Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 147–155
Marseille, 11–16 May 2020

Linguistic, Kinematic and Gaze Information in Task Descriptions

The LKG-Corpus

Tim Reinboth[1,2], Stephanie Gross[1], Laura Bishop[3], Brigitte Krenn[1]
[1]Austrian Research Institute for Artificial Intelligence (OFAI), Freyung 6, 1010 Vienna, Austria;
[2]Faculty of Philosophy and Education, University of Vienna, Universitätsstraße 7, 1010 Vienna, Austria;
[3] RITMO Centre for Interdisciplinary Studies in Rhythm, Time and Motion
University of Oslo, Forskningsveien 3A, 0373 Oslo, Norway.
tim.reinboth@univie.ac.at

Abstract

Data from neuroscience and psychology suggest that sensorimotor cognition may be of central importance to language. Specifically, the linguistic structure of utterances referring to concrete actions may reflect the structure of the sensorimotor processing underlying the same action. To investigate this, we present the Linguistic, Kinematic and Gaze information in task descriptions Corpus (LKG-Corpus), comprising multimodal data on 13 humans, conducting take, put, and push actions, and describing these actions with 350 utterances. Recorded are audio, video, motion and eye-tracking data while participants perform an action and describe what they do. The dataset is annotated with orthographic transcriptions of utterances and information on: (a) gaze behaviours, (b) when a participant touched an object, (c) when an object was moved, (d) when a participant looked at the location s/he would next move the object to, (e) when the participant's gaze was stable on an area. With the exception of the annotation of stable gaze, all annotations were performed manually. With the LKG-Corpus, we present a dataset that integrates linguistic, kinematic and gaze data with an explicit focus on relations between action and language. On this basis, we outline applications of the dataset to both basic and applied research.

Keywords: situated task description, multimodal behaviour, corpus

1. Motivation

The paper presents the LKG-Corpus, comprising linguistic, kinematic and gaze information in task descriptions. Participants changed the positions of objects on a table while they were describing what they were doing. The corpus contains (a) 327 utterances, each describing an action, (b) the accordant motion data, as well as (c) the participants' gaze behaviour. This work is inspired by Knott (2012), and aims towards a better understanding of the role of sensorimotor processes in language. With the LKG-Corpus, we have also extended our previous work by higher resolution methods in both motion capture and eye gaze data collection – c.f. the Multimodal Task Description (MMTD) Corpus (Gross and Krenn, 2016), and the Action Verb Corpus (AVC) (Gross et al., 2018).

The LKG-Corpus was collected to bridge the gap between the AVC-corpus and Knott's (2012) theses about the relation between regularities in (a) the dynamics of concrete actions and their neural representation, and (b) regularities in the structure of utterances about concrete actions – see also Webb (2015). The basic idea of this approach is that an *action episode*, e.g. reaching-to-grasp, is composed of a series of *operations*. In the present paper, we adopt this terminology and refer to action 'episode' and to 'operation', to clearly distinguish two aspects of *action*: (1) the 'episode', indicating complex behaviours such as reaching to grasp a cup, and (2) 'operations', individual processes and their consequences (amongst others visually attending to the target object), organised in a temporally invariant way, that together constitute an episode.

The operations involved in an episode are considered to be *deictic*, meaning that they create representations tied to specific cognitive programs, i.e. the cognitive (inseparable from perceptual and motor) means of achieving some goal (Ballard et al., 1997). One such goal might be grasping a cup. That being her goal, for example, an agent might soon look at the cup. This visual attention to the cup is a deictic operation. Arguing that this fixation is deictic entails that it does two things: it (1) points to the object of the current action episode, and thereby (2) establishes a reference between its own target, i.e. the cup, and the goal of the current episode, i.e. to grasp that cup. Put another way, the deictic operation specifies the concrete referent of the ongoing action episode. Ballard et al. (1997) refer to a series of deictic operations as a *deictic routine*, what we call an action episode.

Deictic operations have been shown to be a computationally simple means for directing behaviour (Ballard, 1991). Moreover, deictic operations are promising in the context of action episodes, because they produce specific sensory consequences, which then define the context for the subsequent operation in the sequence. For example, the deictic operation of visual attention to an object produces, as a sensory consequence, the neural representation of that object.[1] In turn, this representation is the basis on which the agent can then move her arm to grasp the object. Moreover, the neural representation can be thought to trigger the next deictic operation, in this case activation of the 'grasp' oper-

[1] The agent may also covertly attend to the object, rather than visually, through internally directed attention (Chun, 2011). This generates a functionally equivalent representation. However, eye-tracking still remains useful in this case, since a number of eye behaviours have been associated with internally directed attention (Benedek et al., 2017; Walcher et al., 2017)

ation (Knott, 2012). It is this dynamic that establishes the canonical structure of action episodes (see Table 1).

Knott (2012) concentrates on a detailed neuro-scientific account of action episodes using the example of reaching-to-grasp. Taking Knott's conceptualization of action as an inspiration, the LKG-Corpus addresses two of the deictic operations: (a) "attend to object' and (b) "activate motor action". Their sensory consequences are made accessible via eye-tracking (focus on visual saccades, for details see Sect. 2.3.) and motion capture (focus on hand-arm movements, see Sect. 5.1.4.), respectively. In addition, the LKG-Corpus extends the action episodes to lifting up objects, putting them down, and moving them on a table from one place to another. Moreover, linguistic information is added to the picture by letting people verbalise what they are doing. That allows us to empirically investigate the relation between utterances and the sensorimotor variables recorded by eye-tracking and motion capture.

Deictic Operation	**Sensory Consequence**	**Measure in LKG**
Attend to the agent (interoceptive)	Attending to the agent (interoceptive)	*Not applicable*
Attend to the cup (visual)	Attending to the cup (visual)	Eye-tracking
Activate motor action	Reattention to the agent (motor)	Motion Capture
	Reattention to the cup (haptic)	*Not applicable*

Table 1: Knott (2012) proposes an alternating sequence of deictic operations and their sensory consequence, together a *deictic routine*. The sensory consequences necessarily follow from the preceding operation. Moreover, each sensory consequence prepares the agent for the subsequent deictic operation. Together, the processes summarised above constitute a single action episode.

To this end, the LKG-Corpus offers a new initial set of data from which to explore robotic learning of new actions, objects and the accordant words through observation and natural language descriptions. We proceed to explore Knott's suggestion that the structure of utterances concerning concrete action episodes reflects the structure of the sensorimotor representations of the same action episode. In doing so, we continue previous work on action-word mapping – e.g. Krenn et al. (2017), Sadeghi and Scheutz (2017), Hirschmanner et al. (2018) – according to a theory-driven and biologically-inspired framework. Also, we extend previous research regarding this framework by recording data from the perspective of the agent, rather than an observer – cf. Webb (2015).

In Sect. 2. the background we build upon is outlined, and Sect. 3. summarises the related work. Task scenario and technical setup are presented in Sect. 4.. Data and annotation tiers for the LKG-Corpus are described in Sect. 5.. Finally, Sect. 6. offers a closer look at possible applications of the annotated corpus, and an outlook on future work.

2. Background

The framework we draw on is built on research in both neuroscience, e.g. Arbib (1981), and artificial intelligence, e.g. Agre and Chapman (1987). Specifically, the neuroscientific line of research investigates the dynamics of routine activities and searches for regularities in the interaction between a system and its environment. Complementary to this, Agre and Chapman (1987) set out a critique of traditional approaches to planning in AI, in which they developed a theory of activity that likewise emphasises the role of regularities in the interaction between agents and their environment. The present work builds on the idea shared by both sets of research, namely that many routine behaviours are, in fact, constituted by a canonical sequence of actions. Consider again the example of reaching to grasp a cup. Almost at the outset, visual fixations create transient sensory states that enable the execution of a subsequent action (Ballard et al., 1997). According to this account, we fixate (very briefly) on an object to create a stable image on the retina, then classify that object. Fixations are a correlate of visual attention, and the object representation is usually thought to be in the visual domain. However, there is also evidence of motor attention, for example during action preparation (Rushworth et al., 1997). This process also generates a stable neural representation formed through parietal sensorimotor circuits.

In the following, we give some background on the important concepts we were guided by when collecting and analysing the LKG-Corpus.

2.1. Attention

In the context of the presented work, (selective) *visual attention* is to be understood as the ability to quickly direct one's gaze to important objects in our visual surroundings (Itti and Koch, 2001). In the LKG-Corpus, we use eye-tracking to access visual attention. *Motor attention* refers to the same process, however, in the proprioceptive modality. We are able to cover aspects of this modality using motion capture, especially of the hands and arms.

In addition, Knott (2012) also refers to *haptic attention* as being an important feedback indicating that the agent is grasping the object at the end of a reach-to-grasp episode. Haptic attention in this interpretation is exclusively covert (or internal), meaning that the object of attention is not present through the senses, but by being maintained in the agent's working memory (or recalled from long term memory) (Chun et al., 2011). In the absence of neuroimaging data, this modality cannot be represented in the LKG-Corpus. Covert attention processes in the visual and motor domain (e.g. action-preparation (Rushworth et al., 2003)) also remain difficult to address.

2.2. Episode Structure

The representation of the episode, in both the visual and motor domain, may play a central role in determining the linguistic structure of sentences that describe concrete actions (Knott, 2012) – e.g. *the woman grasped the cup*. Knott's proposal is based on a thorough review of psychophysical and neuroscientific literature, and has partially been computationally implemented by Takac et al. (2012).

In theory, and in the computational model, the canonical sequence of actions that constitutes a single episode is crucial. That is because the sequence provides structure. This structure can be exploited as the foundation of other cognitive processes, including linguistic processes.

Specifically, Knott (2012), and subsequently Webb (2015), have investigated a direct structural relationship between sensorimotor processes, the cognitive representation of episodes (e.g. reaching to grasp a cup), and the sentences describing those episodes. The authors argue that part of the structure we observe in language is derived from the canonical structure of sensorimotor processes and representations that underlie action performance and action observation. Such an approach is extremely promising, as it accounts for both the underlying similarities and the surface variability across natural languages. Moreover, the computational implementation reinforces the promise to be applicable across natural languages. In fact, from the invariant sensorimotor representation, it is possible to generate utterances, the components of which can take various positions within the structure of the clause (Takac et al., 2012).

2.3. Processes During Concrete Action Episodes

Thus far, our intention was to capture major individual processes that together make up this structure of action episodes. Processes we are able to address empirically are those related to visual fixations, periods during which the eyes appear immobile while visual information is being collected and processed, also known as fixational eye movements (Rucci et al., 2016). Specifically, visual fixations can indicate visual attention (Weichselgartner and Sperling, 1987). This relationship is particularly strong for voluntary saccadic eye movements (Hoffman and Subramaniam, 1995), i.e. rapid eye movements, initiated voluntarily, that suddenly and quickly change the point of fixation. Hence, these kinds of eye movements also play a central role in our work and in the framework proposed by Knott (2012).

Attention is key because it generates the transitory representations that are operated on during subsequent processes. The generation of such representations via attention appears to be an essential regularity of action preparation (Averbeck and Lee, 2007). That these representations are created precisely by attention (in whichever modality) leads us to refer to the actions that create them as attentional actions (Ballard et al., 1997), or *actions of attention* (Knott, 2012). The term is synonymous with deictic operation. We refer to them here to emphasise that by directing her attention, the agent is doing something. She is creating the sensory state that is necessary for the next operation to be performed. In that way, actions of attention are constitutive of the deictic routine.

This returns us to the issue at hand: that episodes are constituted by a canonical sequence of actions. For reaching to grasp, Knott (2012) outlines a sequence of such actions of attention. We investigated those in the visual domain by eye-tracking. By motion capture, we deepen the exploration of the relation between kinematic and linguistic processes, and of their temporal dynamics – cf. (Rizzolatti and Arbib, 1998). All in all, the extent to which sensorimotor processes provide structure to language still remains an open research question. For example, so-called mirror neurons are referred to in relation to a sensorimotor basis of language and linguistic structure, but the mirror neuron theory remains debatable (Gallese et al., 2011).

3. Related Work

Multimodal corpora that report various forms of linguistic information and visual data are for example (Gross et al., 2018; Gaspers et al., 2014; Panzner, 2016). Some multimodal corpora also include manually annotated or automatically tracked eye gaze data, e.g. (Gross and Krenn, 2016; Stefanov and Beskow, 2016; Ochs et al., 2019). Indeed, studying gaze behaviour is a promising method to gain a better understanding of the role of sensorimotor processes in language (Webb et al., 2010). In most corpora that record eye gaze, actions are conducted on a screen, e.g. (Stefanov and Beskow, 2016; Panzner, 2016; Ochs et al., 2019), rather than in a less constrained setting, as is the case for the MMTD-Corpus (Gross and Krenn, 2016), and even more so the LKG-Corpus. Though, the use of Virtual Reality technologies means this is beginning to change, as corpora are recorded in settings that are more enabling for participants (Ochs et al., 2019). At the same time, data collection is limited by the need to collect high quality material suitable for automatic processing, so the design of data collection experiments reflects a trade-off between less constrained settings on the one hand, and concern about data quality on the other (Vinciarelli et al., 2015). When collecting empirical data to gain insights in human sensorimotor and linguistic processes, it is also a challenge to find the adequate balance to collect data as 'natural' as possible, but constrained enough so that they are still comparable to data collected from other participants. Because of the importance of recording participants behaviour in as high detail as possible for the LKG-Corpus, we opted for a setting less constrained than that of the AVC-Corpus.

Compared to corpora manually annotated for gestures, actions and eye gaze, such as the MMTD-Corpus (Gross and Krenn, 2016), we employed a motion tracking system and an eye-tracker for collecting the LKG-Corpus. These more precise data allow us to conduct in-depth investigations of sensorimotor processes.

Concerning cross-modal and cross-situational word-object and word-action learning, the presented corpus builds upon the AVC-Corpus, cf. (Gross et al., 2018). In both corpora, the positions of three objects are changed on a table and verbally described. There is no interlocutor present. The main differences lie in the recorded data, determined by the overall goal of the two corpora. In addition to speech/audio, hand tracking, object tracking, head pose, and table tracking in the AVC-Corpus, the LKG-Corpus also contains eye-tracking data. Which were not relevant for the AVC-Corpus, as the goal was to collect data that resemble what a robot hears and sees during task descriptions, in order to learn new actions and objects through observation and their natural language descriptions. Goal of the LKG-Corpus additionally is to collect sensorimotor data of actions and their natural language descriptions to extract sensorimotor patterns and compare them to linguistic patterns. Concerning the technical set-up, the use of a head-mounted display in the AVC constrained participants field of view, possibly causing them to move their head more than they would have otherwise. Also, participants received detailed instructions for each action during recordings for the AVC-Corpus, whereas participants in data collection for the LKG-Corpus were able to choose an action to perform. Notably, the LKG-Corpus is relatively small, especially compared to corpora collected for deep machine learning, see for instance (Johnson et al., 2017; Kay et al., 2017). This reflects the diverging goal of such projects and our own. With the LKG-Corpus, we focus on representing the rich multimodal dynamics of concrete actions. We are not only interested in computational applications, but also aim at a better understanding of the nature of language. This reflects the influence of studies such as Suanda et al. (2016) on our own work, who investigate the multimodal dynamics of parent-toddler dynamics. These authors, too, explored the structuring role of sensorimotor processes for language. In general, the LKG-Corpus is also intended to serve as a basis for (robotic) simulations of human-like visual behaviour. Moreover, the understanding of language developed by analysing this kind of data could offer key insights that may ultimately be applied to a developmentally-inspired approach to robot language learning.

4. Setup and Data Collection

4.1. Participants

A convenience sample of 13 participants (Mean Age = 24.08 years, Range 21-28; 8 female, 5 male) was recruited from a population of university students. All participants but one were right-handed. All participants were naive to the purpose of the study. The native languages of the final sample included English (N=1) and German (N=14). All participants were proficient in English (B2 level or above). This study was conducted in English to facilitate the dissemination of the corpus, given limited interest in German-language corpora. Moreover, English was chosen for the sake of consistency with related empirical work, e.g. Webb (2015). All participants gave informed consent to their participation in this study.

4.2. Task Description

The overall goal of the data collection was to gather multimodal data on basic take, put, and push actions. The general set-up is similar to that reported by Gross et al. (2018). In this case too, the participants received instructions to perform actions with three objects, placed on a table in front of them, and to verbalize their actions while carrying them out. In the instructions for the LKG-Corpus, participants were told: "*Your task in this experiment is to change the positions of objects and to describe the actions as you perform them. You are asked to do this as naturally as possible and to pay attention only to the task. Please look through the glasses normally and try not to get distracted. During the task, you will move objects from place to place within the marked area and describe what you are doing.*"

In the present study, participants did not receive specific instructions about which action to perform when, or which aspects of an action to focus on. Possible verbalizations were of the kind: *I take the bottle and move it to the top left of the table.* Participants were instructed to return to a neutral position between actions and to continue performing actions until they were told to stop. They were informed that they would perform around 10 actions during each trial, before being stopped by an experimenter.

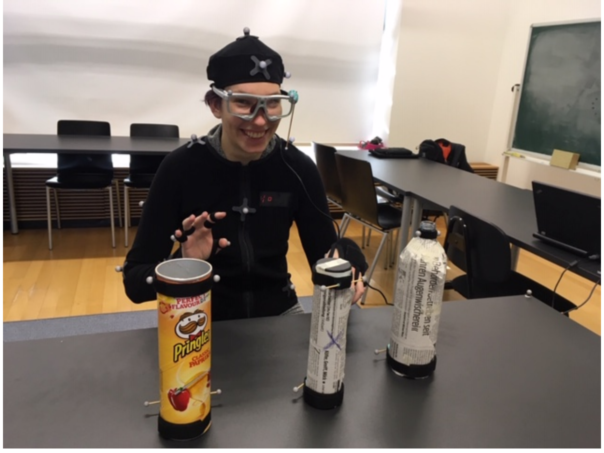

Figure 1: Participants performed the task seated at the table, with the three objects placed on a table in front of them.

The group of participants included left- and right-handers, and both were instructed to perform actions only with their dominant hand. This reflects a concern that hand dominance may result from neural asymmetries in motor planning (Sabaté et al., 2004). Motor planning plays a central role in Knott (2012). To avoid disruptive effects, all actions performed with the non-dominant hand (N = 2) were excluded from the present analysis.

4.3. Technical Setup

The data collected for the present study include audio data (recordings of participants describing their actions), video data (recordings of participants' field of view), as well as eye-tracking and motion capture data.

The user sits in front of a table (see Fig. 2), wearing Sensorimotor Instruments (SMI) ETG 2 wireless glasses, which tracked gaze position at 120 Hz. Magnetic snap-on corrective lenses were available for participants requiring a distance correction (within the range -4.0 to +4.0 dioptres). The glasses featured a custom arrangement of motion capture markers for detection by the motion capture system. One marker was placed at each corner of the top of the frame. A third marker, attached to a small stick, extended down from the left arm of the glasses (see Fig. 3). Eye-gaze data was collected with a laptop connected to an OptiTrack eSync 2 device that transmitted TTL triggers from the motion capture software to the eye-tracking software to mark the start and end of a motion capture recording.

For the motion capture, a 10-camera (Prime 13) OptiTrak motion capture system was used, tracking participants' upper body and head movements at 240 Hz. Each performer was fitted with a custom arrangement of 28 reflective markers. Additional markers (4) were attached to each of three objects the actions were performed with, as well as to the corners of the table. Audio data were recorded using two microphones mounted on the ceiling, via a Focusrite Scarlett 18i8 sound card and recorded in Ableton Live.

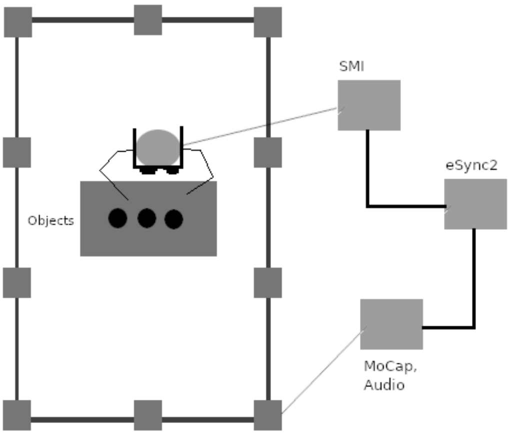

Figure 2: Participants sat at a table in the middle of an array of 10 ceiling-mounted cameras of the OptiTrak system.

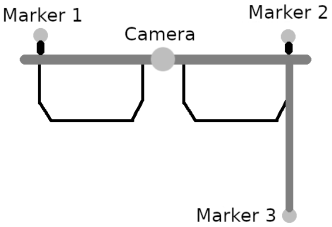

Figure 3: Eye-tracking glasses were fitted with three markers for tracking by the OptiTrak system, and featured a central, front-facing camera.

5. Dataset and Annotation

The present corpus comprises

Recordings	22
Utterances referring to actions	327
Instances of objects being moved	209
Gaze Behaviours	3950

Table 2: Summary of the current state of the LKG-Corpus. Each instance of an object being moved contains at least two actions (first TAKE, second PUT/PUSH), though not necessarily two utterances referring to actions.

5.1. Representation of Information

In addition to raw data from audio recordings, eye-tracking, motion capture, and video recordings from the user's perspective, each instance of a task is represented by:

1. manual orthographic transcriptions of speech (conducted in Praat [2])
2. a video from the participants' perspective, automatically overlaid with a circle indicating the target of the participants' gaze by SMI software
3. schematics of the gaze vector every 5 frames (see Fig. 4), also compiled into an animation of gaze behaviour
4. information on gaze behaviours, for example whether a participant is looking at/close to an object (manually annotated)
5. information on whether the user's dominant hand touches an object (manually annotated)
6. information on when an object is being moved (manually annotated)
7. information on velocity and acceleration of participants' gaze [3]
8. information on velocity and acceleration of a subset of motion capture markers [4]

While the raw data are provided as CSV files, the data from 1. and 4. to 8. (henceforth annotation tiers) is represented as an Elan[5] file (.eaf) and exportable to CSV from Elan. Animations are synchronised with the annotation tiers.

In the following, we describe in more detail (a) the transcribed utterances (annotation tier 1), (b) the calculated gaze vectors (tier 2, 3, 8), (c) the annotated gaze behaviours (tier 4), and (d) the motion capture data (tier 5 to 7).

5.1.1. Utterances

In total, the corpus comprises 350 utterances, of which 327 contain action verbs, such as 'I am grasping the can' and 'I am moving in to the middle of the table'. In these 327 utterances, 15 different verbs are used by the 13 participants: 'move', 'grasp', 'lift', 'put', 'pull','push', 'take', 'hold', 'try', 'grab', 'stand', 'shift', 'pick', 'place', 'reach'. In 23 cases, the verb is missing, as in 'then the can next to the bottle'. In addition to utterances describing actions, there are four scene and sensor descriptions. Two utterances were excluded, as they are only partially audible.

5.1.2. Gaze Vectors

A vector indicating the direction of the gaze for the right eye of the participant was calculated. This gaze vector is based on a precise calibration of the glasses when the participant first puts them on. Given the length of a recording session (15-45 minutes), however, a certain degree of imprecision in the gaze vector coordinates cannot be ruled out because of calibration errors, i.e. when participants bumped or shifted the glasses.

An automated procedure to identify gaze targets was adapted from (Bishop et al., 2019). It involved remapping gaze coordinates into motion capture space. From this,

[2] http://www.fon.hum.uva.nl/praat/

[3] This provides a continuous rather than discretised annotation of gaze. This has the advantage, for example, that it leaves open the parameters of what counts as a visual fixation (see e.g. Rayner (2009) for parameters approximating a typical fixational eye-movement)

[4] This subset includes the markers of the head, and dominant hand and arm.

[5] https://tla.mpi.nl/tools/tla-tools/elan

we generated a schematic representation of each frame (see Fig. 4) and an animation of each trial.

In every case, the direction of the gaze vector was automatically corrected following the data collection, to compensate for imprecision in the placement of markers on the glasses and in the estimation of the gaze vector origin. Automatically adjusted eye gaze vectors were then tested by visual comparison with the video and gaze marker display. The automatically generated schematics and animations are accurate to the recorded data and therefore remain a valuable additional representation.

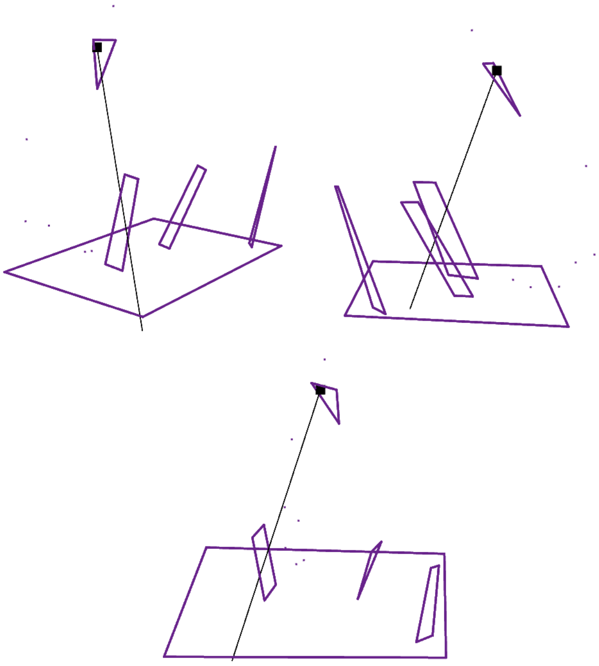

Figure 4: Schematics of the scene at one point in time from three angles, generated by the automated analysis of gaze and motion capture data. The glasses are visible as a triangle, the table as flat plain, and the three objects as upright planes. The black line represents the gaze vector calculated from the eye-tracking data. The individual points represent the markers on the right arm of the participant.

Moreover, automatically extracted is information on when the participants' gaze was stable on one point in space for at least 50ms (annotation tier 8).

5.1.3. Gaze Behaviours

Gaze behaviours (annotation tier 4) were annotated manually because of persistent irregularities that originated from the marker placement on the objects and their resulting representation in motion capture space. (The 3D shapes that represented the objects were not symmetrical across orientations; see Fig. 4.) For manual annotation, we identified distinct patterns of eye movements in the video from the front-facing camera mounted on the glasses. Our label set (see Table 3) resulted from consideration of existing data on eye movements during tasks similar to our own, e.g. Ballard et al. (1997), Webb (2015). The label set was developed iteratively. Three authors of this paper (TR, SG, BK) developed categories as a team, annotated the data independently, tested for interrater agreement, and then revised and retested the label set. A process characteristic for the development of a coding scheme (Weber, 1990).

In the sense that we updated categories according to patterns in the data, our method is partly inductive. Such an approach is useful when, as is the case here, research informs the development of new theories (Rivas, 2018). For example, the challenge of identifying the appropriate level of granularity at which to annotate, had to be addressed iteratively during the development of the label set. Ours is also a common and effective pattern of developing novel coding schemes (Weston et al., 2001). Manual annotation, in this more qualitative manner, allowed us to distinguish between eye-movements in a way that we were better able to represent the rich dynamics of visual behaviour. This produced a detailed annotation tier that could not have been extracted from the raw data automatically.

Primary Label	Reference Object	Secondary Label	Addition	Supplement
on	Object^{1}	stable	trans-Object$^{1\text{-n}}$	
	Object^{1}-Object^{2}	moving	close-Object$^{1\text{-n}}$	
		jumping	alt-Object$^{1\text{-n}}$	
to	Object^{1}-Object^{2}	smooth	trans-Object$^{1\text{-n}}$	
	Object^{1}* *Object^{2}	jumping		
with	Object^{1}	on	smooth	trans-Object$^{1\text{-n}}$
	Object^{1}-Object^{2}	close	jumping	close-Object$^{1\text{-n}}$
		trans		alt-Object$^{1\text{-n}}$

Table 3: Coding scheme for gaze behaviours – Primary labels distinguish three different types of eye movements: gaze on object, from/to a certain location, and with an object/hand being moved. There can be up to two reference objects per annotation. Further labels describe features of the gaze behaviour including whether it is (a) smooth, (b) jumping, (c) stable, (d) crossing an object without change in speed (trans), or (e) alternating between two locations/objects. Aspects of the scene include whether another object is close to the current gaze location.

Given the complexity of the data, the accuracy of labelling was assessed using forced-choice. This meant that all trials in the corpus were annotated by two coders, who had to agree on each labelling instance. Using consensus to ensure interrater agreement in this way is particularly useful when different labels represent qualitatively different constructs (Stemler, 2004). By this we mean that the primary labels *on*, *to*, and *with* reported here, represent different types of eye movements, rather than points on a continuum. Respectively, the three primary labels map onto (a) smooth eye movements, (b) voluntary saccades, and (c) pursuit eye movements (Pruves et al., 2004). Forced choice is a useful method to arbitrate such cases and avoid discrepancies in annotation. This challenge reflects the semantic richness of the gaze behaviour, which made forced-choice consensus an apt method to address interrater agreement.

5.1.4. Motion Capture Data

Motion capture data are made available within the corpus in the form of .csv files of three-dimensional coordinates for each marker. These data offer a number of possibilities. For example, we will align motion capture data with video and eye-tracking to identify where participants' hands were when they started looking at the object they wanted to move. Such an approach also greatly simplifies the annotation of eye-tracking data (Bishop and Goebl, 2018).

In general, motion capture data are an excellent resource as they reflect the dynamics of movement. For instance, velocity, acceleration, and jerk can all be easily calculated for the gesture trajectories. From this, it is possible to investigate the smoothness and speed of movements and whether there are consistent correlations between the utterance and changes in these aspects of participants' movements. In doing so, we move beyond Knott (2012) to investigate the relation between motor action and speech production.

In this regard, our data allow us to investigate dynamics suggested by Glenberg and Gallese (2012), who propose control sharing between language and motor processes. Arguing that 'we say what we do, and do what we say' they suggest a two-fold relation, whereby 1) the act of articulation prepares the associated motor actions, and 2) performing the action primes the articulation. In this framework, sensorimotor and linguistic processes are inseparably coupled, making it essential to explore and understand the relation between the two. In this context, the LKG-Corpus provides a valuable resource for further investigation.

6. Applications and Future Work

The LKG-Corpus is a novel data set that integrates linguistic, kinematic and gaze information with an explicit focus on the basic relations between language and action. As such, it is a step towards better understanding the cognitive processes that subserve language. That also means that it offers two distinct avenues of future research. On the one hand, we expect the data set to contribute to a better understanding of fundamental questions concerning the dynamics of multimodal (inter-)actions which so far have remain unanswered (Vinciarelli et al., 2015). On the other hand, the data support computational modelling of human-robot interaction, notably supporting mutual understanding between human and robot actors.

6.1. Basic Research

The present data allow us to investigate when the participant directs her visual attention to the target object or is looks for a target location. Also, motor actions can be extracted in different granularity. Thus, the data serve as a good basis to identify temporal sequences in (a) visual attention, (b) motor action, and (c) language, and to compare them. For investigating deictic operations going beyond visual attention, further data is required. One approach is to complement the current set-up with neuroimaging data. Electroencephalography (EEG), for example, has been used to decode actions (Schwarz et al., 2017) and as the source for common neural (electrophysiological) features from which to estimate behaviour (Touryan et al., 2016). Simultaneous EEG measures would allow us to explore relations outlined by Knott (2012).

A different path to further develop this work would be to adjust the current design to facilitate automatic annotation of gaze behaviour. Among other improvements for the future, reflective markers will be placed on the objects such that they would be represented as 3D-objects in motion capture space (rather than as surfaces, see 4). To facilitate this, the use of square objects would be more suitable than the use of cylinders. We will avoid other modifications, such as restricting where participants may take hold of an object. Such interventions could improve automatic labelling by the OptiTrack software, but will likely distort participants' actions, and will make the data more artificial.

The fact that the LKG-Corpus comprises motion data at different levels of detail allows us to investigate the relation between manual behaviour and utterances on two levels of granularity at once. On the one hand, manually annotated hand behaviour (e.g. 'hand moving away from/towards object') represents a semantically high-level interpretation of the motion capture data. On the other hand, features of hand movements, such as velocity and acceleration can be automatically extracted from the raw data. Taken together, we can use this information to investigate both the finer dynamics suggested by Knott (2012)'s canonical structure of concrete action episodes, as well as the overlap between speech articulation and motor control suggested by Glenberg and Gallese (2012).

6.2. Applied Research

A pressing question for HRI is how human intentions can be extracted from data. This is essential for robots to successfully act in social situations (Broz et al., 2013), as well as for tackling safety concerns in workplaces shared by humans and robots (Bascetta et al., 2011). To address the problem of what a human is going to do, human speech and/or motor action must be made interpretable to robots (Sciutti et al., 2018). Eye-tracking and motion capture are valuable sources of data, for instance, to exploit the correlation between direction of human gaze and the focus of attention (Frischen et al., 2007). Moreover, the velocity with which an action is performed has been tied to intentions directly (Karg et al., 2013), as well as to emotional status (Sartori et al., 2011). In that respect, the LKG-Corpus is a promising resource for these lines of research in HRI.

Both gaze data (Palinko et al., 2016) and motion data (Vignolo et al., 2017) are already being applied as resources to study and design human-robot interaction. For example, recent research indicated that eye motion and fixations facilitate disambiguation of target objects, while action kinematics have also been directly related to human intention (Sciutti et al., 2015). In fact, in human-robot collaborative settings, motion trajectories allow the prediction of subsequent human action and the anticipatory planning of robot responses (Koppula and Saxena, 2015). To address this, because the LKG-Corpus does not report interactions, we plan to apply our simultaneous motion capture and eye-tracking design to tasks similar to those reported in the MMTD-Corpus (Gross and Krenn, 2016).

7. Acknowledgements

This research is supported by the Vienna Science and Technology Fund (WWTF), project RALLI – Robotic Action Language Learning through Interaction (ICT15-045), the Austrian Research Promotion Agency (FFG), Ideen Lab 4.0 project CoBot Studio (872590), and the Austrian Science Fund (FWF), Project P29427-G24. Data collection was conducted at the Performance Science Lab of the Department of Music Acoustics – Wiener Klangstil, University of Music and Performing Arts Vienna (mdw).

8. Bibliographical References

Agre, P. E. and Chapman, D. (1987). Pengi: An implementation of a theory of activity. In *Proceedings of the Sixth National Conference on Artificial Intelligence (AAAI)*, volume 87, pages 286–272. AAAI Press, Menlo Park, CA, Seattle, WA.

Arbib, M. (1981). Perceptual structures and distributed motor control. In V. Brooks, editor, *Handbook of Physiology: Section II The Nervous System. Part 1: Motor Control*. American Physiological Society.

Averbeck, B. B. and Lee, D. (2007). Prefrontal neural correlates of memory for sequences. *Journal of Neuroscience*, 27(9):2204–2211.

Ballard, D. H., Hayhoe, M. M., Pook, P. K., and Rao, R. P. (1997). Deictic codes for the embodiment of cognition. *Behavioral and Brain Sciences*, 20(4):723–742.

Ballard, D. H. (1991). Animate vision. *Artificial Intelligence*, 48:57–86.

Bascetta, L., Ferretti, G., Rocco, P., Ardö, H., Bruyninckx, H., Demeester, E., and Di Lello, E. (2011). Towards safe human-robot interaction in robotic cells: an approach based on visual tracking and intention estimation. In *2011 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 2971–2978. IEEE.

Benedek, M., Stoiser, R., Walcher, S., and Körner, C. (2017). Eye behavior associated with internally versus externally directed cognition. *Frontiers in Psychology*, 8:1092.

Bishop, L. and Goebl, W. (2018). Performers and an active audience: Movement in music production and perception. *Jahrbuch Musikpsychologie*, 28, 08.

Bishop, L., Cancino-Chacón, C., and Goebl, W. (2019). Eye gaze as a means of giving and seeking information during musical interaction. *Consciousness and Cognition*, 68:73–96.

Broz, F., Nourbakhsh, I., and Simmons, R. (2013). Planning for human–robot interaction in socially situated tasks. *International Journal of Social Robotics*, 5(2):193–214.

Chun, M. M., Golomb, J. D., and Turk-Browne, N. B. (2011). A taxonomy of external and internal attention. *Annual Review of Psychology*, 62:73–101.

Chun, M. M. (2011). Visual working memory as visual attention sustained internally over time. *Neuropsychologia*, 49(6):1407 – 1409.

Frischen, A., Bayliss, A. P., and Tipper, S. P. (2007). Gaze cueing of attention: visual attention, social cognition, and individual differences. *Psychological Bulletin*, 133(4):694.

Gallese, V., Gernsbacher, M. A., Heyes, C., Hickok, G., and Iacoboni, M. (2011). Mirror neuron forum. *Perspectives on Psychological Science*, 6(4):369–407.

Gaspers, J., Panzner, M., Lemme, A., Cimiano, P., Rohlfing, K. J., and Wrede, S. (2014). A multimodal corpus for the evaluation of computational models for (grounded) language acquisition. In *Proceedings of the 5th Workshop on Cognitive Aspects of Computational Language Learning (CogACLL)*, pages 30–37.

Glenberg, A. M. and Gallese, V. (2012). Action-based language: A theory of language acquisition, comprehension, and production. *cortex*, 48(7):905–922.

Gross, S. and Krenn, B. (2016). The OFAI multi-modal task description corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 1408–1414, Portorož, Slovenia, May.

Gross, S., Hirschmanner, M., Krenn, B., Neubarth, F., and Zillich, M. (2018). Action verb corpus. In *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018)*, pages 2147–2151, Miyazaki (Japan), May.

Hirschmanner, M., Gross, S., Krenn, B., Neubarth, F., Trapp, M., and Vincze, M. (2018). Grounded word learning on a pepper robot. In *Proceedings of the 18th International Conference on Intelligent Virtual Agents*, pages 351 – 352. ACM.

Hoffman, J. E. and Subramaniam, B. (1995). The role of visual attention in saccadic eye movements. *Perception & Psychophysics*, 57(6):787–795, Jan.

Itti, L. and Koch, C. (2001). Computational modelling of visual attention. *Nature Reviews Neuroscience*, 2(3):194–203.

Johnson, J., Hariharan, B., van der Maaten, L., Fei-Fei, L., Lawrence Zitnick, C., and Girshick, R. (2017). Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2901–2910.

Karg, M., Samadani, A.-A., Gorbet, R., Kühnlenz, K., Hoey, J., and Kulić, D. (2013). Body movements for affective expression: A survey of automatic recognition and generation. *IEEE Transactions on Affective Computing*, 4(4):341–359.

Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., et al. (2017). The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*.

Knott, A. (2012). *Sensorimotor Cognition and Natural Language Syntax*. MIT Press, Cambridge,MA.

Koppula, H. S. and Saxena, A. (2015). Anticipating human activities using object affordances for reactive robotic response. *IEEE Transactions On Pattern Analysis And Machine Intelligence*, 38(1):14–29.

Krenn, B., Trapp, M., Gross, S., and Neubarth, F. (2017).

Crossmodal cross-situational learning with attention. In *Seventh Joint IEEE International Conference on Development and Learning and on Epigenetic Robotics: Workshop on Computational Models for Crossmodal Learning*, Lisbon, Portugal, Sep. IEEE.

Ochs, M., Blache, P., Montcheuil, G., Pergandi, J.-M., Bertrand, R., Saubesty, J., Francon, D., and Mestre, D. (2019). The acorformed coprus: Investigating multimodality in human-human and human-virtual patient interactions. In *Selected papers from the CLARIN Annual Conference 2018, Pisa, 8-10 October 2018*, 159, pages 108–115. Linköping University Electronic Press.

Palinko, O., Rea, F., Sandini, G., and Sciutti, A. (2016). Robot reading human gaze: Why eye tracking is better than head tracking for human-robot collaboration. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5048–5054. IEEE.

Panzner, M. (2016). TLS dataset. Bielefeld University.

D. Pruves, et al., editors. (2004). *Neuroscience*. Sinauer Associates Inc, Sunderland, MA, 3 edition.

Rayner, K. (2009). Eye movements and attention in reading, scene perception, and visual search. *The Quarterly Journal of Experimental Psychology*, 62(8):1457–1506.

Rivas, C. (2018). Finding themes in qualitative data. In C Seale, editor, *Researching Society and Culture*, pages 431–453. Sage Publications Sage CA: Los Angeles, CA.

Rizzolatti, G. and Arbib, M. A. (1998). Language within our grasp. *Trends in Neurosciences*, 21(5):188–194.

Rucci, M., McGraw, P. V., and Krauzlis, R. J. (2016). Fixational eye movements and perception. *Vision Research*, 118:1 – 4.

Rushworth, M. F., Nixon, P. D., Renowden, S., Wade, D. T., and Passingham, R. E. (1997). The left parietal cortex and motor attention. *Neuropsychologia*, 35(9):1261 – 1273.

Rushworth, M., Johansen-Berg, H., Göbel, S. M., and Devlin, J. (2003). The left parietal and premotor cortices: motor attention and selection. *Neuroimage*, 20:89–100.

Sabaté, M., González, B., and Rodríguez, M. (2004). Brain lateralization of motor imagery: Motor planning asymmetry as a cause of movement lateralization. *Neuropsychologia*, 42(8):1041–1049.

Sadeghi, S. and Scheutz, M. (2017). Joint acquisition of word order and word referent in a memory-limited and incremental learner. In *2017 8th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)*, pages 000357–000362. IEEE.

Sartori, L., Becchio, C., and Castiello, U. (2011). Cues to intention: the role of movement information. *Cognition*, 119(2):242–252.

Schwarz, A., Ofner, P., Pereira, J., Sburlea, A. I., and Mueller-Putz, G. R. (2017). Decoding natural reach-and-grasp actions from human eeg. *Journal of neural engineering*, 15(1):016005.

Sciutti, A., Ansuini, C., Becchio, C., and Sandini, G. (2015). Investigating the ability to read others' intentions using humanoid robots. *Frontiers in Psychology*, 6:1362.

Sciutti, A., Mara, M., Tagliasco, V., and Sandini, G. (2018). Humanizing human-robot interaction: On the importance of mutual understanding. *IEEE Technology and Society Magazine*, 37(1):22–29.

Stefanov, K. and Beskow, J. (2016). A multi-party multimodal dataset for focus of visual attention in human-human and human-robot interaction. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 4440–4444.

Stemler, S. E. (2004). A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. *Practical Assessment, Research & Evaluation*, 9(4):1–19.

Suanda, S. H., Smith, L. B., and Yu, C. (2016). The multisensory nature of verbal discourse in parent–toddler interactions. *Developmental Neuropsychology*, 41(5-8):324–341.

Takac, M., Benuskova, L., and Knott, A. (2012). Mapping sensorimotor sequences to word sequences: A connectionist model of language acquisition and sentence generation. *Cognition*, 125(2):288–308.

Touryan, J., Lance, B. J., Kerick, S. E., Ries, A. J., and McDowell, K. (2016). Common eeg features for behavioral estimation in disparate, real-world tasks. *Biological Psychology*, 114:93–107.

Vignolo, A., Noceti, N., Rea, F., Sciutti, A., Odone, F., and Sandini, G. (2017). Detecting biological motion for human–robot interaction: A link between perception and action. *Frontiers in Robotics and AI*, 4:14.

Vinciarelli, A., Esposito, A., André, E., Bonin, F., Chetouani, M., Cohn, J. F., Cristani, M., Fuhrmann, F., Gilmartin, E., Hammal, Z., et al. (2015). Open challenges in modelling, analysis and synthesis of human behaviour in human–human and human–machine interactions. *Cognitive Computation*, 7(4):397–413.

Walcher, S., Körner, C., and Benedek, M. (2017). Looking for ideas: Eye behavior during goal-directed internally focused cognition. *Consciousness and Cognition*, 53:165 – 175.

Webb, A., Knott, A., and MacAskill, M. R. (2010). Eye movements during transitive action observation have sequential structure. *Acta Psychologica*, 133(1):51–56.

Webb, A. (2015). *Actions of Attention, and Attention to Action: Investigating the Relationship Between Visual Attention, Episodic Representation, and Language*. Phd thesis, University of Otago, Dunedin, NZ.

Weber, R. P. (1990). *Basic Content Analysis*. Sage Publications, Los Angeles, CA.

Weichselgartner, E. and Sperling, G. (1987). Dynamics of automatic and controlled visual attention. *Science*, 238(4828):778–780.

Weston, C., Gandell, T., Beauchamp, J., McAlpine, L., Wiseman, C., and Beauchamp, C. (2001). Analyzing interview data: The development and evolution of a coding system. *Qualitative Sociology*, 24(3):381–400.