Using Virtual Reality Technology in Linguistic Research

Thies Pfeiffer*

A.I. Group, Faculty of Technology, Bielefeld University

Abstract

In this paper, we argue that empirical research on genuine linguistic topics, such as on the production of multimodal utterances in the speaker and the interpretation of the multimodal signals in the interlocutor, can greatly benefit from the use of virtual reality technologies.

Established methodologies for research on multimodal interactions, like the presentation of pre-recorded 2D videos of interaction partners as stimuli and the recording of interaction partners using multiple 2D video cameras have crucial shortcomings regarding ecological validity and the precision of measurements that can be achieved. In addition, these methodologies enforce restrictions on the researcher. The stimuli, for example, are not very interactive and thus not as close to natural interactions as ultimately desired. Also, the analysis of 2D video recordings requires intensive manual annotations, often frame-by-frame, which negatively affects the feasible number of interactions which can be included in a study.

The technologies bundled under the term virtual reality offer exciting possibilities for the linguistic researcher: gestures can be tracked without being restricted to fixed perspectives, annotation can be done on large corpora (semi-)automatically and virtual characters can be used to produce specific linguistic stimuli in a repetitive but interactive fashion. Moreover, immersive 3D visualizations can be used to recreate a simulation of the recorded interactions by fusing the raw data with theoretic models to support an iterative data-driven development of linguistic theories. This paper discusses the potential of virtual reality technologies for linguistic research and provides examples for the application of the methodology.

Index Terms: H.5.2 [Information Interfaces and Presentation]: User Interfaces—Natural Language; I.3.6 [Computer Graphics]: Methodology and Techniques—Interaction Techniques; H.5.1 [Information Interfaces and Presentation]: Multimedia Information Systems—Artificial, Augmented, and Virtual Realities

1 Introduction

The inspiration for this article came from the keynote speech of Maria Kozhevnikov at the 2011 IEEE VR conference titled "Bringing 3D immersive virtual reality technologies to visual-spatial cognition research". In this article, we add an additional perspective to hers. This perspective is shaped by over a decade of cooperation with linguists and psycholinguists. In our group, the Artificial Intelligence Group at Bielefeld University, we run interdisciplinary research projects with linguists which are funded by the Deutsche Forschungsgemeinschaft since 1993 (SFB 360, "Situated Artificial Communicators", 1993 - 2005; SFB 673, "Alignment in Communication", 2006 - 2014). Since then, we use virtual reality technology to conduct experiments, record multimodal interactions, explore and annotate multimodal data and simulate models of human communication behavior in immersive Virtual Reality.

In the following, we present four main areas, where virtual reality technology could be used to support linguistic research.

*e-mail: tpfeiffe@techfak.uni-bielefeld.de

2 Recording Multimodal Interactions

Human-Computer-Interaction in VR environments, such as CAVE™-like installations, makes use of a broad range of input devices. Many VR installations track at least the position and orientation of the head, and some even use data-gloves for gesture input or provide a speech interface. This means that most modalities of interest for linguistic researchers are already being tracked in VR applications. In addition to the devices, VR technology also provides the software infrastructure to access the devices' input in real-time (e.g. VRPN [13] or Open Tracker [11]).

In 2004 the linguists in our research group presented a problem: they wanted to assess the accuracy and precision of manual pointing gestures. In a first study, they had recorded two interlocutors during a demonstration game where one interlocutor was asked to use pointing gestures to communicate a sequence of object references to the other interlocutor. For the recordings they used standard video cameras and – as they saw no other way – they did their measurements using a ruler on the video monitor. Obviously, the accuracy of their measurement method was not very high. In a joint project, we replicated their original study [5], this time using a marker-based tracking system of our VR installation to track the exact positions of the index finger during the pointing tasks. The high accuracy and the high number of observed pointing gestures allowed a data-driven quantitative approach, which finally led to a precise model of pointing direction and extension [9].

3 Annotation of Multimodal Data

The annotation of video recordings of human-human interactions is typically done using an annotation software such as Anvil [3] or ELAN [1]. Anvil has recently been extended to support the visualization of skeletal information [8], but the annotation process is still very elaborate. The linguistic researchers have to develop a coding manual, a detailed description of what to annotate and how to represent it, and the annotators, typically more than one working on the same sequences, follow this manual watching the recorded sequences on frame-level. This procedure is often repeated several times, either because of the number of different aspects to be annotated or the iterative refinement of the coding manual.

A typical VR setup which supports multimodal interaction already integrates the heterogeneous sensory data and produces abstract descriptions of the ongoing interactions on more declarative levels [4, 6]. This means that some annotations that currently have to be done manually by the linguistic researcher based on the video recordings can be replaced by the output produced by the multimodal interaction framework based on the tracking data on-the-fly. This can happen even during the recording of the experiment. Many more annotations could be automated similarly by extending the interaction framework to support the required features.

In addition to that, as the virtual environment is completely described at each moment, some interactions can be put in context automatically. In our own work, we used this to automatically generate statistical information about the accuracy with which objects have been pointed at during a dyadic interaction about reference communication [9].

Taken together, the possibility to detect complex features in the multimodal input and to make exact reference to the virtual envi-

ronment while automatically creating appropriate annotations significantly reduces the workload of the linguistic researcher.

4 Making Multimodal Data Interactively Explorable

Having video recordings and feature descriptions in the annotations still might not enable the linguistic researcher to literally see the big picture. While the multimodal annotation tool MacVisSTA [12] provides a side-by-side synchronous view of all datastreams, it does not integrate between the different streams and sticks to a WIMP desktop interface. The immersive environment of a VR installation can help the researcher to explore the recorded multimodal datasets interactively, moving back and forth in time and changing the viewing perspective from speaker to listener or any other suitable perspective. For the aforementioned study on pointing gestures, we created the Interactive Augmented Data Explorer [10] (IADE), which presents a simulation of the setting used during the recordings, displays the tracking data and the recorded audio and video files and augments these information by the manual annotations. This interactive multimedia presentation of the gathered data provides a genuine view which allowed us to identify problems in the tracking data (obscured markers) and false annotations. The interactive view can be used to support further annotations as well, as the annotators can zoom in on certain aspects and consider all relevant data from the best perspective, e.g. switching to the perspective of the speaker.

5 Interactive Content for Studies on Complex Interactions

The aspects considered so far have addressed issues of data recording, displaying of multimedia data, and annotation and analysis. The fourth aspect, from our perspective, is probably the most important: virtual reality can be used to create highly controlled interactive simulations of communication behavior, which can be used as stimulus material in studies on human communication. This includes the use of embodied conversational agents [2] which achieve a certain level of behavioral realism (e.g. [14, 7]).

Creating stimulus material to address scientific questions regarding dialog is otherwise quite complex. If repeatability is required, the material is basically restricted to pre-recorded 2D video sequences. This method, however, provides only a minimal level of interactivity. In addition to that, there can be no real interaction between the interlocutors, so, e.g., grounding of object references in the proximal interaction space is not properly possible. The only other option would be the observation of two humans communicating, which would make it difficult if only specific adjacency pairs of dialog are of interest, as they cannot be enforced properly. Making one interlocutor a confederate of the experimenter could improve upon this, but then there is still a lot of intra-personal variability in the behavior of the confederate. In addition, it is questionable whether they are able to produce a modified behavior in-line with both the experiment conditions and their own style and believability.

6 Conclusion

We presented four areas in which we believe that linguistic research could benefit from using virtual reality technology. The first two areas address issues where the available technology is already very advanced and can be directly used. While the scientific knowledge in the area of immersive scientific visualization is quite elaborated, there is only sparse evidence of its application to data from the domain of linguistics, especially regarding multimodal data on human-human interactions. The most important advances, however, are expected from the central focus of virtual reality: the generation of plausible worlds where the participants of linguistic experiments could immerse into controlled dialog games that feel natural but still follow the required constraints of experimental design.

To summarize, virtual reality bundles most of the relevant technologies required to work on modern theories of multimodal communication in the field of linguistics. It would be fruitful to combine this with efforts for creating and mining large multimodal corpora. In the end, this line of research would also be beneficial for basic research interests in virtual reality and 3D user interfaces regarding natural interaction with such systems.

References

[1] H. Brugman and A. Russel. Annotating multimedia/multi-modal resources with elan. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation*, pages 2065–2068. Citeseer, 2004.

[2] J. Cassell, J. Sullivan, S. Prevost, and E. Churchill, editors. *Embodied Conversational Agents*. MIT Press, Cambridge, MA, 2000.

[3] M. Kipp. Anvil - a generic annotation tool for multimodal dialogue. In *Seventh European Conference on Speech Communication and Technology*, 2001. Retrieved September 2011.

[4] D. Koons, C. Sparrel, and K. Thorisson. *Integrating simultaneous input from speech, gaze and hand gestures*. AAAI Press, 1993.

[5] A. Kranstedt, A. Lücking, T. Pfeiffer, H. Rieser, and I. Wachsmuth. Deixis: How to Determine Demonstrated Objects Using a Pointing Cone. In S. Gibet, N. Courty, and J.-F. Kamp, editors, *Gesture Workshop 2005*, LNAI 3881, pages 300–311, Berlin Heidelberg, 2006. Springer-Verlag GmbH.

[6] M. Latoschik. Designing Transition Networks for Multimodal VR-Interactions Using a Markup Language. In *Proceedings of the IEEE fourth International Conference on Multimodal Interfaces, ICMI 2002, Pittsburgh, USA, October 2002*, pages 411–416, 2002.

[7] J. Lee, S. Marsella, D. R. Traum, J. Gratch, and B. Lance. The Rickel Gaze Model: A Window on the Mind of a Virtual Human. In C. Pelachaud, J.-C. Martin, E. André, G. Chollet, K. Karpouzis, and D. Pelé, editors, *IVA '07: Proceedings of the 7th international conference on Intelligent Virtual Agents*, volume 4722 of *Lecture Notes in Computer Science*, pages 296–303, Berlin, Heidelberg, 2007. Springer-Verlag.

[8] Q. Nguyen and M. Kipp. Annotation of human gesture using 3d skeleton controls. In *Proceedings of 7th International Conference on Language Resources and Evaluation (LREC), ELDA*, 2010.

[9] T. Pfeiffer. *Understanding Multimodal Deixis with Gaze and Gesture in Conversational Interfaces*. Berichte aus der Informatik. Shaker Verlag, Aachen, Germany, December 2011.

[10] T. Pfeiffer, A. Kranstedt, and A. Lücking. Sprach-Gestik Experimente mit IADE, dem Interactive Augmented Data Explorer. In S. Müller and G. Zachmann, editors, *Dritter Workshop Virtuelle und Erweiterte Realität der GI-Fachgruppe VR/AR*, pages 61–72, Aachen, 2006. Shaker.

[11] G. Reitmayr and D. Schmalstieg. Opentracker-an open software architecture for reconfigurable tracking based on xml. In *Virtual Reality, 2001. Proceedings. IEEE*, pages 285 –286, march 2001.

[12] R. T. Rose, F. Quek, and Y. Shi. Macvissta: a system for multimodal analysis. In *Proceedings of the 6th international conference on Multimodal interfaces*, ICMI '04, pages 259–264, New York, NY, USA, 2004. ACM.

[13] R. Taylor II, T. Hudson, A. Seeger, H. Weber, J. Juliano, and A. Helser. Vrpn: a device-independent, network-transparent vr peripheral system. In *Proceedings of the ACM symposium on Virtual reality software and technology*, pages 55–61. ACM, 2001.

[14] D. Traum and J. Rickel. Embodied agents for multi-party dialogue in immersive virtual worlds. In *Proceedings First Int. Joint Conference on Autonomous Agents and Multiagent systems*, pages 766–773, 2002.