A Roadmap for Technological Innovation in Multimodal Communication Research*

Alina Gregori[1][0000−0003−0143−2000], Federica Amici[2][0000−0003−3539−1067], Ingmar Brilmayer**[3][0000−0001−7227−2054], Aleksandra Ćwiek[4][0000−0002−1513−0188], Lennart Fritzsche[1], Susanne Fuchs[4][0000−0001−6751−9286], Alexander Henlein[1][0000−0002−2611−1417], Oliver Herbort[5], Frank Kügler[1][0000−0001−8101−0005], Jens Lemanski[6][0000−0003−3661−4752], Katja Liebal[2], Andy Lücking[1][0000−0002−5070−2233], Alexander Mehler[1][0000−0003−2567−7539], Kim Tien Nguyen[1], Wim Pouw[7][0000−0003−2729−6502], Pilar Prieto[8], Patrick Louis Rohrer**[8,9][0000−0002−2714−7294], Paula G. Sánchez-Ramón[1,8][0000−0002−3394−1013], Martin Schulte-Rüther[10][0000−0002−7198−9923], Petra B. Schumacher[3][0000−0003−0263−8502], Stefan R. Schweinberger[11][0000−0001−5762−0188], Volker Struckmeier[1], Patrick C. Trettenbrein[12][0000−0003−2233−6720], and Celina I. von Eiff[11] ***

[1] Goethe University Frankfurt/M.
[2] University of Leipzig
[3] University of Cologne
[4] Leibniz Centre General Linguistics, Berlin
[5] Julius-Maximilians-University of Würzburg
[6] WWU Münster and University of Hagen
[7] Donders Institute for Brain, Cognition, and Behaviour, Radboud University Nijmegen
[8] Universitat Pompeu Fabra, Barcelona
[9] Nantes Université, France
[10] University Medical Center Göttingen
[11] Friedrich Schiller University of Jena
[12] Max Planck Institute for Human Cognitive and Brain Sciences, Leipzig

Abstract Multimodal communication research focuses on how different means of signalling coordinate to communicate effectively. This line of research is traditionally influenced by fields such as cognitive and neuroscience, human-computer interaction, and linguistics. With new technologies becoming available in fields such as natural language processing and computer vision, the field can increasingly avail itself of new ways of analyzing and understanding multimodal communication. As a result, there is a general hope that multimodal research may be at the "precipice

* Supported by the DFG priority program *Visual Communication* (ViCom).
** External collaborator
*** For the ViCom Consortium, alphabetical order except lead author; send correspondence to gregori@lingua.uni-frankfurt.de

of greatness" due to technological advances in computer science and resulting extended empirical coverage. However, for this to come about there must be sufficient guidance on key (theoretical) needs of innovation in the field of multimodal communication. Absent such guidance, the research focus of computer scientists might increasingly diverge from crucial issues in multimodal communication. With this paper, we want to further promote interaction between these fields, which may enormously benefit both communities. The multimodal research community (represented here by a consortium of researchers from the Visual Communication [ViCom] Priority Programme) can engage in the innovation by clearly stating which technological tools are needed to make progress in the field of multimodal communication. In this article, we try to facilitate the establishment of a much needed common ground on feasible expectations (e.g., in terms of terminology and measures to be able to train machine learning algorithms) and to critically reflect possibly idle hopes for technical advances, informed by recent successes and challenges in computer science, social signal processing, and related domains.

Keywords: Multimodal communication · Natural language processing · Technical innovation

1 Introduction: Multimodal Communication

When people talk to each other, they naturally communicate with their whole bodies (e.g., [58]). That is, besides speech or sign they use facial expressions, move their hands and arms for gesturing, laugh, gaze, shrug, nod, sigh, among other things. All these signals cohere into social interactions (e.g, [119]) and are interpreted in relation to one another [76]. While some of these signals are perceived in the acoustic modality, others are perceived visually, tactilely or olfactorily. Thus, interactions are not only produced by the whole body, but also perceived by various sense modalities. Hence, communication is multimodal, captured in the eponymous term of *multimodal communication*.

We are researchers with different backgrounds working on multimodal communication, specifically on gestures, sign languages, didactic and clinical aspects of visual communication, animal communication, and human-computer interaction systems. Our work contributes to the Priority Programme *Visual Communication* (ViCom), supported by the German Research Foundation (DFG). ViCom aims at disclosing the specific characteristics of the visual modality as a communication channel and its interaction with other channels (especially the acoustic one) to develop models of human communication and their cognitive and evolutionary foundations. We share an interest in visual communication and are particularly interested in exploring to what extent vocal or signed linguistic communication is multimodal at its core and substantially shaped by (social) interaction [56,77]. We differ in the theoretical frameworks we employ [37], the populations and species we work with, and the methodologies we use. What ties us together is the idea that mutual progress can be made by employing principles

of open science, sharing our expertise on different scientific strands, and paving the way for cooperative science to move beyond the border of a single discipline.

Many aspects of multimodal communication remain unknown, including the mechanisms by which multiple signals are quickly integrated in perception and coordinated in production [187,6]. The study of multiple signals is therefore required, and in need of technological advancement to be explored. The detailed investigation of different signals in isolation, as well as their cross-modal integration in different populations and species, requires tools and methods that differ from those developed for written (and usually digitized) text. Over the past years, tools and methods have been developed for simultaneously keeping track of signals on various channels, but not necessarily in a joint effort by computer science (expertise in data processing), linguistics (expertise in linguistic structure and its contributions to meaning, etc.), and other disciplines investigating principles of communicative behavior.

1.1 Goals of this Article

Not least due to advances in audiovisual technology (i.e., recording and storing ephemeral utterance events by means of camera and microphone for analysis), an empirically grounded theory of multimodal language use and interaction is developing [38] – "a scientific tool whose importance to our discipline equals that of the microscope to biology" [178, p. 275].

We believe that multimodal research is necessary to advance theoretical research on human and non-human animal communication. Here, we survey the state-of-the-art of existing research tools that can and have been applied in multimodal research on communication. We evaluate current approaches, point out short- and long-term aims, and identify the scientific innovations that such aims imply. We further provide suggestions for tools and applications, which we believe might help move the interdisciplinary research dialogue between cognitive and computer science forward. Finally, we highlight the availability of these tools, not only to increase their familiarity among linguists, but also to highlight how some tools are now primarily used by computer scientists due to high technical skills required to wield these tools.

1.2 Overarching Terms

"Multimodal communication" is a broad field of research that considers not merely acoustic signals (which may be grouped into meaningful units) as being able to move communication forward. While most of this paper will revolve around the acoustic and visual modes of communication, note that tactile cues (e.g., in languages of the deaf-blind) or olfactory signals (e.g., in animal communication) are used for communication as well. Typical examples of visual communication are sign languages, gesturing, facial expressions, eye gaze, and orofacial movements, but also diagrams, emojis, and written text, c.f., Figure 1.

Multimodality can refer to the sensory channels (e.g., acoustic, visual) or different content types within a sensory channel (e.g., signs and facial expressions, eye gaze, text and emojis).

Some forms of multimodal communication comprise complex grammar and can express a potentially infinite range of different ideas (e.g., discourse in sign languages), others have lexicalized meanings (e.g., individual signs or symbolic and highly conventionalized gestures), or may refer to specific aspects of the current situation (e.g., deictic gestures indicating spatial relations). In addition, the different forms vary in complexity: For example, lexical items of a sign language are highly complex signals that consist of discrete parameters (i.e., hand shape, place of articulation, movement, and orientation), while symbolic gestures may be formed in a simpler manner. Moreover, some forms of visual communication such as sign languages are highly conventionalized (vertical axis of Figure 1), while others, such as iconic gestures, may be created spontaneously. Within gesture studies, degrees of conventionalization of hand-and-arm movements are located on *Kendon's Continuum*, popularized by [113, p. 37]. Note that the level of conventionalization may differ within each form. For example, a simple drawing of an object, like a hammer, is first and foremost its representation in the iconic sense – it refers to a hammer through a sense of resemblance (but see [25,17,57] for critical discussions of reducing reference to resemblance). However, if a hammer is combined with a sickle in a specific manner, it may stand for a symbolic and conventional representation of Marxist-Leninist philosophy.

Lastly, forms vary in the specific effectors – the medium used for expression (e.g., hands, face, paper/screen, c.f., colors in Figure 1) – they are typically tied to, or their dynamicity (deictic gestures often being relatively static, while beat gestures are dynamic). The use of visual cues in communication – which makes it multimodal – can also set an implicit "tone" to the communicative situation. This can be shown in the expression of emotional states or reversing the meaning of what is said. Note also that multimodal communication may happen without intention, or without consciously controlling effectors to communicate (e.g., in non-human animal research).

2 Development of Multimodal Research

2.1 Establishing Multimodal Research

While the idea of communication being multimodal is not new itself, multimodality has yet to be adequately incorporated and modelled in linguistic theories. In sharp contrast to the well-established interest in written and/or spoken language understood as a unimodal phenomenon, multimodal aspects of communication have received only some initial attention.

Multimodal communication research has emerged on the foundation of work from pioneering scholars from the late 20th century (e.g., [79,80,113]), who claimed that gesture had to be understood as an integral part of communication. Since then, visual cues have been gradually accepted as a central component of

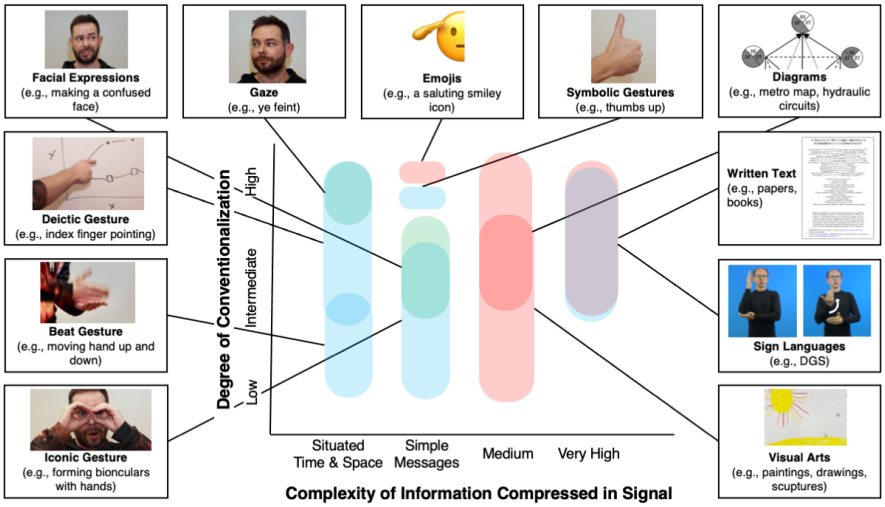

Figure 1. Examples of a variety of different forms of visual communication (non-exhaustive). Different forms of visual communication are organized horizontally according to the complexity of the information that could potentially be conveyed and vertically according to their level of conventionalization. Colors indicate the effectors that are primarily involved (hands: blue, face: green, paper/screens: red). Images are re-used courtesy of Oliver Herbort. Still images of DGS signs are copyright [184] and re-used under Creative Commons Attribution 4.0 International License.

many communicative scenarios from a variety of perspectives. Crucially, this acceptance has been triggered by cutting-edge studies on both signed and spoken languages. This makes important contributions to the assumption that the study of visual aspects of communication should definitely be considered a worthwhile (or even necessary) addition to theories. This relates not only to "communication" in the abstract sense but also more centrally to specifically linguistic theories (e.g., [143] for a review, [43,42] for gesture, [155] for signed languages, [65] for the gesture–sign interface, [105] for the gesture–grammar interface).

It is our belief that the incorporation of multimodal aspects of communication into linguistic theories may well come to resemble the inclusion of prosodic aspects into grammatical models of language. Prosody initially was not deemed important with regard to its effects on other linguistic subfields such as syntax, semantics, lexical analysis, or information structure, c.f., [174,36]. This changed only when the established methods of the neighboring subfields started considering that prosody has a productive impact on the language system as a linguistic subfield in its own right (e.g., [69,179]). Important research by [146], [92] and [60] demonstrated the phonological status of intonation across languages. Nowadays, prosody has been shown to not only constitute an essential module of grammar

but also having an impact on language production and processing (see [153] for a review).

There is some level of agreement that multimodality can be an important factor for communication research. However, from a wider perspective, multimodal communication research has not yet agreed on standard definitions (e.g., for concepts, categories, relations), nor on tools or methodologies required. Two central dimensions have already been investigated in the relation between language use and gesture (see [1] for a review), namely timing and meaning. Issues that have not been fully solved include the transition from pure manual gesture annotation to tools like automatic recognition of specific patterns, segmentation tools, and natural language processing.

2.2 Interaction with Other Fields

Technological innovation in multimodal communication research is relevant not only for understanding human communication but also for a number of other fields, such as communication of emotions, human-computer interaction (HCI), neuroscience, clinical research and animal communication (do specific gestures and vocalizations co-occur to convey novel meanings during communicative interactions [7]? A multimodal approach may be crucial to really understand the meaning entailed by animal communication). For example, principles such as the tight coupling of perception and expression between social agents not only underpin communication using language [145], but also extend to nonverbal facial [172] and auditory [52] communication and benefit from multimodal measurements (e.g., [64]).

The consideration of multimodality in neuroscience may foster theoretical and functional innovations: Brain systems mediating perception and expression in a modality are often overlapping both in humans and non-human animals (see, e.g., [52] for the auditory domain), yet higher-order regions processing abstract information that is not part of the signal (e.g., syntax and semantics) such as the language network seem to be organized in a modality-independent fashion [189,185,90]. Increasingly, "mobile friendly" neuroscience methods such as fNIRS (functional near-infrared spectroscopy) are used for the investigation of more than one person and more than one brain during interaction [126] to assess brain mechanisms and synchronization/coordination processes across brains [88]. Such approaches would greatly benefit from additional fine-grained assessment of multimodal data streams reflecting communicative behavior and integrated neurobehavioral analysis.

Multimodal communication researchers are often in the business of understanding time-varying bodily motions acting together in a referential or indicative way. While the tools and conceptual schemes to categorize these complex communicative objects have advanced, the analyses of the time-varying motion have as such been lacking. There is thereby relatively little integration with human movement science which focuses on non-referential bodily motions, and hitherto a general lack of application of concepts from kinematics or biomechanics, and a lack of integration of tools that deal with high-dimensional time-

varying data. But this is changing too. For example, there is increasingly robust research on the kinematic information about intentions (e.g., [144,188,29]), application of kinematic-acoustic analyses in typical and non-typical populations (e.g., [39,104]), and new ways of compressing high-dimensional data for analyzing multimodal signaling (e.g., [5,149]).

Time-sensitive, kinematic studies can contribute to a debated but difficult to investigate phenomenon of multimodal communication, namely the cross-modal constitution of "ensembles" [79,118]. Such ensembles, *when used repetitively in conversation*, have a statistical effect (as assessed in information theoretic terms for speech–gesture pairs by [115]). A generalization of such ensembles has been suggested to be a cornerstone of (the speed of producing and comprehending) multimodal communication in terms of multimodal gestalts [67]; see also the challenges pointed out in section 1. However, recurrent ensembles or gestalts may lead to a simplification of form – on side of the gesture, on side of speech, or both [107]. It is suggestive to ascribe such simplifications to a balance between production effort and comprehension, facilitated by repetition of use. In order to quantify such recurrent phenomena, combined temporal and spatial measurements are needed. These may also feed into time-dependent, embedding-based approaches as employed, for instance, in semantic change detection [176].

Communicative movement is special and not simply guided by clearly testable performance variables [183]. Rather, for communicative movements "meaning is a performance variable" [95, p. 359]. Yet, meaning needs not always be such an elusive concept, and communicative movements may also have informative value qua movements, for example by deviating from how one usually moves [144] or making use of biomechanical stabilities [151].

Advancement in multimodal communication research may also provide strong benefits to clinical research, for example, to people with sensory impairments or users of a cochlear implant [8,45]. Moreover, many developmental and mental disorders are associated with problems in visual communication and social functioning. Disentangling the respective mechanisms for different disorders is important for advancing diagnostics (e.g., [171]) and intervention. For example, although perception, expression, and imitation of facial emotions are disturbed in the Autism Spectrum Disorder (ASD) [190], it is controversial whether facial motor mimicry is involved as a mechanism that drives such disturbance [41,172]. Research into cross-modal imitation (for instance, a reflexive facial expressive response that matches the emotion perceived in another´s voice) [110] suggests a multimodal nature of emotion processing, but studies tackling this issue explicitly remain quite rare [200].

For clinical contexts and applications in real-world settings (i.e. when ecological validity is particularly important), methods are needed that are easy to apply, non-invasive, and flexible. For example, facial emotion expressions or speech were traditionally measured using electrodes or sensor coils (e.g., facial electromyography [172], real-time magnetic resonance imaging (MRI), or electromagnetic articulography [123], respectively). Although precise and well-established, such methods have the disadvantage of being bound to a lab than more recent contact-

free approaches like video recordings for the analyses of facial expression, acoustic analysis for automatic speech recognition, and recognition of emotions in speech [170,40]. Video-based assessment of body pose is also increasingly used in clinical populations [111], however, developing efficient machine learning algorithms for semantic analysis of body pose and gestures remains a challenge. Ultimately, this depends on the exact definition of semantically interpretable body configurations for the respective situation (e.g., a gesture "vocabulary", categories of the emotional tone of movement, etc.). Furthermore, these new methods also allow studying other species during natural interactions, without having to recur to invasive research, which can be ethically problematic in some cases.

Overall, multimodal data acquisition poses both a challenge and an opportunity in this respect: richer data allows for more precise characterizations, but its mapping to meaningful communicative events requires stringent, theory-based descriptions and theoretical models. Theoretical questions (broadly understood) as well as empirical knowledge of previous studies should lay the foundation for the development of standard procedures and technological development of new tools. A step in the right direction would be, from our point of view, to establish fruitful collaborations across researchers, labs, research institutions, etc. In this way, we could integrate our different needs, develop standard definitions of concepts (e.g., from semantics, prosody, and technology), achieve more systematicity in terms of manual labeling and foster the development of technological tools for the benefit of theoretical and applied research in this field.

3 Available Tools, Methods and Databases

Having assessed the integration of mulitmodal communication research into cognitive science, in this section we provide a state-of-the-art on available tools, methods and databases for empirical multimodal research. The employment of automatic tools and methods for studying multimodal natural language use can be roughly distinguished in terms of an annotation ladder, sketched in Figure 2. To give a brief, stepwise description: (i) First, the physical signals realizing the communicative behavior under observation need to be recorded. Therefore the specificities of the recording techniques and knowledge about previous studies are key factors. Recording multiple signals with different devices imposes a synchronization problem (bottom row). (ii) Within the continuous recording stream, the actual units of observation have to be identified (segmentation). This may require mapping the recorded signals into a meaningful feature space first (e.g., mapping Cartesian movement coordinates onto a skeleton model). Segments from different signals have to be aligned (middle row). (iii) The units of observation can then be classified (e.g., functionally, taxonomically, etc.). Note that the labels of annotation and how to use them should be defined in an annotation scheme. Multi-channel classification basically has to decide whether different signals/segments belong to a single "ensemble" [80], bear a coherence relation, or remain unrelated. Along these lines, results of automatized processing finally feed into qualitative or quantitative theory building and testing (see

section 1). The annotation ladder sketch is useful for assessing the subsequent survey of recording and annotation tools in the discussion in section 4.

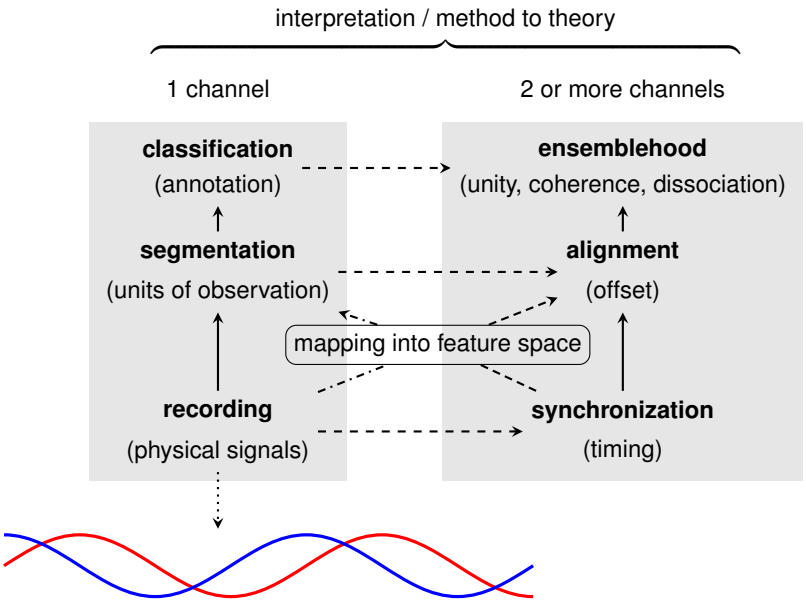

Figure 2. The annotation ladder for multimodal signals on a single channel (left) gives rise to temporal or (dis-)integrative multi-channel relationships on each rung (right). The blue and red lines represent distinct modes of continuous signals. The unidirectionality of arrows indicates a step on the next rung. Depending on the current study, however, knowledge from higher rungs may be used in the opposing direction (top-down) to contribute to solving annotations on lower rungs. Note that in multilogue the annotation ladder applies both within a single speaker and between different speakers.

3.1 Motion Recording Techniques

Data recording is often a crucial step when doing research in multimodal communication and it offers a variety of technologically supported tools. In this section, we will review the main behavioral recording techniques. When multimodal communication research is carried out, human-to-human communication is usually recorded with one or more video cameras in *immobile* (cameras distributed in space [101]) or *mobile* (cameras move with the people recorded; e.g., to reflect the perspectives of the actors [72,201]) recording setups. For good audio quality, it is advisable to use (external) condenser microphones. In the following, we describe different implementation options for data recording, which are not mutually exclusive but can be combined in various ways to balance out their individual advantages and disadvantages. Note, however, that overly complex

setups may come at the expense of truly huge amounts of data, which also may need to be synchronized and processed in a complex fashion during analyses (see section 3.3). The choice of a recording technique is often based on the research objectives, space, price, mobility, technical know-how, scalability, accuracy, frequency, feature, scenario, etc. [51,206].

Recording scenarios can focus on the body as a whole or on specific body parts. Full body tracking can be carried out by *room-scale systems* (which statically record scenes; are very specialized and accurate; require specifications in advance, e.g., [22]), by *AI systems* (which are standalone cameras typically with depth sensors; natively track body motion using specific key points), or by *body trackers* (which are key points placed on the actor; can be either standalone or combined with external stations; typically 3–8 trackers are needed; e.g., [141,204]). Recording strategies focusing on single body parts usually record the movement of the face, hands or eyes, using more specific key points (not allowing full body tracking), and they are usually implemented by using mobile *head-mounted systems*. Examples of these systems are *eye tracking glasses* (which are glasses-type devices indicating the exact point the recorded eye fixates; e.g., [191,75]), *Virtual Reality (VR) glasses* (which usually measure head movements in space, but increasingly also provide direct hand tracking and gesture recognition, native eye tracking, and face tracking; they are inexpensive and widespread; can help to visualize stimuli/scenarios [136]; with the disadvantages that a portion of the face is covered and they do not allow for direct human interaction) or *Augmented Reality (AR) glasses* (which augment the real world for the user; and include hand and eye tracking).

Tracking technology provides versatile means and devices for recording body signals of various kinds. However, from the practical viewpoint of multimodal researchers, their productive use in empirical studies places some obstacles: (i) Special equipment is needed, which also needs to be operated by an expert. That is, many of the above-mentioned methods do not work out of the box for untrained researchers. (ii) The tracking data needs to be post-processed. However, in contrast to written and digitized text, there is a glaring lack of automatic annotation procedures (see section 1). (iii) Even manual annotation poses problems: standard annotation tools are not prepared to handle data types other than digitized text, audio or video files – see section 3.4. Furthermore, there are no agreed-upon annotation schemes or standards for many non-verbal kinematic communicative behaviors.

3.2 Neural & Physiology

In contrast to motion recording techniques (section 3.1) that capture spontaneous, behavioral data from the environment, neuroscientific methods are also used with a focus on a better understanding of the underlying control or processing components in multimodal communication. However, while there is relative robustness of e.g., eye-tracking data against artifacts that originate from freely moving participants – given careful and repeated calibration and firmly fitting eye-tracking glasses – recording and analysis of, for example, electrophysiological

data in interactive settings are more complicated. A number of such methods are introduced in the following.

For example, *electroencephalography* (EEG) is a powerful tool to investigate the time-course of cognitive processes. The electrical brain activity is recorded from an electrode cap and any kind of movement generates artifacts – even blinking. Most previous studies with freely-moving subjects used so-called dual-task experiments, in which participants have to perform a common task from experimental psychology (e.g., oddball detection) while performing a second task (e.g., [100,124]). This type of study compares the response to a target stimulus in different mobility conditions, for instance walking and sitting. Traditional averaging methods to compute event-related potentials (ERPs; i.e. electrophysiological signals time-locked to a stimulus event) have proven quite robust to movement-related artifacts after standard preprocessing [100]. However, the movements in these studies were temporarily and semantically unrelated to the experimental task. Thus, their influence on the ERPs of interest might be small and "canceled out" in the averaging process. In real-world experiments that are, for instance, interested in language use and co-speech gestures, however, the bodily movements (gestures) are related to another modality of the experimental stimuli of interest (speech). Therefore, they systematically occur in critical trials and thus overlap with each other and with the critical language input. As a consequence, the language-related ERPs usually overlap with gesture-related potentials, as well as fixation-related potentials (since participants are looking at gestures while listening to speech). Failing to account for this overlap might lead to a critical distortion of the results. Therefore, calculating traditional, averaged ERPs is not suited for this kind of data. Instead, it is advisable to use time-resolved regression (deconvolution) to calculate regression-based ERPs (rERPs) in real-world studies. There are toolboxes for EEGs that have the flexibility for multimodal analysis (c.f. [166] or [44].

Another common method to measure brain activity is *functional near-infrared spectroscopy* (fNIRS), which is increasingly used in interaction research (e.g., [152,177]) due to its ability to capture multimodal signals. fNIRS is a haemodynamic technique to assess functional activity based on the different optical properties of oxygenated and deoxygenated haemoglobin [169]. fNIRS devices are highly portable and relatively robust to motion artifacts (e.g., [9,147]). For example, accurate signals have been obtained even while dancing [127] or playing table tennis [10].

In addition, the emergence of *hyperscanning* studies [120] (i.e., measuring the activity of multiple brains simultaneously) has started to decipher the brain mechanisms underlying social interaction as a whole (rather than recording only the brain of one involved person) (e.g., [54,137,11]). Opening new ways to measure interaction, hyperscanning provides an experimental means that enables more ecological validity in neuroscientific studies of the social brain. More generally, we can also assess other physiological measures, such as individuals' heart rate or skin conductance, which should be sensitive to individuals' movement.

A challenge of working with multimodal data is to identify the point at which modality-specific information is accessible. For example, the intonational contour provides cues as the speech signal unfolds, and gestures or signs consist of various components. When we are considering time-sensitive data, it is thus crucial to consider the different signal components. ERP research on sign languages has for instance used different time-locking points (triggers) that vary in latency up to around 400 ms: *handshape change* (i.e. the neutral position of the hand between two signs), *target handshape* (point of complete access to target handshape) and *sign onset* (point at which target location is reached) [68].

3.3 Data Synchronization

A peculiarity of synchronization between audio and video is that these data streams tremendously differ in most experiments regarding their sampling rate, with audio having a much higher rate than video. For synchronization purposes, it is advisable that one sampling frequency is an integer multiple of the other (for further recommendations, see [130]). Otherwise, this can lead to a serious timing difference between the two streams (offset). This offset between two streams increases over time, i.e. the longer a recording, the more asynchronous the timing of the streams. In controlled, laboratory experiments, the synchronization of two data streams (e.g., eye-tracking and EEG) is straightforward and usually achieved via software trigger pulses which are simultaneously sent to the two recording devices by the stimulus presentation software (or during preprocessing). Alternatively, a generated audio beep can be recorded on two devices, allowing for synchronization during data post-processing. Synchronization based on image data is less common. In the future, another possibility, especially for non-specialists, might be to move forward with the synchronization of various data streams on multiple computers (which have different internal clocks that can result in delays between signals). Also, since these synchronization events are usually repeated trial by trial, the offset between two streams virtually never reaches a level that might be relevant for analysis. While a high-precision synchronization in the ms range for events between multiple participants or across multiple data streams is not necessary for every application, it is essential for applications such as time-locked neural responses (e.g. event-related potentials), neural synchronization, behavioral synchronization that uses analysis techniques for coherence in the frequency domain (such as wavelet coherence or lagged cross–correlation) and the temporal study of acoustic and visual cues.

In interactive experiments, researchers often do not depend on stationary, wired equipment, and they may even not present stimuli with specific presentation software. One way to facilitate synchronization is then to use a clapperboard like in the film industry, and later align the different streams manually, before processing. A shortcoming of this approach is that, as time unfolds, timing imprecision increases. Another powerful tool is *Lab Streaming Layer (LSL)*, an open source, platform-independent library to send, receive, and synchronize

(neuro-)physiological and behavioral data from a large variety of devices[13]. LSL performs clock synchronization in regular intervals (default 5s), keeps track of the differences between recorded data streams, and stores the data and timing information in the extensible data format .xdf[14] via the app LabRecorder[15]. LSL also offers software for smartphones and is adaptable to specific recording situations.

3.4 Manual Annotation and Existing Multimodal Corpora

Annotation Tools The manual annotation of multimodal data allows researchers to quantitatively analyze multimodal data, by using annotation software such as ELAN [197], EXMARaLDA [168], or Anvil [83]. These types of software allow for the creation of time-aligned annotations on various tiers or tracks, which can in turn be organized hierarchically to show relationship dependencies between different annotations. While these tools are open source, an actual system of coding methods has not been established in the field. As a result, individual researchers, labs, or funded projects develop coding schemes that may be disseminated through the publication of a coding manual.

While multimodal annotation software supports the annotation of both the acoustic signal and the visual signal, a more in-depth and potentially fine-grained annotation of the acoustic signal should preferably be done in Praat [21], for a tutorial on Praat see [20]. Two potential areas of application are conceivable for Praat. First, calculating precise time-aligned measures of gestures-speech interaction requires the accurate demarcation of corresponding domains in speech, e.g., phrase, word, or syllable boundaries which are connected to visually communicative events [113,102]. Second, investigating the prosody-gesture link (e.g., [175,102,48]) requires an analysis of prosodic categories like pitch accents and/or boundary tones. Usually, this type of analysis relies on annotating prosody according to a ToBI system [74] of a given language, which is most conveniently applied in a speech processing tool like Praat. In addition, Praat allows for many kinds of acoustic-phonetic analyses. For gesture research, the individual spectral or temporal parameters can be extracted and related to components of gestures such as the apex, the stroke, the gesture phase or phrase [79]. In particular, with respect to prosody, Praat allows for detailed phonetic analyses of pitch accentuation, concretely their f0-shape or f0-height relations. For instance, measuring the slope of falling pitch accents as a function of the presence or absence of focus can in turn be related to the degree of alignment between a gestures' apex and the pitch accent [59].

Manual annotation is a time-consuming, labor-intensive practice that can highly benefit from technological advances. In terms of gesture annotation, combining (automatic) motion-tracking data with manual annotation allows labelers to achieve consistent measures of time points when an individual gesture begins

[13] https://github.com/sccn/labstreaminglayer
[14] https://github.com/sccn/xdf
[15] https://github.com/labstreaminglayer/App-LabRecorder.git

or ends. Recent effort has focused on using automatic annotation tools to speed up the annotation process (e.g., SPUDNIG [163]; the annotation tool from [71]). While such automatic systems have revealed high reliability with human coders identifying moments of movement (i.e., gesturing) and moments of rest, there is still much work to be done with regard to automatically assessing more nuanced aspects of individual gestures (in terms of type or function with regards to speech).

As previously mentioned, much gesture research accounts for the interaction between gesture and speech, consequently resulting in the need for annotation of multiple modes of communication (i.e., not only gestures but also textual transcriptions and further annotation including, e.g., prosodic annotation or part-of-speech annotation). While researchers have a multitude of automatic tools to facilitate such transcriptions and annotations, it is important that the resulting annotations capture the phenomena of interest. That is, they should maintain information such as hesitations, filled pauses, restarts, etc., as these aspects of speech may be of key interest to researchers who are working on speech fluency, for example.

Annotation schemes To achieve reliability, comparability, and ease of multimodal data processing, several coding schemes have been developed for (gesture) annotation. Today, an internet search will return dozens of proposed coding schemes, such as M3D [165], OTIM [19], LASG [24], NEUROGES [96], MUMIN [4]. While many of these coding systems were designed to establish and develop standard annotation procedures to assess the form and communicative function of co-speech gestures, the theoretical foundations underlying each system vary widely, as well as the aspects of gestures that the system proposes to annotate. For example, a recent review of 10 gesture annotation systems [164] showed how most systems largely agree on approaches to code gestural form (e.g., handshape, palm orientation, etc.); however, only about half of the reviewed systems included guidelines for articulators other than the hands (e.g., head movements or facial expressions, etc.). For these articulators, specific annotation schemes have been developed (e.g., for facial expressions [46], gesture timing [84], phonetics [93], turns [173]), also for species other than humans [26,140,192,194]. Methodological differences in assessing gestural meaning are even more pronounced. For example, in the field of gesture studies, McNeill's (1992) [113] conception of gestures being iconic, metaphoric, deictic, or beat types is widely accepted, yet only one of the reviewed annotation systems directly adopts McNeill's categorization of gestures, and only one system takes a "dimensional" approach to assess gestural meaning. Specifically, M3D labels a gesture's meaning in terms of degrees of iconicity, metaphoricity, or deixis as proposed by McNeill in 2006 [114]. Other systems either do not account for gestural meaning or develop their own taxonomies based on criteria stemming from form and/or function. Similar challenges apply to the annotation of sign languages. While the use of glosses to refer to (the meaning of) signs is consistent and signbanks (sign language resources) may link lemma collections to video corpora (e.g., [33]), there are

several annotation systems for phonological parameters of manual cues (e.g., [63,182,35]) and separate systems for non-manual cues (eyes: [30]; mouth: [34]). In non-human animal research, additional strategies come into play to assess the meaning of a gesture: context and response of the recipients [99,66,7].

The availability of so many labeling options offers the advantage that researchers may choose to adopt a particular labeling system over another, as it may be particularly relevant to answer the types of questions the researcher aims to answer. For example, a researcher interested in the pragmatic functions of gestures may be interested in the CorpAGEst labeling scheme [23]. However, the field may also benefit from adopting more standardized terminology and approaches to the assessment of gestural data (e.g., gesture classification, the classification of the pragmatic meanings associated with gesture, gesture phasing schemes; c.f., section 1.2).

Any approach to labeling gestural data should be widely accessible to the general community and easily adaptable. Indeed most labeling systems merely publish a short manual that briefly describes the annotation values that are to be employed in ELAN. NEUROGES offers occasional training seminars to become official NEUROGES-Certified labelers. The manual for the M3D system offers more detailed examples, workflow tips, and solutions to ambiguous cases and the system will be soon supported by further online training materials. Thus, the community as a whole would benefit greatly from converging on a set terminology and key approaches to assessing gesture, and crucially to making this approach as openly accessible and reproducible as possible. Importantly, this should be taken into account when considering how we can advance in tandem with computer technology specialists.

In this context, it is also important to keep apart *annotation* (labeling an annotation unit) from *segmentation* (identifying an annotation unit) [180]. A well-known example from gesture studies is the individuation of gestures and the demarcation of gesture phases (preparation, stroke, retraction – see [78]). Since identifying annotation units is logically prior to annotation, differences in the identification of annotation units do not only affect labeling, but also impact any time-related analysis, from descriptive figures (e.g., number of gestures, mean stroke length) to time-series analysis (e.g. [150]). Moreover, they also affect the analysis of temporal relations between the relative timing of and some relationships between communicative events on different channels (see e.g., [135]). Different segmentations furthermore lead to different outputs of multimodal behavior-producing systems, where output behaviors are regimented in terms of the *Behavior Markup Language* [193], a representation format that captures the timing of various signals relative to each other. Note that segmentation poses a genuine problem for evaluating annotation schemes, which is usually carried out in terms of agreement studies (see the corresponding chapters in [70]). The reason is that widespread statistical coefficients like Kappa [31] work on different annotators' labels of a common set of items – whereas it is the very items that are in question in segmentation. To this end, researchers have developed

unitizing [87] or segmentation agreement [180] approaches, the latter being also used within the video annotation software ELAN by means of *Staccato* [108].

Hence there are detailed annotation schemes only for a small subset of multimodal communication signals, and there are abstract markup languages for representing structured uni- or multimodal signals. There is still a need for a unified form- and function-based annotation system for the full range of communicative behaviors, a need already addressed but not satisfactorily solved in Birdwhistell's Kinesics [18], which might also be successfully applied in species other than humans.

Multimodal Corpora Annotated multimodal corpora represent a crucial resource for gesture researchers, as a single corpus may be used to answer a whole host of research questions through different analyses or the addition of further annotation. However, in order to make the most of such resources, it is necessary that they be made openly accessible. Online repositories such as TalkBank[16], The Language Archive[17], or Ortolang[18] host a large number of multimodal corpora. A browse through the multimodal corpora available on these websites makes apparent the vast diversity of the types of corpora available. For example, the TalkBank repository hosts multiple subcomponents which host corpora specific to child development (e.g., CHILDES [109]), multilingualism (e.g., BilingBank), or clinical research (e.g., DementiaBank [16]).

In line with what has been previously mentioned for annotation schemes, the development of multimodal corpora has also often been carried out in order to answer very specific research questions or to reach particular objectives. As such, multimodal corpora often present a lot of variation. For instance, they may include spontaneous conversational speech or play (e.g., the Signes et Familles corpus, [121], or the EVA corpus [116]), recorded presentations (e.g., the M3D-TED corpus [165]) to structured task-based corpora (e.g., the SAGA corpus [106]) to a combination thereof (DGS [154]). For a discussion of the specificity of multimodal corpora, as well as a general overview of the goals of multimodal corpus linguistics, see [138]. Importantly, the multitude of diverse multimodal corpora which are openly available in different online repositories represents a rich resource (for research on humans but not other species) that can be exploited to foster joint advancement in technological and multimodal communication research.

3.5 Machine Learning

In addition to manual labeling, there is also the option of automated data processing by appropriately trained systems. The data generated in this way is much more error-prone than human-generated data, but can be used as a basis for the actual annotation so that the data only needs to be corrected (e.g. by filtering

[16] www.talkbank.org
[17] https://archive.mpi.nl/tla/
[18] https://www.ortolang.fr/

them so that only relevant data points are processed, or by extending them with features that are helpful but beyond the scope of the actual annotation; e.g., [202,117]).

Most tools are accessible if video footage of the communication is available. There are countless tools (for a review see: [122]) to recognize (communicating) persons and their pose based on these videos (e.g., OpenPose [28], MMPose [32], PaddleDetection [134], OpenFace [12], although the latter has been reported to require a non-negligible amount of manual verification [62]). Depending on the system and model, the general pose of the actor can be determined based on predefined key points such as the head, elbows, shoulders, or feet, but also the position and posture of the hands and fingers, facial movements, and gaze directions. These data can then be used to quantify a number of parameters such as the amount of motion of an actor or a particular body part [186]. In animal research, tools such as DeepLabCut [112] or SLEAP, revolutionized the ease with which researchers can track morphologically unique body poses in a wide range of animals. DeepLabCut also allows for flexible tracking of objects together with biological objects, e.g., in communicative contexts where there are also interactions of objects.

In addition to Pose Estimation, there are also systems that already perform appropriate classifications at a wide variety of levels. For instance, there are classifiers that determine the action and interactions that people perform [205], what type of hand gestures are performed [131], or the classification of emotions based on facial expressions [98], body language [3], or spoken language [81]. As a current limitation, these systems are usually trained on very specific training data and thus, the target classes are predefined. Further, the machine learning community is increasingly taking up the challenge to employ state-of-the art machine learning architectures for manual gesture detection (e.g., [91]), which thereby goes beyond the current tools that researchers might already use (e.g., Spudnig).

Such supporting Machine Learning tools are not limited to visual information but include also acoustic information. For most analyses, a conversion of spoken words to a text format seems necessary, providing a base for the normalization of spoken language which is needed to identify aspectual differences (e.g., in the intonation of otherwise identical words). Countless tools are available to convert spoken language to text (e.g., Whisper [158]). However, depending on the tool, a lot of information can be lost, because many tools may clean up the speech directly: For example, stuttering, intonation pauses, and overlapping of speakers are usually lost.

Depending on the application, it might be important to use systems that translate from one modality to another, e.g., by reconstructing hand gestures from body movements [125] or facial movements from speech [161]. These models can then be used to analyze the correlation between modalities [27]). However, translations can also contribute to accessibility, e.g., through models that recognize words based on lip movements (lip reading) [49] or translate sign languages into spoken languages [73].

Not only pre-trained models need to be used, but one can train models by using simulations [196] or by understanding the annotation process itself as a bootstrapping approach (a model is trained with a small dataset, the human corrects and improves the model, which results in speeding up the annotation process [198]). Promising developments that increase flexibility are tools such as NOVA [15]. Using a principle of "collaborative machine learning", NOVA provides a general user interface tailored towards manual annotation, with further integration of supervised machine classification methods (such as Support Vector Machines or Neural Networks). These can be trained on initial manual annotations, in turn allowing the user to hand-correct and retrain the classifier. In general, a restriction is that many of these modern machine learning systems are not freely available or easy to set up and use, require expensive hardware (graphics cards), or require specialized programming effort, making these systems very inaccessible to researchers from disciplines other than computer science.

In addition, personal data often raise ethical issues. For example, models that reconstruct speech based on lip movements can be used for people who do not like to be overheard, or there are initial approaches that can track people and movements based on reflected Wi-Fi signals [160,53]. The machine learning community generates many ongoing interdisciplinary implications that go beyond developing tools. A good example in this regard is the recent advance in the artificial recreation of believable human co-speech gestures, or gesture synthesis in short [128]. While it may seem that gesture synthesis might only have implications for human-computer interaction systems such as avatar design in games or other contexts, it also indirectly informs theories in cognitive science and linguistics. For example, machine learning models trained on associations of acoustic signal with body poses occurring as co-speech gestures, become very capable of synthesizing rhythmic beat-like gestures from novel acoustic signals alone [55,128]. Therefore, such models show that there is information in speech sounds that can reliably predict the presence of a gesture [203], and they also allow identifying what features in speech are predictive for specific kinematic properties in gesture [50].

Exciting further research in this direction comes from work that makes use of joint multimodal embedding spaces. Neural networks (transformers) trained on detecting co-regularities between gesture poses and speech content (represented as text), for instance, can make reliable predictions about discourse markers, and can differentiate between the language spoken (Spanish vs. English) based on body poses alone [2]. These findings from the machine learning community thus forward theorizing in multimodal communication research about the information available in multimodal signals, and how they inform one another, and they will also further shape cognitive science research about what information humans use in practice during communication [187].

3.6 Factors that can Accelerate Integration between Disciplines

In order to promote interdisciplinarity in multimodality research and implement automatic tools and methods in the process, cooperation and mutual help are

crucial. This entails understanding the other field's questions and mode of inquiry, to formulate joint research questions and possibly conduct studies in a way that other modes of inquiry may become available. Moreover, this interoperability is a mutual endeavor, in that researchers need to become more literate about each other's core state-of-the-art and the key methods for investigating the same phenomenon.

Data requirements, metadata practices, and tools to overcome privacy challenges to open data If we are moving towards more integrated fields, rather than isolated pockets of specializations [129], bridges between disciplines need to be built. One such bridge is metadata maintenance, i.e. how we archive data so that it is maximally reusable later, possibly also in other fields with different conventions. The fast pace of algorithm development in computer science drives innovation, but may sometimes be at odds with requirements for empirical studies that try to use these algorithms as research tools. For example, replicability, evaluation of validity, and e.g., clinical utility require systematic investigation of larger data samples [85]. The resulting resources (e.g., stimulus databases, tests, algorithms, tutorials, workshops) should ideally be available to the scientific community according to open science principles (e.g., [184,165]).

Open science principles can be a challenge for the protection of privacy. Indeed, especially in multimodal communication research on humans, original data that support one's analyses are often not openly shared, because they often consist of audiovisual recordings of identifiable people. There is, however, an increasing number of tools that allow to partially mask the identities from video and audio automatically, while still extracting non-identifiable information that can support analyses, such as facial, hand, and body pose information [82,133,156].It is important to note that these tools do not count as anonymization tools, because either the transformed sound is still re-transformable to its original (thereby allowing identification in principle) or is still present next to a bodily mask. Indeed, as the yearly voice privacy challenge shows [19], the anonymization of voices while maintaining their rich transmission of information is still an unsolved problem. In any case, new computer vision and signal processing methods allow for decreasing privacy risks when sharing multimodal data, which is a positive development. Hopefully, these practices will be increasingly picked up by researchers working in different domains.

Community of learners We believe that an important way to become more literate as a community of researchers is to take up responsibilities that support a "community of learners". In the most general sense, this means that as researchers, we strive to provide the didactic means to facilitate becoming literate in the particular methods employed. Practically, this can mean a number of things. We need to write more transparent "computationally reproducible" manuscripts ensuring that data and code are well-annotated with additional documentation provided. There are many tools available now to fully

[19] https://www.voiceprivacychallenge.org/

integrate and write one's manuscript in a "computationally reproducible way", enabling readers to follow the computation procedures step-by-step using tools like RMarkdown or Jupyter notebook [142] or more recent platform agnostic tools (Quarto[20]), which depict one strategy to make publications and tools easier to reuse.

Another practical implication of a community of learners is that new tools or terminologies are supported in the literature with hands-on tutorials that are written for either more or less informed audiences. The Huggingface platform [21], for example, offers a wide collection of pre-built and even fine-tuned machine learning models provided by the community, including sample code and "spaces" where they can be tried out directly. This may mean that a tutorial on machine learning or phonology will look very different depending on whether you are tailoring it toward readers in computer science or linguistics. Lastly, of course, there must be undergraduate and graduate curricular integration at universities that ensure that the different fields can, and do cooperate.

3.7 Summary

In this section, we provided insights into the currently available tools, methods, and databases for data collection, enrichment, and analysis in multimodal communication research. A non-exhaustive overview of these resources is given in Figure 3. Focusing on the visual and acoustic modes of communication, we started by presenting possible recording techniques for the visual mode (section 3.1) and communicative modes in general (section 3.2). Before or after recording, the synchronization between different types of data needs to be ensured (section 3.3). Subsequently, data can be enriched by manual segmentation and annotation (section 3.4) or by using natural language processing (section 3.5). Then, datasets can be analyzed, assembled to form multimodal corpora (section 3.4) and used across research disciplines to move multimodal communication research forward (section 3.6).

4 Discussion

This paper aimed to provide an overview of the currently available tools and methods in multimodal communication research. We presented state-of-the-art tools that can facilitate research in this field and expressed specific requirements to achieve feasible technological development that can be integrated into data collection, preparation, and analysis in the visual and acoustic domains. Multimodal data acquisition and digital data analysis are yet relatively new challenges in communication research (although they have been addressed with a different focus in semiotics [148, Part 2], Conversation Analysis [167], and human-computer interaction [132]). Therefore, we highlighted the need to introduce specific terminology and importantly, presented the availability of various

[20] https://github.com/quarto-dev/quarto
[21] https://huggingface.co/ (last visited 27.01.2023)

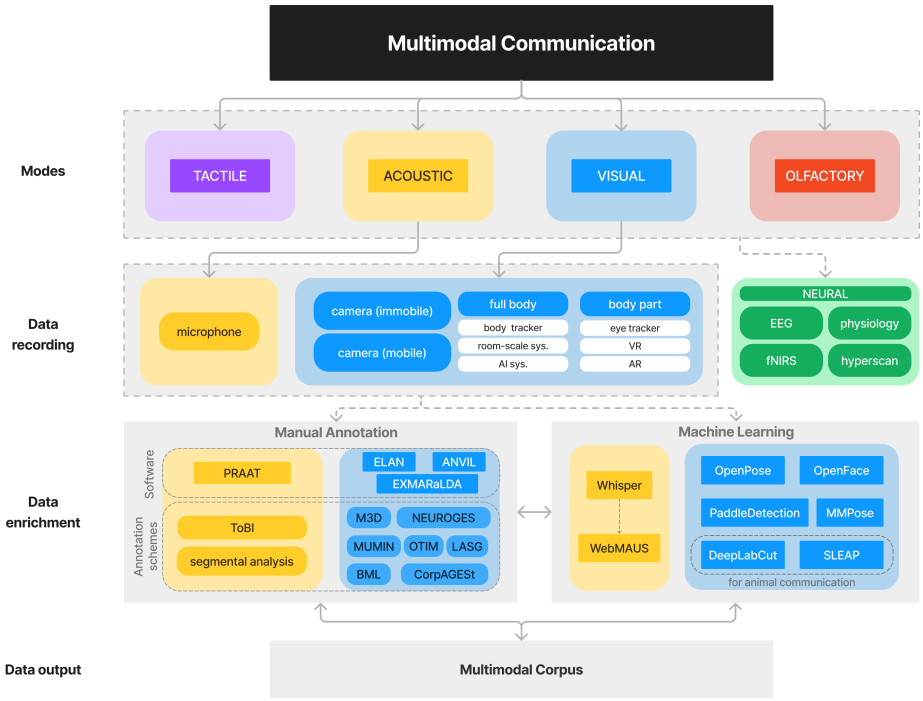

Figure 3. Non-exhaustive overview over available tools for data recording and enrichment.

kinds of systems. These are steps in establishing congruency within the fields of multimodal communication research on the one hand, but also computer science on the other hand.

We argue that increased interdisciplinarity in cognitive science and computer science with regard to multimodal communication will have important implications: On the one hand, increased literacy by linguists/cognitive scientists in computer science implies a better understanding of what machine learning algorithms actually do, and what they can do for multimodal communication in the future (e.g., better understandings of large language models like Lambda or GPT3 [181]). Similarly, while multimodal researchers might hope that annotating minimal meaningful units as metaphoric gesture strokes will soon be something of the past, it is also clear that machine learning systems cannot learn to classify linguistic categories if researchers do not first agree on the definition or application of those categories. On the other hand, computer scientists becoming more literate in aspects of multimodal communication will also prevent downright renderings of the object of study (e.g., sign languages being understood as incomplete languages; [199]) and minimally it will also combine what can be done with what a particular community of users benefits from (c.f., [47,89]).

In short, mutually informed research communities can advance research in their own respective fields with crucial understandings from other disciplines. As a side benefit, interdisciplinarity also provides training data for machine learning for processing of increasingly complex linguistic and multimodal structures (manual annotations lead to the training of models which in turn can improve the underlying annotations; e.g., Multimodal Distributional Semantics [13,157]). Investigating the synchronization of acoustic and visual aspects of communication opens the way to investigate multimodal transfer learning in new ways[22]. One example is how data from one modality are technologically used to segment or disambiguate data in another modality or to reconstruct them in an expectation-driven manner when they have not been manifested or detected [14,139]. This approach would eventually allow us to study the diverse relationships and interdependencies of the modalities involved – whether on the level of signals or their higher-level representations. An application of multimodal interoperability allows the transfer and leveraging of annotations of one modality for the annotation or automatic processing of another modality.

We can imagine the analysis of communication to be broken down into different levels of observation, which can be visualized as follows (cf. Figure 2): **signal** → **event** → **communicative behavior** → **[from quantitative to qualitative] meaning** → **embedding in utterance context**. From left to right, there is a tendency for an increased amount of interpretation, even from a human point of view. In terms of measurements, we are expecting an increase of nominal scale measurements (classifications; cf. the distinction between well-understood type-i measurements and less understood type-ii measurements by [61] – on the right-hand side it is even not obvious what the scales should be, indicated by the square bracket parenthesis ("from quantitative to qualitative"). With the raw signal as input to a machine learning algorithm, it is therefore an increasingly difficult task to automatically classify the respective units of observation: (Acoustic and visual) Signals and communicative events can be identified more easily and automatically than the more complex communicative acts that involve meaning, unless there is a huge amount of data annotated for the various observational levels.

These considerations, against the backdrop of the overview given mainly in sections 2 and 3, point at a number of (near-) future challenges:

- The majority of tools focus on the initial – and fundamental – step of recording signals of various provenance, and in a synchronized way. In order to progress from tracking to parsing, tools for segmentation and classification are of course welcome. As above mentioned, this is not an obligation for computer science in the first place; rather communication researchers are in the need to agree on annotation schemes for individual signals as well as integrated ones for multiple signals (see section 3.4), and provide annotated datasets.
- Computational linguistics has developed powerful algorithms and tools for processing text. However, these devices cannot readily be applied to spoken

[22] For a recent overview of transfer learning, see [195].

or signed languages: the speech or sign stream are signals that do not come in discrete units [159]. Hence, the acoustic or visual signal has to be transcribed first. In the acoustic domain, automatic transcription is brought about by Automatic Speech Recognition (ASR) systems.
However, current ASR systems miss out on a couple of features of the audio file [103]. For instance, the sound string "bob are (.) uh is_sleeping" (using minimal notation following a conversation-analytic transcription system [173]) is transcribed as *Bob is sleeping*. While this output can be input to natural language processing tools, it lacks two peculiarities that are important for multimodal communication studies: the speech error [97] is ignored as is the hesitation marker *uh*, which may both trigger a lexical access-related gesturing [86]. In order to capture these features of spoken language, ASR systems have to be developed "more impurely" from a phonetic and incremental point of view, including that communication researchers agree on a useful transcription system (for humans *and* computers) – and provide big amounts of transcribed data.

– While temporal synchrony is an important aspect of multiple signals in face-to-face interactions, it yet does not fully determine coordination of semantic meaning of those signals [162]. While temporal alignment is an observable, measurable feature of multi-channel communication, semantic integration involves interpretation [94]. That is, with temporal alignment as with any signal, the meaning (also the grammar) is not in the signal but imposed by the one processing the signal. Hence, it remains to be seen how far automatic multi-signal classifications (the top-right node in Figure 2) can be pushed.
– We observe a plurality of methods: there are VR-based tracking methods (see section 3.1), methods that work on the basis of video recordings (currently the most widespread ones; sections 3.1 and 3.5), and physiological and neurological recordings (section 3.2). This pluralism poses the questions (i) whether the approaches should be developed into enhanced stand-alone pipelines, or (ii) how they can be inter-operated despite their *prima facie* incommensurability. In either case, it would surely be beneficial to incorporate some of the advances from computer vision and machine learning (section 3.5).

We believe that interoperability in multimodal communication will play an important role in the further development of multimodal annotation. Thus, we envision a system in which actions are generated by subjects in controlled environments to provide experimenters with controlled access to multimodal data. This can ground the communicative aspects involved in these actions and their manifestations in the form of gestures, gazes, body movements, etc, but also in equally controlled objects, their properties, and relations. One can see in this an alternative to independent approaches to multimodal research, an alternative that is integrating the point of view of combining various views on multimodality.

Concretely, the wishes of our consortium towards the automatization of processes in multimodal research concern multiple steps in data acquisition. Starting with the facilitation of programming communication experiments (partly auto-

mated by functional software) and generating synchronized (acoustic and visual) stimuli, technical innovation can unburden research before data collection. As has been brought up in this paper, technical innovation can improve the processing of data for analysis. This may include the automatic identification of specific factors of visual data annotation (on-/offset and turning points of movements; temporal alignment between acoustic and visual cues; grouping and clustering of e.g., gestures; up to identification of smallest meaning bearing units). Ideally, this can ultimately be achieved for gestures as well as signs and for human and non-human communication. Similarly, the automatic processing of acoustic signals could be facilitated by providing better segmentation of vocalized input or automated prosodic annotation. This can ideally lead to the training of neural networks (as mentioned in section 3.5) which could largely support the annotation of big data sets.

This leads us to a final, self-reflective note: We started out by envisioning "a roadmap for technical innovation in multimodal communication research". On every path on this roadmap we observed, however, the need for well-worked out formats, standards and guidelines, *defining our units of analysis* in the first place. Addressing this, important roadmap ground will already be covered.

Table of contributions (contributions indicated by grey cell color)

Initials	Concept	Sec 1	Sec 2	Sec 3	Sec 4	Rev/Edit
AG	■	■	■	■	■	■
FA						■
IB				■		
AC	■	■				■
LF				■		■
SF	■	■		■		■
AH				■		
OH		■				
FK				■		■
JL		■				
KL						■
AL	■	■		■	■	■
AM					■	
KTN						■
WP	■		■	■	■	■
PP			■			■
PR				■		■

Continued on next page

Table of contributions (contributions indicated by grey cell color) (Continued)

Initials	Concept	Sec 1	Sec 2	Sec 3	Sec 4	Rev/Edit
PSR						
MSR						
PS						
SS						
VS						
PT						
CvE						

References

1. Abner, N., Cooperrider, K., Goldin-Meadow, S.: Gesture for linguists: A handy primer. Language and Linguistics Compass **9**(11), 437–451 (2015). https://doi.org/10.1111/lnc3.12168
2. Abzaliev, A., Owens, A., Mihalcea, R.: Towards understanding the relation between gestures and language. In: Proceedings of the 29th International Conference on Computational Linguistics. pp. 5507–5520 (2022)
3. Ahmed, F., Bari, A.H., Gavrilova, M.L.: Emotion recognition from body movement. IEEE Access **8**, 11761–11781 (2019). https://doi.org/10.1109/ACCESS.2019.2963113
4. Allwood, J., Cerrato, L., Jokinen, K., Navarretta, C., Paggio, P.: The mumin coding scheme for the annotation of feedback, turn management and sequencing phenomena. Language Resources and Evaluation **41**(3), 273–287 (2007). https://doi.org/10.1007/s10579-007-9061-5
5. Alviar, C., Dale, R., Dewitt, A., Kello, C.: Multimodal coordination of sound and movement in music and speech. Discourse Processes **57**(8), 682–702 (2020). https://doi.org/10.1080/0163853X.2020.1768500
6. Alviar, C., Kello, C.T., Dale, R.: Multimodal coordination and pragmatic modes in conversation. Language Sciences p. 101524 (2023). https://doi.org/10.1016/j.langsci.2022.101524
7. Amici, F., Oña, L., Liebal, K.: Compositionality in primate gestural communication and multicomponent signal displays. International Journal of Primatology (2022). https://doi.org/10.1007/s10764-022-00316-9
8. Anderson, C.A., Wiggins, I.M., Kitterick, P.T., Hartley, D.E.H.: Adaptive benefit of cross-modal plasticity following cochlear implantation in deaf adults. Proceedings of the National Academy of Sciences of the United States of America **114**(38), 10256–10261 (2017). https://doi.org/10.1073/pnas.1704785114
9. Aranyi, G., Pecune, F., Charles, F., Pelachaud, C., Cavazza, M.: Affective Interaction with a Virtual Character Through an fNIRS Brain-Computer Interface. Frontiers in Computational Neuroscience **10** (Jul 2016). https://doi.org/10.3389/fncom.2016.00070
10. Balardin, J.B., Zimeo Morais, G.A., Furucho, R.A., Trambaiolli, L., Vanzella, P., Biazoli, C., Sato, J.R.: Imaging Brain Function with Functional Near-Infrared Spectroscopy in Unconstrained Environments. Frontiers in Human Neuroscience **11**, 258 (May 2017). https://doi.org/10.3389/fnhum.2017.00258

11. Balconi, M., Fronda, G., Bartolo, A.: Affective, Social, and Informative Gestures Reproduction in Human Interaction: Hyperscanning and Brain Connectivity. Journal of Motor Behavior **53**(3), 296–315 (May 2021). https://doi.org/10.1080/00222895.2020.1774490
12. Baltrusaitis, T., Zadeh, A., Lim, Y.C., Morency, L.P.: Openface 2.0: Facial behavior analysis toolkit. In: 2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018). pp. 59–66. IEEE (2018). https://doi.org/10.1109/FG.2018.00019
13. Baroni, M.: Grounding distributional semantics in the visual world. Language and Linguistics Compass **10**(1), 3–13 (2016). https://doi.org/10.1111/lnc3.12170
14. Barros, P., Parisi, G.I., Fu, D., Liu, X., Wermter, S.: Expectation learning for adaptive crossmodal stimuli association. In: EUCog Meeting Proceedings. EUCog, EUCog Meeting (Nov 2017). https://doi.org/ARXIV:1801.07654
15. Baur, T., Heimerl, A., Lingenfelser, F., Wagner, J., Valstar, M.F., Schuller, B., André, E.: eXplainable cooperative machine learning with NOVA. KI – Künstliche Intelligenz (Jan 2020). https://doi.org/10.1007/s13218-020-00632-3
16. Becker, J.T., Boller, F., Lopez, O.L., Saxton, J., McGonigle, K.L.: The natural history of Alzheimer's disease: Description of study cohort and accuracy of diagnosis. Archives of Neurology **51**(6), 585–594 (1994). https://doi.org/10.1001/archneur.1994.00540180063015
17. Bierman, A.K.: That there are no iconic signs. Philosophy and Phenomenological Research **23**(2), 243–249 (1962). https://doi.org/10.2307/2104916
18. Birdwhistell, R.L.: Kinesics and Context. Conduct and Communication Series, University of Pennsylvania Press, Philadelphia (1970). https://doi.org/10.9783/9780812201284
19. Blache, P., Bertrand, R., Ferré, G., Pallaud, B., Prévot, L., Rauzy, S.: The corpus of interactional data: A large multimodal annotated resource. In: Handbook of linguistic annotation, pp. 1323–1356. Springer (2017). https://doi.org/10.1007/978-94-024-0881-2_51
20. Boersma, P.: The use of Praat in corpus research. In: Durand, J., Gut, U., Kristoffersen, G. (eds.) The Oxford handbook of corpus phonology, pp. 342–360. Oxford handbooks in linguistics, Oxford University Press, Oxford (2014). https://doi.org/10.1093/oxfordhb/9780199571932.013.016
21. Boersma, P., Weenink, D.: Praat: doing phonetics by computer [computer program] version 6.3.03. http://www.praat.org/ (2022)
22. Bohannon, R.W., Harrison, S., Kinsella-Shaw, J.: Reliability and validity of pendulum test measures of spasticity obtained with the polhemus tracking system from patients with chronic stroke. Journal of neuroengineering and rehabilitation **6**(1), 1–7 (2009). https://doi.org/10.1186/1743-0003-6-30
23. Bolly, C.T.: CorpAGEst annotation manual (ii. speech annotation guidelines) (2016), https://corpagest.wordpress.com/working-papers/
24. Bressem, J.: A linguistic perspective on the notation of form features in gestures. In: Müller, C., Cienki, A., Fricke, E., Ladewig, S.H., McNeill, David und Bressem, J. (eds.) Body – Language – Communication. An International Handbook on Multimodality in Human Interaction, Handbooks of Linguistics and Communication Science, vol. 1, chap. 70, pp. 1079–1089. De Gruyter Mouton, Berlin and Boston (2013). https://doi.org/10.1515/9783110261318.1079
25. Burks, A.W.: Icon, index, and symbol. Philosophy and Phenomenological Research **9**(4), 673–689 (1949). https://doi.org/10.2307/2103298

26. Caeiro, C.C., Waller, B.M., Zimmermann, E., Burrows, A.M., Davila-Ross, M.: OrangFACS: A muscle-based facial movement coding system for orangutans (Pongo spp.). International Journal of Primatology **34**(1), 115–129 (2013). https://doi.org/10.1007/s10764-012-9652-x
27. Caliskan, A., Bryson, J., Narayanan, A.: Semantics derived automatically from language corpora contain human-like biases. Science **356**(6334), 183–186 (Apr 2017). https://doi.org/10.1126/science.aal4230
28. Cao, Z., Hidalgo Martinez, G., Simon, T., Wei, S., Sheikh, Y.A.: OpenPose: Realtime multi-person 2d pose estimation using part affinity fields. IEEE Transactions on Pattern Analysis and Machine Intelligence (2019). https://doi.org/10.1109/TPAMI.2019.2929257
29. Cavallo, A., Koul, A., Ansuini, C., Capozzi, F., Becchio, C.: Decoding intentions from movement kinematics. Scientific reports **6**(1), 1–8 (2016). https://doi.org/10.1038/srep37036
30. Chételat-Pelé, E., Braffort, A., Véronis, J.: Annotation of non manual gestures: Eyebrow movement description. In: sign-lang@ LREC 2008. pp. 28–32. European Language Resources Association (ELRA) (2008)
31. Cohen, J.: A coefficient of agreement for nominal scales. Educational and Psychological Measurement **20**, 37–46 (1960). https://doi.org/10.1177/001316446002000104
32. Contributors, M.: Openmmlab pose estimation toolbox and benchmark. https://github.com/open-mmlab/mmpose (2020)
33. Cormier, K., Crasborn, O., Bank, R.: Digging into signs: Emerging annotation standards for sign language corpora. In: Efthimiou, E., Fotinea, S.E., Hanke, T., Hochgesang, J.A., Kristoffersen, J., Mesch, J. (eds.) Proceedings of the LREC2016 7th Workshop on the Representation and Processing of Sign Languages: Corpus Mining. pp. 35–40. European Language Resources Association (ELRA), Portorož, Slovenia (May 2016)
34. Crasborn, O., Bank, R.: An annotation scheme for the linguistic study of mouth actions in sign languages (2014), https://hdl.handle.net/2066/132960
35. Crasborn, O., Zwitserlood, I., van der Kooij, E., Ormel, E.: Global SignBank manual, version 2 (11 2020). https://doi.org/10.13140/RG.2.2.16205.67045/1
36. Cutler, A., Dahan, D., Van Donselaar, W.: Prosody in the comprehension of spoken language: A literature review. Language and speech **40**(2), 141–201 (1997)
37. Dale, R.: The possibility of a pluralist cognitive science. Journal of Experimental and Theoretical Artificial Intelligence **20**(3), 155–179 (2008). https://doi.org/10.1080/09528130802319078
38. Dale, R., Warlaumont, A., Johnson, K.: The fundamental importance of method to theory. Nature Reviews Psychology **2**, 55–66 (2022). https://doi.org/10.1038/s44159-022-00120-5
39. Danner, S.G., Barbosa, A.V., Goldstein, L.: Quantitative analysis of multimodal speech data. Journal of Phonetics **71**, 268–283 (2018). https://doi.org/10.1016/j.wocn.2018.09.007
40. Dogdu, C., Kessler, T., Schneider, D., Shadaydeh, M., Schweinberger, S.R.: A comparison of machine learning algorithms and feature sets for automatic vocal emotion recognition in speech. Sensors **22**(19) (2022). https://doi.org/10.3390/s22197561
41. Drimalla, H., Baskow, I., Behnia, B., Roepke, S., Dziobek, I.: Imitation and recognition of facial emotions in autism: A computer vision approach. Molecular Autism **12**(1) (2021). https://doi.org/10.1186/s13229-021-00430-0

42. Ebert, C., Ebert, C.: Gestures, demonstratives, and the attributive/referential distinction. Talk at Semantics and Philosophy in Europe 7, ZAS, Berlin (2014)
43. Ebert, C., Ebert, C., Hörnig, R.: Demonstratives as dimension shifters. Proceedings of Sinn und Bedeutung **24**(1), 161–178 (2020)
44. Ehinger, B.V., Dimigen, O.: Unfold: an integrated toolbox for overlap correction, non-linear modeling, and regression-based EEG analysis. PeerJ **7**, e7838 (2019). https://doi.org/10.7717/peerj.7838
45. von Eiff, C.I., Frühholz, S., Korth, D., Guntinas-Lichius, O., Schweinberger, S.R.: Crossmodal benefits to vocal emotion perception in cochlear implant users. iScience **25**(12), 105711 (2022). https://doi.org/10.1016/j.isci.2022.105711
46. Ekman, P., Friesen, W.V.: Facial Action Coding System: A Technique for the Measurement of Facial Movement. Consulting Psychologists Press, Palo Alto, CA (1978). https://doi.org/10.1037/t27734-000
47. Erard, M.: Why sign-language gloves don't help deaf people. The Atlantic, https://www.theatlantic.com/technology/archive/2017/11/why-sign-language-gloves-dont-help-deaf-people/545441/ (2017)
48. Esteve-Gibert, N., Prieto, P.: Prosodic structure shapes the temporal realization of intonation and manual gesture movements. Journal of Speech, Language, and Hearing Research **56**(3), 850–864 (2013). https://doi.org/10.1044/1092-4388(2012/12-0049)
49. Fernandez-Lopez, A., Sukno, F.M.: Survey on automatic lip-reading in the era of deep learning. Image and Vision Computing **78**, 53–72 (2018). https://doi.org/10.1016/j.imavis.2018.07.002
50. Ferstl, Y., Neff, M., McDonnell, R.: Understanding the predictability of gesture parameters from speech and their perceptual importance. In: Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents. pp. 1–8 (2020). https://doi.org/10.1145/3383652.3423882
51. Filippeschi, A., Schmitz, N., Miezal, M., Bleser, G., Ruffaldi, E., Stricker, D.: Survey of motion tracking methods based on inertial sensors: A focus on upper limb human motion. Sensors **17**(6), 1257 (2017). https://doi.org/10.3390/s17061257
52. Frühholz, S., Schweinberger, S.R.: Nonverbal auditory communication – Evidence for integrated neural systems for voice signal production and perception. Progress in Neurobiology **199**, 101948 (2021). https://doi.org/10.1016/j.pneurobio.2020.101948
53. Geng, J., Huang, D., De la Torre, F.: Densepose from wifi. arXiv preprint arXiv:2301.00250 (2022). https://doi.org/10.48550/arXiv.2301.00250
54. Gerloff, C., Konrad, K., Kruppa, J., Schulte-Rüther, M., Reindl, V.: Autism Spectrum Disorder Classification Based on Interpersonal Neural Synchrony: Can Classification be Improved by Dyadic Neural Biomarkers Using Unsupervised Graph Representation Learning? In: Abdulkadir, A., Bathula, D.R., Dvornek, N.C., Habes, M., Kia, S.M., Kumar, V., Wolfers, T. (eds.) Machine Learning in Clinical Neuroimaging, vol. 13596, pp. 147–157. Springer Nature Switzerland, Cham (2022). https://doi.org/10.1007/978-3-031-17899-3_15
55. Ginosar, S., Bar, A., Kohavi, G., Chan, C., Owens, A., Malik, J.: Learning individual styles of conversational gesture. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3497–3506 (2019)
56. Ginzburg, J., Poesio, M.: Grammar is a system that characterizes talk in interaction. Frontiers in Psychology **7**, 1938 (2016). https://doi.org/10.3389/fpsyg.2016.01938
57. Goodman, N.: Languages of Art. An Approach to a Theory of Symbols. Hackett Publishing Company, Inc., Idianapolis, 2 edn. (1976)

58. Goodwin, C.: Pointing as situated practice. In: Kita, S. (ed.) Pointing: Where Language, Culture, and Cognition Meet, chap. 2, pp. 217–241. Lawrence Erlbaum Associates, Inc., Mahwah, New Jersey (2003). https://doi.org/10.4324/9781410607744
59. Gregori, A., Kügler, F.: Multimodal marking of focus: Articulatory and visual hyperarticulation (submitted)
60. Gussenhoven, C.: The phonology of tone and intonation. Cambridge: Cambridge University Press (2004). https://doi.org/10.1017/CBO9780511616983
61. Gwet, K.: Handbook of Inter-Rater Reliability. STATAXIS Publishing Company, Gaithersburg, MD (2001)
62. Hammadi, Y., Grondin, F., Ferland, F., Lebel, K.: Evaluation of various state of the art head pose estimation algorithms for clinical scenarios. Sensors **22**(18), 6850 (2022). https://doi.org/10.3390/s22186850
63. Hanke, T.: HamNoSys – representing sign language data in language resources and language processing contexts. In: LREC. vol. 4, pp. 1–6 (2004)
64. Hartz, A., Guth, B., Jording, M., Vogeley, K., Schulte-Rüther, M.: Temporal behavioral parameters of on-going gaze encounters in a virtual environment. Frontiers in Psychology **12** (2021). https://doi.org/10.3389/fpsyg.2021.673982
65. Herrmann, A., Pendzich, N.K.: Nonmanual gestures in sign languages. In: Müller, C., Cienki, A., Fricke, E., Ladewig, S.H., McNeill, D., Bressem, J. (eds.) Handbook Body -– Language -– Communication, pp. 2147–2160. DeGruyter Mouton, Berlin, Boston (2014)
66. Hobaiter, C., Byrne, R.W.: The meanings of chimpanzee gestures. Current Biology **24**, 1596–1600 (2014)
67. Holler, J., Levinson, S.C.: Multimodal language processing in human communication. Trends in Cognitive Sciences **23**(8), 639–652 (2019). https://doi.org/10.1016/j.tics.2019.05.006
68. Hosemann, J., Herrmann, A., Steinbach, M., Bornkessel-Schlesewsky, I., Schlesewsky, M.: Lexical prediction via forward models: N400 evidence from German Sign Language. Neuropsychologia **51**(11), 2224–2237 (2013). https://doi.org/10.1016/j.neuropsychologia.2013.07.013
69. Höhle, T.N.: Über Komposition und Derivation: zur Konstituentenstruktur von Wortbildungsprodukten im Deutschen. Zeitschrift für Sprachwissenschaft **1**(1), 76–112 (1982). https://doi.org/10.1515/zfsw.1982.1.1.76
70. Ide, N., Pustejovsky, J. (eds.): Handbook of Linguistic Annotation. Springer Netherlands, Dordrecht (2017). https://doi.org/10.1007/978-94-024-0881-2_1
71. Ienaga, N., Cravotta, A., Terayama, K., Scotney, B.W., Saito, H., Busà, M.G.: Semi-automation of gesture annotation by machine learning and human collaboration. Language Resources and Evaluation pp. 1–28 (2022). https://doi.org/10.1007/s10579-022-09586-4
72. Jaimes, A., Sebe, N.: Multimodal human–computer interaction: A survey. Computer Vision and Image Understanding **108**(1), 116–134 (2007). https://doi.org/10.1016/j.cviu.2006.10.019, special Issue on Vision for Human-Computer Interaction
73. Jiang, Z., Moryossef, A., Müller, M., Ebling, S.: Machine translation between spoken languages and signed languages represented in signwriting. arXiv preprint arXiv:2210.05404 (2022). https://doi.org/https://doi.org/10.48550/arXiv.2210.05404
74. Jun, S.A.: The ToBI transcription system: Conventions, strengths, and challenges. In: Barnes, J., Shattuck-Hufnagel, S. (eds.) Prosodic Theory and Practice, pp. 151–181. MIT Press, Cambridge (2022)

75. Kano, F., Tomonaga, M.: How chimpanzees look at pictures: a comparative eye-tracking study. Proceedings of the Royal Society B: Biological Sciences **276**(1664), 1949–1955 (2009)
76. Kelly, S., Healey, M., Özyürek, A., Holler, J.: The processing of speech, gesture, and action during language comprehension. Psychonomic Bulletin & Review **22**(2), 517–523 (2015). https://doi.org/10.3758/s13423-014-0681-7
77. Kempson, R., Cann, R., Gregoromichelaki, E., Chatzikyriakidis, S.: Language as mechanisms for interaction. Theoretical Linguistics **42**(3-4), 203–276 (2016). https://doi.org/10.1515/tl-2016-0011
78. Kendon, A.: Some relationships between body motion and speech. An analysis of an example. In: Siegman, A.W., Pope, B. (eds.) Studies in Dyadic Communication, chap. 9, pp. 177–210. Pergamon Press, Elmsford, NY (1972)
79. Kendon, A.: Gesticulation and speech: Two aspects of the process of utterance. In: Key, M.R. (ed.) The Relationship of Verbal and Nonverbal Communication, pp. 207–227. No. 25 in Contributions to the Sociology of Language, Mouton, The Hague (1980)
80. Kendon, A.: Gesture: Visible Action as Utterance. Cambridge University Press, Cambridge, MA (2004). https://doi.org/https://doi.org/10.1017/CBO9780511807572
81. Khalil, R.A., Jones, E., Babar, M.I., Jan, T., Zafar, M.H., Alhussain, T.: Speech emotion recognition using deep learning techniques: A review. IEEE Access **7**, 117327–117345 (2019). https://doi.org/10.1109/ACCESS.2019.2936124
82. Khasbage, Y., Alcaraz Carrión, D., Hinnell, J., Robertson, F., Singla, K., Uhrig, P., Turner, M.: The Red Hen Anonymizer and the Red Hen Protocol for de-identifying audiovisual recordings. Linguistics Vanguard (0) (2022). https://doi.org/https://doi.org/10.1515/lingvan-2022-0017
83. Kipp, M.: Anvil-a generic annotation tool for multimodal dialogue. In: Seventh European conference on speech communication and technology (2001). https://doi.org/10.21437/Eurospeech.2001-354
84. Kipp, M., Neff, M., Albrecht, I.: An annotation scheme for conversational gestures: How to economically capture timing and form. Journal on Language Resources and Evaluation - Special Issue on Multimodal Corpora **41**(3-4), 325–339 (2007). https://doi.org/10.1007/s10579-007-9053-5
85. Kowallik, A.E., Schweinberger, S.R.: Sensor-based technology for social information processing in autism: A review. Sensors **19**(21), 4787 (2019). https://doi.org/10.3390/s19214787
86. Krauss, R.M., Hadar, U.: The role of speech-related arm/hand gestures in word retrieval. In: Campbell, R., Messing, L.S. (eds.) Gesture, speech, and sign, pp. 93–116. Oxford University Press, Oxford (1999). https://doi.org/10.1093/acprof:oso/9780198524519.003.0006
87. Krippendorff, K.: Content Analysis: An Introduction to Its Methodology. SAGE Publications, Thousand Oaks, CA, 4 edn. (2018)
88. Kruppa, J.A., Reindl, V., Gerloff, C., Oberwelland Weiss, E., Prinz, J., Herpertz-Dahlmann, B., Konrad, K., Schulte-Rüther, M.: Brain and motor synchrony in children and adolescents with ASD—a fNIRS hyperscanning study. Social Cognitive and Affective Neuroscience **16**(1-2), 103–116 (07 2020). https://doi.org/10.1093/scan/nsaa092
89. Kubina, P., Abramov, O., Lücking, A.: Barrier-free communication. In: Mehler, A., Romary, L. (eds.) Handbook of Technical Communication, chap. 19, pp. 645–706. No. 8 in Handbooks of Applied Linguistics, De Gruyter Mouton, Berlin and Boston (2012)

90. Kuhnke, P., Beaupain, M.C., Arola, J., Kiefer, M., Hartwigsen, G.: Meta-analytic evidence for a novel hierarchical model of conceptual processing. Neuroscience & Biobehavioral Reviews **144**, 104994 (2023). https://doi.org/10.1016/j.neubiorev.2022.104994
91. Köpüklü, O., Gunduz, A., Kose, N., Rigoll, G.: Real-time hand gesture detection and classification using convolutional neural networks. In: Proceedings of the 14th IEEE International Conference on Automatic Face & Gesture Recognition. pp. 1–8. FG 2019 (2019). https://doi.org/10.1109/FG.2019.8756576
92. Ladd, D.: Intonational phonology. Cambridge: Cambridge University Press, 2 edn. (2012). https://doi.org/10.1017/CBO9780511808814
93. Ladefoged, P.: The revised international phonetic alphabet. Language **66**(3), 550–552 (1990). https://doi.org/10.2307/414611
94. Lascarides, A., Stone, M.: Discourse coherence and gesture interpretation. Gesture **9**(2), 147–180 (2009). https://doi.org/10.1075/gest.9.2.01las
95. Latash, M.L.: Synergy. Oxford University Press (2008). https://doi.org/10.1093/acprof:oso/9780195333169.001.0001
96. Lausberg, H., Sloetjes, H.: Coding gestural behavior with the neuroges-elan system. Behavior research methods **41**(3), 841–849 (2009). https://doi.org/10.3758/BRM.41.3.841
97. Levelt, W.J.M.: Monitoring and self-repair in speech. Cognition **14**(1), 41–104 (1983). https://doi.org/10.1016/0010-0277(83)90026-4
98. Li, S., Deng, W.: Deep facial expression recognition: A survey. IEEE transactions on affective computing **13**(3), 1195–1215 (2020). https://doi.org/10.1109/TAFFC.2020.2981446
99. Liebal, K., Oña, L.: Different approaches to meaning in primate gestural and vocal communication. Frontiers in Psychology **9**, 478 (2018)
100. Liebherr, M., Corcoran, A.W., Alday, P.M., Coussens, S., Bellan, V., Howlett, C.A., Immink, M.A., Kohler, M., Schlesewsky, M., Bornkessel-Schlesewsky, I.: Eeg and behavioral correlates of attentional processing while walking and navigating naturalistic environments. Scientific reports **11**(1), 1–13 (2021). https://doi.org/10.1038/s41598-021-01772-8
101. Liszkowski, U., Brown, P., Callaghan, T., Takada, A., De Vos, C.: A prelinguistic gestural universal of human communication. Cognitive Science **36**(4), 698–713 (2012). https://doi.org/10.1111/j.1551-6709.2011.01228.x
102. Loehr, D.P.: Temporal, structural, and pragmatic synchrony between intonation and gesture. Laboratory Phonology: Journal of the Association for Laboratory Phonology **3**(1), 71–89 (2012). https://doi.org/10.1515/lp-2012-0006
103. Lopez, A., Liesenfeld, A., Dingemanse, M.: Evaluation of automatic speech recognition for conversational speech in Dutch, English, and German: What goes missing? In: Proceedings of the 18th Conference on Natural Language Processing. pp. 135–143. KONVENS 2022 (2022)
104. Lozano-Goupil, J., Raffard, S., Capdevielle, D., Aigoin, E., Marin, L.: Gesture-speech synchrony in schizophrenia: A pilot study using a kinematic-acoustic analysis. Neuropsychologia **174**, 108347 (2022). https://doi.org/10.1016/j.neuropsychologia.2022.108347
105. Lücking, A.: Gesture. In: Müller, S., Abeillé, A., Borsley, R.D., Koenig, J.P. (eds.) Head Driven Phrase Structure Grammar: The handbook, chap. 27, pp. 1201–1250. No. 9 in Empirically Oriented Theoretical Morphology and Syntax, Language Science Press, Berlin (2021). https://doi.org/10.5281/zenodo.5543318

106. Lücking, A., Bergman, K., Hahn, F., Kopp, S., Rieser, H.: Data-based analysis of speech and gesture: The Bielefeld speech and gesture alignment corpus (SaGA) and its applications. Journal on Multimodal User Interfaces **7**(1), 5–18 (2013)
107. Lücking, A., Mehler, A., Menke, P.: Taking fingerprints of speech-and-gesture ensembles: Approaching empirical evidence of intrapersonal alignmnent in multimodal communication. In: Proceedings of the 12th Workshop on the Semantics and Pragmatics of Dialogue. pp. 157–164. LonDial'08, King's College London (2008)
108. Lücking, A., Ptock, S., Bergmann, K.: Assessing agreement on segmentations by means of *Staccato*, the *Segmentation Agreement Calculator according to Thomann*. In: Efthimiou, E., Kouroupetroglou, G., Fotina, S.E. (eds.) Gesture and Sign Language in Human-Computer Interaction and Embodied Communication, pp. 129–138. No. 7206 in LNAI, Springer, Berlin and Heidelberg (2012). https://doi.org/10.1007/978-3-642-34182-3_12
109. MacWhinney, B.: The CHILDES Project: Tools for analyzing talk. Lawrence Erlbaum Associates, Mahwah, NJ, 3 edn. (2000)
110. Magnee, M., Stekelenburg, J.J., Kemner, C., de Gelder, B.: Similar facial electromyographic responses to faces, voices, and body expressions. NeuroReport **18**(4), 369–372 (2007). https://doi.org/10.1097/WNR.0b013e32801776e6
111. Marschik, P.B., Kulvicius, T., Flügge, S., Widmann, C., Nielsen-Saines, K., Schulte-Rüther, M., Hüning, B., Bölte, S., Poustka, L., Sigafoos, J., Wörgötter, F., Einspieler, C., Zhang, D.: Open video data sharing in developmental and behavioural science (2022). https://doi.org/10.48550/ARXIV.2207.11020
112. Mathis, A., Mamidanna, P., Cury, K.M., Abe, T., Murthy, V.N., Mathis, M.W., Bethge, M.: DeepLabCut: Markerless pose estimation of user-defined body parts with deep learning. Nature neuroscience **21**(9), 1281–1289 (2018). https://doi.org/10.1038/s41593-018-0209-y
113. McNeill, D.: Hand and Mind – What Gestures Reveal about Thought. Chicago University Press, Chicago (1992). https://doi.org/10.2307/1576015
114. McNeill, D.: Gesture: A psycholinguistic approach. In: Brown, K. (ed.) The encyclopedia of language and linguistics, pp. 58–66. Elsevier (2006)
115. Mehler, A., Lücking, A.: Pathways of alignment between gesture and speech: Assessing information transmission in multimodal ensembles. In: Giorgolo, G., Alahverdzhieva, K. (eds.) Proceedings of the International Workshop on Formal and Computational Approaches to Multimodal Communication under the auspices of ESSLLI 2012, Opole, Poland, 6-10 August (2012)
116. Mlakar, I., Verdonik, D., Majhenič, S., Rojc, M.: Understanding conversational interaction in multiparty conversations: the EVA Corpus. Language Resources and Evaluation (2022). https://doi.org/10.1007/s10579-022-09627-y
117. Monarch, R.M.: Human-in-the-Loop Machine Learning: Active learning and annotation for human-centered AI. Simon and Schuster (2021)
118. Mondada, L.: The local constitution of multimodal resources for social interaction. Journal of Pragmatics **65**, 137–156 (2014). https://doi.org/10.1016/j.pragma.2014.04.004
119. Mondada, L.: Challenges of multimodality: Language and the body in social interaction. Journal of Sociolinguistics **20**(3), 336–366 (2016). https://doi.org/10.1111/josl.1_12177
120. Montague, P.: Hyperscanning: Simultaneous fMRI during Linked Social Interactions. NeuroImage **16**(4), 1159–1164 (2002). https://doi.org/10.1006/nimg.2002.1150

121. Morgenstern, A., Caët, S.: Signes en famille [corpus] (2021)
122. Munea, T.L., Jembre, Y.Z., Weldegebriel, H.T., Chen, L., Huang, C., Yang, C.: The progress of human pose estimation: A survey and taxonomy of models applied in 2d human pose estimation. IEEE Access **8**, 133330–133348 (2020). https://doi.org/10.1109/ACCESS.2020.3010248
123. Narayanan, S., Toutios, A., Ramanarayanan, V., Lammert, A., Kim, J., Lee, S., et al.: Real-time magnetic resonance imaging and electromagnetic articulography database for speech production research (TC). The Journal of the Acoustical Society of America **136**, 1307 (2014). https://doi.org/10.1121/1.4890284
124. Nenna, F., Do, C.T., Protzak, J., Gramann, K.: Alteration of brain dynamics during dual-task overground walking. European Journal of Neuroscience **54**(12), 8158–8174 (2021). https://doi.org/10.1111/ejn.14956
125. Ng, E., Ginosar, S., Darrell, T., Joo, H.: Body2hands: Learning to infer 3d hands from conversational gesture body dynamics. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11865–11874 (2021)
126. Nguyen, T., Schleihauf, H., Kayhan, E., Matthes, D., Vrtička, P., Hoehl, S.: The effects of interaction quality on neural synchrony during mother-child problem solving. Cortex **124**, 235–249 (2020). https://doi.org/10.1016/j.cortex.2019.11.020
127. Noah, J.A., Ono, Y., Nomoto, Y., Shimada, S., Tachibana, A., Zhang, X., Bronner, S., Hirsch, J.: fMRI Validation of fNIRS Measurements During a Naturalistic Task. Journal of Visualized Experiments (100), 52116 (Jun 2015). https://doi.org/10.3791/52116
128. Nyatsanga, S., Kucherenko, T., Ahuja, C., Henter, G.E., Neff, M.: A comprehensive review of data-driven co-speech gesture generation. arXiv preprint 2301.05339 (2023). https://doi.org/10.48550/arXiv.2301.05339
129. Núñez, R., Allen, M., Gao, R., Miller Rigoli, C., Relaford-Doyle, J., Semenuks, A.: What happened to cognitive science? Nature Human Behaviour **3**(8), 782–791 (2019). https://doi.org/10.1038/s41562-019-0626-2
130. Offrede, T., Fuchs, S., Mooshammer, C.: Multi-speaker experimental designs: Methodological considerations. Language and Linguistics Compass **15**(12), e12443 (2021). https://doi.org/10.1111/lnc3.12443
131. Oudah, M., Al-Naji, A., Chahl, J.: Hand gesture recognition based on computer vision: a review of techniques. journal of Imaging **6**(8), 73 (2020). https://doi.org/10.3390/jimaging6080073
132. Oviatt, S.: Ten myths of multimodal interaction. Communications of the ACM **42**(11), 74–81 (1999). https://doi.org/10.1145/319382.319398
133. Owoyele, B., Trujillo, J., De Melo, G., Pouw, W.: Masked-Piper: Masking personal identities in visual recordings while preserving multimodal information. SoftwareX **20**, 101236 (2022). https://doi.org/10.1016/j.softx.2022.101236
134. PaddlePaddle: PaddleDetection, object detection and instance segmentation toolkit based on PaddlePaddle. https://github.com/PaddlePaddle/PaddleDetection (2019)
135. Paggio, P., Navarretta, C.: Integration and representation issues in the annotation of multimodal data. In: Navarretta, C., Paggio, P., Allwood, J., Alsén, E., Katagiri, Y. (eds.) Proceedings of the NODALIDA 2009 workshop: Multimodal Communication – from Human Behaviour to Computational Models. pp. 25–31. Northern European Association for Language Technology (2009)
136. Pan, X.N., Hamilton, A.F.D.: Why and how to use virtual reality to study human social interaction: The challenges of exploring a new research landscape. British Journal of Psychology **109**(3), 395–417 (2018). https://doi.org/10.1111/bjop.12290

137. Pan, Y., Cheng, X., Zhang, Z., Li, X., Hu, Y.: Cooperation in lovers: An fNIRS-based hyperscanning study: Cooperation in Lovers. Human Brain Mapping **38**(2), 831–841 (Feb 2017). https://doi.org/10.1002/hbm.23421
138. Paquot, M., Gries, S.T.: A practical handbook of corpus linguistics. Springer Nature (2021)
139. Parisi, G.I., Kemker, R., Part, J.L., Kanan, C., Wermter, S.: Continual lifelong learning with neural networks: A review. Neural Networks **113**, 54–71 (2019). https://doi.org/10.1016/j.neunet.2019.01.012
140. Parr, L., Waller, B., Burrows, A., Gothard, K., Vick, S.J.: Brief communication: MaqFACS: A muscle-based facial movement coding system for the rhesus macaque. American Journal of Physical Anthropology **143**(4), 625–630 (2010)
141. Peer, A., Ullich, P., Ponto, K.: Vive tracking alignment and correction made easy. In: 2018 IEEE conference on virtual reality and 3D user interfaces (VR). pp. 653–654. IEEE (2018). https://doi.org/10.1109/VR.2018.8446435
142. Peikert, A., Brandmaier, A.M.: A Reproducible Data Analysis Workflow With R Markdown, Git, Make, and Docker. Quantitative and Computational Methods in Behavioral Sciences pp. 1–27 (2021). https://doi.org/10.5964/qcmb.3763
143. Perniss, P.: Why we should study multimodal language. Frontiers in psychology **9**, 1109 (2018). https://doi.org/10.3389/fpsyg.2018.01109
144. Pezzulo, G., Donnarumma, F., Dindo, H., D'Ausilio, A., Konvalinka, I., Castelfranchi, C.: The body talks: Sensorimotor communication and its brain and kinematic signatures. Physics of life reviews **28**, 1–21 (2019). https://doi.org/10.1016/j.plrev.2018.06.014
145. Pickering, M.J., Garrod, S.: An integrated theory of language production and comprehension. Behavioral and Brain Sciences (4), 329–347 (2013). https://doi.org/10.1017/s0140525x12001495
146. Pierrehumbert, J.B.: The phonology and phonetics of English intonation. Ph.D. thesis, Massachusetts Institute of Technology (1980)
147. Pinti, P., Tachtsidis, I., Hamilton, A., Hirsch, J., Aichelburg, C., Gilbert, S., Burgess, P.W.: The present and future use of functional near-infrared spectroscopy (fNIRS) for cognitive neuroscience. Annals of the New York Academy of Sciences **1464**(1), 5–29 (Mar 2020). https://doi.org/10.1111/nyas.13948
148. Posner, R., Robering, K., Sebeok, T.A., Wiegand, H.E. (eds.): Semiotik : ein Handbuch zu den zeichentheoretischen Grundlagen von Natur und Kultur = Semiotics. No. 13 in Handbücher zur Sprach- und Kommunikationswissenschaft, de Gruyter, Berlin (1997)
149. Pouw, W., Dingemanse, M., Motamedi, Y., Özyürek, A.: A systematic investigation of gesture kinematics in evolving manual languages in the lab. Cognitive science **45**(7), e13014 (2021). https://doi.org/10.1111/cogs.13014
150. Pouw, W., Dixon, J.A.: Gesture networks: Introducing dynamic time warping and network analysis for the kinematic study of gesture ensembles. Discourse Processes **57**(4), 301–319 (2020). https://doi.org/10.1080/0163853X.2019.1678967
151. Pouw, W., Fuchs, S.: Origins of vocal-entangled gesture. Neuroscience & Biobehavioral Reviews p. 104836 (2022). https://doi.org/10.1016/j.neubiorev.2022.104836
152. Power, S.D., Falk, T.H., Chau, T.: Classification of prefrontal activity due to mental arithmetic and music imagery using hidden Markov models and frequency domain near-infrared spectroscopy. Journal of Neural Engineering **7**(2), 026002 (Apr 2010). https://doi.org/10.1088/1741-2560/7/2/026002
153. Prieto, P.: Intonational meaning. WIRES Cognitive Science **6**(4), 371–381 (2015). https://doi.org/10.1002/wcs.1352

154. Prillwitz, S., Hanke, T., König, S., Konrad, R., Langer, G., Schwarz, A.: DGS corpus project–development of a corpus based electronic dictionary German Sign Language/German. In: sign-lang@ LREC 2008. pp. 159–164. European Language Resources Association (ELRA) (2008)
155. Quer, J., Pfau, R., Herrmann, A.: The Routledge Handbook of Theoretical and Experimental Sign Language Research. Routledge (2021)
156. Rachow, M., Karnowski, T., O'Toole, A.J.: Identity masking effectiveness and gesture recognition: Effects of eye enhancement in seeing through the mask. arXiv preprint 2301.08408 (2023). https://doi.org/10.48550/arXiv.2301.08408
157. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning. pp. 8748–8763. PMLR (2021)
158. Radford, A., Kim, J.W., Xu, T., Brockman, G., McLeavey, C., Sutskever, I.: Robust speech recognition via large-scale weak supervision. arXiv preprint arXiv:2212.04356 (2022). https://doi.org/10.48550/arXiv.2212.04356
159. Ramscar, M., Port, R.F.: How spoken languages work in the absence of an inventory of discrete units. Language Sciences **53**, 58–74 (2016). https://doi.org/10.1016/j.langsci.2015.08.002
160. Ren, Y., Wang, Z., Wang, Y., Tan, S., Chen, Y., Yang, J.: Gopose: 3d human pose estimation using wifi **6**(2) (jul 2022). https://doi.org/10.1145/3534605
161. Richard, A., Zollhöfer, M., Wen, Y., de la Torre, F., Sheikh, Y.: Meshtalk: 3d face animation from speech using cross-modality disentanglement. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 1173–1182 (October 2021)
162. Rieser, H., Lawler, I.: Multi-modal meaning – an empirically-founded process algebra approach. Semantics and Pragmatics **13**(8), 1–55 (2020). https://doi.org/10.3765/sp.13.8
163. Ripperda, J., Drijvers, L., Holler, J.: Speeding up the detection of non-iconic and iconic gestures (spudnig): A toolkit for the automatic detection of hand movements and gestures in video data. Behavior Research Methods **52**(4), 1783–1794 (2020). https://doi.org/10.3758/s13428-020-01350-2
164. Rohrer, P.: A temporal and pragmatic analysis of gesture-speech association: A corpus-based approach using the novel MultiModal MultiDimensional (M3D) labeling system. Ph.D. thesis (2022)
165. Rohrer, P.L., Vilà-Giménez, I., Florit-Pons, J., Gurrado, G., Gibert, N.E., Ren, P., Shattuck-Hufnagel, S., Prieto, P.: The multimodal multidimensional (m3d) labeling system (Jan 2023). https://doi.org/10.17605/OSF.IO/ANKDX
166. Sassenhagen, J.: How to analyse electrophysiological responses to naturalistic language with time-resolved multiple regression. Language, Cognition and Neuroscience **34**(4), 474–490 (2019). https://doi.org/10.1080/23273798.2018.1502458
167. Schegloff, E.A.: On some gestures' relation to talk. In: Atkinson, J.M., Heritage, J. (eds.) Structures of Social Action. Studies in Conversational Analysis, chap. 12, pp. 266–296. Studies in Emotion and Social Interaction, Cambridge University Press, Cambridge, MA (1984)
168. Schmidt, T., Wörner, K.: EXMARaLDA – creating, analysing and sharing spoken language corpora for pragmatic research. Pragmatics **19**(4), 565–582 (2009)
169. Scholkmann, F., Kleiser, S., Metz, A.J., Zimmermann, R., Mata Pavia, J., Wolf, U., Wolf, M.: A review on continuous wave functional near-infrared spectroscopy and imaging instrumentation and methodology. NeuroImage **85**,

6–27 (Jan 2014). https://doi.org/10.1016/j.neuroimage.2013.05.004, https://linkinghub.elsevier.com/retrieve/pii/S1053811913004941
170. Schuller, B.W.: Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends. Communications of the ACM **61**(5), 90–99 (2018). https://doi.org/10.1145/3129340
171. Schulte-Ruether, M., Kulvicius, T., Stroth, S., Wolff, N., Roessner, V., Marschik, P.B., Kamp-Becker, I., Poustka, L.: Using machine learning to improve diagnostic assessment of ASD in the light of specific differential and co-occurring diagnoses. Journal of Child Psychology and Psychiatry **64**(1), 16–26 (2023). https://doi.org/10.1111/jcpp.13650
172. Schulte-Ruether, M., Otte, E., Adiguezel, K., Firk, C., Herpertz-Dahlmann, B., Koch, I., Konrad, K.: Intact mirror mechanisms for automatic facial emotions in children and adolescents with autism spectrum disorder. Autism Research **10**(2), 298–310 (2017). https://doi.org/10.1002/aur.1654
173. Selting, M., Auer, P., Barth-Weingarten, D., Bergmann, J., Bergmann, P., Birkner, K., Couper-Kuhlen, E., Deppermann, A., Gilles, P., Günthner, S., Hartung, M., Kern, F., Mertzlufft, C., Meyer, C., Morek, M., Oberzaucher, F., Peters, J., Quasthoff, U., Schütte, W., Stukenbrock, A., Uhmann, S.: Gesprächsanalytisches Transkriptionssystem 2 (GAT 2). Gesprächsforschung – Online-Zeitschrift zur verbalen Interaktion **10**, 353–402 (2009), www.gespraechsforschung-ozs.de
174. Shattuck-Hufnagel, S., Turk, A.E.: A prosody tutorial for investigators of auditory sentence processing. Journal of psycholinguistic research **25**, 193–247 (1996)
175. Shattuck-Hufnagel, S., Yasinnik, Y., Veilleux, N., Renwick, M.: A method for studying the time-alignment of gestures and prosody in American English: 'Hits' and pitch accents in academic-lecture-style speech. In: Esposito, A., Bratanic, M., Keller, E., Marinaro, M. (eds.) Fundamentals of verbal and nonverbal communication and the biometric issue, pp. 34–44. IOS Press, Amsterdam (2007)
176. Shoemark, P., Liza, F.F., Nguyen, D., Hale, S., McGillivray, B.: Room to Glo: A systematic comparison of semantic change detection approaches with word embeddings. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. pp. 66–76. EMNLP-IJCNLP, Association for Computational Linguistics, Hong Kong, China (2019). https://doi.org/10.18653/v1/D19-1007
177. Sitaram, R., Zhang, H., Guan, C., Thulasidas, M., Hoshi, Y., Ishikawa, A., Shimizu, K., Birbaumer, N.: Temporal classification of multichannel near-infrared spectroscopy signals of motor imagery for developing a brain–computer interface. NeuroImage **34**(4), 1416–1427 (Feb 2007). https://doi.org/10.1016/j.neuroimage.2006.11.005
178. Streeck, J.: Gesture as communication I: Its coordination with gaze and speech. Communication Monographs **60**(4), 275–299 (1993)
179. Struckmeier, V.: Attribute im Deutschen: Zu ihren Eigenschaften und ihrer Position im grammatischen System. No. 65 in studia grammatica, Akademie Verlag, Berlin (2007)
180. Thomann, B.: Oberservation and judgment in psychology: Assessing agreement among markings of behavioral events. Behavior Research Methods, Instruments, & Computers **33**(3), 339–248 (2001)
181. Tiku, N.: The Google engineer who thinks the company's AI has come to life (2022)
182. Tkachman, O., Hall, K.C., Xavier, A., Gick, B.: Sign language phonetic annotation meets phonological corpustools: Towards a sign language toolset for phonetic

notation and phonological analysis. In: Proceedings of the Annual Meetings on Phonology. vol. 3 (2016)
183. Torricelli, F., Tomassini, A., Pezzulo, G., Pozzo, T., Fadiga, L., D'Ausilio, A.: Motor invariants in action execution and perception. Physics of Life Reviews (2022)
184. Trettenbrein, P., Pendzich, N.K., Cramer, J.M., Steinbach, M., Zaccarella, E.: Psycholinguistic norms for more than 300 lexical signs in German Sign Language (dgs). Behavior Research Methods **53**, 1817–1832 (2021). https://doi.org/10.3758/s13428-020-01524-y
185. Trettenbrein, P.C., Papitto, G., Friederici, A.D., Zaccarella, E.: Functional neuroanatomy of language without speech: An ale meta-analysis of sign language. Human Brain Mapping **42**(3), 699–712 (2021). https://doi.org/https://doi.org/10.1002/hbm.25254
186. Trettenbrein, P.C., Zaccarella, E.: Controlling video stimuli in sign language and gesture research: The openposer package for analyzing openpose motion-tracking data in r. Frontiers in Psychology **12** (2021). https://doi.org/10.3389/fpsyg.2021.628728
187. Trujillo, J.P., Holler, J.: Interactionally embedded gestalt principles of multimodal human communication. Perspectives on Psychological Science 17456916221141422 (2023)
188. Trujillo, J.P., Simanova, I., Bekkering, H., Özyürek, A.: Communicative intent modulates production and comprehension of actions and gestures: A Kinect study. Cognition **180**, 38–51 (2018)
189. Uddén, J., Hultén, A., Schoffelen, J.M., Lam, N., Harbusch, K., van den Bosch, A., Kempen, G., Petersson, K.M., Hagoort, P.: Supramodal Sentence Processing in the Human Brain: fMRI Evidence for the Influence of Syntactic Complexity in More Than 200 Participants. Neurobiology of Language **3**(4), 575–598 (2022). https://doi.org/10.1162/nol_a_00076
190. Uljarevic, M., Hamilton, A.: Recognition of emotions in autism: A formal meta-analysis. Journal of Autism and Developmental Disorders **43**(7), 1517–1526 (2013). https://doi.org/10.1007/s10803-012-1695-5
191. Valtakari, N.V., Hooge, I.T., Viktorsson, C., Nyström, P., Falck-Ytter, T., Hessels, R.S.: Eye tracking in human interaction: Possibilities and limitations. Behavior Research Methods **53**(4), 1592–1608 (2021)
192. Vick, S.J., Waller, B.M., Parr, L.A., Smith Pasqualini, M.C., Bard, K.A.: A cross-species comparison of facial morphology and movement in humans and chimpanzees using the facial action coding system (FACS). Journal of Nonverbal Behavior **31**(1), 1–20 (2007)
193. Vilhjálmsson, H., Cantelmo, N., Cassell, J., E. Chafai, N., Kipp, M., Kopp, S., Mancini, M., Marsella, S., Marshall, A.N., Pelachaud, C., Ruttkay, Z., Thórisson, K.R., van Welbergen, H., van der Werf, R.J.: The behavior markup language: Recent developments and challenges. In: Pelachaud, C., Martin, J.C., André, E., Chollet, G., Karpouzis, K., Pelé, D. (eds.) Intelligent Virtual Agents. pp. 99–111. Springer, Berlin and Heidelberg (2007). https://doi.org/10.1007/978-3-540-74997-4_10
194. Waller, B.M., Lembeck, M., Kuchenbuch, P., Burrows, A.M., Liebal, K.: GibbonFACS: A muscle-based facial movement coding system for hylobatids. International Journal of Primatology **33**(4), 809–821 (2012)
195. Weiss, K., Khoshgoftaar, T.M., Wang, D.: A survey of transfer learning. Journal of Big data **3**(1), 1–40 (2016). https://doi.org/10.1186/s40537-016-0043-6

196. Winkler, A., Won, J., Ye, Y.: Questsim: Human motion tracking from sparse sensors with simulated avatars. In: SIGGRAPH Asia 2022 Conference Papers. pp. 1–8 (2022). https://doi.org/10.1145/3550469.3555411
197. Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., Sloetjes, H.: Elan: A professional framework for multimodality research. In: 5th international conference on language resources and evaluation (LREC 2006). pp. 1556–1559 (2006), https://hdl.handle.net/11858/00-001M-0000-0013-1E7E-4
198. Wu, X., Xiao, L., Sun, Y., Zhang, J., Ma, T., He, L.: A survey of human-in-the-loop for machine learning. Future Generation Computer Systems (2022). https://doi.org/10.1016/j.future.2022.05.014
199. Youmshajekian, L.: Springer nature retracts chapter on sign language deaf scholars called "extremely offensive". Retraction Watch, https://retractionwatch.com/2023/01/23/springer-nature-retracts-chapter-on-sign-language-deaf-scholars-called-extremely-offensive/ (2023)
200. Young, A.W., Frühholz, S., Schweinberger, S.R.: Face and voice perception: Understanding commonalities and differences. Trends in Cognitive Sciences **24**(5), 398–410 (2020). https://doi.org/10.1016/j.tics.2020.02.001
201. Yu, C., Ballard, D.H.: A multimodal learning interface for grounding spoken language in sensory perceptions. ACM Trans. Appl. Percept. **1**(1), 57–80 (2004). https://doi.org/10.1145/1008722.1008727
202. Yu, F., Seff, A., Zhang, Y., Song, S., Funkhouser, T., Xiao, J.: Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. arXiv preprint arXiv:1506.03365 (2015). https://doi.org/10.48550/arXiv.1506.03365
203. Yunus, F., Clavel, C., Pelachaud, C.: Sequence-to-sequence predictive model: From prosody to communicative gestures. In: Digital Human Modeling and Applications in Health, Safety, Ergonomics and Risk Management. Human Body, Motion and Behavior: 12th International Conference, DHM 2021, Held as Part of the 23rd HCI International Conference, HCII 2021, Virtual Event, July 24–29, 2021, Proceedings, Part I. pp. 355–374. Springer (2021). https://doi.org/10.1007/978-3-030-77817-0_25
204. Zeng, Q., Zheng, G., Liu, Q.: Pe-dls: a novel method for performing real-time full-body motion reconstruction in vr based on vive trackers. Virtual Reality pp. 1–17 (2022). https://doi.org/10.1007/s10055-022-00635-5
205. Zhang, H.B., Zhang, Y.X., Zhong, B., Lei, Q., Yang, L., Du, J.X., Chen, D.S.: A comprehensive survey of vision-based human action recognition methods. Sensors **19**(5), 1005 (2019). https://doi.org/10.3390/s19051005
206. Zhou, H., Hu, H.: Human motion tracking for rehabilitation—a survey. Biomedical signal processing and control **3**(1), 1–18 (2008). https://doi.org/10.1016/j.bspc.2007.09.001