Running head: Dual eyetracking

Eyetracking for two-person tasks with manipulation of a virtual world

Jean Carletta, Robin L Hill, Craig Nicol, and Tim Taylor,

University of Edinburgh

Jan Peter de Ruiter,

Max Planck Institute for Psycholinguistics, Nijmegen

Ellen Gurman Bard

University of Edinburgh

Abstract

Eyetracking facilities are typically restricted to monitoring a single person viewing static images or pre-recorded video. We describe a system which makes it possible to study visual attention in coordination with other activity during joint action. The software links two eyetracking systems in parallel and provides an on-screen task. By locating eye movements against dynamic screen regions, it permits automatic tracking of moving on-screen objects. Using existing SR technology, the system can also cross-project each participant's eye-track and mouse location onto the other's on-screen work space. Keeping a complete record of eyetrack and on-screen events in the same format as subsequent human coding, it permits analysis of multiple modalities. The software offers new approaches to spontaneous multimodal communication, joint action and joint attention. These capacities are demonstrated using an experimental paradigm for cooperative on-screen assembly of a two-dimensional model. The software is available under an open source license.

Eyetracking for two-person tasks with manipulation of a virtual world

Monitoring eye movements has become an invaluable method for psychologists studying many aspects of cognitive processing, including reading, language processing, language production, memory, and visual attention (Cherubini, Nüssli, & Dillenbourg, 2008; Duchowski, 2003; Griffin, 2004; Griffin & Oppenheimer, 2006; Meyer & Dobel, 2003; Meyer, van der Meulen, & Brooks, 2004; Rayner, 1998; Spivey & Geng, 2001; Trueswell & Tanenhaus, 2005; G. Underwood, 2005; Van Gompel, Fischer, Murray, & Hill, 2007). While recent technological advances have made eyetracking hardware increasingly robust and suitable for more active scenarios (Land, 2006, 2007), current software can register gaze only in terms of predefined, static regions of the screen. To take eyetracking to its full potential, we need to know what people are attending to as they work in a continuously changing visual context and how their gaze relates to their other actions and to the actions of others. While present limitations simplify data collection and analysis and call forth considerable ingenuity on the part of experimenters, they invite us to underestimate the real complexity of fluid situations in which people actually observe, decide, and act. At present, we are only beginning to understand how people handle multiple sources of external information, or multiple communication modalities. Despite growing interest in the interactive processes involved in human dialogue (Pickering & Garrod, 2004), the interaction between language and visual perception (Henderson & Ferreira, 2004), how visual attention is directed by participants in collaborative tasks (Bangerter, 2004; Clark, 2003) and the use of eye movements to investigate problem solving (Charness, Reingold, Pomplun, & Stampe, 2001; Grant & Spivey, 2003; J. Underwood, 2005), the generation of suitably rich, multimodal datasets has hitherto been difficult.

There is certainly a need for research of such breadth. Complex multi-modal signals are available in face-to-face dialogue (Clark & Krych, 2004), but we do not know how often interlocutors actually take up such signals and exploit them on-line. Examining each modality separately will give us an indication of its potential utility but not necessarily of its actual utility in context. Single modality studies may leave us with the impression that, for joint action in dialogue or shared physical tasks, all instances of all sources of information influence all players. In some cases, we tend to under-estimate the cost of processing a signal. For example, we know that some indication of the direction of an interlocutor's gaze is important to the creation of virtual copresence, and has many potential uses (Cherubini et al., 2008; Kraut, Gergle, & Fussell, 2002; Monk & Gale, 2002; Velichkovsky, 1995; Vertegaal & Ding, 2002). Controlled studies with very simple stimuli, however, show that processing the gaze of another is not a straightforward bottom-up process. Instead it interacts with what the viewer supposes the gazer might be looking at (Lobmaier, Fischer, & Schwaninger, 2006). In genuine situated interaction, there are many sources of such expectations and all require some processing on the part of an interlocutor. General studies of reasoning and decision making (Gigerenzer, Todd, & Group, 1999) suggest that people have astute strategies for circumventing processing bottlenecks in the presence of superabundant information. It would be surprising if they did not streamline their interactions in the same way. To know how, we need to record individuals dividing their attention between the fluid actions of others, their own attempts at communication, and the equally fluid results of a joint activity.

Given two eyetrackers and two screens, four breakthroughs are required before the technology can be usefully adapted to study cooperation and attention in joint tasks. First, one central virtual world or game must drive both participants' screens, so that both can see and

manipulate objects in the same world. Second, it must be possible to record whether participants are looking at objects that are moving across the screen. Third, it must be possible for each participant to see indications of the other's intentions, as they might in real face-to-face behavior. For on-screen activities, those intentions would be indicated by icons representing their partner's gaze and mouse location. Finally, to give a full account of the interactions between players, the eyetracking, speech, and physical actions of the two participants must be recorded synchronously. Finding solutions to these problems would open up eyetracking methodology not just to studies of joint action but to studies of related parts of the human repertoire, for example, joint viewing without action, competitive rather than collaborative action or learning what to attend to in acquisition of a joint skill. In this paper, we describe solutions to these problems using the SR Research Eyelink II platform. The resulting software is available under an open source license from http://wcms.inf.ed.ac.uk/jast. We demonstrate the benefits of the software within an experimental paradigm in which two participants jointly construct a figure to match a given model.

Review

Commercial and proprietary eyetracking software reports where a participant is looking, but only in absolute terms or using pre-defined static screen regions that do not change over the course of a trial. The generic software for the Eyelink II tracker supplied with the system (SR-Research, n.d.) provides a good basis for going forward. It can be used to generate raw eye positions, and will calculate blinks, fixations and saccades, associating fixations with static, user-defined regions of the screen. It does not support the definition of dynamic screen regions which would be required to determine whether the participant is looking at an object in motion. Nor

does it support the analysis of eye data in conjunction with alternative data streams such as language, video and audio data, or eye stream data from two separate machines. It will, however, take messages generated by the experimental software and add them to the rest of the synchronized data output, and it can pass messages to other Eyelink trackers on the same network. We use this facility to implement communication between our eyetrackers. SR Research, the makers of the Eyelink II, provided us with sample code that displays the eye cursor from one machine on the display of another by encoding it as a string on the first machine, sending the string as a message to the second machine, parsing the string back into an eye position on the second machine, and using that information to change the display (Brennan, Chen, Dickinson, Neider, & Zelinsky, 2008).

Despite the lack of commercial support for dual eyetracking and analysis against other ways of recording experimental data, some laboratories are beginning to implement their own software. In one advance, a series of eye movements (scanpaths) produced by experts are later used to guide the attention of novices (Stein & Brennan, 2004). In another, two eyetrackers are used in parallel with static displays but without cross-communication between one person's tracker and the other's screen (Hadelich & Crocker, 2006; Richardson, Dale, & Kirkham, 2007; Richardson, Dale, & Tomlinson, in press). In a third, one participant's scanpath is simulated by using an automatic process to display a moving icon while genuinely eyetracking the other participant as he or she views the display (Bard, Anderson et al., 2007). In a fourth, Brennan et al. (2008) projected genuine eye position from one machine onto the screen of another while participants shared a visual search task over a static display. Steptoe et al. (2008) and Murray and Roberts (2006) keyed the gaze of each avatar in an immersive virtual environment to actual

tracked gaze of the participant represented by the avatar, but without a shared dynamic visual task.

Data capture

Dual eyetracking could be implemented using any pair of head-free or head-mounted eyetrackers as long as they can pass messages to each other. Our own implementation uses two head-mounted Eyelink II eyetrackers. Eyetracking studies start with a procedure to calibrate the equipment that determines the correspondence between tracker readings and screen positions, and usually incorporate periodic briefer recalibrations to correct for any small drift in the readings. The Eyelink II outputs data in its proprietary binary format, "EDF", which can either be analyzed with software supplied by the company or converted to a time-stamped ASCII format that contains one line per event. The output normally contains a 500 Hz data stream of eye locations with additional information about the calibration and drift correction results used in determining those locations, plus an online parsed representation of the eye movements in terms of blinks, fixations, and saccades. In addition to this data originating from the eyetracker itself, the eyetracker output will contain any messages that have been passed to the eyetracker from the experimental software, stamped with the time they were received. An Eyelink II natively uses two computers connected via a local network. The host machine drives the eyetracking hardware by running the data capture and calibration routines, and the display machine runs the experimental software. Our installation uses two Pentium 4 3.0 GHz display machines running Windows XP with 1 GB DDR RAM, a 128 MB Graphics card, a 21" CRT monitor, a Gigabit Ethernet card, and a Soundblaster Audigy 2 sound card, and two Pentium 4 2.8GHz host machines running Windows XP and ROMDOS7.1 for the Eyelink II control software, with 512 MB DDR RAM and a Gigabit Ethernet card.

INSERT FIGURE 1 AROUND HERE

Our arrangement for dual eyetracking is shown in Figure 1. Here, because there are two systems, four computers are networked together. In addition to running the experimental software, the display machines perform audio and screen capture using Camtasia (Tech-Smith, n.d.) and close-talking headset microphones. Audio capture is needed if we are to analyze participant speech; screen capture provides verification and some insurance against failure in the rest of the data recording by at least yielding a data version that can easily be inspected, although the resulting videos have no role in our current data analysis.

As usual, the display machines are networked to their respective hosts so that they can control data recording and insert messages into the output data stream, but in addition, the display machines pass messages to each other. These are used to keep the displays synchronized. For instance, if participant A moves his eyes, his mouse, or some on-screen object, the experimental software running on A's display machine will send a message to that effect to the experimental software on participant B's display machine, which will then update the graphics to show the A's gaze cursor, A's mouse cursor, and the shared object in their new locations. Coordinating the game state is simply a matter of passing sufficient messages. There is, of course, a lag between the time when a message is generated on one participant's machine and the time when it is registered on the other's. In our testing, 99% of the lags recorded were less than 20 ms. In pilot experiments, debriefed participants did not notice any lag. Even with the small additional time needed to act on a message to change the display, this degree of lag is tolerable for studies of joint activity. During any trial, the experimental software should loop through checking for local changes in the game and passing messages until both of the participants have signaled that the trial has finished. In our experimental software, each loop

takes around 42 ms. Since eye positions move faster than other screen objects, we pass them twice during each loop. Whenever a display machine originates or receives a message, it sends a copy of that message to its host machine for insertion into its output data stream. As a result, the output of both eyetracking systems contains a complete record of the experiment, although the eye positions are sampled more coarsely on the opposing than on the native machine.

Example

The Experimental Paradigm

In our experimental paradigm, two participants play a series of construction games. Their task is to reproduce a static two-dimensional model by selecting the correct parts from an adequate set and joining them correctly. Either participant can select and move (left mouse button) or rotate (right mouse button) any part not being grasped by the other player. Two parts join together permanently if brought into contact while each is 'grasped' by a different player. Parts break if both participants select them at the same time, if they are moved out of the model construction area, or if they come into contact with an unselected part. Any of these 'errors' may be committed deliberately to break an inadequate construction. New parts can be drawn from templates as required.

These rules are contrived to elicit cooperative behavior: no individual can complete the task without the other's help. The rules can easily be reconfigured, however, or alternative sets of rules can be implemented. Figure 2a shows an annotated version of the initial participant display from our implementation of the paradigm with labels for the cursors and static screen regions. Figure 2b shows a later stage in a trial where a different model (top right) is built. Initial parts are always just below the model. A broken part counter appears at the top left, a timer in the

middle at the top, and new part templates across the bottom of the screen. After each trial, the participants can be shown a score reflecting accuracy of their figure against the given model.

INSERT FIGURE 2 AROUND HERE

This experimental paradigm is designed to provide a joint collaborative task. The general framework admits a number of independent variables, such as the complexity of the model, the number of primary parts supplied, the difficulties presented by the packing of the initial parts, and whether or not the participants have access to each other's speech, gaze cursor, or mouse position. These variables can all be altered independently from trial to trial. The paradigm is suitable for a range of research interests in the area of joint activity. Performance can be automatically measured in terms of time taken, breakages, and accuracy of the constructed figure against the target model (maximum percentage of pixels that are colored correctly when the constructed figure is laid over the model and rotated). In addition, the paradigm allows for an analogous task for individuals that serves a useful control. In the individual "one-player" version, two parts join together when they touch, even if only one is selected.

The experimental software (JCT)

The JAST Joint Construction Task, or JCT, is software for running experiments that fit this experimental paradigm. It is implemented under Windows XP in Microsoft Visual C++.Net and draws on a number of open source software libraries. These include Simple DirectMedia Layer (SDL) support for computer graphics (Simple-DirectMedia-Layer-Project, n.d.); add-ons available for download along with SDL and from the SGE project (Lindström, 1999) that provide further support for things like audio, rotation, collision detection, and text; Apache's XML parser, Xerces (Apache-XML-Project, 1999) and the Simple Sockets Library for network

communication (Campbell Jr. & McRoberts, 2005) which we use instead of the Eyelink II networking support specifically so that the software can also be run without eyetracking. When the experimental software is run without an eyetracker, rather than passing messages to a host machine, the display machine stores the time-stamped messages in an ASCII file that follows the same structure that the eyetracker itself produces, so that the file can be used in the same onward processing. Even without the richness of eye-movement data, collaborative behavior can therefore still be investigated with this software using only two standard networked PCs.

The configuration of an experiment is defined by a set of Extensible Markup Language (XML) files. Each experiment consists of a number of trials using one model per trial. Models are built of polygon parts. Curved surfaces are simulated by a polygon with a high number of vertices. Our software includes a utility to generate these curved parts. The initial and target configuration for each model, and descriptions of the polygon parts used, is stored in a "Stimulus Set" XML file, and the experiment configuration links to a number of these files which are presented in order. For each trial, the experiment configuration file specifies the stimulus set, whether to show the clock, whether this clock counts up or down, the time limit, whether the position of the partner's eye and mouse should be visible, and any text or graphics to show between trials. It also specifies the machines the experiment will run on; whether the experiment is to be run for individuals or pairs of participants; the size and location of the static screen regions; and a number of experimenter-composed text messages to display at certain points in the experiment. When run, the software performs the eyetracker calibration, and then presents the trials in the order specified, performing drift correction between trials.

In this implementation of the experimental paradigm, the messages passed between the eyetrackers, and therefore stored in the output, are:

- Markers showing the time when each trial starts and ends, along with the performance scores for the number of breakages and the accuracy of the final construction;
- Markers near the beginnings of trials showing when a special audio tone and graphical symbol were displayed, to make it easier to synchronize the data against audio and video recordings;
- Participant eye positions;
- Sufficient information to reconstruct the model-building state, including joins, breakages, part creation, and any changes in the locations or rotations of individual parts.

Our video reconstruction utility takes the ASCII data from one of the eyetrackers and uses it to create a video that shows the task with the eye and mouse positions of the two participants superimposed. The videos produced from the two ASCII data files for the same experiment are the same apart from subtle differences in timing: each video shows the data at the time that it arrived at, or was produced by, the individual eyetracker used as its source. The utility, which is again written in Microsoft Visual C++.Net, uses FFMPEG (Anonymous, n.d.) and libraries from the JCT experimental software to produce MPEG2 format videos with a choice of screen resolutions and color depths.

Interpreting Participant Behavior

Capturing the data with the methods described creates a complete and faithful record of what happened during the experiment, but using primitives below the level of abstraction required for analysis. For instance, the messages describe the absolute positions of eyes, mice, and all parts throughout the task, but not in terms of when a participant is looking at a particular part, even though knowing this is essential to understanding how gaze, speech, and action relate.

Therefore, the next step is to add this higher level of analysis to the data set. To do this, we transfer the data out of the ASCII format exported by the Eyelink II and into an XML format. XML format allows us to use existing parsers and makes it easy to check that the data conform to our expectations by validating data files against a document-type definition specifying what sorts of events and objects they should contain.

The resulting XML file contains most of the information gathered so far in the experiment. Drawing data from the output of both eyetrackers and from the experiment configuration file, it includes all of the messages that were passed, parsed eye data for both participants, plus a list of the parts, their movements and their join events, with part properties, such as shape and initial locations, taken from the experiment configuration file. Although it would not be difficult to include the full 500 Hz eye position data in the XML representation, it is now omitted for several reasons. First, this level of detail is better suited for the kind of parsing provided by the eyetracker software itself. Second, including it would increase file size substantially. Third, it is unnecessary for our intended analysis.

In addition to re-encoding these kinds of data from the experimental software, the utility that produces the XML file adds a number of data interpretations:

- *Look events* during which a participant looked at a part, composite, or static region (typically referred to as regions of interest or ROIs). Where gaze is on the a movable object (a dynamic region of interest or DROI), these events cover the entire time that the eye position is within a configurable distance of the moving screen region associated with whatever is being looked at, whether the eye is currently engaged in a fixation, smooth pursuit or a saccade. Since the XML file contains parsed eye movements, class of eye activity can be established by later analysis, if required.

- *Hover events* during which a participant located the mouse over a part or composite without 'grasping' or moving the part by button press. Hover events also cover the entire time that the mouse cursor is within a configurable distance of the moving screen region associated with the part.
- *Construction history* for each trial, a description of how the pair constructed the figure, given as a set of binary trees where the leaves are the parts, each with a unique identifier, and each node uniquely identifies a composite created by joining any other composites or parts that are the children of the node.
- *Composition phases*, for each creation of a composite, a division of the time since the previous composite was created into phases, from the first move of a constituent part to the final join. While the final phase covers the final action of docking the last piece of the composite, two earlier phases simply divide the rest of the composition time in half. Like the construction history, this representation can be used to study construction strategy or to find sub-tasks of particular difficulty.
- *Per-trial breakage scores* calculated over the XML format for use as a parity check against those coming directly from the experimental software.

The JastAnalyzer software that performs this processing requires the same environment as the JCT experimental software. It works by passing the ASCII data back through libraries from the JCT software in order to interpret the task state. Because we exploit the resulting XML data for a number of different purposes, we call this our "General Data Format", or GDF. It is straightforward, given the data format, to produce scripts that show useful analyses, such as the lag between when one participant looks at an object and when the other participant follows. It is

also easy to produce per-trial summary statistics, such as how many times a participant looked at each of the parts and what percentage of the time they spent looking at the clock.

Export to Data Display and Analysis Packages

Thus far, we have seen how the automatically collected JCT data are represented. For more complicated data analysis, we export from our General Data Format into the format for existing data display and analysis packages. These allow the data to be inspected graphically, "playing" it synchronized to the video and audio records. Such packages are useful both for checking that the data are as expected and for browsing with an eye to understanding participant behavior. But real strength of such packages lies in the ability to create or import other sources of information for the same data, such as orthographic transcription, linguistic annotations relating to discourse phenomena, and video coding. Some packages also provide data search facilities that can go beyond the simple analyses generated with our GDF-based scripts, for example, investigations of the relationship between task behavior and speech. Our software includes export to two such packages: ELAN (MPI, n.d.) and the NITE XML Toolkit, or NXT (Language Technology Group, n.d.). Both can display behavioral data in synchrony with one or more corresponding audio and video signals, and both support additional data annotation. ELAN has strengths in displaying the time course of annotations that can be represented as tiers of mutually exclusive, timestamped codes, while NXT has strengths in supporting the creation and search of annotations that relate to each other both temporally and structurally, as is usual for linguistic annotations built on top of orthographic transcription.

Both ELAN and NXT use XML in their data formats. Because our utilities for export to these programs are based on XSLT stylesheets (World Wide Web Consortium, 1999), the

standard technique for transducing XML data from one format to another, they will run on any platform as long as a stylesheet processor is installed.

Example experiment and analyses afforded

Method

Task. We can illustrate the utility of the Joint Construction Task systems via a designed corpus of time-aligned multimodal data (eye movements, actions, and speech) produced while pairs of individuals assembled a series of tangrams collaboratively. Produced as part of the Joint-Action Science and Technology (JAST) project (http://www.euprojects-jast.net/), this corpus was used to explore factors which might benefit human-robot interactions by studying human-human interactions in collaborative practical tasks. In this example, we devised 16 target tangrams, none of which resembled any obvious nameable entity. To engineer referring expressions, we designed each part to represent a unique color-shape combination, with each color represented only once and each shape at most twice. The same initial set of seven pieces was available at the beginning of each construction trial. All had to be used.

Because trials in our paradigm can take over four minutes to complete, drift correction is important. Mid-trial interruptions for calibration are undesirable for collaborative problem solving. Instead, the software package offers a manual correction utility for use on the reconstructed videos which enables optional off-line adjustments to be made.

Design. In order to investigate the relative worth of speech and gaze feedback in joint action, communication modalities were varied factorially: participants could speak to each other and see the other person's eye position; participants could speak but could not see where the other person was looking; participants could not speak but could see where their collaborator was

looking; participants could neither speak nor see the gaze location of their partner. Condition order was rotated between dyads, but each pair of participants built four models under each condition. Additionally, as leadership and dominance factors can modulate how people interact, half the dyads were assigned specific roles: one person was the Project Manager while the other was the Project Assistant. The other half were merely instructed to collaborate. All were asked to reproduce the model tangram as quickly and as accurately as possible while minimizing breakages. To determine the usefulness of a verbal channel of communication during early strategy development, half of the dyads encountered the speaking conditions in the first half of the experiment, and half in the second.

Results.

In the following sections, we indicate how such a corpus may be exploited. In the first two cases, we illustrate types of automatically produced data that could be used to explore some current questions. In the third section, we show how these automatically recorded events can be combined with human coding to provide multi-modal analyses. We cite other work which further exploits the rich data of similar corpora.

Fine-grained action structures. Often resources are provided for joint actors on the assumption that whatever is available to an actor is used. Sometimes (Bard et al, 2007) this proves not to be the case. The data recorded for this experiment permit very fine-grained analyses of action/gaze sequences which would allow us to discover without further coding who consults what information at critical phases of their actions. Figure 2b shows participant B's screen twenty seconds into an experimental trial. In this figure, participant A is has grasped the triangle on the right, and B has grasped the triangle on the left. Two seconds after the point captured in Figure 2b, the game software records that the players successfully joined the two

parts together. Because our software captures continuous eyetracks, and because it records the precise time when parts are joined, we can use the construction and eyetrack records to discover that in the 10s preceding the join, participant A's gaze, which is shown here as a small circular cursor, moved rapidly between the two moving triangles, and briefly rested in the workspace position where the pieces were ultimately joined. Participant B's gaze, which is not displayed on this snapshot of B's screen but would appear on the reconstructed video, also traveled between the two triangles, but with two separate excursions to fixate the target model at the top right corner of the screen. Thus, the players are consulting different aspects of the available information while achieving a common goal. This particular goal – combination of two triangles to construct a larger shape – was successfully achieved. Because all partial constructions have unique identifiers composed of their ancestry in terms of smaller constructions or primary parts, it is possible to distinguish successful acts of construction, which are not followed by breakage and a restart, from those which are quickly broken. It would then be possible to discover whether, as in this example, successful constructions were characterized by differentiation of visual labor in their final delicate phases, while unsuccessful constructions were not. It would also be possible to discover whether the distribution of critical phase visual labor in each dyad could predict overall performance (automatically recorded per trial durations, breakage counts, and accuracy scores). Since the composition history automatically divides the final half of each interval between construction points from two earlier periods, we might also look for differentiation or alignment of gaze during those earlier phases where players must develop a strategy for achieving the next sub-goal.

Trial-wise measures: actions entraining gaze. Our software can also summarize events during a given trial. We illustrate how such figures might be used to determine whether gaze is

largely entrained by the ongoing physical actions. If so, there may be little opportunity for role or strategy to determine attention here. We have several measures to use.

In the following examples, figures for participant A are followed by those for B in parentheses. In the trial seen in Figure 2b, 71% (72%) of the time is spent looking at the model, and 27% (20%), at the construction area, with the rest spent looking in other areas or off-screen. The player's position overlaps with a tangram part (moving or stationary) in the construction area 24% (11%) of the time, divided over 213 (173) different occasions, of which 56 (41) involve stable fixations. There are 29 occasions (of the 213 (173) gaze-part overlaps) when both participants' eyetracks overlap with the same object, totaling 21.5s across the entire trial, but only 14 occasions when one player's gaze overlaps with an object that the other player is currently manipulating. Thus, we have prima facie evidence that for this dialogue the common task on yoked screens does not draw the majority of players' visual activity to the same objects at the same time. Nor does an object that one player moves automatically entrain the other player's gaze. Making individual tests on pilot trials for a new type of joint task would allow the experimenter to determine whether the task had the necessary gaze-drawing properties for the experimental purpose.

In capturing this information about dyadic interaction, our experimental setup reveals a level of detail that is unprecedented in previous studies. The data from our eyetracker software suffice for studies of how the experimental variables interact and influence measures of task success such as speed and accuracy, as well as enabling the analyst, for instance, to calculate the lag between when one person looks at a part and when the partner does, or to determine who initiates actions, as would be required for measuring dominance.

Hand coding. While the system automates some tasks, it is clear that it cannot meet every experimental purpose. Export to NITE or ELAN allow further coding to suit the man experimental goals. As a simple demonstration, we follow the data of this study through transcription and reference coding. We use ChannelTrans (ICSI, n.d.) to create orthographic transcription where the conversational turns are time-stamped against the rest of the data. By importing the transcription into the NITE XML Toolkit, we can identify and code the referring expressions used during the task. Figure A1 in the Appendix illustrates the use of an NXT coding tool to code each referring expression with the system's identifier for its referent. The result is a time-aligned multimodal account of gaze, reference, action, and trial conditions.

This allows us, for instance, to count the number of linguistic expressions referring to each part and relate these references to the gaze and mouse data. As an example of the sorts of measures this system makes it easy to calculate, over the eight speech condition trials for one of the participant dyads, there were 267 instances of speech and 1262 of "looking" at DROIs which lasted over 45 ms duration. On 78 occasions, a player looked at a part while uttering a speech segment containing a reference to it. On 95 occasions, one player looked at a part while the other was uttering a speech segment referring to it. Figure 3 shows the complete data flow used to obtain this analysis.

Using data like these, we have shown how mouse movements coincide particular distributions of referring expressions (Foster et al., 2008) and how the roles assigned to the dyad influence audience design in mouse gestures (Bard, Hill, & Foster, 2008).

INSERT FIGURE 3 AROUND HERE

Empirical investigation: Determining whether dialogue or gaze projection during joint action leads to greater visual alignment.

If dialogue encourages alignment all the way from linguistic form to intellectual content, as Pickering and Garrod (2004) propose, and if there is a shared visual environment in which to establish common ground (Clark, 1996; Clark & Brennan, 1991; Clark, Schreuder, & Buttrick, 1983; Lockridge & Brennan, 2002), then the opportunity for two people to speak should assure that they look at their common visual environment in a more coordinated way. Similarly, having a directly projected visual cue that indicates where another person is looking (perhaps analogous to using a laser pointer) offers an obvious focus for visual alignment. Thus, we would expect conditions allowing players to speak as they construct tangrams to have more aligned attention than those where they play silently. And we would expect conditions where each one's gaze is cross-projected on the other's screen to have more aligned attention than those with no gaze cursors. Using the study described above, we can test these hypotheses via analyses of viewing behavior, visual alignment and joint action which were not previously possible.

First, we examine a new dependent variable that can be automatically extracted from the data: the time spent looking at the tangram parts and the partially built tangrams. These are particularly complex Dynamic Regions of Interest (DROIs), because either player is free to move, rotate or even destroy any of the parts at any time. In this paradigm, however, the parts are the tools and the tangrams are the goals of the game. They must be handled and guided with care. In this example we began by examining the time participants spent looking at any of these ROIs, dynamic or otherwise, as a percentage of the total time spent performing the task using a 2 (speech, no speech) x 2 (gaze visible, invisible) x 2 (dyad member: A, B) Analysis of Variance with dyads as cases. Since we know that in monologue conditions, what people hear someone say will influence how they view a static display (Altmann & Kamide, 1999, 2004, 2007), we would expect the speaking conditions to direct players' gaze to the parts under discussion quite

efficiently. In fact, the proportion of time spent inspecting any of the available ROIs was significantly lower when participants could speak to each other (17.41%) than when they were unable to speak vs. (21.78%) [F (1,31) = 53.49, $p < 0.001$]. We would also expect the direction of one player's gaze to help focus the other's attention on objects in play. We know that gaze projection provides such help for static ROIs (Stein & Brennan, 2004). In fact, there was no discernable difference in players' DROI viewing time as a consequence of being able to see their partner's eye position [$F(1,31) < 1$]: with and without cross-projected gaze, players tracked DROIs about 20% of the time overall. So players were not looking at the working parts as much when they were speaking; and seeing which piece their partner was looking at did not alter the proportion of time spent looking at task-critical objects.

Second, we exploit a measure that is contingent on being able to track both sets of eye movements against potentially moving or shifted targets: how often both players are looking at the same thing at exactly the same time. Here, a 2 (Speech) x 2 (Gaze) ANOVA indicates that the ability to speak in fact reduced the number of instances of aligned gaze by nearly 24% [39.1 vs. 29.7: $F(1,31) = 15.76$, $p < 0.001$]. Reduction in frequency of visual alignment is not an artefact of trial duration. In fact, trials involving speech were an average of seven seconds longer than those without speech [95.99s vs. 89.02s; $F(1,31) = 5.11$, $p < 0.05$] and might be expected to increase the number of eye-eye overlaps, even if just by chance. Again, however, the ability to see precisely where the other person was looking had no influence [$F(1,31) < 1$] and there was no interaction.

The analyses which reveal overlapping gaze can also generate the latency between the point when one player begins to look at a ROI and the point where the other's gaze arrives, the eye-eye lag (Hadelich & Crocker, 2006). Eye-eye lag was shorter when participants could speak

to each other [197ms with speech compared to 230ms without speech; $F(1,31) = 6.44$, $p < 0.05$]. Perhaps surprisingly, the projection of a collaborator's gaze position onto the screen failed to make any difference to these lag times [212ms when gaze position was visible vs. 214ms when it was not; $F(1,31) < 1$] and again there was no interaction between the variables.

Taken together, these results indicate that in a joint collaborative construction task the facility for members of a working dyad to speak to each other reduces the proportion of time spent looking at the construction pieces and reduces the number of times both partners look at the same thing at the same time. Visual alignment appears to be inhibited when a dyad could engage in speech, but when it did occur there was a shorter delay in coordination as the lag between one person looking at a potential target and their partner then moving their eyes onto the same target was smaller. In contrast, a direct indicator of a collaborator's gaze position did not appear to have any effect (facilitatory or inhibitory) on the measures examined. Contrary to expectations, therefore, being able to discuss a yoked visual workspace does not automatically yoke gaze. And there is no evidence that one person will track what another person is looking at more often when their eye position is explicitly highlighted. The combined influence and shape of any interaction between available modalities is obviously of critical importance to the understanding and modeling of multi-modal communication; and as our example demonstrates, the paradigm expounded here offers an ideal method of enhancing research into this topic.

Finally, gaze coordination can be further examined using cross-recurrence analysis. This technique has been used in the investigation of language and visual attention by Richardson and colleagues (Richardson & Dale, 2005; Richardson et al., 2007; Richardson et al., in press) to demonstrate that visual coordination can manifest as being in phase or non-random without necessarily being simultaneous. However Richardson et al. only used shared static images or

replayed a monologue to participants. Bard and colleagues (Bard, Hill, & Arai, 2009; Bard et al., 2008; Bard, Hill, Nicol, & Carletta, 2007) have taken this technique further to examine the time-course of visual alignment during the Joint Construction Task.

Discussion

Our software implements a new method for collecting eyetracking data from a pair of interacting participants, including showing the same game on the screens of the eyetrackers with indicators of the partner's gaze and mouse positions. The method allows us to construct a record of participant interaction that gives us the fine-scale timing of gaze, mouse, and task behaviors relative to each other and to moving objects on the screen.

We have demonstrated our techniques using a pair of Eyelink II eyetrackers and experimental software that implements a paradigm in which two participants jointly construct a model. Although our experimental paradigm is designed for studying a particular kind of joint action, the basic methods demonstrated in the software can be used to advance work in other areas. The software shows how to use one task to drive two eyetracker screens, record gaze against moving screen objects, show a partner's gaze and mouse icons, and synchronize the eyetracking, speech, and game records of two participants. Inventive message passing is the key, since messages can both be passed between the eyetrackers and stored in the output record, providing all of the data communication required. Tracking the relationship between gaze and moving objects, for instance, is simply a case of having the experimental software note the movement as messages in the eyetracker output, so that it can be checked against the gaze track analytically later on. Audio and screen capture can be synchronized to the eyetracker output by having the experimental software provide audible and visual synchronization marks, noting the time of these marks relative to the eyetracker output, again as messages. Coordinating two

screens requires each copy of the experimental software to record its state in outgoing messages and respond to any changes reported to it in incoming ones. In theory, our techniques will work with any pair of eyetrackers that can handle message passing, including pairs of eyetrackers from different manufacturers. The message passing itself could become a bottleneck in some experimental setups, especially if it was used to study groups, but we found it adequate for our purpose. Manufacturers could best support the methodological advances we describe by ensuring that they include message passing functionality that is both efficient and well-documented.

These advances are useful not just for our work, but also for work in other areas where more limited eyetracking methods are currently in use. Work on visual attention could benefit from the ability to register dynamic screen regions. Multiple object tracking (Bard et al., 2009; Bard, Hill et al., 2007; Pylyshyn, 2006; Pylyshyn & Annan, 2006; Wolfe, Place, & Horowitz, 2007), subitization (Alston & Humphreys, 2004), and feature binding (Brockmole & Franconeri, 2009; Luck & Beach, 1998) for instance, would be less constrained if the objects were treated in the way we suggest.

It is joint action that has the most to gain, however, because it requires all of the advances made. The ability to import the data that comes out of the experimental paradigm into tools used for multimodal research, such as ELAN and NXT, offers particular promise for the study of joint action, especially where language is involved. These tools will enable the basic data to be browsed, combined with orthographic transcription, enriched with hand annotation that interprets the participant behaviors, and searched effectively. Using these techniques will allow a fuller analysis of what the participants are doing, saying, and looking at than any of the current techniques afford. The effect of copresence on joint action (Fussell & Kraut, 2004; Fussell et al.,

2004; Horton & Keysar, 1996; Kraut, Fussell, & Siegel, 2003; Kraut et al., 2002) can also easily be manipulated by altering the proximity of the two eyetrackers: the limit only depending on the network connections. Similarly, the effect of perspective on language use (Hanna & Tanenhaus, 2004), including the use of deictic expressions such as "this" or "that", can be investigated, as well as the differences in eye movements during comprehension (Spivey, Tanenhaus, Eberhard, & Sedivy, 2002) vs. production (Horton & Keysar, 1996) and whether speakers engage in "audience design" or instead operate along more ego-centric lines (Bard, Anderson et al., 2007; Bell, 1984). Functional roles, such as instruction giver or follower, can be assigned to members of a dyad to determine how this might influence gaze behavior and the formation of referring expressions (Engelhardt, Bailey, & Ferreira, 2006).

The experimental environment opens up many possibilities for eye-tracking dynamic images and active scene viewing, topics which are still surprisingly under-researched. In particular, the smooth pursuit of objects (Barnes, 2008; Burke & Barnes, 2006) controlled either by the person themselves or by their partner in a purposeful, non-random fashion can be investigated. Data can be output in its raw, sample-by-sample form, permitting customized smooth pursuit detection algorithms to be implemented; as total gaze durations (accumulated time that the eye position coincided with an onscreen object irrespective of eye stability); or as a series of fixations automatically identified by the SR-Research software. As most paradigms only involve static images, eye movements are almost exclusively reported in terms of saccades or fixations, ignoring the classification of pursuit movements. The EyeLink II also offers either monocular or binocular output, both of which can be utilized by our software.

The eyes gather information required for motor actions and are therefore proactively engaged in the process of anticipating actions and predicting behavior (Land & Furneaux, 1997).

These skills are developed surprisingly early, with goal-directed eye movements being interpretable by the age of one (Falck-Ytter, Gredeback, & von Hofsten, 2006). More recently, Gesierich, Bruzzo, Ottoboni, & Finos (2008) studied gaze behavior and the use of anticipatory eye movements during a computerized block-stacking task. The phenomenon of the eyes pre-empting upcoming words in the spoken language stream is, as we have noted, the basis of the well-established "visual world" paradigm in psycholinguistics (Altmann & Kamide, 1999, 2004, 2007).

In summary, we have successfully developed a combined, versatile hardware and software architecture that enables ground-breaking and in-depth analysis of spontaneous, multimodal communication during joint action. Its flexibility and naturalistic game-based format help narrow the distinction between field and laboratory research while combining vision, language and action. The system allows for more or less visual information recording (monocular, binocular data or none at all) as well as for varying symmetrical and asymmetrical cross-projections of action or attention. Although optimized for joint action it can operate in a stand-alone solo mode. This system therefore offers a new means of investigating a broad range of topics, including vision and active viewing, language and dialogue, joint action, copresence, strategy development and problem solving, and hand-eye coordination.

.

References

Alston, L., & Humphreys, G. W. (2004). Subitization and attentional engagement by transient stimuli. *Spatial Vision, 17*(1-2), 17-50. doi:10.1163/15685680432277825

Altmann, G. T. M., & Kamide, Y. (1999). Incremental interpretation at verbs: Restricting the domain of subsequent reference. *Cognition, 73*, 247-264. doi:10.1016/S0010-0277(99)00059-1

Altmann, G. T. M., & Kamide, Y. (2004). Now you see it, now you don't: mediating the mapping between language and the visual world. In J. Henderson & F. Ferreira (Eds.), *The integration of language, vision and action.* (pp. 347–386): Psychology Press.

Altmann, G. T. M., & Kamide, Y. (2007). The real-time mediation of visual attention by language and world knowledge: Linking anticipatory (and other) eye movements to linguistic processing. *Journal of Memory and Language, 57*, 502-518. doi:10.1016/j.jml.2006.12.004

Apache-XML-Project. (1999). *Xerces C++ Parser*. Retrieved June 2009, from http://xerces.apache.org/xerces-c/.

Bangerter, A. (2004). Using pointing and describing to achieve joint focus of attention in dialogue. *Psychological Science, 15*(6), 415-419. doi:10.1111/j.0956-7976.2004.00694.x

Bard, E. G., Anderson, A. H., Chen, Y., Nicholson, H. B. M., Havard, C., & Dalzel-Job, S. (2007). Let's you do that: Sharing the cognitive burdens of dialogue. *Journal of Memory and Language, 57*(4), 616-641. doi:10.1016/j.jml.2006.12.003

Bard, E. G., Hill, R., & Arai, M. (2009). *Referring and gaze alignment: Accessibility is alive and well in situated dialogue.* Paper presented at the Cogsci 2009, Amsterdam, the Netherlands.

Bard, E. G., Hill, R., & Foster, M. E. (2008). *What tunes accessibility of referring expressions in task-related dialogue?* Paper presented at the Proceedings of the 30th Annual Meeting of the Cognitive Science Society, CogSci2008, Washington, D.C.

Bard, E. G., Hill, R., Nicol, C., & Carletta, J. (2007). *Look here: Does dialogue align gaze in dynamic joint action?* Paper presented at the AMLaP2007, Turku, Finland.

Barnes, G. (2008). Cognitive processes involved in smooth pursuit eye movements. *Brain and Cognition, 68*(3), 309-326. doi:10.1016/j.bandc.2008.08.020

Bell, A. (1984). Language style as audience design. *Language in Society, 13*(2), 145-204.

Brennan, S. E., Chen, X., Dickinson, C., Neider, M., & Zelinsky, G. (2008). Coordinating cognition: The costs and benefits of shared gaze during collaborative search. *Cognition, 106*(3), 1465–1477. doi:10.1016/j.cognition.2007.05.012

Brockmole, J. R., & Franconeri, S. L. (Eds.). (2009). *Binding: A Special Issue of the Journal 'Visual Cognition'*. New York: Psychology Press.

Burke, M., & Barnes, G. (2006). Quantitative differences in smooth pursuit and saccadic eye movements. *Experimental Brain Research, 175*(4), 596-608. doi:10.1007/s00221-006-0576-6

Campbell Jr., C., & McRoberts, T. (2005). *The Simple Sockets Library*. Retrieved July 2008, from http://mysite.verizon.net/astronaut/ssl/.

Charness, N., Reingold, E. M., Pomplun, M., & Stampe, D. M. (2001). The perceptual aspect of skilled performance in chess: Evidence from eye movements. *Memory & Cognition, 29*(8), 1146-1152.

Cherubini, M., Nüssli, M.-A., & Dillenbourg, P. (2008). *Deixis and gaze in collaborative work at a distance (over a shared map): a computational model to detect misunderstandings.* Proceedings of the 2008 symposium on Eye tracking research & applications, Savannah, Georgia. doi: 10.1145/1344471.1344515

Clark, H. H. (1996). *Using Language*. Cambridge, UK: Cambridge University Press.

Clark, H. H. (2003). Pointing and placing. In S. Kita (Ed.), *Pointing: Where language, culture, and cognition meet* (pp. 243-268). Mahwah, NJ, US: Lawrence Erlbaum Associates.

Clark, H. H., & Brennan, S. E. (1991). Grounding in communication. In L. B. Resnick, J. Levine & S. D. Teasley (Eds.), *Perspectives on socially shared cognition* (pp. 127-149). Washington, DC: APA.

Clark, H. H., & Krych, M. A. (2004). Speaking while monitoring addressees for understanding. *Journal of Memory and Language, 50*, 62-81. doi:10.1016/j.jml.2003.08.004

Clark, H. H., Schreuder, R., & Buttrick, S. (1983). Common ground and the understanding of demonstrative reference. *Journal of Verbal Learning and Verbal Behavior, 22*, 245-258.

Duchowski, A. T. (2003). *Eye tracking methodology: Theory and practice.* London: Springer.

Engelhardt, P. E., Bailey, K. G., & Ferreira, F. (2006). Do speakers and listeners observe the Gricean Maxim of Quantity? *Journal of Memory and Language, 54*(4), 554-573. doi:10.1016/j.jml.2005.12.009

Falck-Ytter, T., Gredeback, G., & von Hofsten, C. (2006). Infants predict other people's action goals. *Nature Neuroscience, 9*(7), 878-879. doi:10.1038/nn1729

Foster, M. E., Bard, E. G., Guhe, M., Hill, R. L., Oberlander, J., & Knoll, A. (2008). The roles of haptic-ostensive referring expressions in cooperative task-based human-robot dialogue. *Proceedings of the 3rd ACM/IEEE International Conference on Human Robot Interaction*, 295-302. doi:10.1145/1349822.1349861

Fussell, S., & Kraut, R. (2004). Visual co-presence and conversational coordination: commentary on Pickering & Garrod, Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences.*, 27, 196-197. doi:10.1017/S0140525X04290057

Fussell, S., Setlock, L., Yang, J., Ou, J. Z., Mauer, E., & Kramer, A. (2004). Gestures over video streams to support remote collaboration on physical tasks. *Human-Computer Interaction, 19*(3), 273-309. doi:10.1207/s15327051hci1903_3

Gesierich, B., Bruzzo, A., Ottoboni, G., & Finos, L. (2008). Human gaze behaviour during action execution and observation. *Acta Psychologica, 128*(2), 324-330. doi:10.1016/j.actpsy.2008.03.006

Gigerenzer, G., Todd, P. M., & ABC Research Group (1999). *Simple heuristics that make us smart*. Oxford: OUP.

Grant, E. R., & Spivey, M. J. (2003). Eye movements and problem solving: Guiding attention guides thought. *Psychological Science, 14*(5), 462-466. doi: 10.1111/1467-9280.02454

Griffin, Z. M. (2004). Why look? Reasons for eye movements related to language production. In J. M. Henderson & F. Ferreira (Eds.), *The integration of language, vision, and action: Eye movements and the visual world.* New York: Psychology Press.

Griffin, Z. M., & Oppenheimer, D. M. (2006). Speakers gaze at objects while preparing intentionally inaccurate labels for them. *Journal of Experimental Psychology: Learning, Memory and Cognition, 32*(4), 943-948. doi: 10.1037/0278-7393.32.4.943

Hadelich, K., & Crocker, M. W. (2006). *Gaze alignment of interlocutors in conversational dialogues.* Paper presented at the 19th CUNY Conference on Human Sentence Processing, New York, USA.

Hanna, J., & Tanenhaus, M. K. (2004). Pragmatic effects on reference resolution in a collaborative task: evidence from eye movements. *Cognitive Science, 28*, 105-115. doi:10.1207/s15516709cog2801_5

Henderson, J. M., & Ferreira, F. (Eds.). (2004). *The interface of language, vision, and action: Eye movements and the visual worlds.* New York, NY: Psychology Press.

Horton, W. S., & Keysar, B. (1996). When do speakers take into account common ground? *Cognition, 59*, 91-117. doi:10.1016/0010-0277(96)81418-1

Kraut, R., Fussell, S., & Siegel, J. (2003). Visual information as a conversational resource in collaborative physical tasks. *Human-Computer Interaction, 18*, 13-49. doi:10.1207/S15327051HCI1812_2

Kraut, R., Gergle, D., & Fussell, S. (2002). *The use of visual information in shared visual spaces: Informing the development of virtual co-presence.* Paper presented at the CSCW 2002.

Land, M. F. (2006). Eye movements and the control of actions in everyday life. *Progress in Retinal and Eye Research, 25*(3), 296-324. doi:10.1016/j.preteyeres.2006.01.002

Land, M. F. (2007). Fixation strategies during active behaviour: a brief history. In R. P. G. v. Gompel, M. H. Fischer, W. S. Murray & R. L. Hill (Eds.), *Eye Movements: A Window on Mind and Brain* (pp. 75-95). Oxford: Elsevier.

Land, M. F., & Furneaux, S. (1997). The knowledge base of the oculomotor system. *Philosophical Transactions of the Royal Society of London, Series B-Biological Sciences, 352*(1358)), 1231-1239.

Lindström, A. (1999). *SGE: SDL Graphics Extension.* Retrieved July 2008, from http://www.etek.chalmers.se/~e8cal1/sge/index.html.

Lobmaier, J. S., Fischer, M. H., & Schwaninger, A. (2006). Objects capture perceived gaze direction. *Experimental Psychology, 53*(2), 117-122. doi:10.1027/1618-3169.53.2.117

Lockridge, C., & Brennan, S. E. (2002). Addressees' needs influence speakers' early syntactic choices. *Psychonomic Bulletin & Review, 9*(3), 550-557.

Luck, S. J., & Beach, N. J. (1998). Visual attention and the binding problem: a neurophysiological perspective. In R. D. Wright (Ed.), *Visual Attention.* (pp. 455-478). Oxford: Oxford University Press.

Meyer, A. S., & Dobel, C. (2003). Application of eye tracking in speech production research. In J. Hyönä, R. Radach & H. Deubel (Eds.), *The mind's eye: Cognitive and applied aspects of eye movement research* (pp. 253-272). Amsterdam: Elsevier.

Meyer, A. S., van der Meulen, F., & Brooks, A. (2004). Eye movements during speech planning: Speaking about present and remembered objects. *Visual Cognition, 11*, 553-576. doi:10.1080/13506280344000248

Monk, A., & Gale, C. (2002). A look is worth a thousand words: Full gaze awareness in video-mediated conversation. *Discourse Processes, 33*, 257-278. doi:10.1207/S15326950DP3303_4

Murray, N., & Roberts, D. (2006). *Comparison of head gaze and head and eye gaze within an immersive environment.* Paper presented at the 10th IEEE International Symposium on Distributed Simulation and Real Time Applications., Los Alamitos, CA. doi:10.1109/DS-RT.2006.13

Pickering, M., & Garrod, S. (2004). Towards a mechanistic psychology of dialogue. *Behavioral and Brain Sciences, 27*, 169-190. doi:10.1017/S0140525X04000056

Pylyshyn, Z. W. (2006). Some puzzling findings in multiple object tracking (MOT): II. Inhibition of moving nontargets. *Visual Cognition, 14*, 175-198. doi:10.1080/13506280544000200

Pylyshyn, Z. W., & Annan, V. (2006). Dynamics of target selection in multiple object tracking (MOT). *Spatial Vision, 19*(6), 485-504. doi:10.1163/156856806779194017

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin, 124*, 372-422.

Richardson, D., & Dale, R. (2005). Looking to understand: The coupling between speakers' and listeners' eye movements and its relationship to discourse comprehension. *Cognitive Science, 29*(6), 1045-1060. doi:10.1207/s15516709cog0000_29

Richardson, D., Dale, R., & Kirkham, N. (2007). The art of conversation is coordination: common ground and the coupling of eye movements during dialogue. *Psychological Science, 18*(5), 407-413. doi:10.1111/j.1467-9280.2007.01914.x

Richardson, D., Dale, R., & Tomlinson, J. (in press). Conversation, gaze coordination, and beliefs about visual context. *Cognitive Science*. doi: 10.1111/j.1551-6709.2009.01057.x

Simple-DirectMedia-Layer-Project. (n.d.). *SDL: Simple DirectMedia Layer.* Retrieved June 2009, from http://www.libsdl.org/.

Spivey, M., & Geng, J. (2001). Oculomotor mechanisms activated by imagery and memory: Eye movements to absent objects. *Psychological Research, 65,* 235-241. doi:10.1007/s004260100059

Spivey, M., Tanenhaus, M. K., Eberhard, K. M., & Sedivy, J. C. (2002). Eye movements and spoken language comprehension: Effects of visual context on syntactic ambiguity resolution. *Cognitive Psychology, 45*(4), 447-481. doi:10.1016/S0010-0285(02)00503-0

SR-Research. (n.d.). *Complete Eyetracking Solutions.* Retrieved June 2009, from http://www.sr-research.com/EL_II.html.

Stein, R., & Brennan, S. E. (2004). *Another person's eye gaze as a cue in solving programming problems.* Paper presented at the International Conference on Multimodal Interfaces.

Steptoe, W., Wolff, R., Murgia, A., Guimaraes, E., Rae, J., Sharkey, P., et al. (2008). *Eye-tracking for avatar eye-gaze and interactional analysis in immersive collaborative virtual environments.* Paper presented at the CSCW08, San Diego, CA.

Tech-Smith. (n.d.). *Camtasia Studio*. Retrieved July 2008, from http://www.techsmith.com/camtasia.asp.

Trueswell, J. C., & Tanenhaus, M. K. (Eds.). (2005). *Approaches to Studying World-Situated Language Use: Bridging the Language-as-Product and Language-as-Action Traditions.* Cambridge, MA: MIT Press.

Underwood, G. (Ed.). (2005). *Cognitive Processes in Eye Guidance*. Oxford: Oxford University Press.

Underwood, J. (2005). Novice and expert performance with a dynamic control task: scanpaths during a computer game. In G. Underwood (Ed.), *Cognitive Processes in Eye Guidance* (pp. 303-323). Oxford: Oxford University Press.

Van Gompel, R. P. G., Fischer, M. H., Murray, W. S., & Hill, R. L. (Eds.). (2007). *Eye movements: A window on mind and brain.* Oxford: Elsevier.

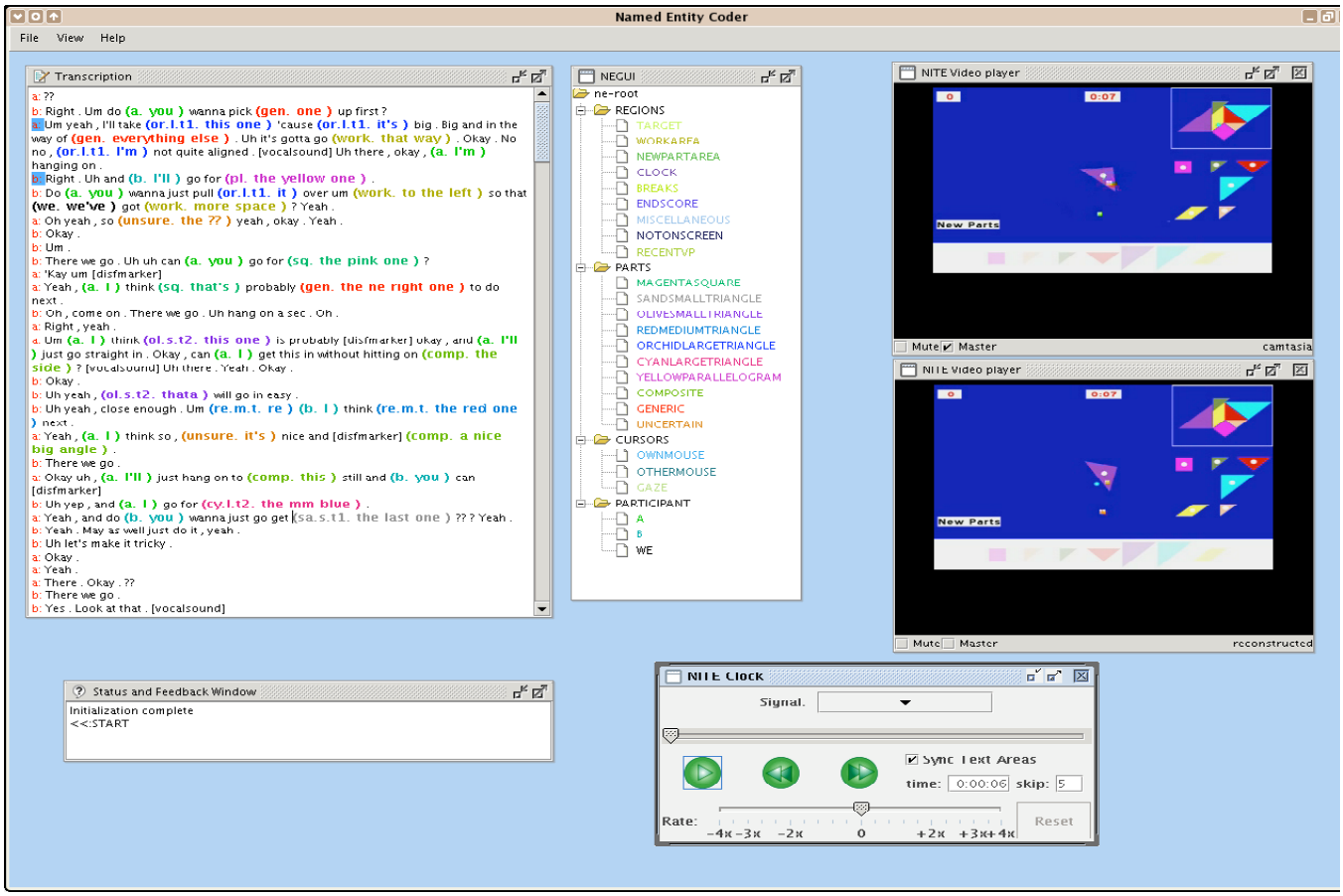
Velichkovsky, B. M. (1995). Communicating attention: Gaze position transfer in cooperative problem solving. *Pragmatics and Cognition, 3*(2), 199-222.

Vertegaal, R., & Ding, Y. (2002). *Explaining effects of eye gaze on mediated group conversations: amount or synchronization?* Paper presented at the CSCW 2002.

Wolfe, J. M., Place, S. S., & Horowitz, T. S. (2007). Multiple object juggling: Changing what is tracked during extended multiple object tracking. *Psychonomic Bulletin & Review, 14*(2), 344-349.

Appendices

Figure A1. Screen shot of NXT's discourse entity coding tool in use to code referring expressions for the example experiment


Named Entity Coder
File View Help
Transcription
a: ??
b: Right . Um do (a. you) wanna pick (gen. one) up first ?
a: Um yeah , I'll take (or.l.t1. this one) 'cause (or.l.t1. it's) big . Big and in the way of (gen. everything else) . Uh it's gotta go (work. that way) . Okay . No no , (or.l.t1. I'm) not quite aligned . [vocalsound] Uh there , okay , (a. I'm) hanging on .
b: Right . Uh and (b. I'll) go for (pl. the yellow one) .
b: Do (a. you) wanna just pull (or.l.t1. it) over um (work. to the left) so that (we. we've) got (work. more space) ? Yeah .
a: Oh yeah , so (unsure. the ??) yeah , okay . Yeah .
b: Okay .
b: Um .
b: There we go . Uh uh can (a. you) go for (sq. the pink one) ?
a: 'Kay um [disfmarker]
a: Yeah , (a. I) think (sq. that's) probably (gen. the ne right one) to do next .
b: Oh , come on . There we go . Uh hang on a sec . Oh .
a: Right , yeah .
a: Um (a. I) think (ol.s.t2. this one) is probably [disfmarker] okay , and (a. I'll) just go straight in . Okay , can (a. I) get this in without hitting on (comp. the side) ? [vocalsound] Uh there . Yeah . Okay .
b: Okay .
b: Uh yeah , (ol.s.t2. thata) will go in easy .
b: Uh yeah , close enough . Um (re.m.t. re) (b. I) think (re.m.t. the red one) next .
a: Yeah , (a. I) think so , (unsure. it's) nice and [disfmarker] (comp. a nice big angle) .
b: There we go .
a: Okay uh , (a. I'll) just hang on to (comp. this) still and (b. you) can [disfmarker]
b: Uh yep , and (a. I) go for (cy.l.t2. the mm blue) .
a: Yeah , and do (b. you) wanna just go get (sa.s.t1. the last one) ?? ? Yeah .
b: Yeah . May as well just do it , yeah .
b: Uh let's make it tricky .
a: Okay .
a: Yeah .
a: There . Okay . ??
b: There we go .
b: Yes . Look at that . [vocalsound]
NEGUI
ne-root
REGIONS
TARGET
WORKAREA
NEWPARTAREA
CLOCK
BREAKS
ENDSCORE
MISCELLANEOUS
NOTONSCREEN
RECENTVP
PARTS
MAGENTASQUARE
SANDSMALLTRIANGLE
OLIVESMALLTRIANGLE
REDMEDIUMTRIANGLE
ORCHIDLARGETRIANGLE
CYANLARGETRIANGLE
YELLOWPARALLELOGRAM
COMPOSITE
GENERIC
UNCERTAIN
CURSORS
OWNMOUSE
OTHERMOUSE
GAZE
PARTICIPANT
A
B
WE
NITE Video player
0
0:07
New Parts
Mute
Master
camtasia
NITE Video player
0
0:07
New Parts
Mute
Master
reconstructed
Status and Feedback Window
Initialization complete
<<:START
NITE Clock
Signal:
Sync Text Areas
time: 0:00:06
skip: 5
Rate:
-4x -3x -2x 0 +2x +3x +4x
Reset

Author Note

Drs. Jean Carletta, Craig Nicol, Robin Hill, Human Communication Research Centre, School of Informatics, University of Edinburgh; Dr. Ellen Gurman Bard, School of Philosophy, Psychology, and Language Sciences, and Human Communication Research Centre, University of Edinburgh; Dr. Jan Peter de Ruiter, Max Planck Institute for Psycholinguistics, Nijmegen.

Professor dr de Ruiter is now in the Department of Linguistics, University of Bielefeld, Germany. Dr Hill is now in the Department of Psychology, School of Philosophy, Psychology, and Language Sciences, University of Edinburgh.

This work was supported by the EU FP6 IST Cognitive Systems Integrated Project "JAST", Joint Action Science and Technology (FP6-003747-IP). The authors are grateful to Joseph Eddy and Jonathan Kilgour for assistance with programming.

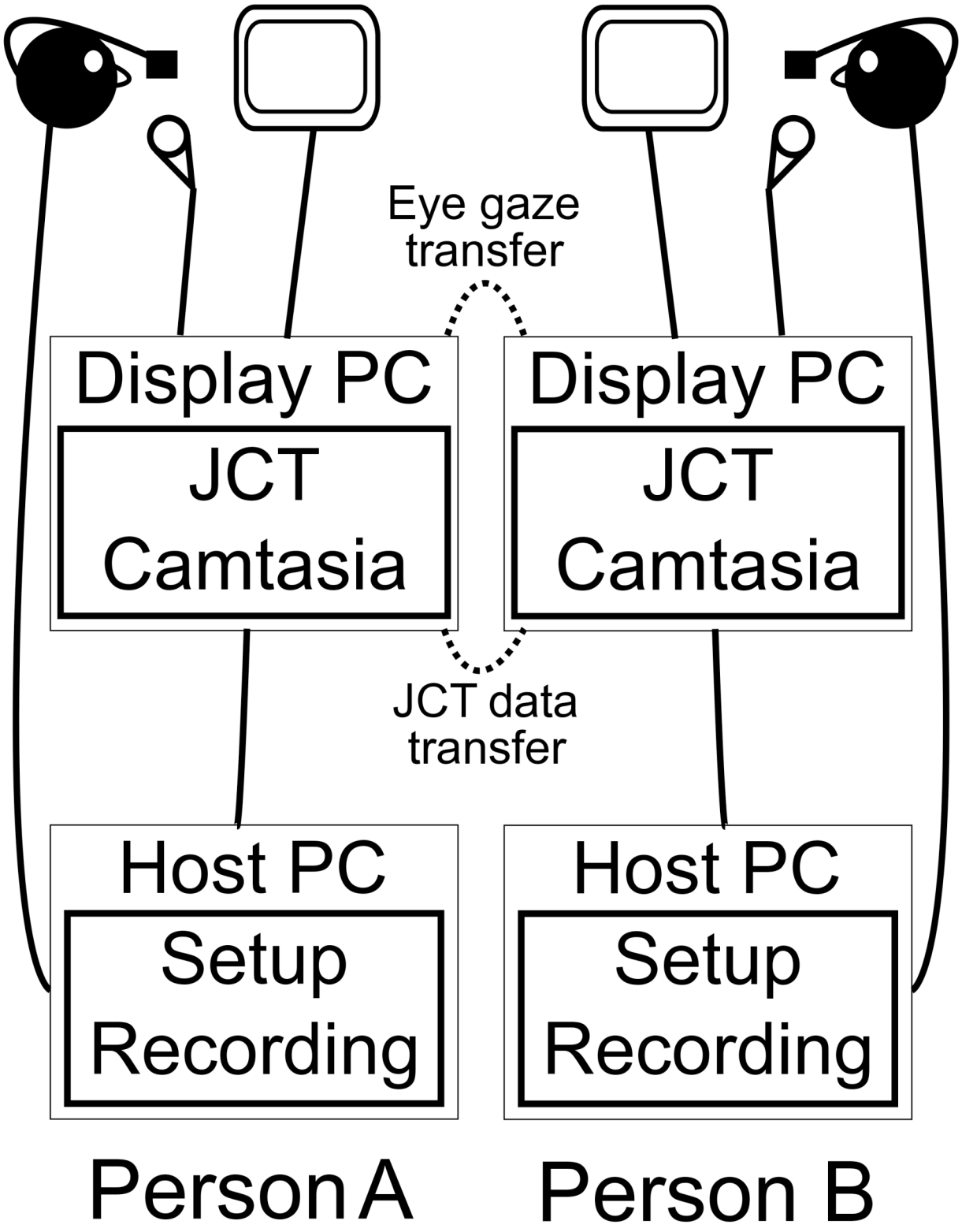
Correspondence concerning this article should be addressed to Dr. Jean Carletta, University of Edinburgh, Human Communication Research Centre, Informatics Forum, 10 Crichton Street, Edinburgh EH8 9AB, UK. *email*: J.Carletta@ed.ac.uk

Figure Captions

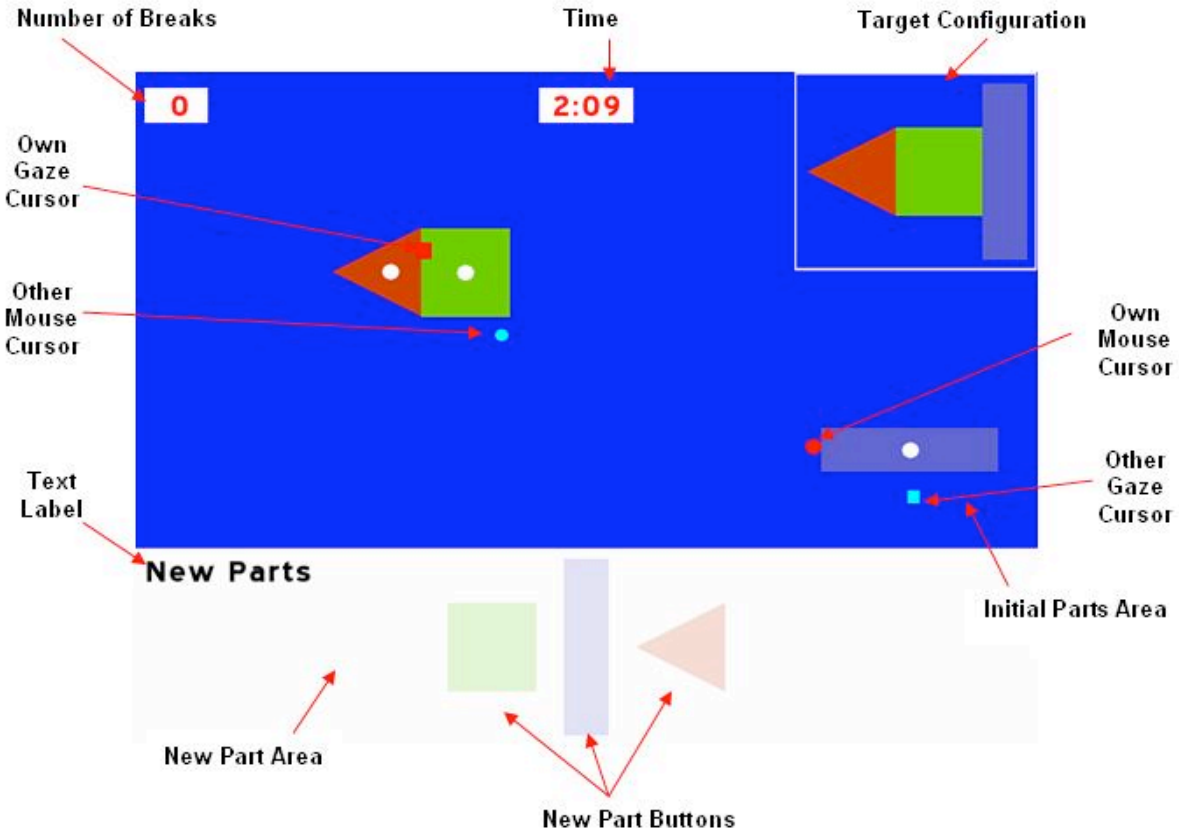
Figure 1. Hardware in the JAST Construction Task (JCT) experimental setup.

Figure 2 Annotated screen shots of two variants of the Jast Construction Task. a) Initial layout of a construction task; b) Screen shot 20s into a trial in a tangram construction task. Readers will find versions as originally colored in the online copy of this paper.

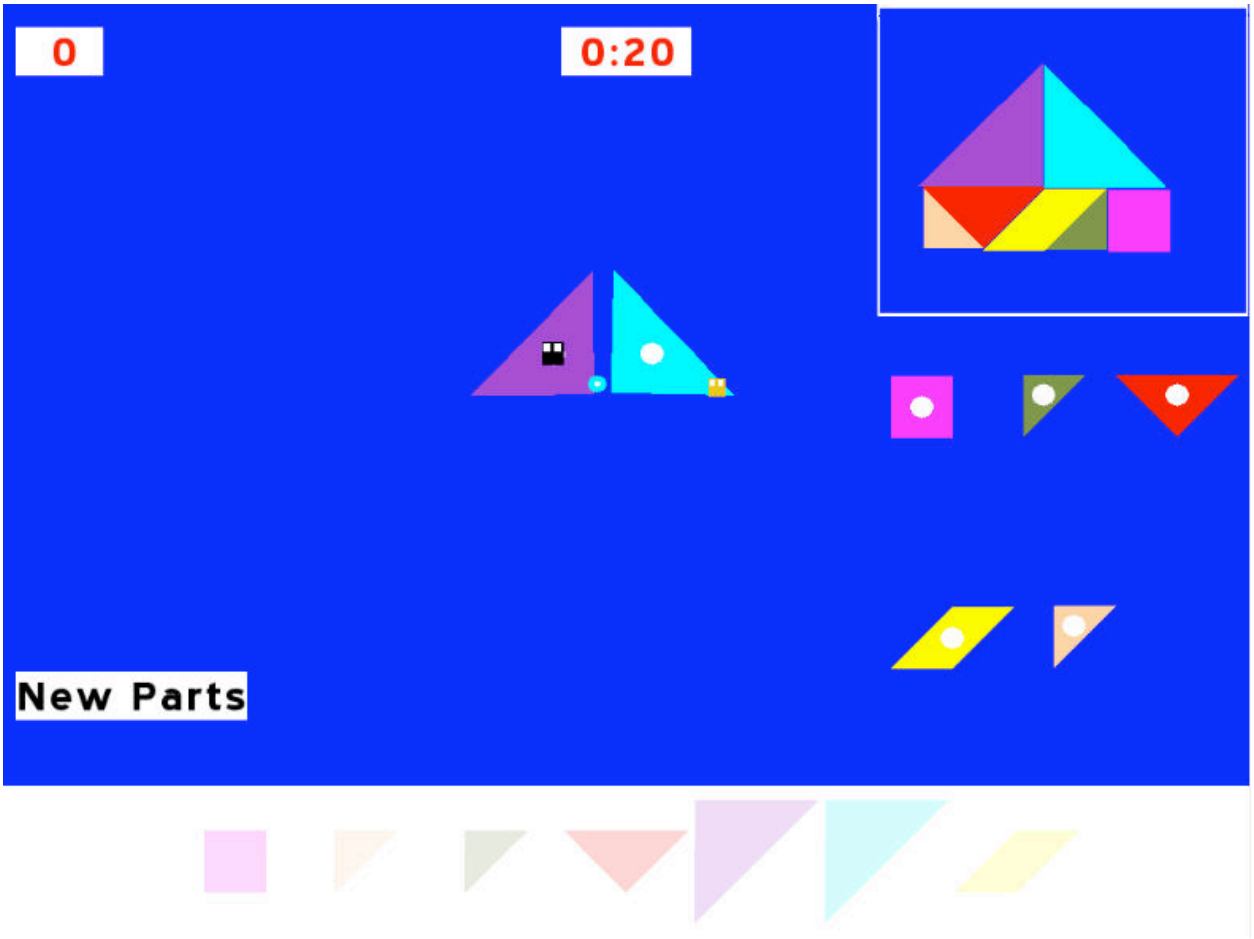
Figure 3. Data flow used to obtain the analysis for the example experiment.

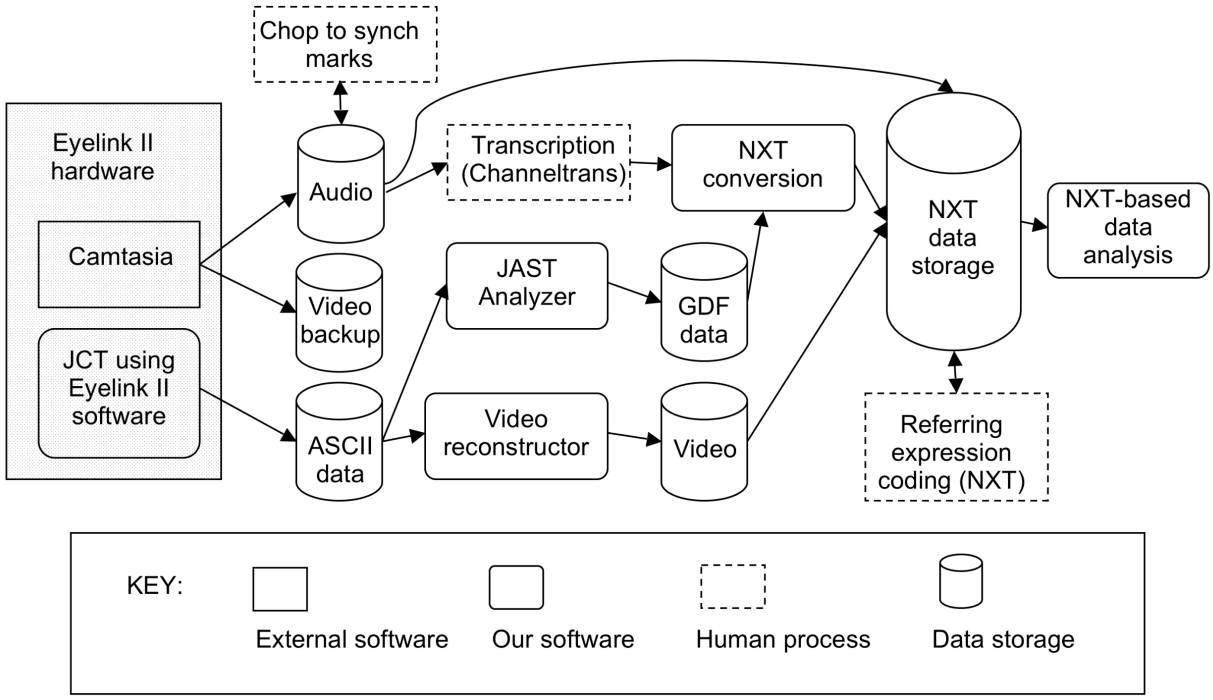

Eye gaze
transfer
Display PC
JCT
Camtasia
Display PC
JCT
Camtasia
JCT data
transfer
Host PC
Setup
Recording
Host PC
Setup
Recording
Person A
Person B

a)


Number of Breaks
Time
Target Configuration
0
2:09
Own Gaze Cursor
Other Mouse Cursor
Own Mouse Cursor
Other Gaze Cursor
Text Label
New Parts
Initial Parts Area
New Part Area
New Part Buttons

b)




Chop to synch marks
Eyelink II hardware
Camtasia
JCT using Eyelink II software
Audio
Video backup
ASCII data
Transcription (Channeltrans)
JAST Analyzer
Video reconstructor
NXT conversion
GDF data
Video
NXT data storage
NXT-based data analysis
Referring expression coding (NXT)
KEY:
External software
Our software
Human process
Data storage