Semantics & Pragmatics Volume 13, Article 8: 1–55, 2020
https://doi.org/10.3765/sp.13.8

Multi-modal meaning — An empirically-founded process algebra approach*

Hannes Rieser
Bielefeld University

Insa Lawler
University of North Carolina at Greensboro

Submitted 2020-04-02 / First decision 2020-06-07 / Revision received 2020-07-06 / Second decision 2020-07-07 / Third decision 2020-07-22 / Revision received 2020-07-27 / Accepted 2020-07-27 / Published 2020-07-30 / Final typeset 2022-05-17

Abstract Humans communicate with different modalities. We offer an account of multi-modal meaning coordination, taking speech-gesture meaning coordination as a prototypical case. We argue that temporal synchrony (plus prosody) does not determine how to coordinate speech meaning and gesture meaning. Challenging cases are asynchrony and broadcasting cases, which are illustrated with empirical data. We propose that a process algebra account satisfies the desiderata. It models gesture and speech as independent but concurrent processes that can communicate flexibly with each other and exchange the same information more than once. The account utilizes the ψ-calculus, allowing for agents, input-output-channels, concurrent processes, and data transport of typed λ-terms. A multi-modal meaning is produced integrating speech meaning and gesture meaning into one semantic package. Two cases of meaning coordination are handled in some detail: the asynchrony between gesture and speech, and the broadcasting of gesture meaning across several dialogue contributions. This account can be generalized to other cases of multi-modal meaning.

Keywords: multi-modal meaning, speech-gesture meaning coordination, process algebra, ψ-calculus, typed λ-calculus, λ-ψ-calculus

* We would like to thank four anonymous reviewers, and (in alphabetical order) Chris Barker, Robin Cooper, Elena Gregoromichelaki, Florian Hahn, Julian Hough, Ruth Kempson, Andy Lücking, the editor Louise McNally, and especially Jim Pryor for valuable comments or discussion. We are also grateful to the audiences in Bielefeld, Bochum, Ghent, Gothenburg, Munich, Stuttgart, and Toulouse. This paper builds on the following conference proceedings: Rieser 2015, 2017, Lawler, Hahn & Rieser 2017.

1 Introduction

Humans do not only communicate by speech. Information can also be communicated with body postures, eye gazes, co-speech gestures, facial expressions, intonation, etc. If any of the latter accompany speech, it seems natural to assume that they build a *meaning unit* for the speaker and the recipient, as McNeill (1992) and others argue. Visual or auditory cues can interact differently with speech (e.g., Ekman & Friesen 1969): They can provide *complementary* information to the information provided by speech. For instance, a deictic gesture can specify the locale in question, or a roundish iconic gesture can indicate the shape of the object described by speech. Visual or audible cues can *enrich* speech information (e.g., Slama-Cazacu 1976, Ladewig 2014, Schlenker 2018), such as gesticulating the shape of an object instead of using an adjective. Gesture information can also *disambiguate* speech information. Consider Figure 1:[1]

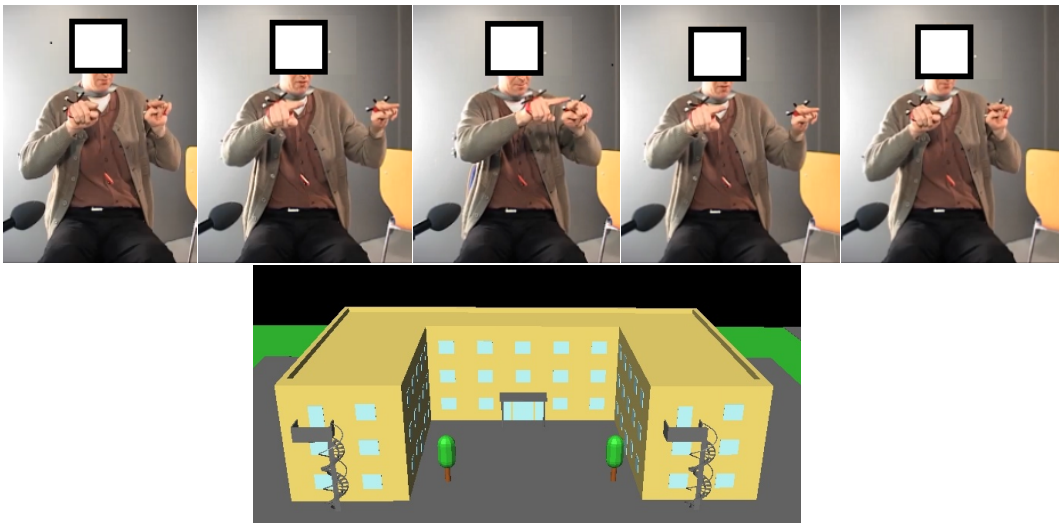

Figure 1 The speaker draws two lines, first straight ahead, and then towards each other. He utters (German): *Das Rathaus ist [dreigeschossig (pause)] wie ein Hufeisen.* English: *The town hall is [three stories tall (pause)] like a horseshoe.* (Brackets mark the gesture overlap.) The town hall is depicted below the stills.

The speaker gesticulates a *cornered* horseshoe while describing a town hall. Since a horseshoe could be round or cornered (as in a classroom), the

1 The video for this datum cannot be made available due to protection of privacy.

gesture disambiguates which kind of horseshoe form is meant. Visual or audible cues can also provide the audience with *meta-information*, such as irony indicated by some intonation patterns (e.g., Kreuz & Roberts 1995, Bryant & Fox Tree 2005, Schlöder 2017) or a skeptical facial expression (e.g., Attardo et al. 2003, Deliensa et al. 2018). We have also explored the contribution of attention and perceptual focus (Velichkovsky, Pomplun & Rieser 1996, Clermont et al. 1998), pointing gestures (Lücking, Pfeiffer & Rieser 2015), and other dialogue relevant gestures (Rieser 2011, Rieser, Bergmann & Kopp 2012) to the overall communicated meaning.

In all these cases, the speaker communicates what we dub a *multi-modal meaning*. The pieces of information communicated via different channels (e.g., visual and audio-acoustic) constitute the overall communicated meaning. To formally model this idea of a multi-modal meaning, one needs a *unified formal framework*. In this paper, we provide a novel process algebra framework for modeling the combination or coordination of speech meaning and non-speech meaning. The key idea is to model the dynamics of this meaning interaction in terms of *independent* but *concurrent* processes that can flexibly interact with each other. As we show, such an approach has important advantages compared to other multi-modal meaning accounts. We illustrate our approach with co-speech gestures, which we take to be a paradigmatic case of multi-modal meaning. But we indicate how our account can be used for other modalities.

Any formal model of non-speech meaning needs to address how non-speech meaning is fixed. For instance, one needs a mapping from annotated eye gaze or gesture information to some (propositional or sub-propositional) semantic representation. In this paper, we work with a standard approach to fixing the meaning of co-speech gestures, as we explain in Section 2. Any account of non-speech meaning also needs to address how non-speech meaning contributes to the overall communicated meaning. In this paper, we focus on the independence and concurrency of speech and gesture. To simplify the illustration of our framework, we assume that gesture meaning and speech meaning combine to a single complex proposition.[2] We also assume that all meaning contributions can be modeled among others with a typed λ-calculus. It is important to stress that the key properties of our framework do

2 We assume that gesture meaning contributes what is called *at-issue* content. Others have argued that gesture meaning contributes *non-at issue* content (for more on this topic see, e.g., Ebert & Ebert 2014, Schlenker 2018, Esipova 2018, 2019). We briefly describe some accounts in Section 5.

not hinge on these assumptions. Our account could be adapted for different sets of hypotheses.

We proceed as follows: In Section 2, we introduce basic information about co-speech-gestures and working hypotheses about gesture meaning. In Section 3, we specify challenges for the hypothesis that the *temporal synchrony* of speech and gesture (plus prosody information) determines how to coordinate their meanings. Challenging cases are *asynchrony cases*, where the gesture stroke comes substantially earlier or later than the suitable speech part, and what we dub *broadcasting cases*, where the gesture meaning is combined with speech meaning more than once. The empirical examples that illustrate these cases also underline that gesture meanings need to be modeled *independently* of speech. In Section 4, we specify desiderata for a speech-gesture meaning coordination account. We need a framework that fully acknowledges both the independence and the concurrency of speech and gesture, that encodes an incremental processing of semantic information, and that enables pieces of information to interact flexibly and more than once. In Section 5, we show that existing co-speech gesture accounts do not deal with all these challenges. In Section 6, we argue that a *process algebra* account, based on the ψ-calculus, fits the bill, and we describe its basics. It treats gesture and speech as independent *processes* that operate concurrently, can communicate with each other flexibly, and can exchange the same information more than once. We illustrate our account by combining the ψ-calculus with an ordinary typed λ-calculus. In Section 7, we apply the process algebra account to two empirical examples, and in Section 8, we indicate how it can be used for other modalities.

2 A case study: Co-speech gestures

Co-speech gestures are spontaneous movements of hands or fingers that do not have a *lexical* meaning. We include all gestures that accompany a speech portion (perhaps interspersed with pauses). Examples of such gestures are pointing gestures and iconic gestures, like the iconic gesture in Figure 1. On Kendon's continuum (Figure 2), such gestures are located at the top. Among other things, this continuum is set up by how much the properties of gesture types resemble linguistic properties. The bottom is made up of sign languages. They have standards of form, a syntax, and so forth.

We work with a speech-gesture corpus that has been annotated guided by annotation manuals and statistically evaluated, namely the Speech and Gesture Alignment (SaGA) corpus. These annotations were shown to be reproducible (Lücking et al. 2013: sect. 2.2). Although we focus in this paper on modeling how to coordinate speech meaning and gesture meaning, we need some working hypotheses about gesture meaning to illustrate our account: We proceed from the popular assumption that the *morphological* features of the gesture stroke (i.e., its kinetic peak) determine the gesture's meaning, such as the handshape used, the shape of trajectory drawn, etc.[3] Such an account was first suggested by Kopp, Tepper & Cassell (2004). The basic assumption is that these features are not arbitrary. The gesture's morphology is described by attribute-value pairs.[4] For an example see Figure 3, which analyzes the gesture in Figure 1.[5] One can compute a gesture's meaning by mapping its attribute-value matrix (AVM) onto a logical formula.

We assume that the meaning of iconic gestures is *sub-propositional*, functioning as a modifier, predicate, full noun phrase, or referring expression. For instance, the gesture in Figure 1 could function as modifying the descriptive information conveyed by the noun phrase *horseshoe*. The town hall looks like a cornered horseshoe rather than a round one. So-called *postholds* of a gesture are holds of the gesture stroke's hand-configuration. We follow McNeill

3 Elsewhere, we argue that the meaning of co-speech gestures depends on the meaning of the accompanying speech (Lawler, Hahn & Rieser 2017, Rieser & Lawler 2020). For more see Footnote 33. For other thoughts on the meaning dependency see, e.g., Lücking 2013: pp. 197-198, Han, Hough & Schlangen 2017.

4 We do not factor in the *gesture space*. Our corpus data show that gesture space varies individually in use, extent and position, and that the extent of a gesture is often not proportional to the object depicted.

5 These annotations are not static because transitions between movements are represented. Their level of fine-grainedness (e.g., one vs. two lines) was tested using computer simulation techniques (Lücking et al. 2013). Dynamic gestures have been successfully simulated using the annotations.

Figure 2 Kendon's continuum (described in McNeill 1992: p. 37)

et al. (2001), Enfield (2004), and Krifka (2007) in assuming that postholds prolong the stroke and its meaning.

Lastly, we focus on *iconic* gestures. In previous research, we examined other kinds of gestures, such as pointing gestures (Lücking, Pfeiffer & Rieser 2015) or gestures that regulate discourse (Hahn & Rieser 2011, Rieser 2011), and we analyzed "mixed" gestures that exhibit iconic and interactive meanings (e.g., postholds that maintain the topic). However, treating all kinds of gestures here would take us too far afield. As will be evident later, our account can in principle accommodate all of them.

3 Challenges for coordinating speech meaning and gesture meaning

A natural starting point for coordinating speech meaning and gesture meaning is the *temporal overlap* of speech and gesture. We first introduce accounts that implement this idea — either in isolation or together with prosody information. Then, we describe two substantial challenges to these accounts that we illustrate with empirical examples.

3.1 Coordinating speech meaning and co-speech gesture meaning via temporal synchrony

In his seminal book *Hand and Mind: What Gestures Reveal about Thought* (1992), McNeill suggests a rule for speech-gesture meaning coordination. Its basic idea is that the *temporal* overlap of speech and gesture determines their

left hand (selected annotation)		right hand (selected annotation)	
attribute	*value*	*attribute*	*value*
gesture kind	drawing	gesture kind	drawing
wrist movement	forward>right	wrist movement	forward>left
path of wrist movement	line>line	path of wrist movement	line>line
two handed configuration	mirror-sagittal	two handed configuration	mirror-sagittal

Figure 3 An AVM of the gesture stroke in Figure 1. The '>' represent the transitions between movements, i.e., the change of hand configurations. The right and the left hand draw two lines in gesture space, while facing each other in a mirror-sagittal manner.

meaning coordination. Gesticulation is aligned with semantically matching speech. McNeill's Semantic Synchrony Rule (SSR) is as follows (1992: p. 27):

> Semantic synchrony means that the two channels, speech and gesture, present [the] same meaning at the same time. The rule can be stated as follows: If gestures and speech co-occur they must cover the same idea unit.

So, McNeill assumes that co-occurring gesture and speech semantically cover the same *idea unit*. An idea unit is a meaning unit above the lexical level, for instance, a verb phrase's meaning. To "present the same meaning" means that the same idea unit is presented. They can do so in a complementary or more redundant way. His example is a speaker who utters *he bends it [i.e., a tree] way back* and in parallel gesticulates the fastening of the tree. The fastening information complements the speech information. Together they express the idea unit that a character is seizing a tree and bending it back (1992: p. 27). According to McNeill, *multiple gestures* represent the idea unit from different perspectives. *Multiple speech clauses* that overlap with a single gesture could be problematic for SSR. But McNeill is confident that the cases are ones where the second clause is semantically a continuation of the gesture stroke (1992: pp. 28-29). If he were right, an answer to the coordination question might be simple: The time span of the speech-gesture overlap determines what speech meaning a given gesture modifies or supplements.

Yet, although SSR works for several paradigmatic cases, subsequent research has challenged it. Upon closer examination, gesture and speech often do not operate in one-to-one temporal synchrony. This becomes clear when one considers annotated data which are time-stamped. A gesture stroke can come substantially earlier or later than its semantically matching speech (for an overview of the literature, see, e.g., Wagner, Malisz & Kopp 2014). For instance, a gesture meaning in the role of a modifier may have to wait until it meets a noun meaning it can combine with. We illustrate asynchrony cases with examples in Section 3.3.

3.2 Coordinating speech meaning and co-speech gesture meaning via temporal synchrony plus prosody information

Several researchers proposed that a temporal constraint together with *prosody* information is decisive for speech-gesture meaning coordination. Kendon

(1972, 1980, 2004), in his work on gestures and natural conversation, observed that gesture strokes are correlated with the onset of a stressed nuclear syllable. McNeill captured this observation in his Phonological Synchrony Rule (1992: p. 26):

> The synchrony rule at this level is that the stroke of the gesture precedes or ends at, but does not follow, the phonological peak syllable of speech (Kendon 1980).

The relation between stroke and nuclear stress has been further explored; for example, McNeill et al. (2001) demonstrate how motion, prosody, intention, and discourse structure are aligned.

This research and work by Johnston (1998) inspired grammar-bound models of speech-gesture integration, such as HPSG approaches (see, e.g., Alahverdzhieva & Lascarides 2010, Lücking 2013), to employ nuclear stress for modeling speech-gesture integration. These accounts combine a temporal constraint with a phonological constraint differently. Alahverdzhieva & Lascarides (2010) employ prosodic word accounts and Klein's (2000) approach to represent phonological structures in an HPSG format to introduce new constraints for speech-gesture coordination (see, e.g., their Situated Prosodic Phrase Constraint).[6] More recently, Alahverdzhieva, Lascarides & Flickinger (2017) introduced options for relaxing this constraint using defeasible inference. Lücking (2013) proposes an alternative: Observing the difference between meter (accent) and rhythm (phonological phrase or tone unit), he stresses a gesture's relation to the information structure of the utterance as manifested in a phonological phrase. The gesture affiliate (the speech portion it is associated with) bears marked accent in the sense of Engdahl & Vallduví (1994, 1996), i.e., it is focused. The accent is on a phonological word.[7]

Exploring the depths and challenges of these approaches would take us too far afield here. For instance, it is controversial whether grammar and phonology can be so closely aligned or whether they enjoy independence

6 Klein (2000) uses prosodic words and metrical trees to represent phonological structures in an HPSG format. Metrical trees model stress assignment in an intonation phrase or tone unit. Using a function mkMtr (make metrical tree), prosodic constituents are set up based on syntactic ones.

7 How intonation structures work in certain types of dialogue is shown in Couper-Kuhlen 2005, 2014.

(cf., e.g., Elordieta 2008, Wagner 2015).[8] There also does not seem to be a full *algorithmic* theory of intonation in sight.[9] However, for our purposes, these controversies do not matter. Even if grammar and phonology could be aligned, there are more substantial problems regarding speech-gesture coordination: As we show in what follows, such accounts cannot do justice to the variety of *asynchrony* cases and to what we call *broadcasting* cases. The upshot of our analysis is that neither the temporal overlap nor the prosodic accent (plus some temporal constraint) fully determines how to coordinate speech meaning and gesture meaning.

3.3 Challenging cases: Asynchrony cases and broadcasting cases

Two substantial challenges for a promising coordination account are asynchrony cases and broadcasting cases. In asynchrony cases, the gesture stroke comes (substantially) earlier or later than the suitable speech part. In broadcasting cases, the gesture meaning is used more than once. In what follows, we illustrate each case with empirical examples.

The initial example is from the SaGA corpus and the subsequent one from an experimental study. SaGA contains 25 route description dialogues generated as follows: In a first step, a so-called Route-Giver "drives" through a virtual reality (VR) town along a route. The second step is to report this ride to a so-called Follower, who is expected to follow the route by her- or himself. In our example, the Route-Giver describes the route into a park and around a pond.[10]

In Tables 1a and 1b, we provide the German wording (n-G), the English close paraphrase (n-E), selective left-hand information (n-LH), and selective right-hand information (n-RH) for the Route-Giver's gestures (see Figure 5) for some number n indicating the order in the sequence. The handshapes named in the left- and right-hand information are depicted in Figure 4. Gesture overlaps are marked with aligned {} or [] brackets.

What can we observe in this transcript?

8 Considering syntax-prosody mismatches as the only relevant datum and thus in contrast to Klein 2000, Haji-Abdolhosseini (2003) develops a calculus of pitch accents and information structure which does not depend on syntax.

9 Loehr (2007)'s data suggest that there is no one-to-one mapping between syntax, tone units, and gesture phrases.

10 The same datum is analyzed in Giorgolo 2010: pp. 98-103 from a Montague grammar perspective. The video for this datum cannot be made available due to protection of privacy.

1-G	Route-Giver:	Wenn du dort eingefahren bist, fährst du [{geradeaus} auf einen Teich zu. Einen Teich.]
1-E		When you have driven in there, you drive [{straight} towards a pond. A pond.]
1-LH		[L-Handshape O]
1-RH		{R-Handshape open O}
2-G	Route-Giver:	[Und an diesem Teich. Du {fährst drauf zu und} du fährst rechts herum.]
2-E		[And at this pond. You {drive towards it and} you drive right around it.]
2-LH		[L-Handshape O]
2-RH		{R-Handshape D}
3-G	Route-Giver:	[Die Hecke, {die geht noch ungefähr} so 50 m.]
3-E		[The hedge, {it runs another roughly} 50 m.]
3-LH		[L-Handshape O]
3-RH		{R-Handshape loose B}
4-G	Route-Giver:	[Und dann sind dort {auch} hin und {wieder} {Sitzbänke}.]
4-E		[Then there are {also} here and {there} {benches}.]
4-LH		[L-Handshape O]
4-RH		{R-Handshape loose D} (first two overlaps) gesture expressing doubt: {wiggling of loose D handshape} (third overlap)
5-G	Route-Giver:	[{Aber du fährst um den Teich herum.} Rechts herum.]
5-E		[{But you drive around the pond.} Right around.]
5-LH		[L-Handshape O]
5-RH		{R-Handshape D}
6-G	Route-Giver:	[Und manchmal ist da auch {nen Eisverkäufer}. Und an dem fährst du rechts ab.]
6-E		[And sometimes there is {an ice-cream man} there. And there you drive off to the right.]
6-LH		[L-Handshape O]
6-RH		{R-Handshape G}

Table 1a A route-description from the SaGA corpus plus key gestures made by the Route-Giver. LH-gesture overlaps are marked with aligned [] brackets. RH-gesture overlaps are marked with aligned {} brackets. Dialogue continues in Table 1b.

7-G	Follower:	Was heißt “manchmal”?
7-E		What does “sometimes” mean?
8-G	Route-Giver:	[Ja, könnte verändert werden. {Auf} meiner Tour war dort ein Eisverkäufer.]
8-E		[Well, could change. {On} my tour there was an ice-cream man there.]
8-LH		[L-Handshape O]
8-RH		{R-Handshape G}

Table 1b Dialogue continued from Table 1a.

(1) The L-Handshape O indicating *rund′* (‘round’) starts well before the word *Teich* (‘pond’) is produced (1-G), namely at *geradeaus* (‘straight towards’). This gesture stroke is too early, so to speak, and thus an *asynchrony* case.

(2) The L-Handshape O is then held until (8-G), i.e., over many contributions. It is a *posthold* of the gesture stroke. A plausible explanation for the long hold is that the pond and the route related to it are the topics of the route description at this stage: The pond is the Route-Giver’s topic from reporting his entering the park and his going toward the pond until he introduces the next landmark (not described). We analyze this posthold as an instance of what we call *broadcasting* cases (see below).[11]

11 As one reviewer remarked, this raises the question of how many tokens are involved across contributions. Following McNeill’s gesture individuation condition (from lap position to lap position), we have two tokens. But if small variations count, we have more.

Figure 4 Basic fingerspelling ASL forms of the handshapes named in Tables 1a and 1b: O, B, D, and G, respectively (images released to Public Domain by user Ds13 in the English Wikipedia on 18th December 2004, https://commons.wikimedia.org/wiki/File:Asl_alphabet_gallaudet.png)

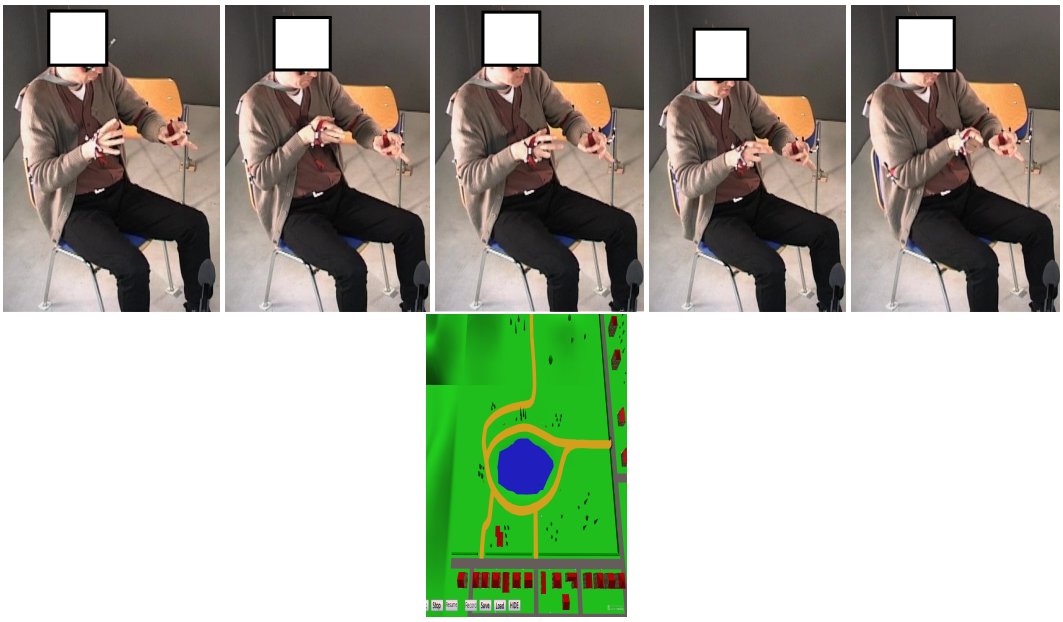

Figure 5 Some of the Route-Giver's gestures (corresponding to the number sequence in Table 1a): 1-LH & 1-RH, 2-LH & 2-RH, 3-LH & 3-RH, 4-LH & 4-RH, 5-LH & 5-RH. The pond is depicted in the picture.

(3) In contrast to the L-Handshape O held constant, the R-Handshape varies among different postures. It delineates the route around the pond (2-RH) using a drawing practice with R-Handshape D. Then it changes to an R-Handshape loose B to indicate a hedge. Afterwards, the R-Handshape loose D is used twice to index two benches followed by a discourse gesture expressing doubt as to the existence of the benches on the Follower's ride (since there could be changes in the Follower's ride through the VR town). In most of these cases, the gesture stroke is not well aligned with the semantically matching speech from a temporal point of view (cf. the brackets). In other words, the datum involves several *asynchrony* cases.

The term *broadcasting* is taken from Gutkovas, Kouzapas & Gay 2016. It means stable information transfer from one source to multiple targets. In our case, this means that *one and the same* gesture meaning can be used for multiple cases of speech-gesture meaning coordination. The posthold described in (2) is arguably such a case. The left hand's (LH) gesture is held throughout many turn-constructional units and turns. The shape formed with the fin-

gers resembles the pond's shape. As McNeill et al. (2001) and Enfield (2004: p. 72) emphasize, in such cases of a gesture posthold, the gesture *meaning* is upheld (see also Krifka 2007: sect. 4). A posthold prolongs the availability of the semantic information. Our example illustrates why the gesture *meaning* is concerned. Without an uphold of LH's gesture meaning, the meaning contributions of the right hand (RH)'s gestures cannot be properly understood. Intuitively, the signing done with RH indicates different objects and the path towards the pond. But RH's gestures *only* represent the driving towards and around the pond if LH still represents the pond. If LH's gesture meaning were no longer available, RH's gestures would be rather meaningless. And the expression *dort* ('there') in (4-G) can only be properly understood with reference to LH's gesture meaning. It cannot be properly resolved considering speech and the indexing of RH alone. *There* refers to a portion of the pond. RH's pointing to the left hand's gesture makes this clear. All this suggests that the information communicated by LH's gesture needs to be re-used. So, we have several cases of multi-modal meaning, featuring one and the same gesture meaning, but different tokens of speech meanings. This is a case of broadcasting because we have a single output-term (gesture meaning) and multiple input-slots with a fitting signature (multiple utterances). This contrasts with RH's gestures. RH supports the introduction of different objects, the path, the hedge, the benches, and the ice-cream man. These are arguably only combined once with speech.

We think that the example illustrates another case of broadcasting: The anaphora in the utterances taking up the multi-modal meaning of *Teich* (i.e., occurrences of *da* in *drauf*, *dort*, *da*) are aligned through the broadcasted information, i.e., various identities are established between the broadcasted meaning and the multi-modal anaphora meaning (say, *runder′-teich′* ('round pond')). That LH's gesture stroke is held across different turns indicates to the Follower: You are still at the pond, whatever the speech says. In (1-G), observe the subtle difference in the function of LH's gesture in *Wenn du dort eingefahren bist, fährst du geradeaus auf einen Teich zu* and in the subsequent *Einen Teich*. We analyze this example in more detail in Section 7.2.

Such broadcasting cases illustrate the *independence* of gesture and speech processes. Gestures can move along with speech, but they need not.[12] A gesture can introduce new meaning or modify an existing one (more than once) *regardless of its precise temporal occurrence*. Moreover, broadcasting points to a fundamental difference between gesture and speech: Gesture informa-

12 McNeill et al. (2001) also observe independency cases.

9-G	Neben {dem Ball} ist eine Kiste.
9-E	Beside {the ball} there is a box.
9-LH	—
9-RH	{R-Handshape D, practice drawing, indicating round.}
10-G	Neben dem Ball {ist eine Kiste.}
10-E	Beside the ball {there is a box.}
10-LH	—
10-RH	{R-Handshape D, practice modeling, indicating box-like}

Table 2 Two cases featuring one and the same utterance but different gestures. In the first case, a roundish gesture stroke overlaps with *dem Ball* and in the second case, a box-like gesture stroke overlaps with *ist eine Kiste*. The gesture overlaps are marked with aligned {} brackets.

tion can be re-used. Speech information cannot be simply re-used (as Asudeh (2012: pp. 95-123) emphasizes).[13]

We also found *asynchrony* cases in our experimental data (Pfeiffer et al. 2019). A special feature of these data is that the *same* gesture occurrence is combined with different utterances and vice versa by re-combining head and torso videos. So, the stimuli are somewhat artificial co-speech occurrences.[14] We consider two cases featuring one and the same utterance but different gestures. The utterance is *Neben dem Ball ist eine Kiste* ('Beside the ball there is a box'). In the first case, a roundish gesture stroke overlaps with *dem Ball* and in the second case, a box-like gesture stroke overlaps with *ist eine Kiste* (see Table 2).[15]

The roundish gesture is depicted in Figure 6. It is classified as an iconic drawing gesture depicting some sort of spiral. So, "roundish gesture" is strictly speaking a misnomer, but we ignore this complication for the moment, and assume that the gesture meaning is $rund'$.

13 Asudeh (2012: pp. 95-123) calls this the *Resource Sensitivity Hypothesis*. We thank one reviewer for pointing this out.

14 All stimuli were tested in a pilot study. No participant rated them as artificial. The study is concerned with whether the subjects ($n > 250$) take into account the gesture shape when selecting objects after a multi-modal input, and whether they judge the same gesture shape differently in different speech contexts. Preliminary results are that gesture shape influences the object selection and that there is a variance in interpreting the same gesture shape (Pfeiffer et al. 2019).

15 The videos are available here: https://doi.org/10.5281/zenodo.3902197

According to traditional semantic theories, both gesture strokes come too *early*. In (9-G), the stroke already overlaps with the definite article. However, the definite article would need a meaning indicating definiteness (if it is represented by an iota-operator). So, the speech meaning cannot straightforwardly fuse with the $rund'$ gesture information. $ball'$ will be the next meaning it can fully integrate with, yielding the multi-modal $(rund'(x) \wedge ball'(x))$. Similarly, in (10-G), the box-like gesture stroke starts with the predication *ist*. It has to let pass ist' and $eine'$ before it can be compositionally combined with $kiste'$. The gesture meaning needs to be temporarily *blocked* or postponed, so to speak, before it can be compositionally combined. So, this example prototypically highlights the independence of gesture and the temporary "blocking" of semantic information.

To sum up, our observations lead to four main results: (I) The precise temporal overlap of speech does not seem to be decisive for speech-gesture meaning coordination. Asynchrony cases are common. (II) Gesture and speech enjoy a considerable independence. They are not always produced simultaneously, and gestures can be held throughout several utterances. (III) In cases of temporal asynchrony of gesture stroke and speech, gesture meaning or speech meaning cannot operate until it can successfully combine with the other. For instance, if a gesture stroke comes too early, its semantic information needs to be suspended or blocked until it can interact with the semantically matching speech part. An account of speech-gesture meaning

Figure 6 Gesture stroke (9-RH, Table 2): a complex trajectory is drawn, resembling a spiral.

coordination must thus capture the introduction, suspension, and interaction of semantic information. (IV) We sometimes need to coordinate one and the same gesture meaning more than once with speech meaning.

4 Desiderata of a satisfying account of speech-gesture meaning coordination

Considering our observations, a satisfying account of speech-gesture meaning coordination should meet the following desiderata:

(a) *Asynchrony*: A satisfying account should accommodate cases where gesture strokes come (substantially) earlier or later than the suitable speech part, i.e., cases where gestures introduce new meaning or modify an existing meaning regardless of their precise temporal occurrence.

(b) *Independence*: A satisfying account should accommodate the independence of gesture and speech, for instance, it must accommodate cases where gesture strokes are held throughout several utterances.

(c) *Blocking*: A satisfying account should allow for the blocking or postponing of semantic information.

(d) *Broadcasting*: A satisfying account should accommodate broadcasting cases, e.g., by allowing for the replication or repetition of meaning pieces.

We also add the desideratum that speech-gesture meaning coordination should be determined *algorithmically*. There should be perspicuity regarding how the gesture meaning coordinates with speech meaning, and the coordination should not be represented in an *ad hoc* fashion but rather be the result of (finite) rule-bound procedures. This enables systematically explaining speech-gesture meaning coordination and generalizing to a variety of data and contexts.

(e) *Algorithmic determination*: A satisfying account should algorithmically determine a gesture's speech relatum and its coordination term.

These desiderata call for a dynamic machinery. Phrased in terms of processes, we need output processes which can give semantic information a

"piggyback ride" and we need input processes which receive this semantic information, get it, and hand it on to the right place, the "right place" being (as a rule) information already existing. Traditional formal accounts in linguistics and philosophy of language analyze whole sentences. However, the phenomena described above require more dynamic models, such as the ones that have been provided by (Segmented) Discourse Representation Theory ((S)DRT) (e.g., Kamp & Reyle 1993, van Eijk & Kamp 2011, Asher & Lascarides 2003), Poesio-Traum Theory (PTT) (e.g., Poesio & Rieser 2010, 2011), Dynamic Syntax (e.g., Kempson et al. 2016), and Type Theory with Records (TTR) (e.g., Cooper 2012, 2017, 2020), where we have incrementally incoming data, structures assigned to these, and updates of information. Dynamic Syntax is especially suited to account for the idea that communicative information is processed by bits, so-called increments. More specifically, incrementality means that syntactic information is read in word-by-word/construction-by-construction and the matching semantic information is considered in a similar way. In this paper, we focus on incrementality in speech-gesture meaning coordination.

Before we present our own account of speech-gesture meaning coordination, we examine whether existing co-speech gesture accounts could meet all specified desiderata.

5 Why existing co-speech gesture accounts do not fully meet the challenges

Although until recently gestures were not widely studied within formal semantics, there are a couple of accounts that inspired or pursued formal modeling. The gesture research initiated by Kendon (1972, 1980, 2004) and McNeill (1992), McNeill et al. (2001) was put on a more systematic footing by computational modeling, where research on four domains was decisive: the collection, annotation, and statistical evaluation of multi-modal corpora; the specification of gesture meaning using formal tools; the use of formal grammars for the description of speech events; and the set-up of models integrating speech meaning and gesture meaning. Since the early 2000s, corpora of multi-modal data have been collected and systematically annotated (e.g., Paggio & Navarretta 2009, Loehr 2007, Lücking et al. 2013). Annotation had to be time-stamped and precise, as far as handshapes and hand postures (palm, back, fingers, wrist) go, to produce life-like avatars. Computational simulation has acted as a testbed for the adequacy of annotation. Observations of

the speech-gesture trade-off and pointing experiments led to the idea of an integrated multi-modal semantics (usually called *multi-modal fusion*).[16] Empirical pointing research acted as a precursor to multi-modal research with a wider empirical coverage. Papers such as Kopp, Tepper & Cassell 2004 cover several of these developments. The idea of an integrated speech gesture semantics appeared independently in a series of papers, see Kopp, Tepper & Cassell 2004, Rieser 2004, Lascarides & Stone 2006, 2009, Lücking, Rieser & Staudacher 2006a,b. In what follows, we show that the most prominent co-speech gesture accounts cannot meet all of the desiderata specified above. Primarily, they don't do enough justice to the fact that speech and gesture are *independent* but *concurrent* processes. This is crucial for modeling asynchrony, blocking, and broadcasting.

5.1 Planners for multi-modal integration

Kopp, Tepper & Cassell (2004) developed a multi-modal micro-planner for iconic gestures and accompanying speech. Their empirical basis was a corpus of route-giving directions. The micro-planner consisted of a novel gesture planner and the system SPUD (Sentence Planning Using Descriptions, Stone et al. 2013). Gesture form features were represented as AVMs based on systematic annotation. An intermediate level of gesture meaning representation was constructed (comparable to the techniques used in SaGA annotations) and mapped onto form features. The planner outputted dynamical lexical entries for gestures. SPUD combined these with lexicalized tree-adjoining grammar (LTAG) entries and generated a multi-modal semantic representation passed on to the surface realization component. So, gestures were analyzed as sub-propositional meaning contributions, and speech meaning and gesture meaning were combined into a single complex proposition. This early account uses an *algorithmic determination* for speech-gesture meaning coordination based on LTAG. Kopp, Tepper & Cassell (2004) concentrate on a speech-gesture *symmetry* case and develop a synchronization solution. They do not deal with phenomena like asynchrony, blocking, or broadcasting, although they seem to be aware of them.

16 Computational pointing research was the first to develop rigid models but was mainly concerned with the topology of the pointing cone (e.g., van der Sluis 2005, Kranstedt et al. 2006, Lücking, Pfeiffer & Rieser 2015) and its computational approximation.

5.2 Grammar-based accounts

Even earlier on, Cohen et al. 1997 is the first work we know of where speech-gesture meaning coordination was modeled. Johnston's (1998) use of typed AVMs, unification, and the temporal constraint for speech-gesture correlation paved the way for subsequent HPSG models. Mainly because of the idea that gesture production depends also on prosody, such as the (nuclear) accent, there was a move from LTAG (following, e.g., Abeillé & Rambow 2000 as used, e.g., in Kopp, Tepper & Cassell 2004, Rieser 2004) to HPSG formalisms, where supra-segmental phonological information can easily be accommodated.

In this manner, Alahverdzhieva & Lascarides (2010) provide an *underspecification* account of gesture meaning, based on Robust Minimal Recursion Semantics' (RMRS) notion of elementary predicates (e.g., Copestake 2007), implemented in an HPSG grammar. Gesture meaning is taken to be derived inferentially via a hierarchy of predicates starting from a root labeled with a gesture form term. They consider prosody, syntax-semantics, and timing as the essential factors determining speech-gesture meaning coordination. They cover cases where speech and gesture do not precisely overlap, achieved with their 'Situated Prosodic Phrase Constraint' rule making use of prosodic constituents. The formalism works with an underspecified speech-gesture relation *vis_rel* to be specified by pragmatic inference as set up in Lascarides & Stone 2009 (see below).

Lücking (2013) advances similar arguments concerning speech-gesture synchrony as Alahverdzhieva & Lascarides (2010), using an annotation-based account of gesture meaning. The affiliate of a gesture (the speech portion it is intuitively associated with) is taken to be marked by *nuclear accent*. The speech-gesture relation is given in an HPSG account for German (Müller 2007) making use of temporal speech gesture overlaps. Gesture semantics is implemented in a vector semantics framework (Zwarts 1997, Zwarts & Winter 2000). An updated version of this theory resolving cases of underspecification with principles of Gestalt theory is provided in Lücking 2016. It is formulated in a Type Theory with Records (TTR) format (cf. Cooper 2012, 2017) and uses information state update technology; semantics and temporal conditions are as in Lücking 2013.

Another move to more complex grammar formats was the implementation of Giorgolo's Montague Grammar approach to speech gesture integration (e.g., Giorgolo 2010) in a Lexical Functional Grammar (LFG) account

(Giorgolo & Asudeh 2011). Both accounts derive the meaning of gestures from annotations. Giorgolo (2010) follows a reconstruction strategy as in Johnston 1998 and Kopp, Tepper & Cassell 2004, and models essentially a synchronous case based on two maps: one goes from linguistic structure to a spatial frame of reference and the other from the observable gesture to the space created by it. Verbal meaning and gestural meaning are fused by a meet operation.

In all these grammar-based accounts, speech and gesture meaning are combined into a single proposition, but the speech-gesture meaning coordination is achieved differently. Giorgolo's (2010), Giorgolo and Asudeh's (2011), and Lücking's (2016) proposals meet the *algorithmic determination* desideratum; they allow for an algorithmic speech-gesture meaning coordination. However, these and the other grammar-based accounts cannot (straightforwardly) meet the other desiderata that we specified. As already noted in Lücking 2013, grammar-bound approaches cannot straightforwardly deal with *independent* co-speech gestures which introduce information that is not affiliated with a speech part, since, e.g., in HPSG-, LFG-, or LTAG-terms, there is no speech element it can be unified with. *Asynchronous* cases are partially captured by some analyses. However, the cited works do not deal with more extreme cases of asynchrony. Some accounts might have the resources to deal with cases where semantic information needs to be (temporally) *blocked*; but it is not clear from their accounts. *Broadcasting*, such as integrating one and the same information more than once, poses the biggest challenge to existing co-speech gesture analyses. For instance, in HPSG analyses, the gesture content attaches to an affiliate in the directly related speech portion. Postholds cannot be (straightforwardly) handled because they overlap with speech they are not directly related to, for example, with a next turn.

5.3 SDRT accounts

To date, Lascarides & Stone 2009 is the most comprehensive study on gesture semantics/pragmatics. It is based on a Segmented Discourse Representation Theory (SDRT) interface (cf. Asher & Lascarides 2003). Roughly, SDRT is characterized by its use of rhetorical relations like Elaboration, Background, or Narration to establish coherence links between discourse units, a right border constraint modeling how in discourse new content can be glued on to an old content, and its information update machinery, needed, for example, for consistency checks and anaphora resolution. Lascarides and Stone develop a

hitherto uncontested logic of gestural space, drawing a distinction between reference within gesture space and external reference, i.e., between a gesticulated entity and an external one. SDRT's rhetorical relations for verbal discourse are extended with new veridical relations Depiction, Replication, and Overlay to establish gesture-gesture meaning coordination as well as speech-gesture meaning coordination. Gestural content itself is determined inferentially by common-sense reasoning allowing for underspecification. The resolution of gesture meaning deploys a hierarchy of 'increasingly specific properties' starting with some gesture form predicate like *hand_shape_asl-a* (following Kopp, Tepper & Cassell 2004) and finally arriving at a property like *sustain*. The content inferred is used to build up units of discourse and to establish rhetorical relations with verbal discourse contributions. In this way, gesture-generated propositions can become part of the hierarchical discourse structure. Furthermore, they provide a dynamic semantics model theory for SDRSs, i.e., SDRT representations, as far as we know, the only one existing in gesture research.

Alahverdzhieva, Lascarides & Flickinger (2017) extend Alahverdzhieva & Lascarides 2010 using SDRT as a coherence-based model of pragmatics and RMRS as the tool for the resolution of underspecification. Instead of their earlier synchrony notion, they have *alignment*, which is not equivalent to temporal simultaneity. They investigate cases where gestures precede or follow the intended speech relatum or where gesture covers more speech material than the intended reading would suggest. In this respect, their section 'Temporal and Prosodic Relaxation' is instructive.

These two SDRT accounts satisfy some of our desiderata. Their formal devices illustrate that they treat co-speech gestures as communicating semantic information that is *independent* of the speech's semantic information, they also allow for an *algorithmic* speech-gesture meaning coordination, and they can analyze some *asynchrony* cases. They might possibly have resources to deal with severe asynchrony cases, the blocking of semantic information, and broadcasting. But they do not analyze such cases, and it is not clear how this should be achieved using their analyses.

5.4 Other formal pragmatic accounts

Recently, the nature of speech-gesture meaning coordination has received increasing attention elsewhere in the formal semantics/pragmatics literature. A shared idea is that gestures often provide so-called *non-at issue* in-

formation. According to Ebert & Ebert (2014), the semantic contribution of co-speech gestures can be treated like the semantic contribution of *appositive* relative clauses. In effect, gesture meaning and speech meaning yield a truth value pair when combined, using Potts (2005)' framework for appositive relative clauses. According to Schlenker (2018)'s approach, the semantic contribution of some iconic co-speech gestures can be treated akin to the semantic contribution of *presuppositions*. An expression with the content p which co-occurs with a gesture with content g comes with the requirement that the local context of p should guarantee that p entails g. According to Schlenker (2018), the timing of a gesture can significantly alter its semantic status. For instance, only co-speech gestures are treated akin to presuppositions. The contribution of post-speech gestures (i.e., gestures that come after the speech portion they modify) is akin to that of appositive clauses. Esipova (2018, 2019) challenges the idea that temporal alignment is decisive. She argues that it depends on syntax-semantics and syntax-prosody interaction whether gesture content is at-issue or non-at-issue.

It is worth discussing whether gestures contribute content that is akin to that of appositive clauses or presuppositions. However, in their current form, such accounts do not meet all our desiderata. Such accounts can model some *asynchrony* cases (e.g., post-speech gestures). But as far as we can see, other kinds of asynchrony cases or the *blocking* of information are not treated. It is not clear whether such accounts can model *broadcasting*. As far as we know, they do not treat postholds. Analyzing gesture meaning in terms of appositive clauses or presuppositions suggests that the meaning of the gestures is heavily dependent on the co-occurring speech. If so, the *independence* of speech and gesture does not seem to be fully accommodated. Finally, speech-gesture meaning coordination is not *algorithmically* determined.

5.5 Upshot

The upshot of our analysis is that while all these accounts have made important progress to understanding speech-gesture meaning coordination, they cannot fully cope with the challenges specified earlier. They don't do enough justice to the fact that gesture and speech are *independent* but *concurrent*, especially regarding broadcasting cases. So, the account that we offer in what follows covers a research field *complementary* to current formal gesture research (cf. Table 3). Although underspecification (as modeled in Alahverdzhieva and Lascarides' and Lücking's works) is not our concern in

this paper, our account has the resources for modeling underspecification, as we indicate further below.

Accounts \ Desiderata	Sub-propos.	Propos.	Ind.	Asyn.	Blocking	Broad-casting	Algor. determ.
Kopp, Tepper & Cassell 2004	✓	✗	✓	✗	✗	✗	✓
Giorgolo 2010, Giorgolo & Asudeh 2011	✓	✗	✓	(✓)	✗	✗	✓
Alahverdzhieva & Lascarides 2010	✓	(✗)	✗	(✓)	✗	✗	(✗)
Alahverdzhieva, Lascarides & Flickinger 2017	(✗)	✓	✓	(✓)	✗	✗	✓
Lücking 2013	✓	✗	✗	✗	✗	✗	(✗)
Lücking 2016	✓	✗	✗	✗	✗	✗	✓
Lascarides & Stone 2009	✗	✓	✓	(✓)	✗	✗	✓
Ebert & Ebert 2014	✗	✓	(✓)	(✓)	✗	✗	✗
Schlenker 2018	✓	✓	(✓)	(✓)	✗	✗	✗
Esipova 2018	✓	✓	(✓)	(✓)	✗	✗	✗

Table 3 Comparison of the accounts: '✓' means 'has been treated,' '(✓)' means 'has been partially treated, '✗' means 'has not been treated,' '(✗)' means 'could perhaps be treated.'

6 A process algebra account of speech-gesture meaning coordination

Standard tools in linguistics, such as the λ-calculus, model phenomena that are either atemporal or *sequential*, as Barendregt (1981–2012: p. 6) explicitly notes. Such a limitation might not be problematic regarding speech meaning (abstracting away from intonation, etc.). Utterances (in an idealized sense) are arguably sequential, for example, an uttered word is followed by another uttered word. However, this limitation renders the λ-calculus not well suited for modeling speech-gesture meaning coordination. As we highlighted, gesture and speech often occur in parallel or partially overlap. Gesture-speech occurrences are non-linear, as Johnston (1998: p. 626) puts it. They are *concurrent* events, hence one needs a model entertaining concurrency. In addition, semantic information in the λ-calculus is *in situ*, once it has been

inserted. By contrast, one and the same gesture information can contribute to multi-modal meaning more than once (broadcasting). That is why we need a calculus that is not limited to sequential events and can processes information more flexibly.

When Barendregt made his comment (1981), research in parallel lambda calculi had already started, perhaps the most comprehensive study to date being still Dezani-Ciancaglini 1997. However, parallel lambda accounts only accommodate concurrency and non-determinism, but not flexible data transport between processes, which we need. We suggest that a *process algebra* account is able to cope with the independence of gesture meaning and speech meaning, cases of asynchrony, cases of blocking of information, cases of broadcasting, and an algorithmic meaning coordination. Process algebras are formal systems working with so-called *concurrent agents*. The basic idea is that these agents exchange information or data using so-called input-output channels. Being fairly abstract, such algebras can be used to model a number of different dynamics. For example, process algebras have been used to describe the goal-oriented behavior of social insects (Tofts 1992; after Fokkink 2000). Another example is Milner (1989)'s model of workers using a common set of tools to iteratively produce workpieces, or a scheduler organizing a recursive succession of actions. Also, everyday devices like pocket calculators or smartphones can be modeled using process algebras; the users and their devices are concurrent systems. For all these applications, process algebra implementations exist. We extend the range of applications to speech and gesture, and we point to a wealth of other multi-modal examples in Section 8.

In what follows, we first informally explain this account, then we provide the formal details, and illustrate our approach with empirical examples. Note that we reserve the notion of speech-gesture *synchrony* for temporal synchrony as read off from time-stamped annotation, and subsume the cases of gesture precedence, sequence, autonomous gestures or non-overlapping cases under speech-gesture *asynchrony*. This allows us to maintain a rigid notion of synchrony. We consider asynchrony to be the normal case favoring the dynamic approach we propose.

6.1 The process algebra account: An informal introduction

A process algebra that allows for several communicating agents seems to be a good fit for modeling speech-gesture meaning coordination. To reflect the *independence* of gesture and speech, we need at least two sets of indepen-

dent meaning carriers, so to speak. We need one set for the speech meaning and at least one for the gesture meaning. To capture the idea that semantic information is *incrementally* built, we suggest modeling speech meaning and gesture meaning as *ongoing dynamic processes* which run concurrently. These processes function as the meaning carriers. We call what they carry the processes' *data*. We use a typed λ-calculus for the semantic analyses of speech meaning. The process for the speech meaning is considered to transport typed λ-terms (which can be incrementally built). The typed λ-terms are the data. Syntax and (supra-segmental) phonology may be conceived as processes that carry phonological or syntax data; but we do not model these.

Our account strictly follows the pace of incoming speech, leading to an *incrementality* analysis. So, as a by-product, we analyze non-regimented speech. Successively incoming bits of speech also determine the speech and gesture processes, scope regularities, and much else.[17]

To implement the idea that gesture and speech can communicate a *joint* meaning, our process algebra account allows for combining or exchanging information from different carriers, so to speak. This is standardly conceived of as a communication between the concurrent processes. The easiest way to realize this communication is to model the processes as transporting terms of the *same formal set-up*. We thus model the gesture process as transporting typed λ-terms. We obtained these λ-terms from rigid annotation of raw data. To implement the exchange of information we use input-output (i/o) processes which operate concurrently. These processes work on a shared channel (see below).

I/o channels should not operate unconstrained. So, we need a mechanism that *restricts* the communication between channels and is defined on sets of concurrent communicating speech and gesture processes. Cases of asynchrony and the blocking of information require a *flexible* mechanism for an exchange of information at the right time. Our desideratum for an algorithmic speech-gesture meaning coordination requires an *algorithmic* mechanism. We cope with these desiderata among others by treating the transported λ-terms (our data) and transporting channels as *typed*. Roughly speaking, our mechanism for meaning coordination examines whether the currently transported λ-term and the transporting channel of the gesture process are of a fitting type. If that is the case — and only if it is — the i/o processes will operate. The data can only go through channels they agree with in

17 The idea tied up with incrementality is similar to Dynamic Syntax, although we focus on semantic matters.

terms of types. So, if a gesture meaning does not fit the simultaneous speech meaning it does not interface with it. Their integration is *blocked*. Recall the case of the roundish gesture that overlapped with the definite description operator. If a fitting speech meaning occurs shortly after, then interfacing takes place through the i/o channels. In this way, cases of asynchrony can be easily dealt with. So, for example, a gesture-meaning output process can send a value to a speech-meaning input process; both then generate a multimodal meaning if no deadlock occurs (e.g., due to incompatible meanings). In Section 7.1, we give a detailed example to show how this mechanism operates and how asynchrony cases and blocking cases are analyzed.

In cases of *broadcasting*, semantic information from one carrier (e.g., the gesture carrier) is used more than once.[18] Our account thus needs to be able to continuously exchange information for some time period and distribute one and the same information to several input slots. We achieve this by employing the replication agent '!' (see below).[19] It operates on a process as a whole. Its main function is to replicate the λ-terms currently transported. The use of the replication operator is empirically constrained. Currently, we use it for cases of gesture postholds. We also observe that speakers copy their gestures or their addressee's gestures. One might use the replication operator for modeling these cases, too.

To implement all these proposals, we use the ψ-calculus, transformed into a λ-ψ-calculus. We explain how it works, from a formal point of view, in what follows. Basically, the input-output system and the concurrency come from the ψ-calculus, and the speech and gesture data transported from the typed λ-calculus. To improve readability, we suppress the types.[20]

18 As Louise McNally pointed out, there might be cases of broadcasting in speech via repetition. In principle, such cases can be modeled with our account, e.g., the indefinite information *a pond* (cf. Table 1a, (1-E)) can be replicated.

19 One reviewer pointed out that the replication operator ! resembles the of course modality (!) in linear logic. There, (!) indicates that a premise can be used arbitrarily often (Asudeh 2012: p. 101, see also the exponential rules in Di Cosmo & Miller 2019). This parallel sets the resource role of data in the λ-ψ-calculus into a broader context. For the difference between '!' in the ψ-calculus and in LFG-based theorizing, cf. Kehler et al. 1999.

20 In the λ-ψ-calculus, the following terms are typed: (a) constants and variables of the λ-calculus, (b) ψ-channels transporting λ-expressions and their parameters, (c) variables/parameters in the interface of λ-expressions and ψ-expressions. See also Footnote 34.

6.2 The process algebra account: A formal introduction

The ψ-calculus (Bengtson et al. 2011) is a version of process algebra (Fokkink 2000) developed out of Milner's π-calculus (see Parrow 2001 for an overview). To our knowledge, the integration of the λ-calculus and the ψ-calculus has not been carried out in computer science or linguistics so far. We use the input-output (i/o) operators and the concurrency operator from the ψ-calculus, and data transported in the typed λ-calculus using i/o channels. What do we have in the ψ-calculus to model the intuitions laid out? We have parameters, operators on these, frames, and agents (see below and Johansson 2010). The data terms can come from any (higher order) logic.[21] In some process algebras, such as the original π-calculus, variables only get associated with i/o channels, but in our algebra, they are also associated with arbitrary data (e.g., our typed λ-terms). (The variables are also called *names*.) Channels help us transport data from one increment of a linguistic utterance to another. The parameters indicating ψ's syntactic categories are given in Definition 1 (adapted from Bengtson et al. 2011: pp. 4-14).[22]

Definition 1

C	the conditions, ranged over by ϕ
A	the assertions, ranged over by ψ
T	the (data) terms or structures, ranged over by N

An example for a condition C would be the antecedent of a conventional if-then-else construction. Assertions A can be used, for instance, to fix the environment of a process operating, for example in a derivation (see Johansson 2010 for details). Data terms T will be exploited in the description of our examples, where the familiar typed λ-calculus is chosen.

The central dynamic elements of the ψ-calculus are so-called *agents* (also called *processes*). Roughly speaking, their function is to embody a variety of information, for instance, semantic information, communication information (i.e., information to give (output) and to expect (input)), and interface information (more in Section 7.1.1). We notate agents using P, Q, … , and channels using M, as illustrated in Definition 2. The deadlock δ is taken from

21 Strictly speaking, two types of semantics are involved here; the model-theoretic semantics of the typed λ-calculus and the operational semantics of the ψ-calculus based on labeled transition systems. The possibility of a λ-ψ-hybrid has been suggested by the developers of the ψ-calculus (cf. Johansson 2010, p. 4). We acknowledge that the integration of the λ-calculus with the ψ-calculus is not a trivial step and demands an in-depth discussion.

22 We use a monotype font for variables or constants of the ψ-calculus to better distinguish them from variables or constants of the λ-calculus.

Fokkink 2000: pp. 7, 25 and is used instead of the Fregean ⊥ from the standard ψ-calculus.[23]

Definition 2

0	Nil, 0-agent, an inactive agent
$\overline{\mathtt{M}}\mathtt{N.P}$	Output
$\underline{\mathtt{M}}\mathtt{x.P}$	Input
$\mathtt{case}\ \phi_1\mathtt{:}\ \mathtt{P}_1 \parallel \ldots \parallel \phi_n\mathtt{:}\ \mathtt{P}_n$	Case construct
$\mathtt{P} \mid \mathtt{Q}$	Parallel/Concurrent
$!\mathtt{P}$	Replication
δ	Deadlock

The syntax '.' separates a prefix from the subsequent agent. 0 and δ can be regarded as *atomic agents*.[24] The 0-agent is inactive. Deadlock δ is used for semantic violation (more in Section 7.1.3). The difference with 0 is that 0 represents non-action, in the sense of idling. By contrast, after δ no further action is possible (in this respect, it is like the Fregean ⊥).

$\overline{\mathtt{M}}\mathtt{N.P}$ (M overbar, N dot P) puts a data structure N onto output channel M, sends it out, and continues with agent P, possibly a 0-agent. One could use this agent to transport a typed λ-term N elsewhere. $\underline{\mathtt{M}}\mathtt{x.P}$ indicates that a data structure (in our implementation, a typed λ-term) is received on the input channel M and substituted for x in P. This construction binds the variable x in P. The role of $\underline{\mathtt{M}}\mathtt{x.}$ is that of a prefix, followed by the agent P.[25]

The case construct $\mathtt{case}\ \phi_1\mathtt{:}\ \mathtt{P}_1 \parallel \ldots \parallel \phi_n\mathtt{:}\ \mathtt{P}_n$ employs the conditions mentioned in Definition 1. The construction will reduce to one of the agents $\mathtt{P}_1 \ldots \mathtt{P}_n$, depending on which of the conditions ϕ_i is true. (The choice is non-deterministic if several $\phi_1 \ldots \phi_n$ are true.). Employing ϕ and $\neg\phi$, this

23 This definition is lacking (| ψ |): an assertion-agent. In contrast to an assertion ψ (see Definition 1), (| ψ |) can be used to insert additional information going along with an agent into a derivation. The definition is also lacking the restriction agent $(\upsilon\ \alpha)\mathtt{P}$. It ensures that the scope of the variable/name α is local to P. This entails that we cannot have an output channel $\overline{\alpha}$ out of P. Hence, one can use this agent to specify purely local information. In our application, these two agents are not needed.

24 This term is not used in the ψ-calculus literature but nicely marks the contrast with, for example, $\overline{\mathtt{M}}\mathtt{N.P}$ and $\underline{\mathtt{M}}\mathtt{x.P}$.

25 In the ψ-calculus literature, $\underline{\mathtt{M}}\mathtt{x.P}$ is given as $\underline{\mathtt{M}}(\lambda\tilde{\mathtt{x}})\mathtt{N.P}$. These λs bind a sequence of variables x in N and P in the ψ-calculus, unlike the λs in the typed λ-terms that we use as data terms. We thus leave out the λs in the ψ-calculus constructions. Although the definition for input is $\underline{\mathtt{M}}(\lambda\tilde{\mathtt{x}})\mathtt{N.P}$, in our application we need only $\underline{\mathtt{M}}\mathtt{x.P}$ with N = empty string; we use only one typed input variable x carrying all the information we need. Further, instead of meta-variables M for channels, we will use $\underline{\mathtt{ch}_\mathtt{i}}$ for input and $\overline{\mathtt{ch}_\mathtt{i}}$ for output, where shared $\mathtt{i} \in \mathtt{N}^+$ indicates identity.

can be used to implement the more conventional `if` ϕ `then P else Q` (i.e., `case` ϕ_1: P_1 || $\neg\phi$: P_2). In what follows, we only make use of this derived construction.

The parallel/concurrent agent '`P | Q`' enables `P` and `Q` to expand independently or to communicate with each other via output and input operators, perhaps after several independent expansions. Here is an example of how agents $\overline{\mathsf{M}}\mathsf{N}.\mathsf{P}$ and $\underline{\mathsf{M}'}\mathsf{x}.\mathsf{Q}$ interact under |:

$$\overline{\mathsf{M}}\mathsf{N}.\mathsf{P} \mid \underline{\mathsf{M}'}\mathsf{x}.\mathsf{Q} \longrightarrow \mathsf{P} \mid \mathsf{Q}[\mathsf{N}/\mathsf{x}], \text{ when } \mathsf{M} \leftrightarrow \mathsf{M}'$$

On the left hand side of the '$\longrightarrow$', agents $\overline{\mathsf{M}}\mathsf{N}.\mathsf{P}$ and $\underline{\mathsf{M}'}\mathsf{x}.\mathsf{Q}$ are coordinated by the concurrent operator | as are `P` and `Q[N/x]` on the right hand side. The arrow $\longrightarrow$ indicates a transition relation, determined by the operational semantics of the ψ-calculus (cf. Johansson 2010). Assume a datum `b` such as a typed λ-expression associated with `N`. The datum exits the left agent $\overline{\mathsf{M}}\mathsf{N}.\mathsf{P}$ via channel $\overline{\mathsf{M}}$; agent `P` (which might be `0`, inactive) remains. Assuming $\mathsf{M} \leftrightarrow \mathsf{M}'$, i.e., the channels are equivalent as in Definition 3 below, `b` enters M' and substitutes for `x` in `Q`. So, we end up with:

$$\mathsf{P} \mid \mathsf{Q}[\mathsf{b}/\mathsf{x}].$$

Below, we usually place type constraints on `M`, `N`, and `x.` | proves central for our concerns. Among others, it is used to model the concurrency between gesture and speech.

Replication `!P` is our replication agent and is understood as equivalent to `P|!P`, which means that `P` can be emitted arbitrarily often.[26]

In addition to the agents, we also have *operators*. The equivariant operators (*equivariance* defined by α-equivalence, see Johansson 2010: p. 40) are given in Definition 3.[27]

Definition 3

$\leftrightarrow: \mathsf{T} \times \mathsf{T} \to \mathsf{C}$	Channel Equivalence
$\otimes\ \mathsf{A} \times \mathsf{A} \to \mathsf{A}$	Composition
$\vdash\ \subseteq \mathsf{A} \times \mathsf{C}$	Entailment

Channel equivalence is used to identify input- and output-channels. We will express it by sub-indexing (e.g., ch_i). The channels to be identified receive

26 Consequently, the λ-ψ-system we set up is not resource sensitive as discussed in, e.g., Asudeh 2012.

27 α-equivalence means that bound variables can be renamed in a term without changing its meaning. So, basically, Definition 3 specifies equivalence classes.

the same name and sub-index. Composition is equivalent to conjunction, and entailment is comparable to ordinary entailment (but it relates an assertion A to a condition C).

For our descriptive aims, only a thinned-out version of the ψ-calculus is needed; using `if-then-else` instead of the more general case construct; and of the equivariant operators, using only channel equivalence, which we represent more simply using sub-indexing. So, we work with the following fragment:

Definition 1 (reduced)

C	the conditions, ranged over by ϕ
T	the (data) terms or structures, ranged over by N

Definition 2 (reduced)

0	Nil, 0-agent
$\overline{M}$N.P	Output
$\underline{M}$x.P	Input
case ϕ_1: P_1 $\|\|$ $\neg\phi$: P_2	Case construct, here as canonical `if-then-else`
P | Q	Parallel/Concurrent
!P	Replication
δ	Deadlock

Definition 3 (reduced)

$\leftrightarrow$: T $\times$ T $\rightarrow$ C	Channel Equivalence, represented as 'ch' plus sub-index marking identity

In what follows, we simplify the λ-ψ representations in the following way:

1. 0-agents terminating a derivation are sometimes omitted.
2. The syntax device '.' separating prefix and follow-up agent P is usually omitted.

Let us go back to our desiderata of speech-gesture meaning coordination. Our account ensures the desideratum of *independence* by formalizing speech meaning and gesture meaning as independent agents communicating via their i/o facilities. *Asynchrony* is captured by input and output processes "crossing" the concurrency operator in a type constrained way, as explained

in the commentary to Definition 2. This is also closely related to the desideratum of *blocking*, i.e., postponing information. Blocking can easily be handled by sub-indexing, the typing of i/o channels, and the data they transport: if the types do not agree, data discharge is blocked. In the next section, we illustrate blocking with a case where the gesture comes 'too early.' But we are aware of data where speech comes before a matching gesture (e.g., post-speech gestures). Such cases can be modeled using the same techniques. *Broadcasting*, such as in the case of postholds, where one kind of information is repeatedly emitted for some time, can be modeled via replication: replication outputs one agent after another looking for a corresponding input channel which might be turns away. Observe that this is different from the handling of linguistic antecedent-anaphora resolution. Speech-gesture meaning coordination is determined *algorithmically* by the λ-ψ-machinery, i.e., the choice of agents, the ψ- and λ-constructions that contain them (see below), and the types. We do not model broadcasting in a technical way in this paper, but as far as trans-propositional anaphora is concerned see Section 7.2.[28]

7 The process algebra account: Applications

We illustrate our account by modeling some of the empirical examples discussed in Section 3.3. In doing so, we show how we use ψ's output channels, input channels, the concurrency operator '|', and how they can be combined with typed λ-structures and λ-techniques.

7.1 The round-ball example

Recall the round-ball example from our recent experimental study (Figure 6). The intuition to be modeled about speech-gesture meaning coordination is shown in Figure 7.[29]

28 The agents in our account exhibit underspecification, since they only have channel information at the beginning for providing information, receiving it, or both, which means that their i/o channels implemented for composition are underspecified. For the underspecification of constants dealing with hyponymy or missing arguments we would resort to underspecification accounts that are germane to our λ-categorial representations (e.g., Egg & Koller 2001).

29 As Andy Lücking emphasized, we idealize here. We omit co-articulation facts between the phonemes /n/ and /d/ and /m/ and /b/, respectively. The sign '$<$' for 'earlier than' presupposes that a segmentation in words can be given.

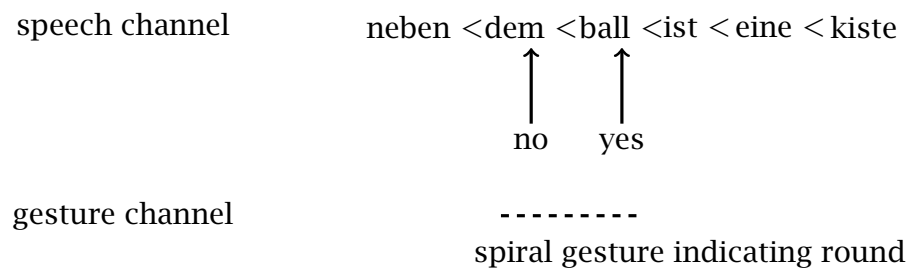

Figure 7 Intuition about speech-gesture meaning coordination for the round-ball example: *Neben dem Ball ist eine Kiste.* (English: *Beside the ball there is a box.*) The gesture stroke overlap is marked with a dashed line.

The gesture meaning $rund'$ must be blocked from interacting with dem' and made to interact with $ball'$. We achieve this by exploiting the fact that transported λ-terms and transporting channels are typed.

We begin by illustrating the general rendering of λ-ψ-interaction in our example (Section 7.1.1). Then, we zoom in on two details: We explain the λ-ψ-agents of our example and how they interact (Section 7.1.2), and the role of δ (Section 7.1.3).

7.1.1 General rendering of the λ-ψ-interaction

The main idea of a process algebra is that one can specify agents or processes according to Definitions 1 to 3. In the following, $\underline{\mathsf{ch}}_{\mathsf{i}}$ represent input channels, $\overline{\mathsf{ch}}_{\mathsf{i}}$ the corresponding output channels as indicated by i, and we use the usual higher-order λ-techniques abstracting over typed variables. In applications of the ψ-calculus, *agents* can be different entities, interacting buffers, schedulers, timers, sliding windows or complex machines. In our application (to our knowledge, undertaken for the first time), agents are semantic contributions of words and gestures. In our example, the incoming words are *neben*<*dem*<*Ball*<*ist*<*eine*<*Kiste* as well as the incoming spiral gesture overlapping the words *dem* and *Ball*. Recall that an agent can encode a variety of information. In our case, every word- or gesture agent gets three sorts of information:

- its compositional meaning information expressed in terms of the typed λ-calculus,

- its communication potential for input, output, case, parallel/concurrent, replication or deadlock expressed by the ψ-agents, and
- interface information for λ-variables and ψ-variables.

In the λ-ψ-interface, we make frequent use of function composition. Use of function composition in non-interface expressions is according to λ-categorial standards. Interfaces link the typed λ-calculus with the ψ-calculus. More on these below. Since all relevant information is typed, we do not need parentheses for the ψ-calculus layers, but only for the λ-terms. To increase readability, λ-expressions acting as functions are enclosed in '$\langle \ldots \rangle$'.

The parallel/concurrent construction, the input-output channels, and the specification of expressions by the type system are the regulating mechanism for information flow. Here is an example: Using the ι-operator in a Russell-Reichenbach style, the λ-representation for *dem* (in *dem Ball*) is $\lambda F(\iota x(F(x)))$ (types left implicit here). It needs a one-place predicate to form a term. The communication potential of $\lambda F(\iota x(F(x)))$ is achieved by the following input-output constellation of channels $\underline{\mathsf{ch}_3}$ and $\overline{\mathsf{ch}_4}$: $\underline{\mathsf{ch}_3}$ is accompanied by what the ψ-literature calls a *name* (i.e., a variable). Let this name be br (inspired by *Ball* and *rund*). It is supplied as an argument to λF. The resulting datum is then transported out by $\overline{\mathsf{ch}_4}$. That accomplished, $\mathsf{0}$ remains. In the λ-ψ-calculus, we write this as follows:

$$\underline{\mathsf{ch}_3}\ \mathsf{br}\ \overline{\mathsf{ch}_4}\langle \lambda F(\iota x(F(x)))\rangle\ (\mathsf{br}).\mathsf{0}$$

The interface between the λ-variables and the ψ-variables is given by br coming from the ψ-calculus, F coming from the typed λ-calculus, and the application of $\lambda F(\iota x(F(x)))$ to br. Given that, we get, for example, *rund*$'$ and *ball*$'$ via $\underline{\mathsf{ch}_3}$ substituting for br and finally for F. We end up with $\iota x(ball'(x) \wedge rund'(x))$. $\iota x(ball'(x) \wedge rund'(x))$ can in turn be exported to the outside to combine with some other word agent. It exits by the output channel $\overline{\mathsf{ch}_4}$. In order for this analysis to work, we must assume the following type structure (Definition 4):

Definition 4

λ-terms:	$F \in \mathrm{T}_{\langle e,t\rangle}$
	$\iota x\phi \in \mathrm{T}_e$ (when $\phi \in \mathrm{T}_t$)
ψ-channels and names:	$\underline{\mathsf{ch}_3} \in \mathrm{T}_{\langle e,t\rangle}$
	$\mathsf{br} \in \mathrm{T}_{\langle e,t\rangle}$
	$\overline{\mathsf{ch}_4} \in \mathrm{T}_e$

Figure 8 shows the value passing of all word agents. On the x-axis, we see the incoming words and the spiral gesture, and on the y-axis the time intervals. States are represented by interacting word agents, here simply abbreviated as primed word tokens. The arrows relating the semantic terms correspond to actions among agents, mimicking the relations of the operational semantics. Thus, the arrows represent the channels transporting the information between agents. The arrow head identifies the target of the information. For instance, the channel ch_2 transports the *rund′* information to *ball′* (target).

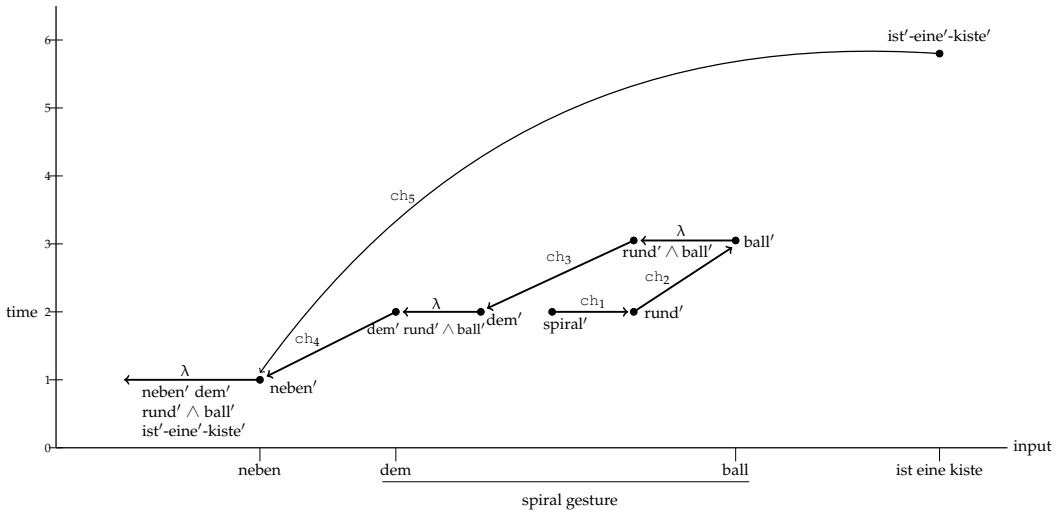

Figure 8 State space of interacting word agents and their values

At the time interval 1, the agent *neben′* is produced. The incoming spiral gesture overlaps with *dem* < *Ball* (these are the words) and produces *rund′* communicated via ch_2 to *ball′*. ch_3 passes the content of *ball′* and *rund′* to the definite-article-agent *dem′* adding up to a definite description agent. The definite description agent in turn interacts with *neben′* via ch_4 and this yields roughly *neben′ dem′ rund′* ∧ *ball′*. Observe that all interaction is first concentrated on generating the multi-modal meaning. The *neben′*-agent has to "wait" for input until this has been achieved.[30] The *ist′-eine′-kiste′*-agent is built up and interacts with *neben′* using ch_5. For simplicity, the input

30 As one reviewer remarked, in a strictly incremental account of language one would expect that the gesture meaning *rund′* does not have to wait for the fully specified noun phrase to be correctly attached. One would expect a linguistic string of *the round …* to be well-formed

ist<*eine*<*Kiste* is represented in Figure 8 as one agent *ist′-eine′-kiste′* and does not conform to the stepwise input.

As Figure 8 shows, both the linguistic and the gestural input are read in by increments. Based on that, the transport of the λ-expressions is guided by two aims: first, the integration of the gesture meaning using *rund′* to build up the multi-modal definite description *dem′ rund′* ∧ *ball′* and to integrate it with the preposition *neben′* to get *neben′ dem′ rund′* ∧ *ball′*. Secondly, the combination of this meaning piece with *ist′-eine′-kiste′*. The compilation of the sentence meaning is due to the fronted constituent (an original corpus datum) *neben′ dem′ rund′* ∧ *ball′*. A more standard German word order would be *Die Kiste ist neben dem Ball* (English: *The box is beside the ball.*) with the subject first. In the latter case, the integration point could be at the end of the utterance, requiring different actions among the agents and, consequently, a different constellation of channels: then, the box-information would communicate channel-wise with the property-information *ist-neben-dem-Ball* via an output-input-facility.

Let us now delve further into the formal analysis. Recall that *rund′* is supposed to interact with *ball′*. As Figure 8 indicates, *rund′* is communicated via ch_2 to *ball′* and yields *rund′* ∧ *ball′*. Here is how we achieve this formally: The function $\langle \lambda F \lambda x (ball'(x) \wedge F(x)) \rangle$ is applied to the argument (ru) ('ru' inspired by *rund*),[31] which must agree type-wise with the ψ-input channel $\underline{\mathsf{ch}_2}$ and the λ-variable F:

(1) $$\underline{\mathsf{ch}_2}\ \mathsf{ru}\ \overline{\mathsf{ch}_3}\langle \lambda F \lambda x (ball'(x) \wedge F(x)) \rangle (\mathsf{ru}).0$$

Assume further that *rund′* has been computed by some λ-ψ-agent, was sent out, and enters into (1) via $\underline{\mathsf{ch}_2}$. Given that, we have exactly the input-structure $\underline{\mathsf{M}}\mathsf{x}.\mathsf{P}$ of Definition 2 above. The property *rund′* is substituted for the variable ru and ends up replacing F according to λ-β-conversion. So, we get $\lambda x (ball'(x) \wedge rund'(x))$, which in turn can leave via $\overline{\mathsf{ch}_3}$, looking

but partial (i.e., underspecified) in a way that other strings would not be (e.g., *Brianna round* …). Thus, it should be feasible that *rund′* does not have to wait to be correctly attached (though as yet incomplete). We agree that this is a viable alternative. We can also establish a solution along these lines (left out here for reasons of space). The solutions are equivalent from a truth-conditional semantics perspective. If one wants to have both solutions, this can be accommodated by ψ's non-deterministic choice (not discussed in this paper).

31 In the following, λ-variables are abbreviated as usual. For ψ-names (i.e., variables) we use the first letter(s) of the constant(s) which will finally become the value of the name. We do this to facilitate the reader's understanding of the formulas. However, there is no inherent connection between the ψ-name string and its final value.

for an appropriate identical input-channel. As shown in (1 (steps)) we have now, since $\underline{\mathsf{ch}}_2$ ru is used up, ~~$\underline{\mathsf{ch}}_2$~~ ~~ru~~ $\overline{\mathsf{ch}}_3$ $\langle\lambda x(ball'(x) \wedge rund'(x))\rangle$.0, and the resulting structure instantiates $\overline{\mathsf{M}}\mathsf{N}.\mathsf{P}$ in Definition 2 above, given suitable typing.

(1 (steps))

$$\begin{aligned}
&\underline{\mathsf{ch}}_2\ \mathsf{ru}\ \overline{\mathsf{ch}}_3\langle\lambda F\lambda x(ball'(x) \wedge F(x))\rangle(\mathsf{ru}).0 && \\
&\overline{\mathsf{ch}}_3\langle\lambda F\lambda x(ball'(x) \wedge F(x))\rangle(rund').0 && \text{by } \underline{\mathsf{M}}\mathsf{x}.\mathsf{P} \\
&\overline{\mathsf{ch}}_3\langle\lambda x(ball'(x) \wedge rund'(x))\rangle.0 && \text{by } \lambda\text{-}\beta\text{-conversion} \\
&.0 && \text{by } \overline{\mathsf{M}}\mathsf{N}.\mathsf{P}
\end{aligned}$$

Recall the representation of *dem*:

(2) $$\underline{\mathsf{ch}}_3\ \mathsf{br}\ \overline{\mathsf{ch}}_4\langle\lambda F(\iota x(F(x)))\rangle(\mathsf{br}).0$$

It says: Via the input channel $\underline{\mathsf{ch}}_3$ some property must come along, substitute for br, and finally for F. Consequently, as shown in (2 (steps)), $\lambda x(ball'(x) \wedge rund'(x))$ can now enter via $\underline{\mathsf{ch}}_3$ and substitute in the end for F, so that we get $\iota x(ball'(x) \wedge rund'(x))$ for *dem runden Ball*, as desired. It is transported out via $\overline{\mathsf{ch}}_4$ and can cooperate with *neben'* (for details see Section 7.1.2).

(2 (steps))

$$\begin{aligned}
&\underline{\mathsf{ch}}_3\ \mathsf{br}\ \overline{\mathsf{ch}}_4\langle\lambda F(\iota x(F(x)))\rangle(\mathsf{br}).0 && \\
&\overline{\mathsf{ch}}_4\langle\lambda F(\iota x(F(x)))\rangle(\lambda x(ball'(x) \wedge rund'(x))).0 && \text{by } \underline{\mathsf{M}}\mathsf{x}.\mathsf{P} \\
&\overline{\mathsf{ch}}_4(\iota x(ball'(x) \wedge rund'(x))).0 && \text{by } \lambda\text{-}\beta\text{-conversion} \\
&.0 && \text{by } \overline{\mathsf{M}}\mathsf{N}.\mathsf{P}
\end{aligned}$$

Observe that the λ-terms in (1 (steps)) and (2 (steps)) function in the end as data entering output channels.

So far, we have demonstrated some aspects of the λ-ψ-calculus' input-output facility. We use it, for instance, to compute the property of a referential term and to send it to its main function, here the definite article. In the next section, we demonstrate the use of the concurrency operator '|' and we show how the multi-modal meaning *dem' rund' ball'* is fused with the predication *ist' eine' kiste'*. Due to incrementality, the information for the prepositional phrase *neben dem Ball* must "wait" until it can be so combined.

In our account, all channels and variables are typed. This plays a key role in avoiding overgeneralization concerning speech-gesture meaning coordination. For instance, one might worry that we cannot deal with a gesture overlap

like, say, *dem Ball ist*.[32] Here, the gesture meaning *rund′* "fires" parallel to different speech constituents (using the replication agent '!': $\overline{\mathsf{ch}}_i.\mathit{rund}'.0 \mid !\overline{\mathsf{ch}}_i.\mathit{rund}'.0$). *rund′* should only fuse with *ball′* but with neither *dem′* nor *ist′*. We achieve this by typing of the data and i/o channels. The gesture meaning would find no appropriate input channels for a fusion with *dem′* or *ist′*. In Section 7.2, we show that an overgeneralization is also avoided if a gesture stroke is held across turns, as, for instance, in the case of broadcasting.

7.1.2 The λ-ψ-agents and how they interact

In what follows, we illustrate the λ-ψ-agents, how they interact, and give a complete analysis of the example. Recall that the roundish gesture intuitively expresses *rund′*, but is depicting some sort of spiral. Building on our work in Pfeiffer et al. 2013, we assume that if the spiral approximates a circle (relative to a contextually specified threshold), then we get the gesture meaning *rund′* (else, as we discuss below, we get δ, which indicates semantic inconsistency). Elsewhere, we argue that the if-else needs to be more complex (Lawler, Hahn & Rieser 2017, Rieser & Lawler 2020). But to simplify the illustration of our process algebra account, we work with the simpler if-else.[33]

In our analysis of the round-ball example, we number the agents according to the multi-modal input. Importantly, '/*' (after the parallel/concurrent operator |) is *not* an operator but indicates the beginning of a comment. The indexing of the channels indicates the information flow and determines the communication between channels allowed. As we mentioned before, the typing of the channels and variables is important, but since the typing regime is straightforward, we specify it only in Footnote 34.[34]

Let us start with the *neben′*-agent (Agent 1).

32 One reviewer provided this as a test case.

33 The guiding idea of the more complex if-else is that the gesture's interpretation depends on the extension of the accompanying speech part, for instance, the roundness interpretation depends on the *Ball* extension. What we call the *final* meaning of a co-speech gesture is thus speech-dependent. For instance, if the spiral gesture overlapped with *Treppenhaus* ('staircase'), its final meaning would be *spiralig′* rather than *rund′*. Our proposal amounts to integrating bits of the semantic model into the semantic representation, i.e., treating meta-language information as a parameter value in the object language. Using structured lexical entries, for example for *Ball* or *Treppenhaus*, the approximation functions could be provided in a more fine-grained manner by extensions using AVMs in the obvious way, and the relevant extension could be tied to some specific (dialogue) context. Lawler, Hahn & Rieser 2017 and Rieser & Lawler 2020 develop these ideas in more detail.

34 The following channels and variables are of T_e: $\overline{\mathsf{ch}}_4$, $\mathtt{drb}$, r, u, v, x.

(Agent 1)
$\underline{\mathsf{ch}_4}\ \mathsf{drb}\ \underline{\mathsf{ch}_5}\ \mathsf{ek}\ \overline{\mathsf{ch}_6}\langle\lambda x\lambda\mathscr{P}.\mathscr{P}(\lambda u(neben'(x,u)))\rangle(\mathsf{drb})(\mathsf{ek}).0 \mid /^*\ neben'$

The comment illustrates that the verbal meaning input is *neben′*. The prefix of the *neben′*-agent begins with the input channels $\underline{\mathsf{ch}_4}$ and $\underline{\mathsf{ch}_5}$. (Agent 1) cannot yet distribute information, but only internally work on incoming pieces of information. It needs input via the input channels $\underline{\mathsf{ch}_4}$ and $\underline{\mathsf{ch}_5}$, before it can output something via the output channel $\overline{\mathsf{ch}_6}$. We use the variable drb inspired by *dem runden Ball* and ek inspired by *eine Kiste*. For instance, *dem′ rund′ ball′* will substitute for drb and (thanks to λ-β-conversion) for x.

Due to the temporal overlap of *dem* and the spiral gesture, the spiral gesture agent works concurrently with the *dem′*-agent, as (Agent 2) shows.

(Agent 2)
$(\underline{\mathsf{ch}_3}\ \mathsf{br}\ \overline{\mathsf{ch}_4}\langle\lambda F(\iota x F(x))\rangle(\mathsf{br}).0) \mid \overline{\mathsf{ch}_1}(spiral').0 \mid /^*\ dem'$ and spiral gesture

The left-hand side of (Agent 2) is the already familiar *dem′*-agent (cf. (2) above). The right-hand side of (Agent 2) is the spiral gesture agent and begins by outputting *spiral′* on $\overline{\mathsf{ch}_1}$.

spiral′ will enter $\underline{\mathsf{ch}_1}$ of (Agent 3), which is the approximation if-else agent, to substitute for sp (inspired by *spiral*).

(Agent 3)
$$\underline{\mathsf{ch}_1}\ \mathsf{sp}.\ \text{if}\langle\lambda z\exists x\exists r(z = spiral' \wedge (\text{approximates}(f_c(spiral'), c, x) = r) \wedge r \geq \text{threshold}_c \wedge circle'(x))\rangle(\mathsf{sp})\ \text{then}\ \overline{\mathsf{ch}_2}\ rund'.0\ \text{else}\ \delta \mid /^*\ spiral'\ \text{going to}\ rund'\ \text{or}\ \delta.$$

(Agent 3) contains an if-then-else construction working as follows: If the input to $\underline{\mathsf{ch}_1}$, instantiating sp, yields $spiral' = spiral'$, and the projection of *spiral′*, $f_c(spiral')$, approximates *circle′* in context c to degree $r \geq \text{threshold}_c$ in c, then *rund′* is output on $\overline{\mathsf{ch}_2}$, else we get the deadlock δ, and there is no follow-up action.

The following channels, variables, and constants are of $\mathrm{T}_{\langle e,t\rangle}$: $\overline{\mathsf{ch}_1}$, $\underline{\mathsf{ch}_1}$, $\overline{\mathsf{ch}_2}$, $\underline{\mathsf{ch}_2}$, $\overline{\mathsf{ch}_3}$, $\underline{\mathsf{ch}_3}$, F, G, H, *ball′*, *circle′*, *rund′*, ru, *spiral′*, sp, br, z.
The following constant is of $\mathrm{T}_{\langle e,\langle e,t\rangle\rangle}$: *neben′*.
The following channels and variables are of $\mathrm{T}_{\langle\langle e,t\rangle,t\rangle}$: $\overline{\mathsf{ch}_5}$, $\underline{\mathsf{ch}_5}$, ek, $\mathscr{P}$.
The following variable is of $\mathrm{T}_{\langle\langle e,t\rangle,\langle\langle e,t\rangle,t\rangle\rangle}$: $\mathscr{Q}$.
The following channel is of T_t: $\overline{\mathsf{ch}_6}$.

In our example, the if-clause is satisfied and thus *rund′* is output via $\overline{\mathsf{ch}}_2$. *rund′* can now enter via $\underline{\mathsf{ch}}_2$ and ru into (Agent 4), which is the already familiar *ball′*-agent (cf. (1) above).[35]

(Agent 4) $\quad \underline{\mathsf{ch}}_2\ \mathsf{ru}\ \overline{\mathsf{ch}}_3\langle\lambda F\lambda x(\mathit{ball}'(x) \wedge F(x))\rangle(\mathsf{ru}).0 \mid /^* \mathit{ball}'$

rund′ substitutes for ru and (thanks to λ-β-conversion) for the λ-variable F, and we get $\lambda x(\mathit{ball}'(x) \wedge \mathit{rund}'(x))$ as the multi-modal meaning of *ball* and the concurrent spiral gesture (which was interpreted as *rund′*).

$\lambda x(\mathit{ball}'(x) \wedge \mathit{rund}'(x))$ can now move onto the output channel $\overline{\mathsf{ch}}_3$ and communicate the property to the *dem′*-agent (Agent 2). As seen in (2 (steps)), $\lambda x(\mathit{ball}'(x) \wedge \mathit{rund}'(x))$ substitutes for br. After λ-β-conversion, we get $\iota x(\mathit{ball}'(x) \wedge \mathit{rund}'(x))$, which can leave via $\overline{\mathsf{ch}}_4$.

The result so far is the multi-modal representation of a round ball. The *neben′*-agent (Agent 1) is now in a position to accept input because the other agents have done their duties. Via $\overline{\mathsf{ch}}_4$, $\iota x(\mathit{ball}'(x) \wedge \mathit{rund}'(x))$ substitutes for drb and moves into the argument slot of *neben′*. This yields:

$$\underline{\mathsf{ch}}_5\ \mathsf{ek}\ \overline{\mathsf{ch}}_6\langle\lambda\mathscr{P}.\mathscr{P}(\lambda u(\mathit{neben}'(\iota x(\mathit{ball}'(x) \wedge \mathit{rund}'(x)), u)))\rangle(\mathsf{ek}).0 \mid$$

The input channel $\underline{\mathsf{ch}}_5$ needs input before $\overline{\mathsf{ch}}_6$ can output something. The *ist′-eine′-kiste′* representation can enter $\underline{\mathsf{ch}}_5$, substitute for *ek*, and, finally, move into the λ-variable position $\mathscr{P}$.

Let us turn to the *ist′-eine′-kiste′*-agent (Agent 5).

(Agent 5) $\quad \overline{\mathsf{ch}}_5\langle\lambda\mathscr{Q}.\mathscr{Q}(\lambda u\ \exists v(v = u))\rangle(\langle\lambda FGH\ \exists r(F(r) \wedge G(r) \wedge H(r))\rangle(\mathit{kiste}')).0 \mid /^* \mathit{ist}'\text{-}\mathit{eine}'\text{-}\mathit{kiste}'$

(Agent 5) must yield a formula conforming to the output channel $\overline{\mathsf{ch}}_5$. The datum which $\overline{\mathsf{ch}}_5$ will in the end have to transport is computed from the term involving the identity for the predication *ist* $(\lambda\mathscr{Q}.\mathscr{Q}(\lambda u\ \exists v(v = u)))$ and its argument $(\langle\lambda FGH\ \exists r(F(r) \wedge G(r) \wedge H(r))\rangle\, \mathit{kiste}')$. The argument has itself a functor-argument structure. Its functor represents the existential quantification for *eine* and the argument *kiste′* encodes its restriction. We first evaluate the argument and get *kiste′* inside the formula substituting for F, which results in $\lambda GH\ \exists r(\mathit{kiste}'(r) \wedge G(r) \wedge H(r))$. We now have

35 The gesture information is treated as a modifier of the noun-semantics *ball′*. Although our system differs substantially from Giorgolo (2010)'s, our solution is similar to the one Giorgolo (2010: pp. 93-98) proposed for *two* in *two towers*.

two λ-bound predicate positions left, G and H, and thus a formula of the type we need for λ-β-conversion with $(\lambda\mathcal{Q}.\mathcal{Q}(\lambda u\ \exists v(v = u)))$, given suitable typing. Applying the functor $(\lambda\mathcal{Q}.\mathcal{Q}(\lambda u\ \exists v(v = u)))$ to the argument results in $\lambda GH\ \exists r(\mathit{kiste}'(r) \wedge G(r) \wedge H(r))$ being a functor for the argument $(\lambda u\ \exists v(v = u))$. The argument is of a type matching that of G. Hence, we arrive at $\lambda H\ \exists r(\mathit{kiste}'(r) \wedge (\lambda u\ \exists v(v = u))(r) \wedge H(r))$, which is equivalent to $\lambda H\ \exists r(\mathit{kiste}'(r) \wedge \exists v(v = r) \wedge H(r))$.

Assuming consistent typing, $\lambda H\ \exists r(\mathit{kiste}'(r) \wedge \exists v(v = r) \wedge H(r))$ is a suitable expression to go on the output channel $\overline{\mathsf{ch}_5}$. This representation of *ist eine Kiste* enters the input-channel $\underline{\mathsf{ch}_5}$ (on the *neben*′-agent, (Agent 1)), substitutes for ek, and finally moves into the λ-position $\mathcal{P}$. This concludes the channel-communication; after standard λ-β-conversions, we end up with the following λ-expression, a single type-theoretical proposition for a multi-modal utterance:

$$\overline{\mathsf{ch}_6}\exists r(\mathit{kiste}'(r) \wedge \exists v(v = r) \wedge \mathit{neben}'(\iota x(\mathit{ball}'(x) \wedge \mathit{rund}'(x)), r)).0\ |$$

This proposition is the multi-modal representation of *Neben dem Ball ist eine Kiste* accompanied by the roundish gesture. The proposition is ready for output via the output channel $\overline{\mathsf{ch}_6}$, for instance, to interact with other (multi-modal) meaning contributions. Importantly, the multi-modal representation entails the speech meaning $\exists r(\mathit{kiste}'(r) \wedge \exists v(v = r) \wedge \mathit{neben}'(\iota x(\mathit{ball}'(x)), r))$.

This concludes our formal analysis of the *Neben dem Ball ist eine Kiste* example. In the next subsection, we say more about the role of δ.

7.1.3 The role of deadlock δ

Finally, let us take a closer look at the approximation if-else agent (Agent 3).

(Agent 3)

$$\underline{\mathsf{ch}_1}\ \mathsf{sp}.\ \text{if}\langle\lambda z\exists x\exists r(z = \mathit{spiral}' \wedge (\text{approximates}(f_c(\mathit{spiral}'), c, x) = r) \wedge r \geq \text{threshold}_c \wedge \mathit{circle}'(x))\rangle(\mathsf{sp})\ \text{then}\ \overline{\mathsf{ch}_2}\ \ \mathit{rund}'.0\ \text{else}\ \delta|/^*\ \mathit{spiral}'\ \text{going to}\ \mathit{rund}'\ \text{or}\ \delta.$$

In our example analysis, we assume that the antecedent of the if-else is true; the spiral gesture approximates a circle. The consequent then outputs *rund*′ for the gesture. However, if the approximation of the projection of *spiral*′, $f_c(\mathit{spiral}')$, is below a threshold$_c$, i.e., the gesture is not round

enough, then deadlock δ results. Recall that δ is used for semantic violation. After δ no further action is possible. Consequently, the flow of information gets stuck, and we do not obtain a multi-modal meaning because of a failed co-speech gesture interpretation.

Another case where deadlock would arise can be observed in our corpus data. There are cases where gestures *contradict* speech. In the following datum (Table 4), it is discussed what the fountain depicted in Figure 9 looks like.[36]

12:42-G	Router-Giver	Das sieht dann aus wie [zwei umgedrehte Tassen.]
12:42-E		This then looks like [two upside-down cups].
12:42-LH & RH		[Each hand shapes a *right-way-up* cup.]
13:28-G	Follower:	Brunnen … ist hellblau und [besteht aus zwei umgedrehten Tassen.]
13:28-E		The fountain … is light blue and [has two upside-down cups.]
13:28-LH & RH		[Each hand shapes an *upside-down* cup.]
13:34-G	Route-Giver:	[Säule mit so ner umgedrehten Untertasse]
13:34-E		[Pillar with such an upside-down cup]
13:28-LH		[LH shapes a *right-way-up* cup.]

Table 4 An excerpt from a route-description from the SaGA corpus (min. 12:42-13:34) plus the relevant gestures. Gesture overlaps are marked with aligned [] brackets.

That *umgedrehte Tassen* is accompanied with a gesture shaping a right-way-up cup would yield δ. Contradictory information cannot be joined. Interestingly, when the Follower repeats the fountain description, he gesticulates upside-down cups. The Route-Giver apparently notices that. When he corrects the Follower's repetition in (13:34-E), the Route-Giver reproduces the original speech-gesture mismatch: an upside-down cup expressed in speech, combined with a gesture indicating a right-way-up cup. In fact, the fountain features objects that look like right-way-up cups (cf. Figure 9). So, the Route-Giver's speech is incorrect but not the gesture. We don't know how to model such cases yet.

36 The video for this datum cannot be made available due to protection of privacy.

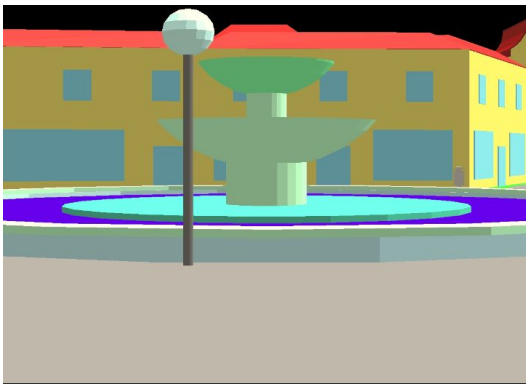

Figure 9 Fountain described in Table 4.

7.2 Broadcasting and multi-modal anaphora

So far, we were concerned with intra-propositional matters observable from the study data. However, in the transcript in Tables 1a and 1b (pp. 10–11) we have other phenomena, like the re-use of gesture information and multi-modal anaphora, that the λ-ψ-calculus can adequately model. Perhaps the most conspicuous trait in the transcript is the holding of a round shape, associated with *Teich* ('pond') across seven contributions and two turns. We have argued that the meaning of LH's gesture needs to be re-used. Without upholding its meaning, the meaning contribution of RH's gestures cannot be properly understood. For instance, RH's gesture only represents the driving around the pond if LH continues to represent the pond. Without going into all technical details, which are reserved for a follow-up paper, we sketch how our account can be used to model such a case of broadcasting.

From (1-LH), we get the $rund'$ information conveyed by the L-Handshape O. It arguably modifies *einen Teich*. It is communicated that the pond is round. The left hand is held in this O shape across the following turns. As we suggested earlier, the round information is not just communicated once. Due to the gesture hold, the $rund'$ information is provided across dialogue contributions and turns. We can model this by applying the replication operator '!' to the semantic information $rund'$.[37] The replication operator emits a copy of an agent and continues with a replication. So, using $\overline{\mathsf{ch}}_i$ for an

37 A reviewer pointed out that Lascarides & Stone 2009 use the notion *replication* as follows: "[…] Replication, which relates successive gestures that use the body in the same way to depict the same entities." (p. 406) This probably covers McNeill (2000)'s *catchments*, not mentioned in their paper. Despite the terminological coincidence, this is different from our

output channel and *rund′* for the replicated information, we get (Agent 1 (Broadcasting)).

(Agent 1 (Broadcasting)) $!\overline{\mathsf{ch}}_i\ \lambda z.\, rund'(z).0$

(Agent 1 (Broadcasting)) enables the *rund′* information to be distributed. This expands to $\overline{\mathsf{ch}}_i\ \lambda z.\, rund'(z).0 \mid !\overline{\mathsf{ch}}_i\ \lambda z.\, rund'(z).0$.

For (1-G), we can provide (Agent 2 (Broadcasting)), using ru ('ru' inspired by *rund*).

(Agent 2 (Broadcasting)) $\underline{\mathsf{ch}}_i\ \mathsf{ru}\ \overline{\mathsf{ch}}_j\langle\lambda F \exists x(teich'(x) \wedge F(x))\rangle(\mathsf{ru}).0$

If we input $\lambda z.\, rund'(z)$ as substitute for ru, we arrive at $\exists x(teich'(x) \wedge rund'(x))$. By an anaphora rule, as suggested in Poesio & Rieser 2011: sec. 5.4, we introduce the definite $\iota x(teich'(x) \wedge rund'(x))$ for *da* in *drauf* in (2-G). Furthermore, *da* will be rendered as a constant n, where $n = \iota x(teich'(x) \wedge rund'(x))$.

Now we have two options: We can say that $!\overline{\mathsf{ch}}_i\lambda z.\, rund'(z).0$ continues to generate *rund′*, but its role has been semantically fulfilled. It is only there for information-structural reasons (e.g., to hold the topic, which we do not deal with here). The other option is that we use another $\lambda z.rund'(z)$ as an input test for the definite description. It, so to speak, controls whether the *rund′* information is correct, and that is all it has to do. This amounts to using (Agent 3 (Broadcasting)) for anaphora.

(Agent 3 (Broadcasting))
$\underline{\mathsf{ch}}_i\ \mathsf{ru}\ \overline{\mathsf{ch}}_j\langle\lambda F \iota x(teich'(x) \wedge rund'(x) = F(x))\rangle(\mathsf{ru}).0$

In (2-G), we have the expression *an diesem Teich*. Having used the replication operator and eliminating the test-*rund′* due to the equivalence $rund'(x) = F(x)$, we get (Agent 4 (Broadcasting)).

(Agent 4 (Broadcasting)) $\overline{\mathsf{ch}}_j\langle\iota x(teich'(x) \wedge rund'(x))\rangle(\mathsf{ru}).0$

This yields *diesem runden Teich*, again confirming to the *rund′* gesture. In (2-G), we also have one occurrence of the anaphora *da* in *drauf*. The *da-*

notion which is bound to post-holds of a single gesture. Also, the technical realization is entirely different.

anaphora gets the *rund′*-information from the multi-modal antecedent *diesem runden Teich*.[38]

In (3-G) and (4-G), the *rund′*-property is held suspended without being used in a linguistic construction, but we still need a copy of it for further use. This is again a replication case. We thus have (Agent 5 (Broadcasting)) for *dort′* in (4-G).

(Agent 5 (Broadcasting))

$$\underline{\mathsf{ch}}_{\mathsf{i}}\ \mathsf{ru}\ \overline{\mathsf{ch}}_{\mathsf{j}}(dort' = \langle\lambda F \iota x(teich'(x) \wedge rund'(x) = F(x))\rangle(\mathsf{ru})).0$$

In (5-G), we also get *rund′* for *den Teich* and *Rechts herum* from its nearest multi-modal antecedent in (2-G) and from the L-Handshape O. This is modeled by an identity $rund'_{antecedent} = rund'_{current\ L\text{-}Handshape\ O}$ in the definite description. It presupposes that we compute the L-Handshape O-semantics. We get (Agent 6 (Broadcasting)).

(Agent 6 (Broadcasting))

$$\underline{\mathsf{ch}}_{\mathsf{i}}\ \mathsf{ru}\ \overline{\mathsf{ch}}_{\mathsf{j}}\langle\lambda g(fahren'(du', \langle\lambda F(um'(\iota x(teich'(x) \wedge rund'(x) = F(x)))))\rangle(g))))\rangle(\mathsf{ru}).0$$

For *da′* in (6-G), roundness is implicitly used, meaning a position at the pond, coming from the L-Handshape O-semantics being still active. So, we get (Agent 7 (Broadcasting)) — overloading our notation which we have used so far with a vector semantics concept.

(Agent 7 (Broadcasting))

$$\overline{\mathsf{ch}}_{\mathsf{i}}\ (da' = n \wedge n \in\ \ |\ \underline{\mathsf{ch}}_{\mathsf{i}}\ \mathsf{ru}\ \overline{\mathsf{ch}}_{\mathsf{j}}\ \langle\lambda g(fahren'(du', \langle\lambda F(um'(\iota x(teich'(x) \wedge rund'(x) = F(x)))))\rangle(g))))\rangle(\mathsf{ru}).0\ \ |).0$$

Here, $|\ \underline{\mathsf{ch}}_{\mathsf{i}}\ \mathsf{ru}\ \overline{\mathsf{ch}}_{\mathsf{j}}\ \langle(\lambda g\ [\ldots]\ |$ — which needs to be distinguished from the replication operator — indicates the finite set of sections of a trajectory around the pond starting with the beginning of the *driving′*.

In (8-G), we have an autonomous occurrence of *rund′* by L-Handshape O again for *dort*, which is further broadcasted for use later on depending on the replication operator.

38 We agree with a reviewer that *rund′* might, after the first integration, itself represent the round pond. We view this as a metonymy of the sort 'property for specific object,', i.e., *rund′* for $dem'(runden'(teich'))$, which we cannot deal with in this paper. It would imply a map from *rund′* to $dem'(rundem'(teich'))$ triggered by an anaphora rule.

To sum up, *rund′* communicated with a posthold held across several contributions is used for several multi-modal semantic representations. In these representations, *rund′* can either stem from the replicated gesture meaning or from anaphora resolution. The latter amounts to integrating *rund′* in the anaphora first and then using it as a test as in Agent 3 (Broadcasting), Agent 5 (Broadcasting), Agent 6 (Broadcasting), and Agent 7 (Broadcasting).

This concludes our presentation of the process algebra account and its application to speech-gesture meaning coordination. In the final section, we point to other cases of multi-modal meaning that could, in principle, be modeled with our account.

8 Concluding remarks and future research

In this paper, we have identified substantial challenges for speech-gesture meaning coordination via a temporal constraint (combined with prosody information). Gesture strokes can come too early or too late or may not be affiliated with any speech parts, and we might need to integrate gesture information more than once. We have proposed to implement a process algebra account for modeling the meaning coordination. This account analyzes speech and gesture as independent concurrent processes that can flexibly communicate with each other and more than once. It enables the incremental analysis of both speech and gesture and their interaction. Importantly, our account is tied neither to our working hypotheses about co-speech gestures nor to the λ-(ψ-)calculus. Other analyses of gesture meaning could be used to specify the data terms for gestures. The ψ-meaning carriers could be used for transporting data terms other than λ-terms (provided that an alternative to λ-β conversion is given), and the interfacing of gesture meaning and speech meaning could be constructed differently. This makes our approach a powerful general modeling of multi-modal meaning.

We used the λ-ψ-calculus to model multi-modal meaning in the case of co-speech iconic gestures. But the calculus is not limited to this case. Other suitable domains are *pointing gestures* (e.g., for reference resolution), communicative *eye-movements*, *eye gazes*, or *eye blinks* (e.g., for reference resolution or other communicative functions), *facial expressions* (e.g., for communicating emotions or attitudes), *nuclear accents* (e.g., for emphasis), *intonation* (e.g., for indicating irony), or *laughter*, and perhaps much more. Another possible application of our account might be the interaction of manual

signs and non-manual markers in *sign languages*.[39] For instance, in American Sign Language lifting one's eyebrows is aligned with asking questions. In principle, the meaning communicated by all these modalities could be represented in a λ-ψ-rendering as agents. The asynchrony, blocking, and broadcasting problem for these forms of embodied communication is very similar to speech-gesture meaning coordination, and their modeling would also depend on exact time-stamped and annotated data. These domains together open up a perspective of considering natural communication as always involving multi-modality of various sorts and treating it as a dynamic network of communicating processes.

In the course of the paper, we have indicated potential future research. We will focus on three topics in our future research, namely, the *speech-dependence of gesture*, *anaphora involving gesture*, and *new application domains* for our framework: The first topic is concerned with modeling the claim that gesture meaning is dependent on or constrained by its accompanying speech meaning. We are developing an account which implies a considerable extension and further empirical grounding of the theory presented here (Rieser & Lawler 2020). This work on gesture meaning dependency will be followed by a theory dealing with broadcasting and anaphora resolution in multi-modal dialogue. Last but not least, the additional application domains of our process algebra account outlined above are worth exploring, such as an application to sign languages.

References

Abeillé, Anne & Owen Rambow (eds.). 2000. *Tree adjoining grammars: Mathematical, computational and linguistic properties*. Stanford, CA: Center for the Study of Language & Information Publications.

Alahverdzhieva, Katya & Alex Lascarides. 2010. Analysing language and coverbal gesture in constraint-based grammars. *17th International Conference on Head-Driven Phrase Structure Grammar*. 5–25. https://doi.org/10.21248/hpsg.2010.1.

Alahverdzhieva, Katya, Alex Lascarides & Dan Flickinger. 2017. Aligning speech and co-speech gesture in a constraint-based grammar. *Journal of Language Modelling* 5(3). 421–464. https://doi.org/10.15398/jlm.v5i3.167.

Asher, Nicholas & Alex Lascarides. 2003. *Logics of conversation*. Cambridge, UK: Cambridge University Press.

39 We owe this suggestion to Chris Barker.

Asudeh, Ash. 2012. *The logic of pronominal resumption*. Oxford, UK: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199206421.001.0001.

Attardo, Salvatore, Jodi Eisterhold, Jennifer Hay & Isabella Poggi. 2003. Multimodal markers of irony and sarcasm. *Humor* 16(2). 243–260. https://doi.org/10.1515/humr.2003.012.

Barendregt, Henk P. 1981–2012. *The lambda calculus, its syntax and semantics*. London: College Publications.

Bengtson, Jesper, Magnus Johansson, Joachim Parrow & Björn Victor. 2011. Psi-calculi: A framework for mobile processes with nominal data and logic. *Logical Methods in Computer Science* 7(1). 1–44. https://doi.org/10.2168/LMCS-7(1:11)2011.

Bryant, Gregory A. & Jean E. Fox Tree. 2005. Is there an ironic tone of voice? *Language and Speech* 48(3). 257–277. https://doi.org/10.1177/00238309050480030101.

Clermont, Thomas, Hendrik Koesling, Marc Pomplun, Elke Prestin & Hannes Rieser. 1998. Eye-movement research and the investigation of dialogue structure. *13th Twente Workshop on Language and Technology*. 61–75.

Cohen, Philip R., Michael Johnston, David McGee, Sharon Oviatt, Jay Pittman, Ira Smith, Liang Chen & Jos Clow. 1997. QuickSet: Multimodal interaction for distributed applications. *5th ACM International Conference on Multimedia*. 31–40. https://doi.org/10.1145/266180.266328.

Cooper, Robin. 2012. Type theory and semantics in flux. In Ruth Kempson, Tim Fernando & Nicholas Asher (eds.), *Philosophy of linguistics. Handbook of philosophy of science*, 271–323. Amsterdam: Elsevier. https://doi.org/10.1016/B978-0-444-51747-0.50009-3.

Cooper, Robin. 2017. Adapting type theory with records for natural language semantics. In Stergios Chatzikyriakidis & Zhaohui Luo (eds.), *Modern perspectives in type-theoretical semantics* (Studies in Linguistics and Philosophy (SLAP) 98), 71–94. Berlin: Springer. https://doi.org/10.1007/978-3-319-50422-3_4.

Cooper, Robin. 2020. From perception to communication: An analysis of meaning and action using a theory of types with records (TTR). Manuscript. https://github.com/robincooper/ttl/blob/master/ttl.pdf (June 2020).

Copestake, Ann. 2007. Semantic composition with (robust) minimal recursion semantics. *ACL 2007 Workshop on Deep Linguistic Processing*. 73–80. https://www.aclweb.org/anthology/W07-12.

Couper-Kuhlen, Elizabeth. 2005. Intonation and discourse: Current views from within. In Deborah Schiffrin, Deborah Tannen & Heidi E. Hamilton (eds.), *The handbook of discourse analysis*, 11–34. Oxford, UK: Blackwell Publishers. https://doi.org/10.1002/9780470753460.ch2.

Couper-Kuhlen, Elizabeth. 2014. Prosody as dialogic interaction. In Dagmar Barth-Weingarten & Beatrice Szczepek Reed (eds.), *Prosodie und Phonetik in der Interaktion — Prosody and phonetics in interaction*, 221–251. Mannheim: Verlag für Gesprächsforschung.

Deliensa, Gaétane, Antonioua Kyriakos, Elise Clinab, Ekaterina Ostashchenkoa & Mikhail Kissinea. 2018. Context, facial expression and prosody in irony processing. *Journal of Memory and Language* 99. 35–48. https://doi.org/10.1016/j.jml.2017.10.001.

Dezani-Ciancaglini, Mariangiola. 1997. *Logical semantics for concurrent lambda-calculus*. Nijmegen: Nijmegen University dissertation. https://www.ru.nl/publish/pages/682191/dezani-ciancaglini_m.pdf.

Di Cosmo, Roberto & Dale Miller. 2019. Linear logic. In Edward N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy*, Summer 2019. Stanford, CA: Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/sum2019/entries/logic-linear/.

Ebert, Cornelia & Christian Ebert. 2014. Gestures, demonstratives, and the attributive/referential distinction. Semantics and Philosophy in Europe 7. https://semanticsarchive.net/Archive/GJjYzkwN/EbertEbert-SPE-2014-slides.pdf (June 2014).

Egg, Marcus & Alexander Koller. 2001. The constraint language for lambda structures. *Journal of Logic, Language and Information* 10(4). 457–485. https://doi.org/10.1023/A:1017964622902.

van Eijk, Jeremy & Hans Kamp. 2011. Discourse representation in context. In Johan van Benthem & Alice ter Meulen (eds.), *Handbook of logic and language*, 2nd edn., 181–252. Amsterdam: Elsevier. https://doi.org/10.1016/B978-0-444-53726-3.00003-7.

Ekman, Paul & Wallace V. Friesen. 1969. The repertoire of nonverbal behavior: Categories, origins, usage, and coding. *Semiotica* 1(1). 49–98. https://doi.org/10.1515/9783110886009.819.

Elordieta, Gorka. 2008. An overview of theories of the syntax-phonology interface. *International Journal of Basque Linguistics and Philology (ASJV)* 42(1). 209–286.

Enfield, Nick J. 2004. On linear segmentation and combinatorics in co-speech gesture: Lao fish trap descriptions. *Semiotica* 149(1-4). 57–123. https://doi.org/10.1515/semi.2004.038.

Engdahl, Elisabet & Enric Vallduví. 1994. Information packaging and grammar architecture: A constraint-based approach. In Elisabet Engdahl (ed.), *Integrating information structure into constraint-based and categorial approaches. Volume 1.3.b of DYANA-2 report*, 41–78. Amsterdam: Institute for Logic, Language & Computation.

Engdahl, Elisabet & Enric Vallduví. 1996. Information packaging in HPSG. In Claire Grover & Enric Vallduví (eds.), *Edinburgh working papers in cognitive science*, vol. 12, 1–31. Edinburgh: Centre for Cognitive Science.

Esipova, Maria. 2018. Focus on what's not at issue: Gestures, presuppositions, appositives under contrastive focus. *Sinn und Bedeutung 22*. 385–402. https://doi.org/10.21248/zaspil.60.2018.473.

Esipova, Maria. 2019. Acceptability of at-issue co-speech gestures under contrastive focus. *Glossa: A Journal of General Linguistics* 4(1). 1–22. https://doi.org/10.5334/gjgl.635.

Fokkink, Wan. 2000. *Introduction to process algebra*. Berlin: Springer. https://doi.org/10.1007/978-3-662-04293-9.

Giorgolo, Gianluca. 2010. *Space and time in our hands*. Utrecht: Utrecht University dissertation. https://www.lotpublications.nl/space-and-time-in-our-hands-space-and-time-in-our-hands.

Giorgolo, Gianluca & Ash Asudeh. 2011. Multimodal communication in LFG: Gestures and the correspondence architecture. *LFG11 Conference*. 257–277. http://web.stanford.edu/group/cslipublications/cslipublications/LFG/16/lfg11.html.

Gutkovas, Ramūnas, Dimitrios Kouzapas & Simon J. Gay. 2016. A session type system for unreliable broadcast communication. Manuscript. http://user.it.uu.se/~ramgu264/papers/sessiondraft.pdf (October 2016).

Hahn, Florian & Hannes Rieser. 2011. Gestures supporting dialogue structure and interaction in the Bielefeld Speech and Gesture Alignment corpus (SaGA). *15th Workshop on the Semantics and Pragmatics of Dialogue*. 182–184.

Haji-Abdolhosseini, Mohamma. 2003. A constraint-based approach to information structure and prosody correspondence. *10th International Conference on Head-Driven Phrase Structure Grammar*. 143–162. https://doi.org/10.21248/hpsg.2003.9.

Han, Ting, Julian Hough & David Schlangen. 2017. Natural language informs the interpretation of iconic gestures: A computational approach. *8th International Joint Conference on Natural Language Processing*. 134–139. https://www.aclweb.org/anthology/I17-2023.pdf.

Johansson, Magnus. 2010. *Psi-calculi: A framework for mobile process calculi: Cook your own correct process calculus – Just add data and logic*. Uppsala: Uppsala University dissertation. http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-123139.

Johnston, Michael. 1998. Unification-based multimodal parsing. *36th Annual Meeting on Association for Computational Linguistics (ACL)*. 624–630. https://doi.org/10.3115/980845.980949.

Kamp, Hans & Uwe Reyle. 1993. *From discourse to logic: Introduction to model-theoretic semantics of natural language, formal logic and Discourse Representation Theory* (Studies in Linguistics and Philosophy (SLAP) 42). Dordrecht, Netherlands: Kluwer Academic Publishing. https://doi.org/10.1007/978-94-017-1616-1.

Kehler, Andrew, Mary Dalrymple, John Lamping & Vijay Saraswat. 1999. Resource sharing in glue language semantics. In Mary Dalrymple (ed.), *Semantics and syntax in lexical functional grammar: The resource logic approach*, 191–208. Cambridge, MA: MIT Press. https://doi.org/10.7551/mitpress/6169.003.0007.

Kempson, Ruth, Eleni Gregoromichelaki, Ronnie Cann & Stergios Chatzikyriakidis. 2016. Language as mechanism for interaction. *Theoretical Linguistics* 42(3-4). 203–276. https://doi.org/10.1515/tl-2016-0011.

Kendon, Adam. 1972. Some relationships between body motion and speech: An analysis of an example. In Aron Wolfe Siegman & Benjamin Pope (eds.), *Studies in dyadic communication*, 177–210. Oxford, UK: Pergamon Press. https://doi.org/10.1016/B978-0-08-015867-9.50013-7.

Kendon, Adam. 1980. Gesticulation and speech: Two aspects of the process of utterance. In Mary Ritchie Key (ed.), *The relationship of verbal and nonverbal communication* (Contributions to the Sociology of Language (CSL) 25), 207–227. Hague: de Gruyter Mouton. https://doi.org/10.1515/9783110813098.

Kendon, Adam. 2004. *Gesture: Visible action as utterance*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511807572.

Klein, Ewan. 2000. Prosodic constituency in HPSG. In Ronnie Cann, Claire Grover & Philip Miller (eds.), *Grammatical interfaces in HPSG*, 171–203. Stanford, CA: CSLI Publications.

Kopp, Stefan, Paul Tepper & Justine Cassell. 2004. Towards integrated microplanning of language and iconic gesture for multimodal output. *6th International Conference on Multimodal Interfaces*. 97–104. https://doi.org/10.1145/1027933.1027952.

Kranstedt, Alfred, Andy Lücking, Thies Pfeiffer, Hannes Rieser & Ipke Wachsmuth. 2006. Deixis: How to determine demonstrated objects using a pointing cone. In Sylvie Gibet, Courty Nicolas & Jean-François Kamp (eds.), *Gesture in human-computer interaction and simulation* (Lecture Notes in Computer Science (LNCS) 3881), 300–311. Berlin: Springer. https://doi.org/10.1007/11678816_34.

Kreuz, Roger K. & Richard M. Roberts. 1995. Two cues for verbal irony: Hyperbole and the ironic tone of voice. *Metaphor and Symbol* 10(1). 21–31. https://doi.org/10.1207/s15327868ms1001_3.

Krifka, Manfred. 2007. Functional similarities between bimanual coordination and topic/comment structure. In Shin Ishihara, Stefanie Jannedy & Anne Schwarz (eds.), *Working papers of the SFB 632, interdisciplinary studies on information structure (ISIS) 8*, 61–96. Potsdam: Universitätsverlag Potsdam.

Ladewig, Silva H. 2014. Creating multimodal utterances. The linear integration of gestures into speech. In Cornelia Müller, Alan Cienki, Ellen Fricke, Silva H. Ladewig, David McNeill & Jana Bressem (eds.), *Body-language-communication: An international handbook on multimodality in human interaction*, vol. 2 (Handbücher zur Sprach- und Kommunikationswissenschaft / Handbooks of Linguistics and Communication Science (HSK) 38), 1662–1677. Berlin/Boston: de Gruyter Mouton. https://doi.org/10.1515/9783110302028.

Lascarides, Alex & Matthew Stone. 2006. Formal semantics for iconic gesture. *10th Workshop on the Semantics and Pragmatics of Dialogue*. 64–71. https://publishup.uni-potsdam.de/frontdoor/index/index/docId/964.

Lascarides, Alex & Matthew Stone. 2009. A formal semantic analysis of gesture. *Journal of Semantics* 26(4). 393–449. https://doi.org/10.1093/jos/ffp004.

Lawler, Insa, Florian Hahn & Hannes Rieser. 2017. Gesture meaning needs speech meaning to denote — A case of speech-gesture meaning interaction. *Workshop on Formal Approaches to the Dynamics of Linguistic Interaction 2017 co-located within the European Summer School on Logic, Language and Information (ESSLLI)*. 42–46. http://ceur-ws.org/Vol-1863/paper_4.pdf.

Loehr, Daniel. 2007. Aspects of rhythm in gesture in speech. *Gesture* 7(2). 179–214. https://doi.org/10.1075/gest.7.2.04loe.

Lücking, Andy. 2013. *Ikonische Gesten: Grundzüge einer linguistischen Theorie*. Berlin: De Gruyter. https://doi.org/10.1515/9783110301489.75.

Lücking, Andy. 2016. Modeling co-verbal gesture perception in type theory with records. Annale of Computer Science and Information Systems (ACSIS) 8. 383–392. https://doi.org/10.15439/2016F83.

Lücking, Andy, Kirsten Bergmann, Florian Hahn, Stefan Kopp & Hannes Rieser. 2013. Data-based analysis of speech and gesture: The Bielefeld Speech and Gesture Alignment corpus (SaGA) and its applications. *Journal on Multimodal User Interfaces* 7. 5–18. https://doi.org/10.1007/s12193-012-0106-8.

Lücking, Andy, Thies Pfeiffer & Hannes Rieser. 2015. Pointing and reference reconsidered. *Journal of Pragmatics* 77. 56–79. https://doi.org/10.1016/j.pragma.2014.12.013.

Lücking, Andy, Hannes Rieser & Marc Staudacher. 2006a. Multi-modal integration for gesture and speech. *10th Workshop on the Semantics and Pragmatics of Dialogue (SemDial-10)*. 106–113. https://publishup.uni-potsdam.de/frontdoor/index/index/docId/964.

Lücking, Andy, Hannes Rieser & Marc Staudacher. 2006b. SDRT and multi-modal situated communication. *10th Workshop on the Semantics and Pragmatics of Dialogue (SemDial-10)*. 72–79. https://publishup.uni-potsdam.de/frontdoor/index/index/docId/964.

McNeill, David. 1992. *Hand and mind: What gestures reveal about thought*. Chicago, IL: Chicago University Press.

McNeill, David. 2000. Catchments and contexts: Non-modular factors in speech and gesture production. In David McNeill (ed.), *Language and gesture*, 312–328. Cambridge, UK: Cambridge University Press. https://doi.org/10.1017/CBO9780511620850.019.

McNeill, David, Francis Quek, Karl-Erik McCullough, Susan D. Duncan, Nobohiro Furuyama, Robert Bryll, Ma Xin-Feng & Rashid Ansari. 2001. Catchments, prosody and discourse. *Gesture* 1(1). 9–33. https://doi.org/10.1075/gest.1.1.03mcn.

Milner, Robin. 1989. *Communication and concurrency*. Upper Saddle River, NJ: Prentice Hall International.

Müller, Stefan. 2007. *Head-Driven Phrase Structure Grammar: Eine Einführung*. 1st edn. Tübingen: Stauffenburg Verlag.

Paggio, Patrizia & Costanza Navarretta. 2009. Integration and representation issues in the annotation of multimodal data. *NODALIDA 2009 Workshop of Multimodal Communication*. 25–31. http://hdl.handle.net/10062/9832.

Parrow, Joachim. 2001. An introduction to the π-calculus. In James A. Bergstra, Alban Ponse & Stephan A. Smolka (eds.), *Handbook of process algebra*, 479–543. Amsterdam: Elsevier. https://doi.org/10.1016/B978-044482830-9/50026-6.

Pfeiffer, Thies, Florian Hofmann, Florian Hahn, Hannes Rieser & Insa Röpke. 2013. Gesture semantics reconstruction based on motion capturing and complex event processing: A circular shape example. *14th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*. 270–279. http://www.aclweb.org/anthology/W/W13/W13-4041.

Pfeiffer, Thies, Insa Lawler, Hannes Rieser & Florian Hahn. 2019. Gesture beats prototype. Manuscript. (January 2019).

Poesio, Massimo & Hannes Rieser. 2010. Completions, coordination, and alignment in dialogue. *Dialogue and Discourse* 1(1). 1–89. https://journals.uic.edu/ojs/index.php/dad/article/view/10669.

Poesio, Massimo & Hannes Rieser. 2011. An incremental model of anaphora and reference resolution based on resource situations. *Dialogue and Discourse* 2(1). 235–277. https://journals.uic.edu/ojs/index.php/dad/article/view/10717.

Potts, Christopher. 2005. *The logic of conventional implicatures*. Oxford, UK: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199273829.001.0001.

Rieser, Hannes. 2004. Pointing in dialogue. *8th Workshop on the Semantics and Pragmatics of Dialogue*. 93–101.

Rieser, Hannes. 2011. Gestures indicating dialogue structure. *15th Workshop on the Semantics and Pragmatics of Dialogue*. 9–18.

Rieser, Hannes. 2015. When hands talk to mouth: Gesture and speech as autonomous communicating processes. *19th Workshop on the Semantics and Pragmatics of Dialogue*. 1–9. http://semdial.org/anthology/Z15-Rieser_semdial_0017.pdf.

Rieser, Hannes. 2017. A process algebra account of speech-gesture interaction. *Workshop on Formal Approaches to the Dynamics of Linguistic Interaction 2017 co-located within the European Summer School on Logic, Language and Information (ESSLLI)*. 67–71. http://ceur-ws.org/Vol-1863/paper_8.pdf.

Rieser, Hannes, Kirsten Bergmann & Stefan Kopp. 2012. How do iconic gestures convey visuo-spatial information? Bringing together empirical, theoretical, and simulation studies. *Gesture and Sign Language in Human-Computer Interaction and Embodied Communication*. 139–150. https://doi.org/10.1007/978-3-642-34182-3_13.

Rieser, Hannes & Insa Lawler. 2020. Modeling the speech dependency of co-speech iconic gestures. Manuscript. (June 2020).

Schlenker, Philippe. 2018. Gesture projection and cosuppositions. *Linguistics and Philosophy* 41(3). 295–365. https://doi.org/10.1007/s10988-017-9225-8.

Schlöder, Julian. 2017. Towards a formal semantics of verbal irony. *Workshop on Formal Approaches to the Dynamics of Linguistic Interaction 2017 co-located within the European Summer School on Logic, Language and Information (ESSLLI)*. 55–59. http://ceur-ws.org/Vol-1863/paper_2.pdf.

Slama-Cazacu, Tatiana. 1976. Nonverbal components in message sequence: "Mixed syntax". In William C. McCormick & Stephan A. Wurm (eds.), *Language and man: Anthropological issues*, 217–228. Hague: Mouton Publishers. https://doi.org/10.1515/9783112321454.

van der Sluis, Ielka Francisca. 2005. *Multimodal reference: Studies in automatic generation of multimodal referring expressions*. Tilburg: Tilburg University dissertation. https://pure.uvt.nl/ws/portalfiles/portal/699745/173990.pdf.

Stone, Matthew, Christine Doran, Bonnie Webber, Tonia Bleam & Martha Palmer. 2013. Microplanning with communicative intentions: The SPUD system. *Computational Intelligence* 19(4). 311–381. https://doi.org/10.1046/j.0824-7935.2003.00221.x.

Tofts, Christopher. 1992. Describing social insect behaviour using process algebra. *Transactions of the Society of Computer Simulation* 9. 227–283.

Velichkovsky, Boris, Marc Pomplun & Hannes Rieser. 1996. Attention and communication: Eye-movement-based research. *Advances of Psychology* 116. 125–154. https://doi.org/10.1016/S0166-4115(96)80074-4.

Wagner, Michael. 2015. Phonological evidence in syntax. In Tibor Kiss & Artemis Alexiadou (eds.), *Syntax — Theory and analysis*, vol. 2 (Handbücher zur Sprach- und Kommunikationswissenschaft / Handbooks of Linguistics and Communication Science (HSK) 42), 1154–1198. Berlin: Mouton de Gruyter. https://doi.org/10.1515/9783110363708-011.

Wagner, Petra, Zofia Malisz & Stefan Kopp. 2014. Gesture and speech in interaction: An overview. *Speech Communication* 57. 209–232. https://doi.org/10.1016/j.specom.2013.09.008.

Zwarts, Joost. 1997. Vectors as relative positions: A compositional semantics of modified PPs. *Journal of Semantics* 14(1). 57–86. https://doi.org/10.1093/jos/14.1.57.

Zwarts, Joost & Yoad Winter. 2000. Vector space semantics: A model-theoretic analysis of locative prepositions. *Journal of Logic, Language, and Information* 9(2). 169–211. https://doi.org/10.1023/A:1008384416604.

Hannes Rieser
Faculty for Linguistics and Literary Studies
Bielefeld University
Universitätsstraße 25
33615 Bielefeld
Germany
hannes.rieser@uni-bielefeld.de

Insa Lawler
Department of Philosophy
University of North Carolina at Greensboro
PO Box 26170
Greensboro, NC 27412-5001
United States of America
irlawler@uncg.edu