A Process Algebra Account of Speech-gesture Interaction

Hannes Rieser,
Bielefeld University, Germany
Hannes.Rieser@Uni-Bielefeld.de

Abstract

The paper is based on extensive corpus work dealing with the interaction of gesture and speech in natural route-description dialogues. The issue discussed is how non-regimented gesture and speech processes can be modelled in a formal system. The main argument is that this cannot be achieved in structural paradigms currently in use. The proposal is to turn instead to process algebras in the tradition of Milner's π-calculus. The special algebra discussed is a newly developed hybrid λ-ψ calculus which can transport typed λ-expressions over communicating input-output channels. Central for the account is the notion of agent. Speech-gesture interaction is implemented via i-o-channel interactions. Interactions are allowed, postponed or blocked using a typing system. Terminating communication among agents leads to a multi-modal meaning representation.

1 Relevance for the workshop

The key-concepts in the workshop title "Formal approaches to the dynamics of linguistic interaction" are implemented in this paper in the following way: The type of linguistic interaction handled is the interface between speech processes and co-verbal iconic gestures. The dynamics comes in due to the incremental modelling of the speech and gesture processes and the interaction among these which results in multi-modal meaning. (Concerning interaction in the CA sense, as for example treated in Kempson et al. 2016, see Section 4.) Finally, the formal side is provided by the hybrid λ-ψ-calculus used. The paper builds upon former work on a λ-π-account of speech-gesture coordination in Rieser (2014, 2015, 216).

2 Speech-gesture Interaction

The paper starts from the corpus-based observation that gestures are semantically related to the speech they accompany. In light of this, the question arises how they interact with speech and how this interaction can be modelled. Virtually all gesture research assumes that gestures have form and meaning. Following Kendon and McNeill (Kendon 2004, McNeill 1992, 2005), a gesture's structure is characterised by the three consecutive stages preparation, stroke, and retraction. Gestures span from rest position to rest position. Rest positions hence determine a gesture's individuation. The stroke extends over a time span measured in systematic annotation. Only the gesture's stroke must be present to represent a gesture; the meaning of a gesture resides in its stroke. In natural conversation, however, rest positions vary. Moreover, the stroke position is often held by communication participants producing so-called "post-stroke-holds" which can be operative within and across turns, see Example 1. The effect of this is usually that given information is kept and hence visually present on the gesture channel while currently new information is produced on the speech channel: Next speaker may already produce her turn, possibly accompanied by her own gesticulation while the previous speaker still holds stroke information. So gesture and speech have their own modes of encoding and yielding information. Gesture information as understood here is encoded in a formal language specifying topological entities like points, lines, planes or solids and intersections of these. It is drawn from systematic annotation of gesture occurrences using hand positions, palm- wrist- and back-of-hand orientation etc. (as developed in Rieser 2010). Gesture information is always partial. The partiality feature is not treated in this paper, a first account of it is given in (Lawler, Hahn,

and Rieser 2017) contained in these proceedings.

3 Gesture Speech Asynchrony

Systematic annotation of multi-modal data shows that when interacting with speech, gestures do not perfectly synchronise with their "privileged" semantic coordination point. Although received knowledge, this is still a research problem for current descriptive and formal gesture research (Alahverdzhieva and Lascarides 2010, Giorgolo 2010, Lascarides and Stone 2006, 2009, Lücking 2013, Oviatt *et al* 1997, Rieser 2014, Röpke *et al* 2013) as the discussion of gesture-attachment issues shows. Gestures can come entirely before the aligned speech, entirely after it or overlap it. Gesture information can be totally independent of speech information, thus providing additional content as in the examples sketched below. Especially this last case is taken as evidence for the independence of the gesture system from the speech system and will largely determine the style of modelling. As a consequence, the description of speech-gesture coordination cannot be given fitting the gesture meaning representation into the speech meaning representation in some naïve compositional way using e.g. unification. Doing so would violate the independence of gestural information and unduly regiment natural data; especially its non-perfect synchronisation with speech would then escape reconstruction. To clarify this last point, assume that a gesture indicating a square comes entirely before or entirely after an utterance of "window" which does not provide the square-information. Then fusing the square property directly with the "window"-representation would avoid to reconstruct non-perfect synchrony. Motivated by corpus data (Lücking *et al* 2012) and concentrating on referential and iconic gestures, we propose to view gesture and speech as independent processes which interact if it is semantically apt, expressed more technically, if their typings fit. Seen from one point of view, speech is gesture's main companion: gesture may "offer" its information to speech and speech may take it up. If taken up, we get multi-modal information, information assembled from two different sources. If rejected, the gesture stroke can be held waiting for a more appropriate communication opportunity, which, however, could fail to arise: Gesture was put on an outgoing channel but could not enter an ingoing port. There are also more subtle types of gesture-speech communication where speech provides the immediate context for gesture interpretation and the result then again interfaces with speech. It is an open question whether we always have this dependence on the speech context. This will not be discussed in this paper (however, see Lawler *et al*. 2016, where that is the central topic).

4 Outline of Process Algebra Used

Before we give some indication of how to model the gesture speech asynchrony described above, we briefly sum up the empirical findings: Empirical data suggest the need for

- channels on which information (data, agents or procedures) can be sent,
- procedures operating concurrently,
- interfaces enabling communication among processes,
- active and non-active processes, and
- communication among agents organised *via* an i-o-mechanism.

The shift to considering communicating processes necessitates the move to a methodology featuring a process ontology instead of a purely domain-of-objects one as usual in linguistics, logics and philosophy. The one we will use is the ψ-calculus (Bengtson et al. 2011, Johansson, 2010), a recent extension of Milner's π-calculus (Milner, 1999, Parrow, 2001, Sangiorgi and Walker, 2001), belonging to the field of Process Algebra (Fokkink 2000, Hennessy 1988, Bergstra et al. 2000). The ψ-calculus works with processes (so-called agents) and data structures which can be transmitted among agents *via* structured channels using an i-o-facility. Essentially, gesture and speech are viewed as such ψ-agents in this paper.

We provide here and comment upon the central definition for the behaviour of ψ-agents P, Q, … , following (Bengtson et al. 2011):

22 Foll: Hold on. Well, you-CUTOFF. Well, you walk now into this
23 WINDING GESTURE
24 street and then where is the sculpture? Is it at the front or to
GESTURE GESTURE
25 the left or to the right
GESTURE GESTURE
26 WINDING GESTURE HELD

Example 1: English translation of a German transcript from the Bielefeld SaGA corpus (Follower). Right-hand winding gesture in green, left-hand indexing gestures in yellow. The winding gesture (stroke and post-hold) extends throughout turns 22 to 26.

Definition:

0	Nil, the empty agent
$\overline{M}$N.P	Output
$\underline{M}$(λx)N.P	Input
τ	Silent agent
case φ_1: P_1 ‖ … ‖ φ_n: P_n	Case construct
(*νa*)P	Restriction
P | Q	Parallel
!P	Replication
(|Ψ|)	Assertion
"."	Sequential composition

The 0 agent is inactive. "$\overline{M}$N.P" (M overbar, N dot P) puts a data structure N onto an outgoing channel M, and continues with process P, possibly a 0 process. "$\underline{M}$(λx)N.P" (M under-bar) indicates that a data structure is received on the input channel M and substituted for the λ-variable *x* in N and P. In the case construct one alternative P_i is chosen given that φ_i is true. The case construct is also used to model the non-deterministic *or*. The restriction *ν* means that the scope of "a" is local to "P". The parallel operator "|" enables P and Q to expand independently or to communicate with each other via the i-o-operators, possibly after several independent expansions. Replication is defined as P|!P which means that P can be repeated arbitrarily often.

Before we present an informal description of how the λ-ψ- calculus can be put to operation, example 1 shows the English translation of a German transcript from the Bielefeld Speech-and-Gesture-Alignment corpus (SAGA, Lücking *et al*, 2012) used for this purpose.

The example is a section of a multi-modal dialogue between a route-giver and a follower. We briefly sketch how the dialogue excerpt can be modelled using the ψ-technology: The follower uses a winding gesture when starting her contribution with "well". On one reading, she wants to modify "street", so the gesture stroke precedes the optimal interface point. Other possible integration points not discussed here are "walk", "into", and most notably, the event of walking-into itself. After, e.g., interaction with "street" and production of a multi-modal meaning "bendy street" the winding gesture is still held. In the end, ψ's i-i-i-o-facility is taken to model speech-gesture coordination. Due to the incremental grammar hypothesized, the logic of the data structures involved (typed λ-calculus) and the logic of ψ we arrive at a complex hybrid tool, the λ-ψ-calculus.

5 Definition of the Speech-gesture Interaction Agent SGIA in the λ-ψ-Calculus

The λ-ψ-agent SGIA that

- handles incrementality,
- implements the intuitively correct scopes, and
- achieves the speech-gesture integration

is defined in the following protocol (0-agents being sometimes omitted):

SGIA $=_{def}$!$\overline{ch1}$<λf λu (f(u) ∧ bendy'(u))>
|$\overline{ch7}$ (we'). <λp(well'(p))(we')>
|$\overline{ch4}$ (w'). $\underline{ch5}$ (i'). $\underline{ch3}$ (ts'). $\underline{ch6}$ (nw'). $\overline{ch7}$. nw'<<< λf λru (λx (f(x, you') ∧ r(x, u))e)w'>i'>ts'>
| $\overline{ch4}$. < walk'>.0
| $\overline{ch6}$. <λp now'(p).0>
| $\overline{ch5}$. into'.0
| $\underline{ch2}$(s').$\overline{ch3}$. <<λg (this x (g(x)) s'>>
| $\underline{ch1}$(b'). $\overline{ch2}$. <b' <λx(street'(x))>>.0

The agent consists of eight concurrent processes, indicated by "|" of which only the gesture-simulating one is recursive due to !. Sequentiality (order among constituents) is achieved by types, not given here. It is helpful to keep in mind that we have o-i-channels indicated by overbar and under-bar, respectively: A winding gesture is produced concurrently with the words <"well", "you", "walk", "now", "into", "this", "street">. Using $\overline{ch1}$ it sends its information to "street", yielding thus "bendy street". The property "bendy street" in turn sends its information via $\overline{ch2}$ to "this" and we get the referring expression "this bendy street". This information is set aside for a while, since the output channel does not immediately find a matching input channel. The information tied to "you" is a propositional function and needs several constants inserted *via* channels $\underline{ch4}$ (w'), $\underline{ch5}$ (i') and $\underline{ch3}$ (ts′), respectively, in particular a relation "walk" defined on an event e and a subject "you" and a relation "into" defined on the same event and the multi-modal referring expression "this bendy street" already compiled. The resulting term is the proposition "There is an event of you walking into this street" that "now" looks for due to its $\overline{ch6}$ and with which it combines moving into $\underline{ch6}$ to yield another proposition, "Now there is an event of you walking into this bendy street", in more colloquial terms (cf. the annotation of the dialogue-part in Example 1), "Now you walk into this bendy street". This new proposition is put on an outgoing channel $\overline{ch7}$ and combines with "well" using input channel $\underline{ch7}$, again generating a proposition "Well, now you walk into this bendy street" while the winding gesture continues to be held due to $!\overline{ch1}<\lambda f\, \lambda u\, (f(u) \wedge bendy'(u))>$. Hence, the formula to be interpreted is in the end

$!\overline{ch1} <\lambda f\lambda u(f(u) \wedge bendy'(u))>.0 \mid well'(now'(walk'(e, you') \wedge into'(x, this' \text{ x } (street'(x) \wedge bendy'(x))))).0$

of which only the second closed process $well'(now'(walk'(e, you') \wedge into'(x, this' \text{ x } (street'(x) \wedge bendy'(x))))).0$ is satisfied.

6 Future Research

The account given handles the property + noun semantics case using λ-ψ-processes. The shortcoming of this particular example is that the initial introduction of the bendy-street-gesture combination into the dialogue is not shown. This opens up the question at which level existing dialogue theories can be married with the process architecture. In talks I already sketched that the basic λ-ψ-i-o-facility can also be used to model split utterances with in-turn-acknowledgements as discussed, e.g., in Eshghi et al. (2015):

A. The doctor.
B. Chorlton?
A. No, Fitzgerald.
B. uh-huh.

In order to do so, one establishes "turn channels" transporting the respective dialogue contributions of A and B. These have to satisfy A's and B's tests modelled with the case-construct. Furthermore, by way of generalisation it can be argued that ψ can be used to model any type of multi-modal information which was subjected to rigid annotation.

Acknowledgements

Thanks for the comments of three ESSLLI-reviewers which helped to improve the Ms. Due to space restrictions not all of the reviewers' suggestions could be taken up. Some would also require much additional research.

References

Alahverdzhieva, K. and Lascarides, A. (2010). Analysing language and co-verbal gesture in constraint-based grammars. In Müller, St., editor, *Proceedings of the 17th International Conference on Head-Driven Phase Structure Grammar (HPSG)*, pp. 5–25, Paris.

Bengtson, J., Johansson, M., Parrow, J., and Björn, V. (2011). Psi-Calculi: A framework for mobile processes with nominal data and logic. *Logical Methods in Computer Science.* Vol. 7 (1:11), 2011, pp. 1-44.

Bergstra, J. A., Fokkink, W. J., Ponse, A. (2000). Process algebra with recursive operations. In Bergstra et al. (editors), *Handbook of Process Algebra.* Amsterdam: Elsevier, pp. 333-391.

Eshghi, A., Howes, C., Gregoromichelaki, E., Hough, J., Purver, M. (2015). Feedback in conversation as incremental semantic update. In *Proceedings of the 11th International Conference on Computational Semantics*, pp. 261-271.

Fokkink, W. (2000). *Introduction to Process Algebra*. Berlin, Heidelberg: Springer.

Giorgolo, G. (2010). *Space and Time in Our Hands.* UIL-OTS, Universiteit Utrecht, 2010.

Hennessy, M. (1988). *Algebraic Theory of Processes.* Cambridge, Mass.: The MIT Press.

Johansson, M. (2010). *Psi-calculi: a framework for mobile process calculi*. Diss. from the Faculty of Science and Technology 94. 184 pp. Upsala.

Kempson, R., Gregoromichelaki, E., Cann, R., Chatzikyriakidis, S., (2016). Language as mechanisms for interaction. *Theoretical Linguistics*, Bd. 42, Heft 3-4.

Kendon, A. (2004). *Gesture: Visible Action as Utterance*. Cambridge: CUP.

Lascarides, A. and Stone, M. (2006). Formal semantics of iconic gesture. In Schlangen, D. and Fernández R., editors, *Proceedings of the 10th Workshop on the Semantics and Pragmatics of Dialogue, (Brandial)*, pp. 64–71, Potsdam.

Lascarides, A. and Stone, M. (2009). A formal semantic analysis of gesture. *Journal of Semantics*, 26(4), pp. 393-449.

Lawler, I., Hahn, F., and Rieser, H. (2016). Multi-modal context-dependency. Extended abstract. *Workshop on Situations, Information, and Semantic Content*, LMU Munich.

Lawler, I., Hahn, F., and Rieser, H. (2017). Gesture meaning needs speech meaning to denote - A case of speech-gesture meaning interaction. In *Proceedings of the Workshop on Formal Approaches to Linguistic Interaction; ESSLLI 2017*.

Lücking, A. (2013). *Ikonische Gesten. Grundzüge einer linguistischen Theorie.* De Gruyter Mouton, Germany.

Lücking, A., Bergmann, K., Hahn, F., Kopp, S., and Rieser, H. (2012). Data-based analysis of speech and gesture: The Bielefeld speech and gesture alignment corpus (SaGA) and its Applications. *Journal on Multimodal User Interfaces* 7(1-2), pp. 5-18.

McNeill, D. (1992). *Hand and Mind: What Gestures Reveal about Thought.* Chicago: Chicago University Press.

McNeill, D. (2005). *Gesture and Thought.* Chicago: Chicago University Press.

Milner, R. (1999). *Communicating and mobile systems: the π-calculus.* Cambridge: CUP.

Oviatt, S., DeAngeli, A., and Kuhn, K. (1997). Integration and synchronisation of input modes during multimodal human-computer interaction. *CHI*, pp. 415-422.

Parrow, J. (2001). An introduction to the π-calculus. In Bergstra, J. A., Ponse, A., and Smolka, S. A., editors, *Handbook of Process Algebra*. Amsterdam: Elsevier, pp. 479–545.

Rieser H. (2010). On factoring out a gesture typology from the Bielefeld speech-and-gesture-alignment corpus (SAGA). In: Kopp S., Wachsmuth I. (eds) *Proceedings of GW 2009*, pp 47–60.

Rieser, H. (2014). Gesture and speech as autonomous communicating processes. Talk at *Workshop on Embodied meaning goes public,* Stuttgart University.

Rieser, H. (2015). When hands talk to mouth. Gesture and speech as autonomous communicating processes. *Proceedings of the 19th Workshop on the Semantics and Pragmatics of Dialogue, (goDIAL),* Gothenburg.

Rieser, H. (2016). A process-algebra account of speech-gesture interaction. Extended abstract, *Workshop on Situations, Information, and Semantic Content*, LMU Munich.

Röpke, I., Hahn, F., and Rieser, H. (2013). Interface constructions for gestures accompanying verb phrases. In *Abstracts of the 35th Annual Conference of the German Linguistic Society*, Potsdam, pp. 295–296.

Sangiorgi, D. and Walker, D. (2001). *The π-calculus. A Theory of Mobile Processes.* Cambridge: CUP.