Comparative Analysis of Think-aloud Methods for Everyday Activities in the Context of Cognitive Robotics

Moritz Meier, Celeste Mason, Felix Putze, Tanja Schultz

Cognitive Systems Lab, University of Bremen, Germany

moritz.meier@uni-bremen.de, celeste.mason@uni-bremen.de

Abstract

We describe our efforts to compare data collection methods using two think-aloud protocols in preparation to be used as a basis for automatic structuring and labeling of a large database of high-dimensional human activities data into a valuable resource for research in cognitive robotics. The envisioned dataset, currently in development, will contain synchronously recorded multimodal data, including audio, video, and biosignals (eye-tracking, motion-tracking, muscle and brain activity) from about 100 participants performing everyday activities while describing their task through use of think-aloud protocols. This paper provides details of our pilot recordings in the well-established and scalable "table setting scenario," describes the concurrent and retrospective think-aloud protocols used, the methods used to analyze them, and compares their potential impact on the data collected as well as the automatic data segmentation and structuring process.

Index Terms: think-aloud, activities of daily living, multimodal, biosignals, cognitive robotics

1. Introduction

The goal of the collaborative research center EASE ("Everyday Science and Engineering," http://ease-crc.org) is to facilitate robotic mastery of everyday activities. Within the EASE Biosignals Acquisition Space and Environment (BASE), we record a large dataset of human everyday activities to study how humans plan, structure, and execute such activities, like setting a table. With this purpose of robotic agent enhancement in mind, development of narrative-enabled episodic memories (NEEMs) [1], i.e. data structures derived from recorded observations, experiences, and activities is the primary goal. We record a large variety of biosignals, such as speech, muscle and brain activity measurements. Figure 1 shows a snapshot of the resulting partitur file created in the ELAN annotation tool [2], containing raw biosignal data as well as automatic and manual annotations. One of the most important modalities in the partitur is speech, produced during the activity in the form of a *think-aloud*, in which participants are asked to continuously say what they are looking at, thinking, doing, and feeling at each moment. The goal of this paper is to compare different think-aloud protocols, namely concurrent and retrospective think-aloud methods. As time for commenting is limited, we assume that different protocols will yield different amounts of information, different types of information, and have different impact on the execution of the activity. Here, we analyze how the biosignal data is recorded in a table setting scenario and compare the generated spoken language.

The EASE Human Activity Data (EASE-TSD) collection in the Biosignals Lab at the Cognitive Systems Lab at University Bremen focuses on household chores in an artificial kitchen setup, where participants perform activities while being

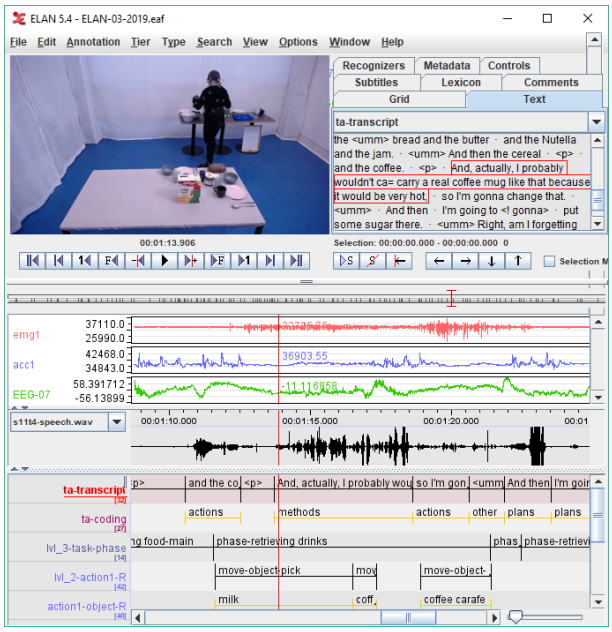

Figure 1: *Partitur file of a table setting scene with multiple biosignals and annotations created in ELAN.*

recorded by a multitude of sensors. The Biosignals Acquisition Space and Environment (BASE) is a 5 x 4 meter interaction space equipped with recording devices including cameras, microphones, a motion capture system, as well as body-worn sensors for inertia, electrodermal activity (EDA), electromyography, electroencephalography, and eye gaze. For more information regarding the sensors and experiment scenario, refer to the detailed technical descriptions in [3, 4]. The EASE-TSD data will be made available to the community after completion.

2. Related Work

2.1. Multimodal Data Sets of Everday Activities

Studies of everyday activities often focus on modeling of kitchen activities. For example, Gehrig et al. [5] focused on the joint recognition of intention, activity, and motion during a cooking task. Their video dataset of ten participants performing ten repetitions of seven unique tasks in a kitchen setting, included 700 image sequences and, after manual annotation, 60 motion primitives, and inferred objects from related activities. The TUM Kitchen Dataset [6] includes video, full-body motion capture, RFID tag readings and magnetic sensor readings from objects used during activities in a natural kitchen setting. They produced manual motion tracker labels and automatic seman-

tic segmentation from this data. The 50 Salads dataset [7] was produced in a non-natural kitchen setting using a third person top-down point of view camera, where participants prepared a meal according to a recipe. Sensors used included RGB-D cameras, and accelerometer-mounted objects. The EPIC Kitchens dataset [8] includes annotations using 125 verb classes and 352 noun classes in varied languages, based on video from a head mounted camera of kitchen activities performed by 32 participants in their homes. Only EPIC-Kitchens used verbal annotation of the executed actions during an activity, but in a highly standardized and formal way. For the EASE Human Activity Data (EASE-TSD), we concentrate on participants' natural speech to better capture the full spectrum of cognitive processes and behavior annotation.

2.2. Think-aloud Methodologies

When deciding upon the potential think-aloud methods that might best serve the purposes of data collection of everyday activities for instructing robots, we weighed a combination of factors: balancing data quality and quantity versus unintended effects on task completion time or other recorded modalities (e.g. minimizing artifacts in EEG), minimizing participant priming for specific behavior and cognitive load, maintaining participant focus and engagement throughout the experiment, and the limitations of memory to produce accurate descriptions, while trying to maintain as natural as possible a setting. Gaining insight from observation of participants' thoughts and perceptions while performing activities requires understanding of influences on the participants' performance while verbalizing in sufficient detail [9, 10].

In particular, we must encourage spoken language that describes the aspects of the activity which would be critical to robotic function in a natural setting, which includes but is not limited to tasks, broken down into actions, the objects manipulated during those actions, and the movement of the participant during the task. Other aspects of interest might be the methods of object manipulation, perceptual details that may affect task completion, and the reasons for deciding to take individual actions. Thus, we are interested in whether different think-aloud protocols elicit different types of information.

Numerous investigations of various think-aloud protocols, including concurrent and retrospective methods, have been performed [11, 12, 13, 14]. It should be noted that concurrent methods used generally do not strictly enforce concurrency during physical activities, as is the case in this study, so conclusions on their validity may not be applicable within this context. Consequently, we must also examine whether concurrent speech might slow or other wise discourage normal levels of activity. Think-aloud methods in HCI studies often allow experimenters to interact with participants to varying degrees in order to gain more frequent or detailed information. As we strive to maintain as natural as possible context when performing table setting, we have chosen to provide instruction only before trials begin and during trial intermissions, as needed. For these reasons, an investigation into the effects of different think-aloud protocols specifically in the area of everyday activities was warranted.

3. Experimental Setup and Data Collection

To compare the output of the think-aloud methods produced during the primary trial variations of interest using audio and video data, we recorded data of 18 participants (six of whom were female), aged 19 to 40 performing table setting with at least 6 trial variations. Participants performed think-aloud trials in German or English language. A total of 234 minutes of audio data have been produced thus far, from which a total of 85 transcripts for retrospective trials and 33 transcripts for concurrent trials have been produced, comprised of 254 and 193 average number of words per trial, respectively. A trial takes an average of 2.6 minutes, with a standard deviation of 1.4.

3.1. Trial Variants

Trial variations were devised to allow observation of a wide array of behaviors and strategies under different conditions. Variation in table setting approaches may emerge due to difference in meal type (breakfast vs. lunch), formality based on diners attending the meal, and differences in group size 'n', see Figure 2 for the breakdown of variants. For example, an informal breakfast table may entail many shared object between diners, whereas setting the table for lunch in a formal setting may be slower as the person displays more attention to detail.

Breakfast

	n=2	n=4
formal		
informal	x	x

Lunch

	n=2	n=4
formal	x	x
informal	x	x

Figure 2: *Trial variants by meal, formality and number of diners.*

Each experimental session follows a set sequence, where participants read information about the study, are provided additional details if requested, give their written consent to study participation, have sensors set up and calibrated, and receive trial and think-aloud instructions (including a training run in a different toy task). Then, they perform a series of table setting trials in random order, and are debriefed. Pre-trial and post-trial questionnaires are also administered.

Before each trial, an identical selection of objects such as cutlery, dishes, and food are arranged on a counter. Some objects are meant to be shared between meals (e.g. bread, butter, water), some were intended for lunch (e.g. soup bowls, salad cutlery, forks), and some for breakfast (e.g. jam, cereal, mugs). Participants were not instructed on item use or placement. Besides the counter, a cleared dining table is available to be set during the trial.

Instruction for each trial followed the same form:

"You and your companions stay in a vacation apartment. It is your first day. The others are out shopping; you agreed to set the table. You want to set this table for a [formal/informal] [breakfast/lunch] for [2/4] people. You want to put all the necessary objects from the counter onto the table and have everything ready to eat, as your friends will return shortly. Please 'think-aloud' about what you are doing as you perform the activity. There are no time constraints (perform at a natural pace). Assume no contextual knowledge from one trial to the next."

3.2. Think-aloud Protocols

The think-aloud methods evaluated in this study include a concurrent think-aloud protocol and a retrospective reporting protocol. Participants either speak while performing activities (concurrent), or while observing their behavior through recorded video from the top camera perspective (retrospective), as shown in Figure 3. For both protocols, speech was recorded through a close-talking microphone. Each trial was either commented using the concurrent or the retrospective protocol, or both. The as-

signment of trial variants to protocols was randomized for each participant. The same basic instructions were given during every trial, regardless of think-aloud method:

"Describe your actions (at each moment) as you interact with the environment. Talk constantly, so long as it does not interfere with the ability to perform the activity. Treat each trial independently, repeat words as needed."

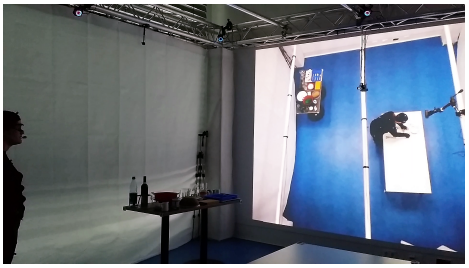

Figure 3: *Participant performing retrospective report*

4. Analysis

4.1. Data Transcription and Annotation

Through transcription with the ELAN [2] annotation tool, we transcribed and annotated the German and English speech recordings. Transcriptions were performed according to a set of rules based on the Verbmobil transcription rules and modified to fit the needs of automatic speech recognition and semantic analysis. Speech segments were chosen based on speech activity such that the timing of spoken words were derived with high precision. Hesitations, non-speech and non-verbal parts, slang, and other variations were denoted with specific symbols.

Table 1: *Think-aloud (TA) utterance coding scheme*

TA code	**Description**
Perception	Describe what they perceive (see, hear, etc)
Actions	Talk about actions (planned, current, or future)
Plans	Describe future goal states (aside from actions)
Methods	Talk about how they achieve a state
Issues	Indicate difficulties or confusion
Reasoning	Talk about why something occurs
Task evaluation	Talk about the state of the task
Memory	Talk about what they do or don't remember
Thoughts	Talk about thought processes
Opinions	Talk about opinions, feelings, etc
Questions	Ask a question
Other	Talk about subjects not pertaining to the task

To gain understanding of whether the described think-aloud protocols provide insight into different cognitive processes, such as planning, evaluation, or behavior execution that will be of most value within the full dataset, one person manually coded utterances using the scheme depicted in Table 1 for this pilot study. These categories serve as the basis for comparison of the content of trial verbalizations in the protocol types as well as through the sequence of actions within a given task. They are based on think-aloud utterance categories common to the relevant papers referenced, as well as to corresponding terms in the EASE Ontology. For the full dataset, multiple individuals will encode each trial, then these annotations will be used to measure inter-rater reliability.

Table 2: *Mean think-aloud speech/non-speech values per trial*

	Concurrent	**Retrospective**	**All**
Words (#)	198	255	228
Words/minute (#)	75	95	86
Hesitations (#)	2	5	4
Pauses (#)	27	26	26
Pause duration (s)	3.90	4.04	3.98
Vocabulary Size	88	100	95

To assess whether participants produced more detailed information in either condition, a measure of their activity is needed. The final frames of the trial video provide a visual check of the objects set during the activity, so the count of objects for each trial was also recorded. This information was used to determine which trials should be used for further analysis, in this case those trials expected to require the fewest/most objects–'informal breakfast for two' and 'formal lunch for four.'

4.2. Think-Aloud Protocol Comparison

Through analysis of the data recorded during these trials, we aim to gain understanding of the relative levels of verbal output, complexity, and activity versus indicators of factors detrimental to successful completion of the task, e.g. pauses and hesitations.

Using these data and annotations when comparing the table setting trials, on the basis of trial types and think-aloud methods, we determine task completion times, average number of words and speaking rates for each trial type, the number and duration of pauses and hesitations, the vocabulary size, complexity, parts-of-speech distribution, and the average number of objects and object types used. Detailed speech content analysis through Linguistic Inquiry and Word Count (LIWC) score comparisons [15] allows us to better understand the participants' focus as they perform their tasks. The LIWC software provides counts of the words and parts of speech, divided into psychology-relevant categories, used in speech produced during think-aloud trials. Frequency of terms and think-aloud utterance category types are then compared based on the think-aloud methods used during the trial, which may be used as a basis to compare each methods' relative value in this table setting context.

5. Results

A summary of the finding produced based on the analysis of the transcriptions of the think-aloud protocols is contained in Table 2. The mean number of words spoken during retrospective reporting is 254 vs. 198 in concurrent speech. Consequently, the mean words per minute for retrospective reports is also comparably high. The mean number of pauses for both conditions was 26 per trial. However, the mean duration of pauses was 0.14 seconds longer during retrospective reports. The mean number of hesitations was also higher for retrospective reports, at 5, compared to 2 for concurrent speech. For the LIWC analysis in Figure 4, the results are normalized to account for the increased speaking rate.

The average object counts, shown in Table 3, refer to the following trial variations: 'informal breakfast for two' (BI2), 'informal lunch for two' (LI2), 'formal lunch for two' (LF2), 'breakfast for four' (BI4), and 'informal lunch for four' (LI4) and 'formal lunch for four' (LF4), with a breakdown of trials

Table 3: *Average object counts per trial variation*

	BI2	LI2	LF2	BI4	LI4	LF4
Silent	19	28	21	27	31	35
Concurrent	21	19	23	28	32	35
All	21	22	23	27	31	35

Table 4: *Average trial duration per variation, in seconds*

	All	Speech	Silent	n=2	n=4	Formal	Informal
Duration	153.8	154.2	173.6	137.3	183.4	173.3	154.8

based on silent and concurrent think-aloud procedures followed during the table-setting activity. The mean trial durations by trial condition are shown in Table 4. The average object counts per trial display the expected pattern–fewer diners, 'breakfast,' and 'informal' setting conditions correspond to fewer objects required per trial. The silent condition trials conform to the same pattern overall, with slightly fewer objects used.

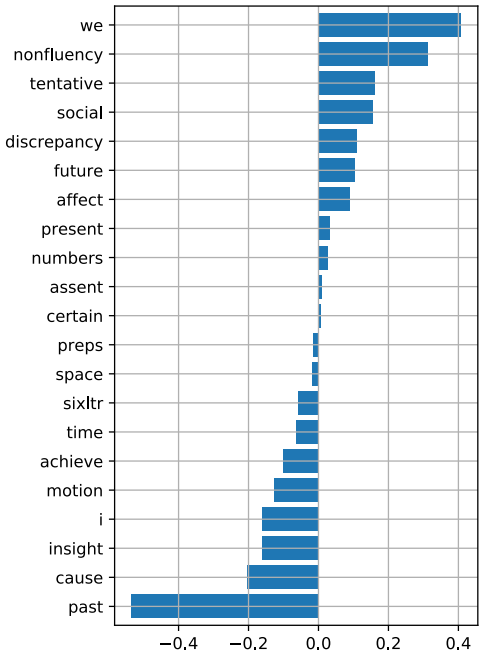

Figure 4: *Normalized difference of relative frequency of LIWC categories for retrospective and concurrent trials. Categories that occur proportionally more in the concurrent think-aloud trials have bars that extend to the right side, and vice versa.*

The plot of LIWC terms in Figure 4 shows relative frequencies of the categories for retrospective r and concurrent c think-aloud trials calculated as $(c - r)/(c + r)$. Duplicate concepts were removed, e.g., 'positive' and 'negative' emotions that were also categorized as 'affect.'

In Figure 5, think-aloud categories are compared based on relative frequency during either of the think-aloud trial types. Participants spoke about their thoughts, reasoning, and actions more frequently while watching the trial they just performed (retrospective think-aloud). However, during the task they were more likely to describe their perceptions, issues they encountered, give their opinions, or to evaluate the state of the task (concurrent think-aloud).

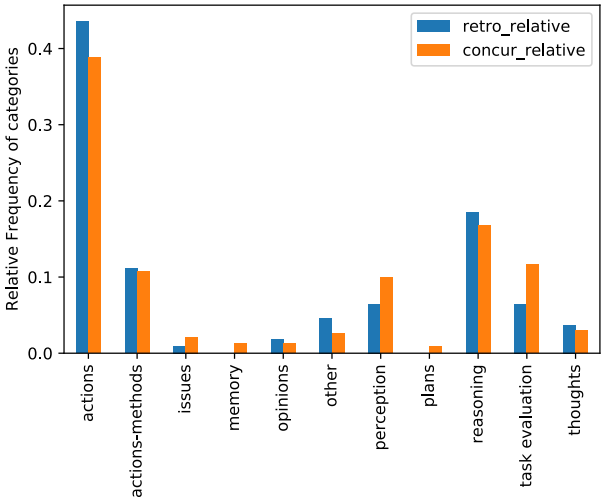

Figure 5: *Comparison of relative frequency of think-aloud categories in retrospective and concurrent trials*

6. Discussion and Conclusion

While the amount of speech produced through retrospective think-aloud procedures may be moderately larger than that of concurrent think-aloud procedures, the type of information produced differs greatly. Shorter duration pauses during concurrent conditions may indicate that participants are better able to align their speech to their actions. Contrary to expectations, the number of hesitations is lower during concurrent speech, and the mean number of pauses are only slightly higher. Based on the relative frequency of LIWC categories, non-fluent and tentative speech is more common during concurrent trials, indicating that participants may have encountered situations that made it difficult to think-aloud while performing tasks. Similar to the think-aloud category comparison, terms related to 'motion,' and 'reasoning' were more common during retrospective trials. Self-reference was also more common, which may be attributed to using a third person perspective as reference.

Some terms followed expected patterns, with 'future' and 'present' related terms more common in concurrent trials and 'past' terms more common in retrospective trials. Speech related to the backstory context, such as 'social' terms, pronouns, and 'affect' related terms were more prevalent when people talked about the task during the concurrent trial. 'Task evaluation' terms were more common in concurrent trials, but speech shifted to 'achievement' terms in the retrospectives, based on the LIWC frequency analysis.

Based on this analysis, it may be concluded that the concurrent think-aloud protocol does not slow or otherwise decrease activity. Both think-aloud methods provide unique, useful information for creation of NEEMs that enhance cognition-enabled robots.

7. Acknowledgements

The research reported in this paper has been supported by the German Research Foundation DFG, as part of Collaborative Research Center (Sonderforschungsbereich) 1320 EASE - Everyday Activity Science and Engineering, University of Bremen (http://www.ease-crc.org/). The research was conducted in subproject H03 Descriptive models of human everyday activity.

8. References

[1] M. Beetz, D. Beßler, A. Haidu, M. Pomarlan, A. K. Bozcuoglu, and G. Bartels, "Knowrob 2.0 – a 2nd generation knowledge processing framework for cognition-enabled robotic agents," in *International Conference on Robotics and Automation (ICRA)*, Brisbane, Australia, 2018.

[2] P. Wittenburg, H. Brugman, A. Russel, A. Klassmann, and H. Sloetjes, "Elan: a professional framework for multimodality research," in *5th International Conference on Language Resources and Evaluation (LREC 2006)*. Max Planck Institute for Psycholinguistics, The Language Archive, Nijmegen, The Netherlands, 2006, pp. 1556–1559. [Online]. Available: https://tla.mpi.nl/tools/tla-tools/elan/

[3] M. Meier, C. Mason, R. Porzel, F. Putze, and T. Schultz, "Synchronized multimodal recording of a table setting dataset," in *IROS 2018: Workshop on Latest Advances in Big Activity Data Sources for Robotics & New Challenges, Madrid, Spain*, 2018.

[4] C. Mason, M. Meier, F. Ahrens, T. Fehr, M. Herrmann, F. Putze, and T. Schultz, "Human activities data collection and labeling using a think-aloud protocol in a table setting scenario," in *IROS 2018: Workshop on Latest Advances in Big Activity Data Sources for Robotics & New Challenges, Madrid, Spain*, 2018.

[5] D. Gehrig, P. Krauthausen, L. Rybok, H. Kuehne, U. D. Hanebeck, T. Schultz, and R. Stiefelhagen, "Combined intention, activity, and motion recognition for a humanoid household robot," in *2011 IEEE/RSJ International Conference on Intelligent Robots and Systems*, Sept 2011, pp. 4819–4825.

[6] M. Tenorth, J. Bandouch, and M. Beetz, "The TUM Kitchen Data Set of Everyday Manipulation Activities for Motion Tracking and Action Recognition," in *IEEE International Workshop on Tracking Humans for the Evaluation of their Motion in Image Sequences (THEMIS), in conjunction with ICCV2009*, 2009.

[7] S. Stein and S. J. McKenna, "Combining embedded accelerometers with computer vision for recognizing food preparation activities," in *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing*. ACM, 2013, pp. 729–738.

[8] D. Damen, H. Doughty, G. M. Farinella, S. Fidler, A. Furnari, E. Kazakos, D. Moltisanti, J. Munro, T. Perrett, W. Price, and M. Wray, "Scaling egocentric vision: The epic-kitchens dataset," in *European Conference on Computer Vision (ECCV)*, 2018.

[9] K. A. Ericsson and A. S. Herbert, "Verbal reports as data," vol. 87, no. 3, pp. 215–251, 1980.

[10] T. Boren and J. Ramey, "Thinking aloud: reconciling theory and practice," *IEEE Transactions on Professional Communication*, vol. 43, no. 3, pp. 261–278, Sept 2000.

[11] M. Hertzum, K. D. Hansen, and H. H. Andersen, "Scrutinising usability evaluation: does thinking aloud affect behaviour and mental workload?" *Behaviour & Information Technology*, vol. 28, no. 2, pp. 165–181, 2009. [Online]. Available: https://doi.org/10.1080/01449290701773842

[12] T. Zhao and S. McDonald, "Keep talking: An analysis of participant utterances gathered using two concurrent think-aloud methods," in *Proceedings of the 6th Nordic Conference on Human-Computer Interaction: Extending Boundaries*, ser. NordiCHI '10. New York, NY, USA: ACM, 2010, pp. 581–590. [Online]. Available: http://doi.acm.org/10.1145/1868914.1868979

[13] J. C. Welsh, S. A. Dewhurst, and J. L. Perry, "Thinking aloud: An exploration of cognitions in professional snooker," *Psychology of Sport and Exercise*, vol. 36, pp. 197 – 208, 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1469029217306763

[14] T. Towne, K. Ericsson, and A. Sumner, "Uncovering mechanisms in video game research: suggestions from the expert-performance approach," *Frontiers in Psychology*, vol. 5, p. 161, 2014. [Online]. Available: https://www.frontiersin.org/article/10.3389/fpsyg.2014.00161

[15] Y. R. Tausczik and J. W. Pennebaker, "The psychological meaning of words: Liwc and computerized text analysis methods," *Journal of language and social psychology*, vol. 29, no. 1, pp. 24–54, 2010.