Interaction Analysis and Joint Attention Tracking in Augmented Reality

Alexander Neumann
Ambient Intelligence Group
Bielefeld University
Bielefeld, Germany
alneuman@techfak.uni-bielefeld.de

Christian Schnier
Interactional Linguistics & HRI
Bielefeld University
Bielefeld, Germany
cschnier@techfak.uni-bielefeld.de

Thomas Hermann
Ambient Intelligence Group
Bielefeld University
Bielefeld, Germany
thermann@techfak.uni-bielefeld.de

Karola Pitsch
Interactional Linguistics & HRI
Bielefeld University
Bielefeld, Germany
karola.pitsch@uni-bielefeld.de

ABSTRACT

Multimodal research in human interaction has to consider a variety of factors, ranging from local short-time phenomena to complex interaction patterns. As of today, no single discipline engaged in communication research offers the methods and tools to investigate the full complexity continuum in a time-efficient way. A synthesis of qualitative and quantitative analysis is required to merge insights about micro-sequential structures with big data patterns. Using the example of a co-present dyadic negotiation analysis to combine methods offered by Conversation Analysis and Data Mining, we show how such a partnership can benefit each discipline and lead to insights as well as new hypotheses evaluation opportunities.

Categories and Subject Descriptors

H.5.1 [**Information Interfaces and Presentation**]: Multimedia Information Systems–*Evaluation/methodology;* J.5 [**Computer Applications**]: Arts and Humanities–*Linguistics*

Keywords

Interaction studies; data mining; conversation analysis; multimodality

1. INTRODUCTION

Human-human cooperation is a complex multimodal phenomenon and subject to research in linguistics, computer science, social sciences etc. In co-present interaction, the participants have a range of communicational resources at their disposal, such as verbal utterances, gestural signals, facial expressions, deictic gestures, body and head orientation. Obviously co-presence and cooperation couples the interaction partners. This coupling, ranging from the communicative surface of observable behavior to the couplings of internal representations is investigated in the context of a research center on *Alignment in Communication*. In this context we developed a new paradigm for the examination of cooperation, which offers a new form of control and manipulations to investigate the particular role of signals and perceptions: the *Interception and Manipulation Interface* (ARbInI) [1].

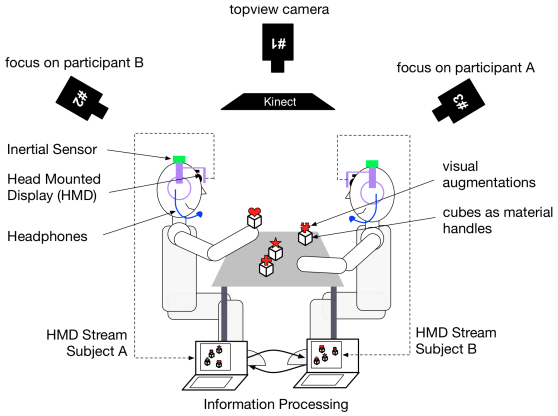


Figure 1: The system contains sensors and devices worn by the participants, a set of static cameras and a depth sensor. Data recorded during a trial is partially processed in real-time and fed back to the participants.

This interface uses Augmented Reality through head-mounted displays (HMD) to intercept a participant's view on a scene, who sees the video signal captured by a camera in the HMD instead. Likewise we can intercept and manipulate the auditory signal by using microphones and closed headphones. This allows us to register which signals (video/audio) the interaction partners have at their disposal at any time, i.e. for the first time in co-present interaction we can truly get access to relevant perceptual cues. In addition we attach sensors to the users' heads to register head orientation, movements and gestures, and use external DV cameras as well as a Microsoft Kinect depth sensor to measure more details of the actions. The collected multimodal interaction data can be investigated from different perspectives to gain insights about the organization of human cooperation.

Conversation Analysis (CA) and *Data Mining (DM)* are both highly developed and established research methods [2] [3] but are so far not or rarely used in combination [4]. We suggest to explore how they can mutually cross-fertilize insights from the other angle, or allow new synergies. For such a linking of methods particular corpora are required which encompass different data streams needed for different analyses. Also, new ways of linking different research approaches need exploring to define how systematization and formalization can bridge the methodological gaps and how one approach might be able to provide analytical support for the other. This includes the challenging task to describe sequential structures in a systemized and formalized way for technical systems to process these structures and to be able to find them within the data [5].

In this paper we will demonstrate how DM methods allow overview rendering of the macro-structure of interaction, which are useful for CA to select interaction episodes to be examined in detail, and how CA identifies patterns, which inspire the feature extraction for DM. We start our presentation with an overview of the AR-system as research instrument and the scenario used to elicit the phenomena of interest. These are in this paper the analysis of joint attention and the coordination between users to establish and maintain it over interaction using various semiotic fields. The discussion of study trial segments both from the sides of CA and DM will allow us to see how conceptually different the approaches are, and where we see intersections where the method link is particularly promising. The subsequent discussion will focus on our experiences with this method link.

2. SYSTEM OVERVIEW

The foundation of our system is the *Augmented Reality based Interception Interface* (ARbInI) [1]. As depicted in Figure 1 our interface contains several devices worn by the user for data recording and feedback. Every user wears a video see-through head-mounted display. Additionally, all users are equipped with an inertial sensor for head movement tracking from the BRIX toolkit [6] and a headset as depicted in Figure 2. The input of the cameras inside the HMD is augmented by our AR-core system [7] and fed back to the user. For marker tracking we rely on the ARToolkitPlus[1].

Besides the worn components, the system's configuration includes static sensors and cameras. The scenario is monitored by three DV-Cameras from the top and from both participants' "shoulder perspectives". Pointing downwards from right above the table there is a *Microsoft Kinect*[2] which tracks depth information and sends them to a computer where the information is stored as a stream of depth images.

All data monitored and collected by the system are available for on- and offline processing. The concept of ARbInI also contains headsets to filter and/or augment what participants hear. This feature was not used during the study discussed in this paper and therefore headsets were not required.

Video and sound files are synchronized with the help of a prototyped sync detection based on BRIX and several open source video and image editing tools [8]. The merged video is the point of destination for the initial annotation process in Conversation Analysis.

[1] handheldar.icg.tugraz.at/artoolkitplus.php

[2] kinectforwindows.org

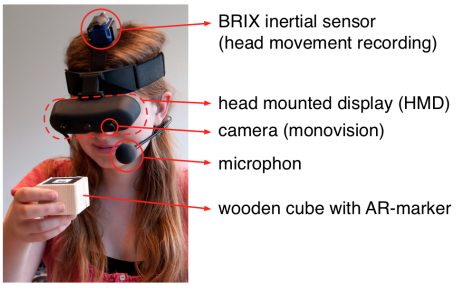


Figure 2: Subject's trial equipment. Participants are equipped with a video see-through head mounted display, a microphone and a BRIX inertial sensor. The video stream is enriched so that subjects see virtual 3D-models placed on the wooden cubes.

3. OBERSEE CORPUS: EXPERIMENT & DATA

The Obersee Scenario was designed to foster negotiation between two participants. Originally created for [4], we redesigned the concept to fit the specific conditions of our AR-Setting. In this scenario two participants are asked to plan a fictional recreation area around the Obersee, a lake in the northwest of Bielefeld. To increase engagement and an easier access, the participants are asked to argue from either the perspective of a conservationist or from an investor's point of view. The only other rule is to get to an agreement in roughly 20 minutes.

We provide 18 mediation objects of possible structures, which can be placed around and in the lake. A mediation object consist of a wooden block which is supposed to be used as a handle and a marker (attached to the top surface) which is used to track an object position and augment the right virtual object into the participant's video stream.

These objects include profit-oriented structures – like a hotel or a water ski installation – and preservation objects like a water protection sign or a nature reserve symbol. Besides these very opposing objects, the majority was designed to be located *somewhere* on the line between profit and natural reservation.

4. METHODS

To study the benefits of data mining methods for CA we conducted the analysis of a study trial separately using methods inspired of Conversation Analysis in its recent multimodal developments and Exploratory Data Analysis and compared the results. This parallel work is necessary to have a technologically unbiased view on the trial so that assumptions made during the data crunching have no effect on the results of the qualitative analysis process.

4.1 Conversation Analysis

Conversation Analysis (CA) describes a qualitative analytical approach, which aims to reveal the underlying orderliness and sequential patterns of everyday social interaction. In this vein, CA is interested in how interlocutors organize their multimodal actions in a meaningful way and in close coordination to both each others behavior and to the material environment, in which their actions are situated. It is in the scope of the analyst to reconstruct the procedures participants use in order to reach particular interactional tasks and subtasks (e.g. perceiving sb., establishing co-orientation, getting the right to speak etc.). Thereby, the procedures' reconstruction is based on the action's interactive organization by analyzing how the respective co-participant co-designs a current turn and/or reacts to a prior turn.

This allows a reconstruction of how the interlocutors have interpreted each other's multimodal projections (gaze, gesture, posture etc.) in situ [2]. Originally developed on the basis of audiotaped recordings of telephone conversations, it has been further developed for the study of multimodal phenomena (Goodwin 2000, Mondada 2006, Heath & Luff 2012 in [9]).

4.2 Data Mining

Data mining is defined as the science of extracting useful knowledge from a large set of data in an automatic or (more common) semi-automatic way [10]. Useful knowledge means patterns or any kind of reorganization, which makes it easier for humans to process this information. The term *data mining* is a bit fuzzy. It does not refer to brand new approaches but covers methods from statistics, machine learning and other information extraction fields.

Besides a wide range of hypothesis validation tests, one also refers to data mining when *Exploratory Data Analysis* (EDA) is chosen due to the lack of a-priori hypotheses about the (often unknown) nature of the data in question. During the exploration process one tries to get a better understanding of the data. This often includes a variety of data visualization attempts like graphs, plots and tables [11]. Sometimes a short look at the raw data already provides cues for a further, more precise analysis. Results can be insights about correlations between features, interesting subsets of collected data or initial hypotheses about the relation of attributes to a target variable [3].

4.3 Quantifying approaches in the study of conversation – state of the art

According to Schegloff [12], the quantification of conversation analytical results is problematic as interactional phenomena are highly context-sensitive and embedded in the sequential organization of talk-in-interaction. Taking the example of "laughter" he stresses that objective measuring units (e.g. "laughter per minute"), used in descriptive statistics, are not the decisive factor for the phenomenon's relevance, but rather that its positioning in the sequential organization of talk and its relevance for the respective co-participant matters. This means that the quantity of the phenomenon's occurrence is not significant to show e.g. sociability, but rather depends on how the interlocutors interactively organize their talk in situ. In this vein, Schegloff treats quantification as a challenging task and suggests that some phenomena, like e.g. repairs with its definable sequential properties, are more suitable to quantify than others [13] [14].

Recently, Schegloff's core statement "the proper grounding and payoffs of quantification have not yet been thoroughly explored" (ebd.) has been addressed by various researchers in the anthology "Conversational Informatics" [5] in order to develop knowledgeable embodied conversation agents (ECAs). However, mostly guidelines for the formulation of interactional strategies are based on numerical characterizations and are divided in context-free class memberships. Nevertheless, some promising approaches implement rules for ECAs as a form of "action-to-action mapping" (Den & Enomoto 2007 in [5]), which addresses the idea of CA to consider "*rules as practices*".

Only a few approaches in conversational research have the objective to quantify multidimensional sequential structures rather than single events, which allow to discover complex hidden repeated patterns on large data corpora (cf. "T-Patterns" in [15]). Our method synthesis of CA and EDA is similar, but profits from its bidirectional approach. Both CA and EDA can verify their results on different levels (CA on the micro-sequential structure; EDA on the macro level) to improve gradually the accuracy of the described phenomenon.

4.4 Method Synthesis

The way in which phenomena observed in Section 5 & 6 below correlate will give evidences about how the data driven approaches can ease analyst's work in the future. CA lacks the possibility to get a rough overview about a huge amount of data in an acceptable amount of time, which EDA offers. Initial findings of data driven analysis can provide pre-structuring of data sets or – in trial-based studies – which data recordings have a high chance of offering a rich set of (pre-defined) potentially promising phenomena.

Additionally, with the help of the results of CA we hope to improve the performance of our data mining mechanisms. Hypotheses developed during data mining processes can be reviewed qualitatively. This might help to identify appropriate algorithms and parameters. The inverted hypothesis verification also provides cues about how well interaction patterns are defined and where more detailed descriptions are required to attempt to generalize of such patterns.

5. CONVERSATION ANALYSIS

In what follows we present a short analysis inspired by Conversation Analytic methods exploring the interaction of a group of participants from our corpus. We will focus on one particular procedure, used by participants to introduce new objects in order to discuss their placement on the map. Thereby, one essential subtask consists in how the initiating party establishes a common focus of attention to the relevant next object. In this vein, we are interested in the interlocutors' procedures of how physical objects are systematically used in the sequential structuring of joint attention activities. In the following we will focus on one particular procedure, which is frequently found in our corpus: Participants pre-configure the material environment and establish mutual orientation at a suitable point in time. From an interactional point of view, analysis suggests that the phenomenon of "joint attention" is not just limited to the human's ability of gaze following [16], but is essentially a multimodal interactional process [17], which is closely interleaved with the participants' orientation to the current task [18].

5.1 Introducing a new object and establishing joint attention

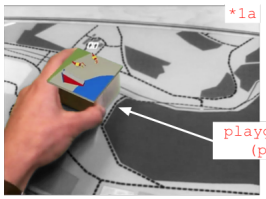
In order to solve the given task described in Section 3, the participants have to suggest objects and negotiate if and where they want to place them. We enter the interaction at a moment in time where a transition between the interlocutors' current task – the conceptual and physical integration of the object "playground" – and an upcoming next task – a negotiation about the object "hotel" – takes place. At the beginning of the following fragment we can recognize that both participants do not accomplish a coordinated end of the object's negotiation "playground". Notations like "01" refer to particular tiers in the transcript. Notations like *1a+b refer to particular pictures, marked with *1a and *1b, in the transcript. The further annotation style, used in addition to the GAT-conventions [19], is explained in the Section 12.

Participant B is about placing the previously discussed object "playground" on the map and thereby projects the completion of this phase (01+*1a+b). Participant A reacts to it by asking "THERE:: you would like to have it," (01) and visually scrutinizes the map checking for potential other locations where the "playground" could be placed (*2b). B briefly answers "YES;"

and orients to the stack of objects located next to the map (*2a). While B thus orients to starting a new task, A continuous to inspect the map for other potential locations to place the "playground". Thus, both participants are oriented to different parts of the map at the end of the previous object's negotiation (cf. *2a+*2b).

Fragment F1.1: 12:09 – 12:53

```
01 B-act: |place(pla)@map|
   A-ver:                |DA:: wollen=sie=es HIN haben,
                         THERE:: you would like to have it,
                           *1a+b
```

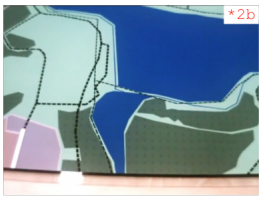


B's field of view

A's field of view

```
02 B-ver:  JA;|
           YES;
   B-gaz:     |››››››stack|@stack..................
   A-ver:                  |ähm::: JA ich denke AUCH
                            uhm::: YES i   think ALSO
   A-gaz:     |@map›››››››››››››››pla..............
                           *2a+b
```

B's field of view



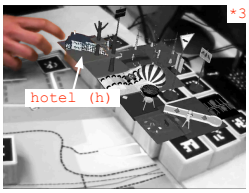
A's field of view

```
03 A-ver:  das es DA ganz gut platziert is;|
           that it is placed well THERE;
   A-gaz:  @stack.........................|
   B-gaz:  @pla...........................|
```

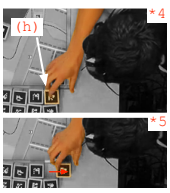
The further course of interaction reveals which multimodal procedures lead to a common focus of reference. Considering line 04 and 05, we can recognize that B uses A's closing statement in 03 to open a new interactional task by addressing the question "SO:; what is particularly !IMPOR!tant for YOU?" and thereby verbally invites A to suggest a next object. Both, the starting marker "SO:;" together with B's right hand in ready-to-grasp-position at the stack (cf. *3) leads to A's re-orientation from the object "playground" (03) to the stack of objects alongside the map in 04. At this moment in time both participants are co-orientated to the stack and A can easily follow B's action: He grabs the object "hotel" from the stack and moves it a few centimeters away from the grouped objects (cf. *4 & *5). Thus, in parallel to his invitation, he chooses himself a next candidate by physically restructuring the interaction space.

```
          12:20
            |
04 B-ver: |SO:; was is IHNEN denn noch besonders
          SO:; what is for YOU       particularly
   B-act: |grasp(h)›››››››››››››››››››pull-out(stack)
   A-gaz: |follows B-act.............................
          *3                          *4
05 B-ver:  !WICH!tig,|
           !IMPOR!tant,
   A-ver:            |°h (2.4) JA: also wie gesagt
                      °h (2.4) WELL: as already stated
   A-gaz:            |@stack›››››››››››››››››››››››
          *5

06 A-ver:  wir dürfen hier die: GRÜNflächen natürlich
           we  ca not here  the:  GREEN space  of course
   A-gaz:  ›››››››››››››››››@stack(foc:h).........
                                *6
```



A's field of view



topview on B

A's field of view

```
07 A-ver:  nich beRÜHREN- und ä|hm
           not   TOUCH-    and  uhm
   B-ver:                      |JA::=aber das ist ja
                                WELL::=but this is

08 B-ver:  kein naTURschutzgebiet-
           not a CONSERVATION area-

      (...) ➤ongoing dispute, not solution-oriented
```

When we consider the gaze-annotation in the following lines 05 and 06, it becomes clear that B's pure action of object reconfiguration is not a relevant orientation-hint for A at that point in time. Instead, A reacts to B's question by firstly scrutinizing the left objects at the stack (marked in green). At the same time, he answers B's question not as an invitation to decide about the next object, but rather with a verbal declaration of principles (05-07), which finally leads to an ongoing, not solution-oriented dispute. Only within this declaration of principles, he explores B's prepared object in passing (06: marked in red & *6). Consequently, the used procedure is different to what is termed by Clark as a "placing-for-procedure" where objects are placed in special sites for addressees in order to make them interpretable [16]. In our case, the object "hotel" is not constituted in a way to offer a connection for A's next action. Instead, B addresses a question to A, which move away the attention of his preparing activity. Thus B prepares a possible next candidate for himself.

```
09 B-ver:  ALSO:: was natürlich noch |!GANZ! WICHtig is
           SO::   what of course also !VERY! IMPORtant is
   B-act:                           |grasp(h)›››››››››
                                                  *7

09 B-ver:  is nen DICKES hotel; (1.4)
           is a    BIG    hotel;  (1.4)
   B-act:  lift(h)›››››››hold.......
                          *8
```

A's field of view

A's field of view

Indeed, B introduces the object "hotel" 27 seconds after the initial object's preparation. Thereby we observe that the actual reference to the object orients to *transition relevance places* (TRP). After the non-solution oriented dispute a pause of two seconds arises. At this point in time, both participants are free for a new interactional task because the crosstalk has come to an end. Interestingly, both participants are oriented to the object "hotel" during that time so that we can assume that A anticipates B's following object's introduction. That shows that both participants have interactively developed a plan for the next action – B by actively preparing these action and A by recognizing that B has prepared a next action. Indeed, B uses this pause and refers

to the initial prepared object “hotel” with an account “SO:; what is of course also !VERY! IMPORtant”, followed by the object’s identification “is a BIG hotel;”. Simultaneously to his verbal introduction, he grasps the isolated object out of the stack (cf. *7) and brings it into a shared scope of action (cf. *8). The found pattern can be gradually described as a 4-step procedure:

1. Spatial reconfiguration, constituted as an action of preparing a next object for oneself (B)
2. Recognizing spatial alternations (A)
3. Exploring the semantic content at a suitable point in time (A)
4. Refer to the prepared object at a suitable point in time (after TRPs) for establishing joint attention and introducing the object as next interactional task (B)

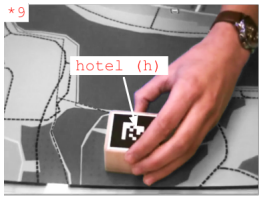
5.2 Proposing the prepared object

Shortly after joint attention to the object “hotel” is successfully established, we observe that B suggests a concrete placement of the object (cf. *9).

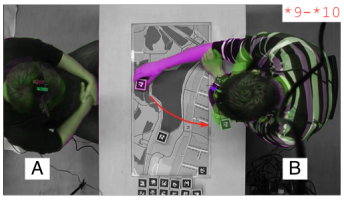
Fragment F1.2: 12:55 – 13:00

```
10 B-ver: am !BES!ten jetzt |hier auf diesen WALD(1.2)|
          most !SUIT!able now here on   that FORREST(1.2)
   B-act:                  |move(h)››››place@map›››HP|
                                                *9

11 A-ver: '!DA::! |auf GAR keinen fall;
   B-act:         |grasp(h)›››››››››››››››move››››››HP
                                                  *10
```



A's field of view

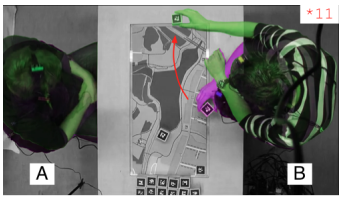


topview

A reacts to it by rejecting the object’s placement with “!DA::! Auf GAR keinen fall;” (“THERE::! Not on ANY account;”) (11). B immediately reacts to A’s rejection by taking the object away of its placement (*9-*10: purple overlay) and holding it in readiness at the edge of the map (*9-*10: green overlay).

Our analyses reveals that it is not just the case that the proposing party tends to talk more to the proposal in cases where the recipient “do-not-yet-accept” the proposal [17]. Moreover the proposing party tends to manipulate the proposed object again and leads it back to a position, which marks the “open status” and readiness for further negotiation.

In fact, in the further course of interaction, both participants fail to reach agreement about where the object should be placed. As a result, the interlocutors decide to put the object temporarily aside. Again, the object’s status is manifested in its positioning alongside the map (cf. *11) at minute 13:50.



topview

5.3 Resumption of the object

In the further course of interaction the interlocutors refer several times to the object “hotel” before they agree upon a joint solution for its placement. Thereby, we observe that the interlocutors have elaborated a semantic association between the “hotel” and the “car park”, leading to the hotel’s final placement. In 21:15 participant A suggests that the hotel should be placed near a street in order to offer parking facilities for the guests. B reacts by taking the object “hotel” from its prior location and places it together with the “car park” at a new location nearby a street.

6. DATA MINING

To analyze negotiation strategies we use information obtained by our system’s marker tracking. The mechanism returns a screen position – where the marker is found within the video image – and a position in 3D space of the camera coordinate system for every marker detected. Screen position values are returned in pixel while 3D position data were calculated with the help of camera calibration data and returned in centimeter. First, we post-processed the video data from the camera located right above the table to estimate the screen position of objects on the table. One marker was glued on the map on table and is used as a reference point. The camera is attached normal to the table so that using a 3D-pose-estimation does not add significant more information to the data. In fact using only screen positions prevents errors caused by facing the marker directly, which is a disadvantageous perspective for pose estimation. Even slightest changes in the perspective could cause a significant change of orientation and lead to false positive object movement events. Second, we use the screen position of objects from the first person view of the subjects’ HMDs, which were recorded during the experiments.

6.1 Object Movement

Object movement is calculated from the top-view’s data and converted into a time series representation for plotting and a qualitative more semantic form represented by tiers, which usually are annotated manually by researchers.

The summary of all movement of an object during the trial showed significant differences in the subjects’ object handling: The least moved object was – as expected – the compass, which is a static marker on the map itself. From the non-static markers we recorded a moved distance from about 19 cm for the miniature golf to about 621 cm for the hot air balloon. A look at the three most moved objects, which are “hot air balloon”, “hotel”, and “car park” could verify if object movement correlates with acts of negotiation (see Table 1 below). A summary of object movement within a certain time period may also identify segments where ongoing negotiating is very likely.

Table 1: Most moved objects (summary). The 5 most moved objects during the investigated trial. The distance in centimeter (third column) was approximated from the pixel-based movement within the video image (second column).

Object	Distance [px]	Distance [cm]
Hot air balloon	6524	621
Hotel	5216	496
Waterski	3490	332
Car park	2859	272
Ropes course	2690	256

6.2 Visual Attention and Joint Attention Estimation

We use the screen position data from the subjects' HMDs to indicate the change of visual attention. To prevent information overload we reduced this feature dimension to three states: *not visible*, *peripherally visible* and *in focus area*. The central focus area was initially chosen according to experience gained during a study with the HMDs used in the discussed study and an EyeLink II eye tracker [20]. However, this parameter is meant to stay changeable if further analysis requires adjustment.

The tracked marker information from the HMDs cannot be analyzed as immediate as the object movement data. Therefore, we focused on two scopes. First, the overall pattern of how visual attention to objects is spread throughout the trial. That means we just have a look at the numbers of visible objects and ignore which objects are in the field of view. Second, the individually produced visual attention data was merged to get an indication for *joint attention*.

The overview of markers within the field of view reveals interesting peaks as depicted in Figure 3. These peaks are landmarks in the ongoing conversation. Every peak is a look back to the stack of objects and indicates the end of an ongoing object negotiation process.

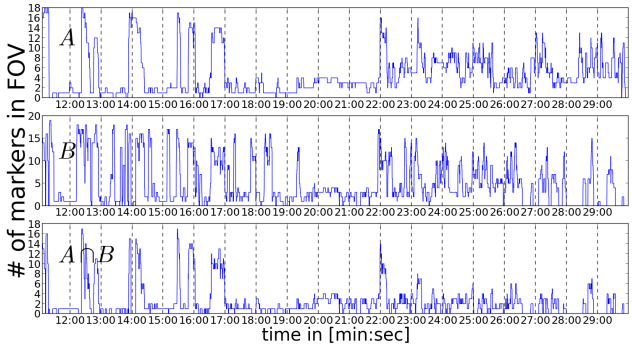


Figure 3: Objects in sight (overview). The amount of objects in the users' field of view shows repeating yet decreasing peaks. Looking at the stack of objects, which gets smaller over time, causes these peaks.

To be more precise, it may indicate the transition between two object handling processes since the first one in ended by choosing the next object to talk about. These peaks occur in 8 out of 10 investigated trials. However, not every look back at the stack needs to be a transition though.

The data-driven approaches delivered several cues and entrance points for further qualitative analysis such as object movement, object focus and visual joint attention cues.

7. FROM DATA MINING TO CONVERSATION ANALYSIS

We have seen that the negotiation of particular objects is closely related to the object's movement within our setting. With this relation in mind, object movement information visualization as seen in Figure 4 can be used to identify sequences where negotiations about certain objects are very likely.

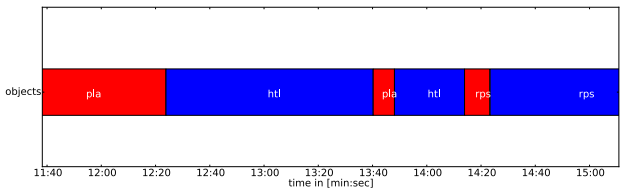


Figure 4: Most moved objects (period). Participants tend to only move the current discussed item or at least move the most interesting one the most. Using this kind of information reveals movement pattern for initial corpus structuring.

As joint attention activities occur more often and in greater detail at transition points of interactional tasks, where the interlocutors have to coordinate attention shifts from one task to another, these interims are of particular importance for our research project.

Thereby, a plot of objects in the interlocutors' field of view over time builds a fruitful analytical resource to identify borders of interactional tasks in even greater detail. Figure 5 shows that participant B is already oriented to the stack, indicated by more than 15 objects in his field of view, whereas participant A is still focused on a single object (as shown, the "playground"). A's shifting gaze to the stack is shortly delayed. The transition from the current interactional task I to the upcoming next task II is displayed by the huge amount of markers in both participants' field of view at the beginning of II. This indicates that both interlocutors are co-oriented to the stack alongside the map now in order to clarify "what's next?".

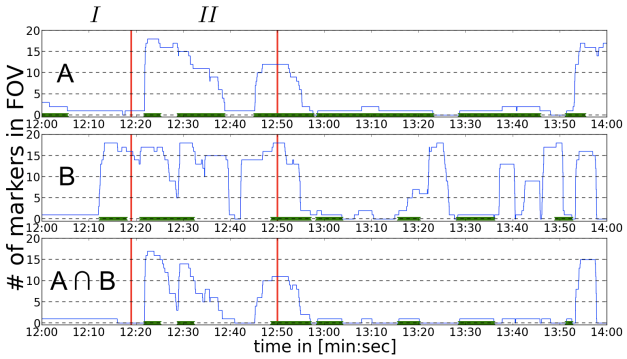


Figure 5: Participant B shifts his focus to the stack first – indicated by a huge amount of objects in sight – after he considered "playground" (phase I) to be set. During phase II, "hotel" is negotiated. The green line indicates the object's visibility to the users.

The analysis in Section 5 reveals that the first phase of negotiation ends without a joint solution of the object's placement. A quick view on the distance covered by the object "hotel" till minute 13:50 shows that the object's movement has reached 4871px (463cm), whereas the overall movement amounts 5216px (496cm) as depicted in Figure 6.

In fact, in the further course of interaction the interlocutors refer several times to the object "hotel" before they agree upon a joint solution for its placement. As the identification of all positions where the open object's status results in further object's manipulations would be a time-consuming task for a conversation analyst, the automatically generated object's movement represents a timesaving analytical resource.

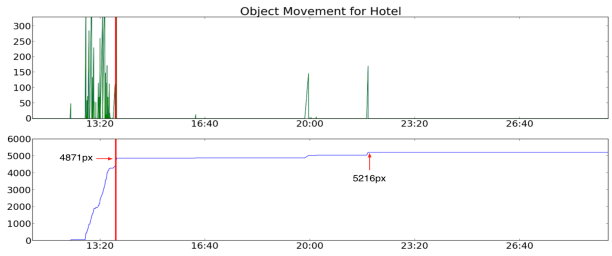


Figure 6: Object's movement for "hotel". Information about velocity (top) and movement distance can be used to determine that the placement at 13:50 was not final.

Moreover, a comparison between the movement data of different objects reveals hints about possible reasons why an object is negotiated once again. Considering Figure 7 we can recognize that the final movement of the object "hotel" takes place shortly after the interlocutors play with the object "car park" (green line, phase 1). Additionally, the hotel is finally moved together with the before handled object "car park" (phase 2). This suggests, that the hotel's final movement is related to the prior negotiation of the "car park". This helps analysts to pre-structure the corpus without analyzing the whole course of interaction.

Additionally, corpus pre-structuring often includes annotation tasks, which consists of time consuming and repetitive steps that require no or only few analytical background. For instance, for our analysis the information about objects' visibility is also relevant on a sequential level. To support this procedure, tracking information were also exported into ELAN's XML format. Due to high risk of cluttering we limit annotations of acronyms to three in a time range. If there are more objects visible in this time period we annotate the number of visible objects only (see Figure 8).

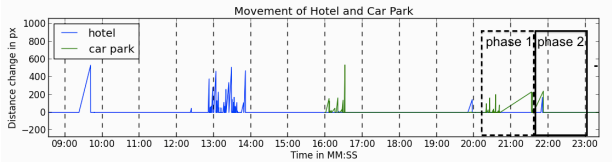


Figure 7: Object's movement for "hotel" in comparison to the "car park". Both objects are placed at the same time, which suggests correlation.

8. DISCUSSION: LINKING THE METHODS

CA studies small segments from a high-level semantic view whereas DM offers methods to interrelate large (or even the whole) corpus yet from the signal-near sub-symbolic perspective. We strive for an interrelation of both methods as seen in Figure 9.

The data-driven hypotheses about relations of object movement, joint attention and objects in sight were evaluated with CA methods. We have focused on the occurrences where these assumptions were valid even though in many cases data-driven assumptions were incorrect. However, the "search space" for interesting phenomena could be narrowed down to a small data subset, which is easier and faster to evaluate than a whole trial.

The following general experiences could be made:

(i) The quality of the annotation data achieved through data mining did not match the results of a manual annotation. ELAN annotations were useful as rough landmarks but did not fulfill the required accuracy for seamless analysis. To reduce the workload for annotators, sensor data quality as well as data (pre-) processing have to be improved.

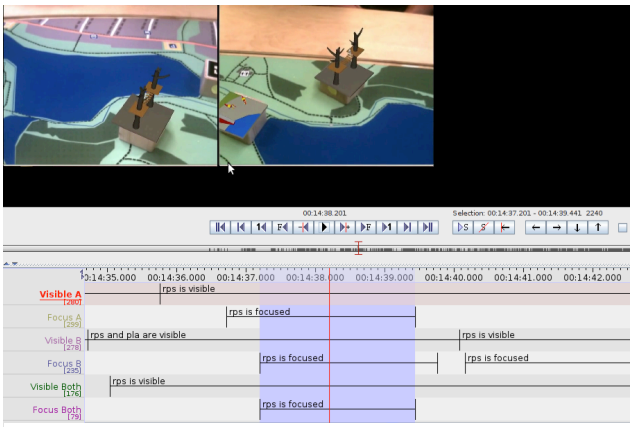


Figure 8: Generated Annotation. The automatically generated annotations were included into ELAN to support the annotator's work.

(ii) Basic information like possibly interesting object negotiations was covered. The hotel was chosen for further analysis cause of a rich set on communication relevant features. Most interesting features were not visually exclusive but multi-modal. Even though the available information channels for the data mining approach were limited to visual information only, the result was similar. Since this is the result for just one trial future analysis has to validate these initial findings.

Additionally, we could verify the subjects' negotiating strategy during CA with the help of DM. Thanks to information gained about the subjects' field of view, the different negotiation phases can be easily determined, and their duration can be measured. However, we only observed a small subset of the available data collected during studies to have a point of destination. In the future we expect better results concerning robustness and accuracy. In combination with features from depth information data streams this will be a valuable asset for negotiation strategy comparison over multiple trials. This increase in direct accessible features will also allows us to test more sophisticated methods from both research disciplines.

We plan to test the CA hypotheses gained from analysis results about object interaction pattern with the help of classifiers, which offers a significant increase in evaluation speed compared to common practices. In the current very early prototypes we could already find several similar occurrences on the movement level. These occurrences are analyzed as a DM hypothesis to validate or to find additional criteria for a more detailed classifier version, which will be tested with data from different trials. This loop will be continued until a sufficient result is achieved.

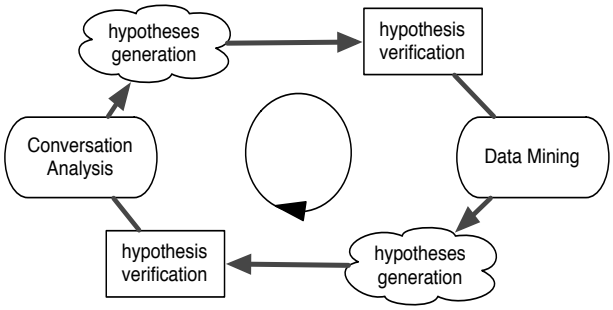


Figure 9: Both CA and DM lead to the formulation of hypotheses (on different levels). The other discipline offers complementary approaches to verify or review findings.

Other investigations show differences in negotiation strategies depending on system stability and accuracy. These strategies differ in head movement and interaction space for instance. With the help of depth image and inertial sensors we plan to investigate when and how specific strategies occur and if they can be identified by our system. Comparing these findings with data from the system might lead to new insights about when system errors lead to users adapting their behavior and when and how system malfunction is acceptable.

9. CONCLUSION

There is much potential in combined data mining and qualitative analysis approaches. It is important to stress that just initial findings were presented here and are meant to show that even with little effort much value can be attained. Improving the interrelation of CA and DM methods is a key target, which we hope to achieve in the near future. However, results show that the method mix offers interesting opportunities for hypotheses generation and validation.

We could show that pattern identified in exploratory data analysis about object manipulation and visual attention during dyadic negotiation can be used to support conversation analysis and also help to validate CA's findings. However, features such as automated annotation have a high requirement on accuracy, which needs to be fulfilled to be an asset.

Interaction methods reconstructed by CA could be used to detect certain negotiation scenarios based on object movement to find recurrent patterns. Improvement with the help of the CA/DM evaluation loop might lead to novel approaches for situation awareness, which can detect human interaction situations and maybe even predict next actions.

10. ACKNOWLEDGEMENTS

This work has partially been supported by the Collaborative Research Center (SFB) 673 Alignment in Communication and the Center of Excellence for Cognitive Interaction Technology (CITEC). Both are funded by the German Research Foundation (DFG).

11. REFERENCES

[1] Dierker, A., Bovermann, T., Hanheide, M., Hermann, T. and Sagerer, G. A Multimodal Augmented Reality System for Alignment Research. In *Proceedings of the 13th International Conference on Human-Computer Interaction* (2009), 1-5.

[2] Have, P. *Doing conversation analysis*. SAGE, 2007.

[3] Larose, D. *Discovering Knowledge in Data - Introduction to Data Mining*. JohnWiley & Sons, Inc, New Jersey, 2005.

[4] Pitsch, K., Brüning, B., Schnier, C., and Dierker, H. Linking Conversation Analysis and Motion Capturing: How to robustly track multiple participants? In *Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality* (Malta, 2010), 63-69.

[5] Nishida, T., ed. *Conversational Informatics. An Engineering Approach*. John Wiley & Sons, 2007.

[6] Zehe, S. BRIX - An Easy-to-Use Modular Sensor and Actuator Prototyping Toolkit. In *The 4th International Workshop on Sensor Networks and Ambient Intelligence* (Lugano, Switzerland 2012), 823-828.

[7] Neumann, A. *Design and Implementation of Multi-modal AR-based Interaction for Cooperative Planning Tasks*. Master's thesis, Bielefeld University (2011).

[8] Pitsch, K., Neumann, A., Schnier, C., and Hermann, T. *Augmented Reality as a Tool for Linguistic Research: Intercepting and Manipulating Multimodal Interaction*. 2013. In Press.

[9] Sidnell, J. and Stivers, T. *The handbook of conversation analysis*. Wiley-Blackwell, 2012.

[10] Witten, I., Frank, E., and Hall, M. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann Publishers, Inc., Burlington, USA, 2011.

[11] Compieta, P., Di Martino, S., Bertolotto, M., Ferruci, F. and Kechadi, T. Exploratory spatio-temporal data. *Journal of Visual Languages*, 18, 3 (June 2007), 255-279.

[12] Schegloff, E. Reflections on quantification in the study of conversation. In *Research on Language and Social Interaction* (1993), Taylor & Francis, 88-128.

[13] Pitsch, K., Vollmer, A., Rohlfing, K., Fritsch, J. and Wrede, B. Tutoring in adult-child-interaction. On the loop of action modification and the recipient's gaze. *Interaction Studies*.

[14] Heritage, J. and Robinson, J. The Structure of Patients' Presenting Concerns: Physicians' Opening Questions. *Health Communication*, 19, 2 (2006), 89-102.

[15] Magnusson, M. Discovering hidden time patterns in behavior: T-patterns and their detection. *Behavior Research Methods, Instruments, & Computers*, 32, 1 (2000), 93-100.

[16] Bakeman, R. and Adamson, L. Coordinating attention to people and objects in mother-infant and peer-infant interactions. *Child Development*, 55 (1984), 1278-1289.

[17] Kidwell, M. and Zimmerman, D. Joint attention as action. *Journal of Pragmatics*, 39 (2007), 592-611.

[18] Yarbus, A. *Eye movements and vision*. New York : Plenum, 1967.

[19] Selting, M. Gesprachsanalytisches Transkriptionssystem (GAT). *Linguistische Berichte* (1998), 91-122.

[20] Kollenberg, T. et al. Visual search in the (un) real world: how head-mounted displays affect eye movements, head movements and target detection. In *Proceedings of the Symposium on Eye-Tracking Research & Applications* (2010), ACM, 121-124.

[21] Clark, H. Pointing and Placing. In *Pointing. Where Language, Culture and Cognition Meet*. L. Erlbaum Associates, 2003.

[22] Houtkoop-Steenstra, H. *Establishing agreement: An analysis of proposal-acceptance sequences*. Mouton De Gruyter, 1987.

12. APPENDIX

`-ver`	`verbal-tier`
`-act`	`acitivites-tier`
`-gaz`	`gaze-tier`
`place(pla)@map`	`A/B places the object "playground" at the map`
`lift/grasp(h)`	`A/B lifts/grasps the object "hotel"`
`›››stack`	`movement to the stack`