Cognitive Science Society. Cognitive Science Society, pp. 1326–1331.

Binder, J.R., Conant, L.L., Humphries, C.J., Fernandino, L., Simons, S.B., Aguilar, M., Desai, R.H., 2016. Toward a brain-based componential semantic representation. Cogn. Neuropsychol. 33 (3–4), 130–174. doi:10.1080/02643294.2016.1147426.

Bracci, S., Caramazza, A., Peelen, M.V., 2015. Representational similarity of body parts in human occipitotemporal cortex. J. Neurosci. 35 (38), 12977–12985. doi:10.1523/JNEUROSCI.4698-14.2015.

Brennan, S.E., Clark, H.H., 1996. Conceptual pacts and lexical choice in conversation. J. Exp. Psychol. Learn. Mem. Cogn. 22 (6), 1482–1493. doi:10.1037/0278-7393.22.6.1482.

Brône, G., Oben, B., 2015. InSight interaction: a multimodal and multifocal dialogue corpus. Lang. Resour. Eval. 49 (1), 195–214. doi:10.1007/s10579-014-9283-2.

Cao, Z., Simon, T., Wei, S.E., Sheikh, Y., 2017. Realtime multi-person 2d pose estimation using part affinity fields. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7291–7299.

Cappellini, M., Holt, B., Hsu, Y.Y., 2022. Multimodal alignment in telecollaboration: a methodological exploration. System, 102931 doi:10.1016/j.system.2022.102931.

Chui, K., 2014. Mimicked gestures and the joint construction of meaning in conversation. J. Pragmat. 70, 68–85. doi:10.1016/j.pragma.2014.06.005.

Clark, H.H., 1996. Using Language. Cambridge University Press.

Clark, H.H., 1997. Dogmas of understanding. Discourse Process. 23 (3), 567–582. doi:10.1080/01638539709545003.

Clark, H.H., Brennan, S.E., 1991. Grounding in communication. In: Resnick, L.B., Levine, J.M., Teasley, S.D. (Eds.), Perspectives on Socially Shared Cognition. American Psychological Association, pp. 127–149. doi:10.1037/10096-006.

Clark, H.H., Wilkes-Gibbs, D., 1986. Referring as a collaborative process. Cognition 22 (1), 1–39. doi:10.1016/0010-0277(86)90010-7.

Clayman, S.E., 2013. Turn-constructional units and the transition-relevance place. In: The Handbook of Conversation Analysis. John Wiley & Sons, Ltd, pp. 151–166. doi:10.1002/9781118325001.ch8.

Couper-Kuhlen, E., Selting, M., 2017. Interactional Linguistics: Studying Language in Social Interaction. Cambridge University Press.

Dideriksen, C., Fusaroli, R., Tylén, K., Dingemanse, M., Christiansen, M.H., 2019. Contextualizing Conversational Strategies: backchannel, Repair and Linguistic Alignment in Spontaneous and Task-Oriented Conversations. In: Goel, A.K., Seifert, C.M., Freksa, C. (Eds.), Proceedings of the 41st Annual Conference of the Cognitive Science Society. Cognitive Science Society, pp. 261–267.

Dobs, K., Isik, L., Pantazis, D., Kanwisher, N., 2019. How face perception unfolds over time. Nat. Commun. 10 (1), 1–10. doi:10.1038/s41467-019-09239-1.

Duran, N.D., Paxton, A., Fusaroli, R., 2019. ALIGN: analyzing linguistic interactions with generalizable techNiques - a Python library. Psychol. Methods 24 (4), 419–438. doi:10.1037/met0000206.

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 LREC, pp. 2992–2995.

Eijk, L., Ernestus, M., Schriefers, H., 2019. Alignment of pitch and articulation rate. In: Proceedings of the 19th International Congress of Phonetic Sciences. Melbourne, Australia, pp. 2690–2694.

Ernestus, M., Baayen, R.H., 2011. Corpora and exemplars in phonology. In: The Handbook of Phonological Theory. Wiley-Blackwell, pp. 374–400.

Fusaroli, R., Tylén, K., Garly, K., Steensig, J., Christiansen, M.H., Dingemanse, M., 2017. Measures and mechanisms of common ground: backchannels, conversational repair, and interactive alignment in free and task-oriented social interactions. In: Gunzelmann, G., Howes, A., Tenbrink, T., Davelaar, E. (Eds.), Proceedings of the 39th Annual Conference of the Cognitive Science Society. Cognitive Science Society, pp. 2055–2060.

Gigant-Molex (Version 1.0) (2019) [Data set]. Available at the Dutch Language Institute: http://hdl.handle.net/10032/tm-a2-p9

Gorgolewski, K.J., Auer, T., Calhoun, V.D., Craddock, R.C., Das, S., Duff, E.P., Poldrack, R.A., 2016. The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments. Sci. Data 3 (1), 1–9. doi:10.1038/sdata.2016.44.

Hartsuiker, R.J., Bernolet, S., Schoonbaert, S., Speybroeck, S., Vanderelst, D., 2008. Syntactic priming persists while the lexical boost decays: evidence from written and spoken dialogue. J. Mem. Lang. 58 (2), 214–238. doi:10.1016/j.jml.2007.07.003.

Hawkins, R.D., Frank, M.C., Goodman, N.D., 2020. Characterizing the dynamics of learning in repeated reference games. Cogn. Sci. 44 (6), e12845. doi:10.1111/cogs.12845.

Haxby, J.V., Guntupalli, J.S., Connolly, A.C., Halchenko, Y.O., Conroy, B.R., Gobbini, M.I., Hanke, M., Ramadge, P.J., 2011. A common, high-dimensional model of the rep-

resentational space in human ventral temporal cortex. Neuron 72 (2), 404–416. doi:10.1016/j.neuron.2011.08.026.
Hoetjes, M., Koolen, R., Goudbeek, M., Krahmer, E., Swerts, M., 2015. Reduction in gesture during the production of repeated references. J. Mem. Lang. 79–80, 1–17. doi:10.1016/j.jml.2014.10.004.
Holler, J., Bavelas, J., 2017. Multi-modal communication of common ground: a review of social functions. In: Church, R.B., Alibali, M.W., Kelly, S.D. (Eds.), Why Gesture? How the Hands Function in Speaking, Thinking and Communicating. Benjamins, pp. 213–240.
Holler, J., Bavelas, J.B., Woods, J., Geiger, M., Simons, L., 2022. Given-new effects on the duration of gestures and of words in face-to-face dialogue. Discourse Process..
Holler, J., Wilkin, K., 2011. Co-speech gesture mimicry in the process of collaborative referring during face-to-face dialogue. J. Nonverbal Behav. 35 (2), 133–153. doi:10.1007/s10919-011-0105-6.
Hutchins, E., Hazlehurst, B., 1995. How to invent a shared lexicon: the emergence of shared form-meaning mappings in interaction. In: Goody, E.N. (Ed.), Social Intelligence and Interaction. Cambridge University Press, Cambridge, pp. 189–205.
Jefferson, G., 1984. Notes on a systematic deployment of the acknowledgement tokens “yeah”; and “mm hm. Pap. Linguist. 17 (2), 197–216. doi:10.1080/08351818409389201.
Knutsen, D., Bangerter, A., Mayor, E., Zwaan, R., Dingemanse, M., 2019. Procedural coordination in the matching task. Collabra Psychol. 5 (1). doi:10.1525/collabra.188.
Krauss, R.M., Weinheimer, S., 1964. Changes in reference phrases as a function of frequency of usage in social interaction: a preliminary study. Psychon. Sci. 1 (1), 113–114. doi:10.3758/BF03342817.
Louwerse, M.M., Dale, R., Bard, E.G., Jeuniaux, P., 2012. Behavior matching in multimodal communication is synchronized. Cogn. Sci. 36 (8), 1404–1426. doi:10.1111/j.1551-6709.2012.01269.x.
Mahowald, K., James, A., Futrell, R., Gibson, E., 2016. A meta-analysis of syntactic priming in language production. J. Mem. Lang. 91, 5–27. doi:10.1016/j.jml.2016.03.009.
Mandera, P., Keuleers, E., Brysbaert, M., 2017. Explaining human performance in psycholinguistic tasks with models of semantic similarity based on prediction and counting: a review and empirical validation. J. Mem. Lang. 92, 57–78. doi:10.1016/j.jml.2016.04.001.
McNeill, D., 1992. Hand and Mind: What Gestures Reveal about Thought. University of Chicago Press.
Mills, G.J., 2011. The emergence of procedural conventions in dialogue. In: Bar-el, L.A., Hölscher, C., Shipley, T. (Eds.), Proceedings of the 33rd Annual Conference of the Cognitive Science Society.
Oben, B., Brône, G., 2016. Explaining interactive alignment: a multimodal and multifactorial account. J. Pragmat. 104, 32–51. doi:10.1016/j.pragma.2016.07.002.
Oostdijk, N. (2000). Het Corpus Gesproken Nederlands. 5, 280–284.
Pardo, J.S., 2006. On phonetic convergence during conversational interaction. J. Acoust. Soc. Am. 119 (4), 2382–2393. doi:10.1121/1.2178720.
Pickering, M.J., Garrod, S., 2004. Toward a mechanistic psychology of dialogue. Behav. Brain Sci. 27 (2), 169–190. doi:10.1017/S0140525X04000056.
Pouw, W., De Wit, J., Bögels, S., Rasenberg, M., Milivojevic, B., Ozyurek, A., 2021a. Semantically related gestures move alike: towards a distributional semantics of gesture kinematics. In: Duffy, V.G. (Ed.), Proceedings of the Digital Human Modeling and Applications in Health, Safety, Ergonomics and Risk Management. Human Body, Motion and Behavior. HCII 2021, 12777 doi:10.1007/978-3-030-77817-0_20, Lecture Notes in Computer ScienceSpringer, Cham.
Pouw, W., Dingemanse, M., Motamedi, Y., Özyürek, A., 2021b. A systematic investigation of gesture kinematics in evolving manual languages in the lab. Cogn. Sci. 45 (7), e13014. doi:10.1111/cogs.13014.
Rasenberg, M., Özyürek, A., Bögels, S., Dingemanse, M., 2022. The primacy of multimodal alignment in converging on shared symbols for novel referents. Discourse Process. 59 (3), 209–236. doi:10.1080/0163853X.2021.1992235.
Rasenberg, M., Özyürek, A., Dingemanse, M., 2020. Alignment in multimodal interaction: an integrative framework. Cogn. Sci. 44 (11), e12911. doi:10.1111/cogs.12911.
Rasenberg, M., Pouw, W., Özyürek, A., Dingemanse, M., 2022. The multimodal nature of communicative efficiency in social interaction. Scientific Reports doi:10.1038/s41598-022-22883-w, in press.
Rauchbauer, B., Nazarian, B., Bourhis, M., Ochs, M., Prévot, L., Chaminade, T., 2019. Brain activity during reciprocal social interaction investigated using conversational robots as control condition. Philos. Trans. R. Soc. B 374 (1771), 20180033. doi:10.1098/rstb.2018.0033.
Reitter, D., Moore, J.D., 2014. Alignment and task success in spoken dialogue. J. Mem. Lang. 76, 29–46. doi:10.1016/j.jml.2014.05.008.
Sacks, H., Schegloff, E.A., Jefferson, G., 1974. A simplest systematics for the organization of turn-taking for conversation. Language 50 (4), 696–735. doi:10.2307/412243, (Baltim).
Schegloff, E.A., 2000. When “others” initiate repair. Appl. Linguist. 21 (2), 205–243. doi:10.1093/applin/21.2.205.
Schegloff, E.A., 2007. Sequence Organization in Interaction: A Primer in Conversation Analysis, 1. Cambridge University Press.
Schegloff, E.A., Jefferson, G., Sacks, H., 1977. The preference for self-correction in the organization of repair in conversation. Language 53 (2), 361–382. doi:10.1353/lan.1977.0041, (Baltim).
Schoffelen, J.M., Oostenveld, R., Lam, N.H.L., Uddén, J., Hultén, A., Hagoort, P., 2019. A 204-subject multimodal neuroimaging dataset to study language processing. Sci. Data 6 (1), 17. doi:10.1038/s41597-019-0020-y.
Shockley, K., Santana, M.V., Fowler, C.A., 2003. Mutual interpersonal postural constraints are involved in cooperative conversation. J. Exp. Psychol. Hum. Percept. Perform. 29 (2), 326–332. doi:10.1037/0096-1523.29.2.326.
Smith, S.M., Nichols, T.E., 2009. Threshold-free cluster enhancement: addressing problems of smoothing, threshold dependence and localisation in cluster inference. Neuroimage 44 (1), 83–98. doi:10.1016/j.neuroimage.2008.03.061.
Stolk, A., Bašnáková, J., Toni, I., 2022. Joint epistemic engineering: the neglected process of context construction in human communication. In: Ibanez, A., Saravia, S.S. (Eds.), Routledge Handbook of Neurosemiotics, eds.. Taylor & Francis.
Stolk, A., Verhagen, L., Toni, I., 2016. Conceptual alignment: how brains achieve mutual understanding. Trends Cogn. Sci. (Regul. Ed.) 20 (3), 180–191. doi:10.1016/j.tics.2015.11.007.
Taylor, J.R., Williams, N., Cusack, R., Auer, T., Shafto, M.A., Dixon, M., Tyler, L.K., Cam, C., Henson, R.N., 2017. The Cambridge Centre for Ageing and Neuroscience (Cam-CAN) data repository: structural and functional MRI, MEG, and cognitive data from a cross-sectional adult lifespan sample. Neuroimage 144, 262–269. doi:10.1016/j.neuroimage.2015.09.018.
Torreira, F., Adda-Decker, M., Ernestus, M., 2010. The Nijmegen corpus of casual French. Speech Commun. 52 (3), 201–212. doi:10.1016/j.specom.2009.10.004.
Trujillo, J.P., Simanova, I., Bekkering, H., Özyürek, A., 2018. Communicative intent modulates production and comprehension of actions and gestures: a Kinect study. Cognition 180, 38–51. doi:10.1016/j.cognition.2018.04.003.
Trujillo, J.P., Vaitonyte, J., Simanova, I., Özyürek, A., 2019. Toward the markerless and automatic analysis of kinematic features: a toolkit for gesture and movement research. Behav. Res. Methods 51 (2), 769–777. doi:10.3758/s13428-018-1086-8.
Uğurbil, K., Xu, J., Auerbach, E.J., Moeller, S., Vu, A.T., Duarte-Carvajalino, J.M., Lenglet, C., Wu, X., Schmitter, S., Van de Moortele, P.F., Strupp, J., Sapiro, G., De Martino, F., Wang, D., Harel, N., Garwood, M., Chen, L., Feinberg, D.A., Smith, S.M., Yacoub, E., 2013. Pushing spatial and temporal resolution for functional and diffusion MRI in the human connectome project. Neuroimage 80, 80–104. doi:10.1016/j.neuroimage.2013.05.012.
Van Essen, D.C., Smith, S.M., Barch, D.M., Behrens, T.E.J., Yacoub, E., Ugurbil, K., 2013. The WU-Minn human connectome project: an overview. Neuroimage 80, 62–79. doi:10.1016/j.neuroimage.2013.05.041.
Van Son, R., Wesseling, W., Sanders, E., van den Heuvel, H., 2008. The IFADV corpus: a free dialog video corpus. In: Proceedings of the LREC, pp. 501–508.
Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P., 2004. Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. 13 (4), 600–612. doi:10.1109/TIP.2003.819861.
Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., Sloetjes, H., 2006. ELAN: a professional framework for multimodality research. In: Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006), pp. 1556–1559.
Yngve, V.H., 1970. On getting a word in edgewise. In: Campbell, M.A. (Ed.), Proceedings of the Sixth Regional Meeting, Chicago Linguistics Society. Department of Linguistics, University of Chicago, pp. 567–578.
Zeman, D., Popel, M., Straka, M., Hajič, J., Nivre, J., Ginter, F., Luotolahti, J., Pyysalo, S., Petrov, S., Potthast, M., Tyers, F., Badmaeva, E., Gokirmak, M., Nedoluzhko, A., Cinková, S., Hajič jr, J., Hlaváčová, J., Kettnerová, V., Urešová, Z., Li, J., 2017. CoNLL 2017 Shared Task: multilingual Parsing from Raw Text to Universal Dependencies. In: Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies, pp. 1–19. doi:10.18653/v1/K17-3001.
Zwiers, M.P., Moia, S., Oostenveld, R., 2021. BIDScoin: a user-friendly application to convert source data to the brain imaging data structure. Front. Neuroinform. 65. doi:10.3389/fninf.2021.770608.