This is a preprint. Please cite as:
Henlein, A., Bauer, A., Bhattacharjee, R., Ćwiek, A., Gregori, A., Kügler, F., Lemanski, J., Lücking, A., Mehler, A., Prieto, P., Sánchez-Ramón, P., Schepens, J., Schulte-Rüther, M., Schweinberger, S. & Von Eiff, C. (2024). An Outlook for AI Innovation in Multimodal Communication Research. In: Duffy, V.G. (eds) Digital Human Modeling and Applications in Health, Safety, Ergonomics and Risk Management. HCII 2024. Lecture Notes in Computer Science, vol 14711. Springer, Cham. https://doi.org/10.1007/978-3-031-61066-0_13

An Outlook for AI Innovation in Multimodal Communication Research*

Alexander Henlein[1][0000−0002−2611−1417], Anastasia Bauer[2][0000−0002−4630−4590], Reetu Bhattacharjee[9][0000−0003−0720−4973], Aleksandra Ćwiek[3][0000−0002−1513−0188], Alina Gregori[1][0000−0003−0143−2000], Frank Kügler[1][0000−0001−8101−0005], Jens Lemanski[9,10][0000−0003−3661−4752], Andy Lücking[1][0000−0002−5070−2233], Alexander Mehler[1][0000−0003−2567−7539], Pilar Prieto[4,5], Paula G. Sánchez-Ramón[1,5][0000−0002−3394−1013], Job Schepens[2][0000−0003−1271−2526]**, Martin Schulte-Rüther[6,7][0000−0002−7198−9923], Stefan R. Schweinberger[8][0000−0001−5762−0188], and Celina I. von Eiff[8][0000−0002−9662−0751] ***

[1] Goethe University Frankfurt/M., Germany
[2] University of Cologne, Germany
[3] Leibniz-Centre General Linguistics, Berlin, Germany
[4] ICREA (Institució de Recerca i Estudis Avançats), Barcelona, Spain
[5] Universitat Pompeu Fabra, Barcelona, Spain
[6] University Medical Center Göttingen, Germany
[7] Department of Child and Adolescence Psychiatry, University Hospital Heidelberg, Germany
[8] Friedrich Schiller University Jena, Germany
[9] University of Münster, Germany
[10] FernUniversität in Hagen, Germany

Abstract. In the rapidly evolving landscape of multimodal communication research, this follow-up to Gregori et al. (2023) [71] explores the transformative role of machine learning (ML), particularly using multimodal large language models, in tracking, augmenting, annotating, and analyzing multimodal data. Building upon the foundations laid in our previous work, we explore the capabilities that have emerged over the past years. The integration of ML allows researchers to gain richer insights from multimodal data, enabling a deeper understanding of human (and non-human) communication across modalities. In particular, augmentation methods have become indispensable because they facilitate the synthesis of multimodal data and further increase the diversity and richness of training datasets. In addition, ML-based tools have accelerated annotation processes, reducing human effort while improving accuracy.

Continued advances in ML and the proliferation of more powerful models suggest even more sophisticated analyses of multimodal communication,

* Supported by the DFG SPP 2392 *Visual Communication* (ViCom)

** External collaborator

*** For the ViCom Consortium, alphabetical order except lead author; send correspondence to henlein@em.uni-frankfurt.de

e.g., through models like ChatGPT, which can now “understand” images. This makes it all the more important to assess what these models can achieve now or in the near future, and what will remain unattainable beyond that.
We also acknowledge the ethical and practical challenges associated with these advancements, emphasizing the importance of responsible AI and data privacy. We must be careful to ensure that benefits are shared equitably and that technology respects individual rights.
In this paper, we highlight advances in ML-based multimodal research and discuss what the near future holds. Our goal is to provide insights into this research stream for both the multimodal research community, especially in linguistics, and the broader ML community. In this way, we hope to foster collaboration in an area that is likely to shape the future of technologically mediated human communication.

Keywords: Multimodal communication · Artificial Intelligence · Large Language Models · Multimodal Computing · Generative AI

1 Introduction

Text-based systems such as ChatGPT [158], Bard [11], or Llama 2 [202] – so-called Large Language Models (LLM) – show remarkable results in a variety of applications. The zero-shot, one-shot, and few-shot capabilities of these models are particularly notable (c.f. [212,200,119]). This means that even if the model is presented with an unknown prompt or task for input, the results are usually still conclusive. They are even considered by some people to be the first step towards Artificial General Intelligence (“AGI”; [148]). All of these models are currently based on the transformer architecture [207], which is now also being used very successfully in other areas of computer science, such as computer vision [104] or audio processing [47]. The current development now consists of transferring these models to multimodality, which currently means that the models are trained on both text and image data [168]. However, there are already transformer-based architectures that support far more modalities [222].

As successful as these transformer-based models seem to be in generalizing over the training data, current research suggests that the true quality lies primarily in the training data [77,196,221,88]. It seems like any architecture, given enough reason, number of parameters, and training effort, can deliver the same results. This is currently most evident in the field of computer vision, where different model architectures, like convolutional models vs. transformers, can achieve competitive results given sufficient training effort [129]. The results of Yadlowsky et al. [220] also point to a similar interpretation. Their study suggests that there is limited evidence supporting the notion that the in-context learning behavior of the models extends beyond their pretraining data in terms of generalization. Concerning multimodality, this finding is of course particularly interesting and allows us to draw the following two conclusions: 1. The models cannot generalize over modalities based on which they have not been trained.

2. With our current machine learning-based methods, we cannot learn things for which we cannot generate training data or for which we cannot measure.

Based on this basic premise and the findings to date as well as the current interpretations of AGI,[11] we would like to provide an overview of the following points and the associated problems in this paper:

(i) Explore the existing state of multimodal systems, detailing their current functionalities and the scope of tasks they can effectively address (**Current Capabilities**).
(ii) Current difficulties in the effective application of these systems in various scientific fields that deal in some way with multimodality of communication (**Current Obstacles**)
(iii) Delve into aspects of multimodal integration that are anticipated to require more extensive research or technological advancements, outlining the challenges and complexities (**Long-Term Obstacles**).

We do not want to examine these points simply from the perspective of computer science, but, like the previous work by Gregori et al., from the broad field of multimodality communication research (see Section 1.2). This point becomes particularly clear in Section 3, which specifically examines multimodal communication analysis from the perspective of several individual disciplines. Mostly based on concrete examples, the corresponding discussions pinpoint various processing needs of multimodal AI tools. Based on this survey, Section 3.10 summarizes some challenges and outlooks for AI innovation in multimodal communication research. Section 4 attempts a cautious assessment of possible long-term obstacles of AI systems, including likely future developments and conditions for potential limitations.

1.1 Goals of this Article

Not least due to advances in Artificial Intelligence (AI), processes are simplified for the research community, if not made possible in the first place, which were previously associated with considerable time and costs [128]. 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 using AI in this kind of research. In this context, we also want to evaluate what is already possible today with the help of AI, but also why it is not yet being used in certain areas. But also what will probably be possible soon and which problems cannot yet be solved in the long term.

1.2 About Us

We are researchers with different backgrounds working on multimodal communication, specifically on gestures, sign languages, didactic and clinical aspects of

[11] See also the aims of potential ChatGPT successors such as Q* (https://www.technologyreview.com/2023/11/27/1083886/unpacking-the-hype-around-openais-rumored-new-q-model/, accessed 6th December 2023).

visual communication, animal communication, socio-emotional interaction, 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 [71].

2 AI in Multimodal Communication Research

The intersection of AI and multimodal communication has ushered in a new era of possibilities, revolutionizing how we understand and facilitate human interaction. This trend is being spurred on by the current development of so-called generative models (or generative AI) [34]. In the realm of text-based models like ChatGPT [158], the generation of relevant output hinges on formulating a textual prompt that describes a specified task. This prompt serves as an instruction for the model, prompting it to generate an appropriate response. This can create a dialog between the user and the model, e.g. to further refine the result or to make new queries based on the answer (prompt chaining).This process can be further optimized by so-called prompt engineering, i.e. the question of how a prompt must be formulated so that the system can provide the best possible response [183]. New models such as GPT4 [158] are not only dependent on text but can also process and generate images using DALL-E [23]. Probably the most powerful model at the moment is GEMINI [199]. This model processes input from the following four modalities as input: text, images, audio, and video, and is capable of generating text and images itself. What is characteristic of all these generative models, is that all models are trained using large amounts of training data. (For GEMINI, no exact data on the data set size is currently known, but it is assumed that the number of tokens is in the trillions.[12]) In the context of text data, the training process often involves predicting the next word in a sequence based on the preceding words. This sequential prediction task helps the model learn the patterns and structures within the data [219]. This leads us to several problems with these models and at the same time with current AI in general. It has been proven that the models do not understand the content of the training data, but only what a good result looks like [99]. As a result, the systems like to invent new facts or hallucinate in general [4]. With the learned structures, these models not only learn what good answers look like but also (tend to) reproduce the biases that are inherent to some degree to the training data. For example, racist or sexist content is not uncommon [169]. Even though the developers of these systems are making more and more efforts to prevent this (or to eliminate biases ex-post), these barriers can often be circumvented with clever jailbreaking [127,212,37].

[12] https://www.cnbc.com/2023/05/16/googles-palm-2-uses-nearly-five-times-more-text-data-than-predecessor.html, accessed 13th February 2024.

It has also become extremely difficult to evaluate and compare the models on standard public tasks [223]. This is because the procurement and sources of the training data make it practically impossible to rule out the possibility that this evaluation data is not already contained in the training data. This also relativizes statements such as: "Here's a list of difficult exams the ChatGPT and GPT-4 have passed[13]". Since it could be proven that this test data is already present in the training data [2].

Finally, it's important to address the training data itself, acknowledging that the process of acquiring and using it is not without criticism. This debate is currently most pronounced in the area of image generation, where millions of images from the internet are used to train these models. The problem is, for example, that these images are used without explicit consent from the creators and artists. As a result, these models enable the generation of images in a similar style, which is then labeled as plagiarism by the original creators [139], whereby such images are sometimes used for fraud [219]. This debate can be transferred one-to-one to the problematic use and reproduction of works in other modalities (music, literature, film, etc.), which has already led to many discussions.

The problems mentioned so far extend to the application of AI-based systems, especially in scientific contexts [211], where scientific findings should ideally be reproducible and comprehensible [162]. The return formats are not always adhered to, which makes it difficult to work with the results. Prompts that appear to work for one model may not work properly after it has been updated or replaced with a newer model. In addition, the best current models (GPT4, GEMINI) are closed-source and run on external servers. Depending on the type of data involved, it is not compatible with data protection law to store or process it on external servers. Alternative open-source models are slowly catching up, but are not yet truly mature in the area of multimodality (e.g. FALCON [5] or Alpaca [198]). And often require resources (both personnel and hardware) that are often not available. Thus, even the use of these models can reach a cost point that is no longer affordable for most research projects.

The last few paragraphs sounded quite negative about generative AI and the current LLMs such as GEMINI. We therefore present examples that are only possible thanks to these systems, as well as applications in which multimodality is an essential component and will probably soon be a core component of Generative AI. Some of these examples have nothing to do with language/communication but are intended to provide a brief overview of where multimodality is still relevant. Systems that prove successful for one application are often transferred to other areas (see e.g. Transformer [80], originally for translation and now also as the basis for Computer Vision (e.g. ViT [36]).

Multimodal Assistant: A multimodal assistant is an AI with which you can not only interact verbally (e.g. Alexa or Siri), but can, for example, also cover the visual context [122]. Application examples for this would be, for example,

[13] https://www.businessinsider.com/list-here-are-the-exams-chatgpt-has-passed-so-far-2023-1, accessed 13th February 2024.

assistance with cooking, initialized by a picture of the available ingredients [199], or the creation of personal training plans, supported by a picture of the available training equipment or other additional information [49]. These are currently the use cases that are most often presented in tech demos[14]

Multilingual Communication: As the systems are generally not only multimodal but also multilingual, they are also suitable for live translations between different languages. The additional advantage is that the models could translate not only the spoken language in face-to-face communication but also the corresponding gestures [12], facial expressions [96] and body movements, which can differ across diverse cultural contexts and could therefore lead to misunderstandings. One of the best-known examples is the usage of head movements in different cultures. In German or US culture, the vertical head movement communicates positivity and horizontal head movement denotes a negative response. In Bulgarian cultural tradition this pattern is reversed [9]. It can also be helpful when translating comics, for example, where images and text often share a common context [29].

Educational Tools: Different people learn best through different modalities (combinations) [147,182]. However, knowledge is often only available in one modality (usually text). Such generative models can be used to generate suitable audio descriptions and explanatory images based on the texts, which can help with understanding. Or they can directly help to convey this knowledge in age-appropriate language.

Healthcare: Multimodal data and systems are particularly valuable in the medical field. Medical images (e.g. X-rays), conversations (e.g. medical history), and various signals (e.g. long-term ECG), to name a few, are combined here. And in the future, intelligent devices such as watches will also be able to track everything. And this is already being done in parts [149,144,210]

Environmental Monitoring: By merging visual, textual, and auditory data, and leveraging satellite imagery, photographs, and sensor inputs, these AI systems offer a comprehensive understanding of ecosystems. They can analyze visual data to track changes in vegetation, assess pollution levels, and identify biodiversity. Additionally, by processing spoken or written reports from field researchers and integrating sensor data, these assistants facilitate real-time, context-aware assessments of environmental conditions [62].

Realtime News Feed: In the realm of real-time news delivery, by incorporating verbal, visual, and potentially other sensory modalities, could AI generate news on the fly. They can process and interpret not only textual news content but also images and videos, providing a more comprehensive and contextually rich understanding of unfolding events. The real-time analysis of multimodal data ensures that users receive up-to-the-minute updates, making these assistants indispensable tools for staying abreast of current affairs in a rapidly evolving media landscape.

[14] c.f. https://youtu.be/UIZAiXYceBI, accessed 16th January 2024.

Multimodal Programming Assistance: In the realm of coding, multimodal assistants redefine the landscape by offering not just verbal but also visual support, changing the development process[85]. These AI systems, trained by code snippets, images, and textual descriptions, assist programmers in understanding and writing code more effectively. Visual cues, such as flowcharts or diagrams, can supplement traditional text-based explanations, aiding in the comprehension of complex algorithms and code structures. Accessibility is also improved, e.g. when complete web pages can be created from simple drawings[15].

Mental Health Support: By incorporating multimodal emotion detection, these assistants can analyze facial expressions, voice tonality, and written text to gauge the user's emotional state [144]. This nuanced understanding enables more empathetic and tailored responses, enhancing the overall therapeutic experience. Visual elements, such as calming images or guided relaxation videos, can be seamlessly integrated to provide a holistic and personalized approach to mental health assistance. The combination of verbal and visual modalities allows for a more accurate assessment of the user's well-being, fostering a supportive environment for individuals seeking help.

Urban Planning: Generative AI can help urban planning by analyzing textual documents, satellite imagery, and citizen feedback. It can help city planners visualize potential changes, understand the impact on communities, and create more sustainable and livable urban environments.

Content Creation: The influences in the area of content creation have been noticeable for some time now, and it has never been so easy to create your own content with AI support without any special prior knowledge. AI can generate art [138], realistic graphics and videos or audios, sometimes recognized as so-called deep fakes [222], create entirely new podcasts (https://podcast.ai/) or books (https://aumgolly.com/). There exist AI-assisted text-to-speech tools, where one can create a voice with a skim sample of own voice recordings (https://speechify.com/), and conversely, speech-to-text multi-language transcription tools (https://trint.com/ or https://speechtext.ai/).

Assist Scientific Work: In the realm of academia, AI assistance has become instrumental in refining the intricacies of scientific communication. Non-native speakers benefit from these tools in crafting more polished manuscripts, addressing stylistic nuances, and eliminating grammar errors. The systems extend beyond mere language refinement, aiding researchers in formulating novel research questions when provided with comprehensive background information. Or even help directly with the preparation or evaluation of studies [184]. In the peer review process, AI streamlines the time-consuming task of writing and refining reviews, optimizing style and tone for more efficient and effective communication within the scientific community. However, the scientific community

[15] https://youtu.be/outcGtbnMuQ?t=980, accessed 14th February 2024.

itself is still very much debating what constitutes good scientific practice in times of large language models, including questions of authorship and responsibility/accountability [94,181,180].

3 Current Obstacles

It is anticipated that many challenges currently faced by the systems described in Section 2 will soon become less significant. These challenges include prompt engineering [183], contextual constraints [22], consideration of mathematical subtleties [60], and hallucinations [4], among others. We therefore refrain from examining these challenges, as we consider them to be more technical than fundamental. A related review is given by Liu et al. [128].

In the realm of multimodal AI, current efforts focus on integrating various sensory modalities – such as visual and auditory (including written, spoken, and sign language), and even olfactory data – to create a comprehensive understanding of the world. So far, however, this has mostly been done by considering small groups of (2-3) modalities. For example, there are approaches to the foundation of semantics on modalities such as vision [15], audition [106], or olfaction [105]. Other approaches address the fusion of different modalities, such as text, video, audio [3], and gestures [118], or fMRI data [1,33]. Advanced multimodal attention mechanisms [152,126] alongside enhanced multimodal transformer models [222,218] have the potential to augment and empower AI systems, enabling dynamic prioritization of relevant modalities for improved outcomes.

Contemporary Large Language Models (LLMs), which are predominantly trained on textual data but may also be exposed to images, demonstrate impressive abilities in understanding visual content. For example, ChatGPT can create an HTML page from a simple hand-drawn sketch, develop training programs by analyzing images of a home gym, suggest recipes after examining the contents of a refrigerator, and interpret the humor in a cartoon illustration. However, many of these functions lack sufficient systematic research and evaluation, making it challenging to assess their scalability. Despite their limitations, these examples offer a preview into the future possibilities that the integration of more modalities could unlock.

In this section, we examine several examples of multimodal analysis that are likely to be realized in the near future. We discuss this in the context of a number of use cases from the Visual Communication research group. The examples range from gesture modeling, to the study of feedback behavior and prosodic features, to the difficulties of studying sign language. Common to all these examples is the tension between the extended capabilities of generative AI and thus data-driven approaches on the one hand, and theory-driven approaches on the other. The section concludes by systematizing this tension in order to explore a bridge between data, representation, and theory-based approaches.

3.1 Manual Gesture

Since an LLM such as ChatGPT [157] is not embodied (in the sense that it lacks vision for perceiving and extremities for producing body movements), one could expect that it does not have anything to say about manual gesture. However, there are plenty of online gesture resources,[16] and there is a growing number of gesture-related scientific publications[17], which can potentially be used as part of a language model's training data. Accordingly, a language model might yet be able to handle multimodal data. As an example, consider (1), taken from SaGA [133], dialogue V11, 2:32.

(1) dann ist das Haus halt so:/*then the house is like this*: 

Intuitively, the speaker outlines the layout of the building talked about. From the gestural demonstration, one can obtain the information that the shape of the building is rectangular, like the symbol "⊔".

What does ChatGPT say to this example? We have to provide a verbal description of the gesture, however:

In the following scenario, Jane meets someone. The person says "The house is like this" and simultaneously gestures as follows: The index finger of the right hand is stretched, the other fingers are curled; the hand with the stretched finger is moved away from body, then towards left, and finally towards the body. Provide a short answer, as to what kind of gesture this is and what the gesture can mean in the context of the utterance.

The gesture described is a "demonstrative" or "illustrative" gesture, used to visually represent or mimic the shape or form of something. In the context of the utterance "The house is like this," the gesture likely represents the shape or layout of the house being described, indicating its size, orientation, or structure.

Provide a short answer as to what shape the house has.

[16] See, for instance, https://paperswithcode.com/search?q_meta=&q_type=&q=gesture, last accessed 6th February 2024.

[17] Such as https://www.gesturestudies.com and the publications and conferences listed there.

Based on the described gesture, the house has a rectangular or elongated shape.

This is just one example that does not allow any general conclusions to be drawn;[18] more systematic study on LLMs and gestures are needed, as pointed out by Wicke (2024) [213]. Nonetheless, this example demonstrates that multimodality, as represented by speech and gesture, can already be captured by unimodal LLMs. We return to this apparent puzzle in Section 3.3 when dealing with another kind of gesture, namely pointing gestures.

3.2 Making sense of the (bodily) multimodal ensemble

Human communication fundamentally embodies multimodality [163]. Previous research presents substantial evidence that gesture and speech are connected [209,208], demonstrating that multimodal signals tend to be temporally aligned at critical points in time (e.g., [114,111,166] see [167] for a review). From an early age, we engage with this multimodal ensemble [56], with the interplay between gestural dynamics and acoustic peaks further motivated by physical impulses of the respiratory system [167]. Notably, advanced tools now enable the automatic detection of these significant temporal peaks across various signals, prompting the question: How can we derive meaning from these discrete moments in time?

A notable instance in speech that generates a peak – in terms of fundamental frequency and/or intensity – is the prominent syllable (cf. Section 3.5); and within a syllable, its nucleus (usually a vowel). Even in a long-standing software for speech analysis like Praat [26], there exists a script for the automatic detection of nuclei [41]. Furthermore, for languages with available resources, automatic speech segmentation at the phoneme level, such as offered by WebMAUS [107] enables precise segmentation, from which we can easily establish syllable boundaries.

Having a syllable as a unit of information allows for the identification of meaningful peaks in the streams of acoustic or kinematic information. However, it is crucial to make informed decisions about which of these peaks should be paired together to potentially create a meaningful aligned signal. Kadavá et al. [102] utilize an 80 ms window to pair acoustic and kinematic peaks, motivated by mechanical coupling and, more importantly, by

the timing of anticipatory or reactionary muscle movements that occur before or after a deceleration peak [13] (after [102, 4191]). The layers of acoustic and kinematic information, along with the procedure used by Kadavá et al. [102], are illustrated in Figure 1. This pipeline provides tangible proof of our advancement in collaboration between linguistics and computer science, enabling us to identify units of information (such as syllables) in a computer-assisted manner and, within those units, automatically find pivotal points that may be carriers of information.

[18] Furthermore, the example was analyzed in [131] and was therefore possibly part of training data.

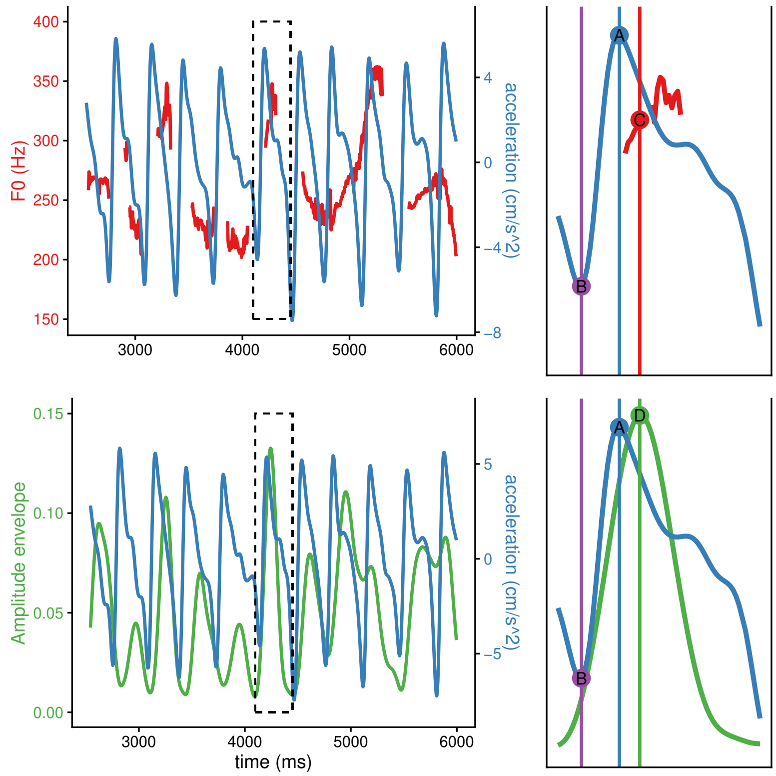


Fig. 1. Representation of amplitude envelope, fundamental frequency (F0), and acceleration over time. The diagrams to the right illustrate the methodology for isolating synchronously occurring peaks within segments demarcated by 80 ms. Identified peaks include: A for acceleration, B for deceleration, C for F0, and D for amplitude envelope. The figure is taken from Kadavá et al. [102, 4192].

When working with speech, whether it is formed into single morphemes, words, or full sentences, we are dealing with meaning. Even when examining phones, which may not be meaningful on their own, we are dealing with segmentable units based on spectral characteristics. As linguists and speech scientists, we can make informed decisions and use labels and annotations to provide data that, in turn, can be used to produce automated tools for segmentation and annotation. The challenge arises when we are faced with vocalizations rather than speech. How do we segment the units if we do not know whether they distinguish meaning? How do we segment units whose spectral characteristics seem to flow into one another (i.e., how do we find a boundary)? To pave the way toward automation, we first need to establish rules that help us answer such questions and, based on these rules, label a significant amount of data. In recent work, [61] proposed a method to label novel vocalizations and test the similarity of neighboring segments. In the future, using their paradigm, we could establish a direct line of investigation and comparison between human and non-human vocalizations.

By combining efforts between linguists, speech scientists, and computer scientists, we can further establish a link that would allow us to connect prelinguistic vocal information with kinematic information. If it is crucial for linguistic signals (e.g., co-speech gestures); why should we not suppose it was crucial before the dawn of language? We know that primates make great use of gesture, although there is contradictory evidence about whether apes can expand their gestural repertoire (e.g., [48,8]). Nevertheless, previous evidence highlights the importance of multimodality in primates [46], even though it is still rarely studied [123]. Developing tools to automate the processing of prelinguistic multimodal signals would enable us to study both human cognition (e.g., through the emergence of communicative systems) and non-human communication. As scientists studying communication systems, we can strive to provide units of meaning, which, in turn, can be utilized by computer scientists to build robust models.

3.3 Pointing (Deixis)

"[M]ost linguistic expressions are based on the perception of objects or situations in the real world" ([206, p. 191], quoted after [204, p. 378]). Such situated language use is characterized by the use of deictic acts, most prominently pointing gestures. Since pointing is bound up with reference, it has – in contrast to other kinds of manual gestures – received early attention from semantics and the philosophy of language (e.g., [59]). In fact, deixis is the hinge between the symbolic realm of language and the indexical realm of the perceptible environment [120]. Accordingly, understanding pointing is an important part of analyzing, interpreting, and taking part in referential communication – regardless of being concerned with human–human (HHI), human–computer (HCI), or computer–computer interaction (CCI).

Successful pointing (deictic behaviour in general) can be construed in terms of *triangulation* [40]: if successful, pointing brings about a three-place relation between a speaker ("pointer"), an addressee, and a common focal object or situation (the thing pointed at). Therefore, AI technology used to study interactions involving deixis must be able to identify the pointer, their pointing device, the addressee, and the object being pointed at. As usual in the domain of language processing, two perspectives have to be kept apart: analysis and synthesis. The former is concerned with the *understanding* of multimodal behavior, the latter deals with its *generation*. Within HCI, powerful algorithms for generating multimodal referring expressions in context have been developed [112,193]. The underlying rationale is that a referential expression is designed in such a way that verbal information from speech and locational information from a pointing gesture uniquely singles out the intended referent within a set of distractor items. Understanding multimodal deixis was arguably the first application of HCI, namely, the "put-that-there" system of Bolt [28], which processed verbal input and pen strokes on a display. This early system thereby circumvented what turned out to be a serious challenge, however, namely identifying the object pointed at (the so-called index), in particular if no descriptive aid from speech is provided [14,113,136,87]. The reason for this processing difficulty is that a pointing gesture does not directly single out an index; rather, it projects a pointing cone that directs the attention of the addressee (e.g., [113,43,132]). Accordingly, processing a pointing gesture and "guessing" its index is a fuzzy perceptual task, which is influenced by various factors such as the object density within the pointing domain or the interlocutors' perspectives [86]. There is no good reason to assume that AI technology fares better in this respect. But in any case, AI technology which is to be used for analyzing referential communication, or want to be able to take part in situated interactions, needs to be "embodied": deixis rests on sensory, visual scene processing (understanding) and mastering a pointing device (such as an index finger; generation). Hence, the successful use of AI in this kind of multimodal research requires an interplay of speech processing, visual analysis, and robotics.

However, we want to mention two phenomena that seem to be particularly involved and (to our knowledge) yet largely ignored by multimodal AI, namely deferred reference and addressee or discourse pointing.

Deferred reference happens when one points to one thing (the index) and refers to another one (the referent). For this to work, index and reference have to stand in a salient relation to each other. As an example, consider (2), where a painting is pointed at, but the painter is referred to ([38, p. 168]):

(2) This [*pointing at a painting*] is my favourite painter.

In (2), the index – the painting – is just given: it has been identified *qua* example. However, a more realistic, pointing cone-based scenario is shown in Figure 2. Here, it cannot be decided by the pointing alone whether the pointer points at the cup or the egg at s_2 – the egg in s_3 is out. However, the pointing gets disambiguated due to affiliated speech:

(3) a. Could you pass me the coffee?
    b. Could you pass me this egg?

This resolution works with deferred references, too [132]:

(4) a. I like this type of china.
    b. This is a hard-working hen.

(4a) selects the cup, (4b) selects the egg (at s_2) as index. Analyzing pointing and its metonymic uses therefore requires an interplay of perceptual scene processing and world knowledge, which, to our knowledge, has not yet been addressed in computational approaches.

It should be emphasized, however, that world knowledge is not a problem for AI. Resolving deferred references *via* descriptions is easily brought about by ChatGPT (GPT 3.5), as exemplified in the following exchange:

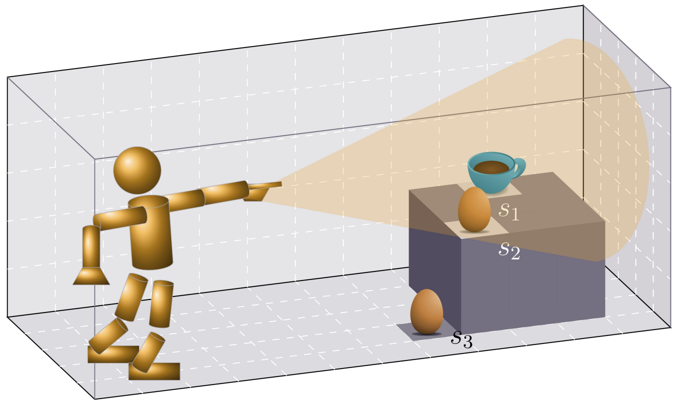


Fig. 2. The pointing cone "highlights" the cup and the egg at s_2 but excludes the egg at s_3 (taken from [132, p. 278]).

In the following scenario, George is pointing at a table with a cup, an egg, a fork, a knife, a book, a table, a spoon, a newspaper, and car keys on it. He says: "This is the best china". Provide a short answer as to which object George is referring to.

George is likely referring to the cup on the table as "china" can refer to fine porcelain or ceramic dinnerware, often used for drinking vessels like cups.

This leads to issues of symbol grounding [81]: How does the LLM "know" such things without having ever seen a cup or an egg and interacted with these objects? Very likely, this puzzle of current AI (namely, communicating without words grounded in direct or indirect experience) is due to "parasitic grounding" of symbolic cycles of large data within human understanding [83]. Given this admittedly sparse evidence, AI does seem to be able to process data that results from human participation in multimodal interactions to a certain level; they are just lacking a physical, robotic incarnation to perceive and to produce bodily modalities. Speaking in terms of the Turing test, this boils down to the difference between purely verbal T2 systems and robotic, sensorimotor T3 systems [82].

We started with a brief overview of what it means to refer and defer by pointing, and then asked whether current AI produces distinguishable behavior, *modulo* the trivial (?) fact of lacking a "body". This is the perspective of the Turing test, and not the perspective of AI used as a tool for preprocessing or annotating multimodal data. Accordingly, we have to distinguish at least three uses of AI in multimodal research:

– AI (narrow sense): **tool** to automatize some task. See other sections for examples of AI tools for preprocessing and annotating multimodal data.
– **Companion**: AI for social or information-seeking interaction. This is the focus of the Turing test.
– **Cognitive modeling**: AI as reverse engineering models of cognition

The examples above indicate that AI in the second sense, namely as companions, are already useful, albeit still being mainly[19] verbal (no T3 systems).

Besides such exophoric uses, pointing gestures are also involved in information management in discourse [67]. Consider, for example, (5), which is taken from the SaGA corpus [133], dialogue V5, starting at 13m58s.

(5) *Context: F is recapitulating the route that has just been described to him by R. Now he is trying to recall the landmark at a certain point of the route (turn 1). Due to his hesitation ("die (.) die"), R completes the utterance (turn 2) while discourse pointing at F. The completion was successful since it got accepted by F (turn 3).*

a. **F:** da steht die (.) die / there is the the

b. **R:** die SKULptur ((pointing at F)) / the sculpture

c. **F:** die skulptur drauf / the sculpture on top

R's pointing gesture in turn b. indicates that the discourse referent F is looking for in turn a. is already part of the common ground (cf. [90]). Accordingly, F accepts it in turn c. Thus, understanding the gesture in (5) requires a model of the ongoing discourse and the facts shared among the participants – on top of recognizing discourse pointing and detaching it from plain exophoric uses. This should be too difficult for an LLM, shouldn't it?

In the following scenario, Paul is recapitulating the route that has just been described to him by Lisa. Now he is trying to recall the landmark at a certain point of the route. He is starting an incomplete utterance: "Then there is the the ...". Due to his hesitation, Lisa completes the utterance while pointing at Paul: "... the sculpture". Provide a short answer as to what kind of gesture this is and what the gesture can mean in the context of the utterance.

[19] LLMs such as ChatGPT do produce pictorial output.

The gesture described is a "completion" or "prompting" gesture, where Lisa assists Paul in recalling information by providing the missing word or phrase through a physical cue. In the context of the utterance, it indicates that Lisa is helping Paul remember the landmark mentioned earlier in their conversation.

Pretty close, isn't it? One has to provide some context, of course. And the context in the above-given example is quite verbose and "semantically loaden". Accordingly, the answer might be shaped by collateral associations evoked by the detailed prelude. And indeed, a shortened context leads to a different, considerably worse exchange:

In the following scenario, Paul is talking to Lisa. Paul: "Then there is the the ...". Lisa, while pointing at Paul: "The sculpture". Provide a short answer as to what kind of gesture this is and what the gesture can mean in the context of the utterance.

In this scenario, Lisa's pointing gesture serves as a "clarification" or "confirmation" gesture, indicating agreement with or correction of Paul's statement by identifying the missing word ("sculpture") in his utterance.

Of course, Lisa's turn is a completion, and the gesture does not point at the missing word. This does not replace systematic gesture probings in LLMs (cf. [213]), but nonetheless indicates some variance or even boundaries depending on the elaborateness of prompt context.

To summarize: We have distinguished three uses of AI in multimodal research (tool, companion, cognitive modeling). From a companion point of view, multimodality emphasizes that current AI systems are mainly T2 systems, lacking T3 sensorimotor, robotic capabilities. However, symbolic descriptions of multimodal interactions enable AI to process even the data resulting from such interactions in a meaningful, interpretable way. This suggests the conclusion that the main challenges of perceptual fuzziness of pointing gestures (which makes it difficult to identify indices in the first place), deferred reference and discourse pointing for AI systems (the multimodal computing gap) rest in sensorimotor, robotic engineering. This seems to be the avenue that AI technology needs to go down to engage physically in deictic behavior.

3.4 Multimodal Ensembles and Gestalts

Speech processing and turn-taking is a tightly coupled system [121]. Adding temporally offset, multimodal signals should increase the massiveness of the binding

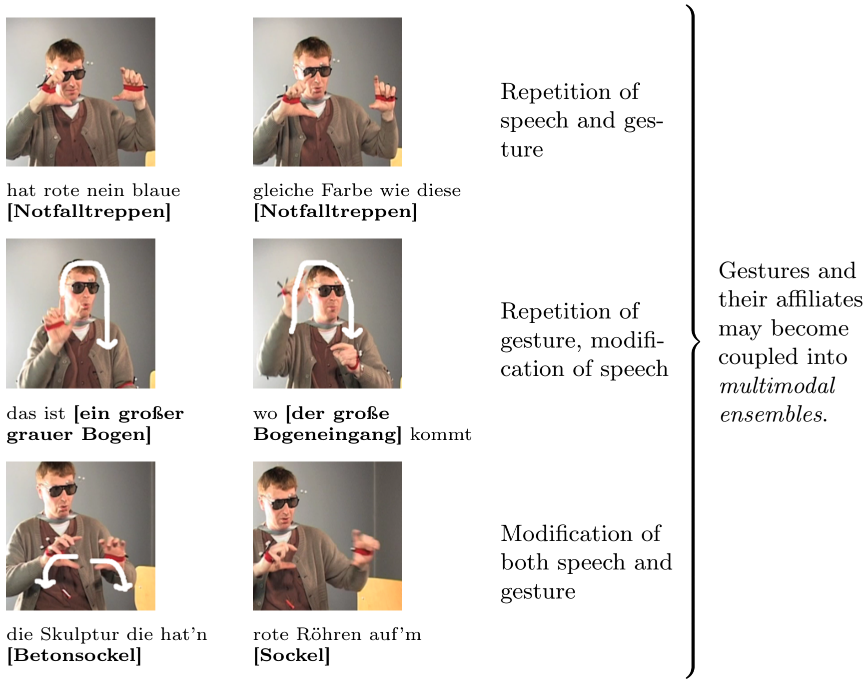


Fig. 3. Routinization of speech–gesture ensembles by repetition [135,143] (figure is taken from the presentation slides).

problem [97,57] and impact this system to the effect that multimodal interaction is much more difficult to process than unimodal communication [91]. But the opposite is the case. One reason for explaining this paradox is that multimodal discourse gives rise to channel-crossing, higher-order processing. Basically, (features of) signals from one channel can inform the interpretation of signals on other channels. This has been observed in terms of the *unity* of speech–gesture pairs, or composite utterances, which cohere into a channel-crossing *ensemble* [103] (cf. the notion of *idea unit* [141]). Empirical evidence was found in the routinization of *recurrent* speech–gesture ensembles [143,135] – see Figure 3 for some examples. On repetition, speech–gesture unity allows for a simplification of the form of the components of an ensemble. A generalisation to multimodal "local gestalts" (early work on multimodal ensembles focused on speech and co-verbal gesture), but without recurrence, has been argued for by [146], leading to a notion of recurrent multimodal gestalts [91]. In recent work, ensembles have been construed within interactionally embedded gestalt perception for rapid processing of multimodal signals [203]. An illustration of a multimodal gestalt is given in Figure 4, simplifying the model of [203] somewhat. The utterance-level interpretation of the speech signal is shaped by both, bottom-up multimodal signal features and top-down gestalt perception. Multimodal signal features and mul-

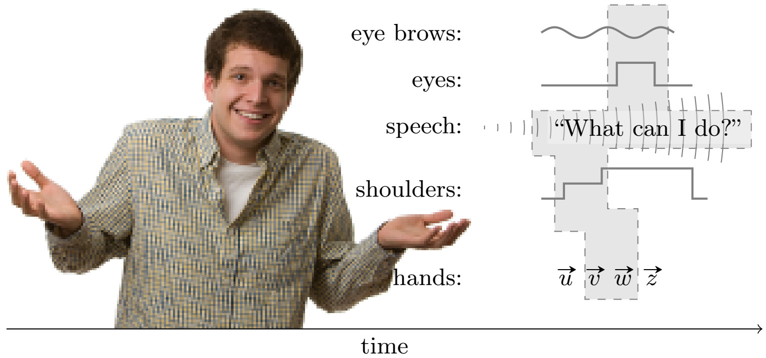


Fig. 4. Shrugging as whole body behavior (Wikimedia Foundation, Inc. Original uploader was Cbarr (WMF), CC BY-SA 3.0, File:RobGrindes-shrug-143px.png). The dashed, gray area indicates a multimodal gestalt.

timodal gestalts inform the interpretation of future multimodal utterances (as captured, for instance, in probabilistic frameworks such as predictive processing [89]; but see [125] for some critical discussion).

From the perspective of processing, the challenge of multimodality is that the *immediacy assumption* [75] holds for multimodal interaction, too (of course!). Processing happens not only bottom-up (the traditional perspective of linguistic grammars and parsing), but also top-down, and involves an immediate integration of processes as diverse as facial recognition [155], emotional tracking [68], perceptual classification [131], social cognition [150], and verbal utterance interpretation (e.g., [72,66]). It remains to be seen whether the *multimodal serialization hypothesis* is sustainable, namely that multimodal language processing is driven by "vertical relevance" which supervenes on sequential relevance [134].

3.5 Temporal, Kinematic and Pragmatic Integration of Gesture and Prosody

David McNeill, a pioneering scholar in the field of gesture studies, claimed that speech-accompanying gestures, also called "co-speech gestures" (a) are closely temporally coordinated with speech (the phonological synchrony rule); and that (b) work together with speech to convey the same pragmatic meaning (the pragmatic synchrony rule [141]). See Figure 5 for an example of temporal coordination of gesture and prosody, based on [55].

In this sense, the combination of verbal, prosodic, and gestural strategies to convey meaning [164] can be described as a "multimodal ensemble" (see Section 3.4) in communication. In the last decades, research has shown clear evidence of a close temporal coordination between prosodic structure (e.g., prominence and phrasing patterns) and the organization of gestural movements. However, less is known about the kinematic coordination between gesture and prosody

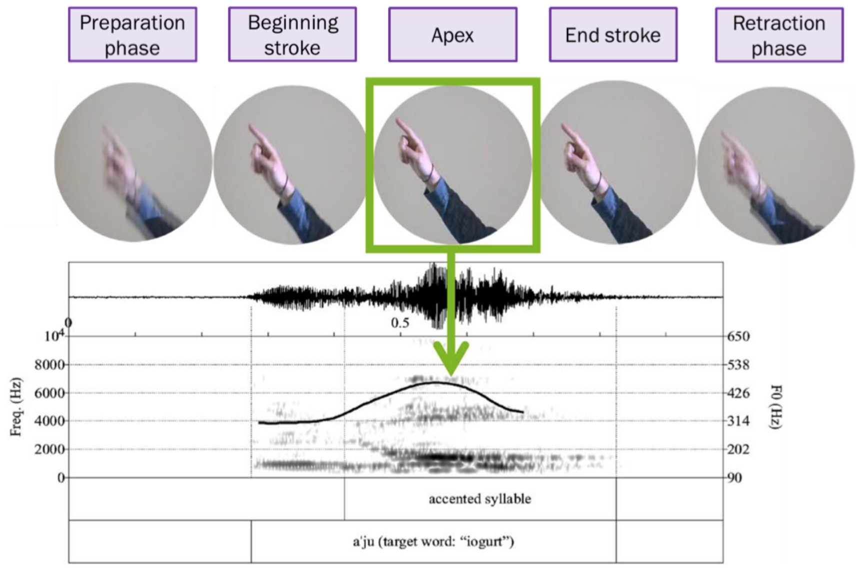


Fig. 5. Graph showing a pointing gesture in its temporal phases (top), time aligned with sound wave and F0 contour (middle) and text annotation (bottom). Adapted from [55].

(e.g., [6]) and the interaction between the two in the marking of pragmatic meaning [172,32]. Following these lines of research, as multimodal language researchers we are interested in assessing three complementary aspects of the relationship between prosody and gesture, specifically their temporal, kinematic, and pragmatic integration.

To do this, multimodal language researchers have typically relied on human annotation. Regarding prosody, the challenge in prosodic annotation lies in translating a continuous speech signal into discrete linguistic entities, such as the assumed categorical elements of intonational phonology [116]. In the realm of intonational phonology theory, models for prosodic categories have been proposed, of which the most widely adopted systems are based on the autosegmental-metrical approach to intonation (Tone and Break Indices, [20]). Human transcription of speech is a time-consuming task, and inter-annotator agreement is, at best, moderate (e.g., [74,197]). Consequently, several attempts have been made to automatically classify or transcribe prosodic categories (e.g., [214,159,188,187,174,175] for an overview). See Figure 6 for an example on pitch accent categories of GToBI training materials, cf. [73]. However, automatic approaches proposed so far have a common limitation – while they can reliably classify coarse prosodic categories, i.e. two classes of pitch accents, classifying more fine-grained pitch accent categories or differences in prominence levels fails to yield reliable results (e.g., [30,174,188]). Moreover, to our knowledge, no standard annotation proposal has

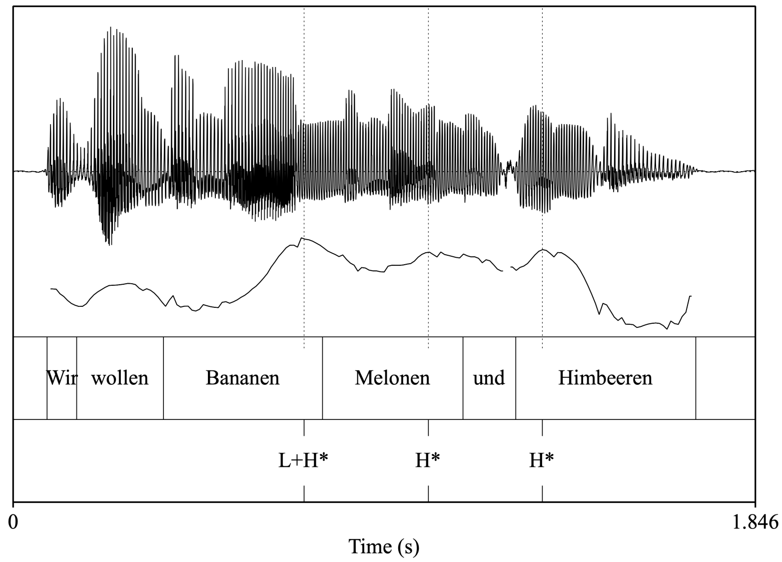


Fig. 6. Sound wave (top), F0-contour (middle), and text and prosodic annotation (bottom) for a German sentence (translating to *"We want bananas, melons and raspberries"*), with annotated GToBI pitch accent categories L+H* and H*.

been made up until now for the labeling of prosodic meaning. One consequence of those weaknesses in manual and automatic classification is the lack of appropriate large-scale and automatically annotated databases and resources for prosody research.

Regarding gestural data, annotations have mainly been done manually by human annotators, with the use of different software and annotation schemes (see [71] for an overview). However, a few tools such as motion-tracking suits or post-processing applications (e.g., OpenPose [35], MediaPipe [137], etc.) can track bodily movements automatically, allowing researchers to process the gestural kinematic signals into measuring aspects such as gesture trajectories, amplitude, or velocity. In combination with manual annotations, these tracking tools are able to identify gestures (see [95]). While these tools are already able to accelerate the manual annotation processes and can fairly easily be incorporated into AI, their approach to the semantic or pragmatic dimensions of gesture is limited. Similar to prosody, the interpretation of gestural behavior is key in order to assess decisions such as referential gesture types or pragmatic functions. These classifications are primarily applied manually (through annotation systems such as M3D [173], LASG [31] or CorpAGEst [27]). For instance in the M3D system, the meaning dimension allows the classification of gesture referentiality as well as pragmatic domains like speech acts, stance-taking, or discourse organization (e.g., information structural marking). Some of the manual annotation systems reach acceptable to good inter-annotator reliability (cf. [172]). Such categorical classifications could provide a challenge for their implementation into processing models for the following reasons: (a) they are sometimes based on subjective interpretations, which can diverge even between experts, and (b) most of these linguistic function-based decisions require a communication-related context to be made reasonably. Thus, while the processing of kinematic signals is on a good path to be a facilitation in gestural research, assessing the discourse-structural and meaning contribution of prosody and gesture will likely remain an annotation challenge for AI tools in the near future. One of the reasons for this is that machine learning requires a large amount of manually annotated data for accurate training, yet this data is not expected to be generated in the required breadth and depth. Given the role of AI as a companion tool (see Section 3.3), one might expect new ways of automatically generating appropriate annotations; however, prosodic information presents particular challenges in this regard.

Given the similarities that can be found between prosody and gesture, both in the temporal and the pragmatic domains, automatic classification of multimodal events may pose a similar obstacle to building large-scale corpora. An important issue for both temporal coordination and pragmatic coordination of prosody and gesture is the synchronization of the acoustic and visual channels to provide accurate assessments of the data. When possible, data synchronization should be considered in advance, by using an appropriate technological recording setup, but it can also be done in post-processing. While temporal synchronization might be covered by AI tools soon, semantic integration of speech and gesture will most likely be a bigger challenge. With prosody-gesture research being a relatively young field in linguistics, major theoretical developments regularly occur, which poses another challenge for up-to-date training of AI tools. Therefore, communication researchers need to be in accordance and very specific about their classification criteria in order to be able to train annotation algorithms. This can pose a challenge to linguists due to perceptual subjectivity in linguistic interpretation.

With innovative AI technology, there is hope to increase classification accuracy in the domains of prosody-gesture research. A more precise mapping between theoretically motivated form-based categories (in prosody and gesture) and the continuous speech signal could be achieved in the future. However, especially the functional mapping of prosodic and gestural categories to the acoustic and visual signal might continue to be a challenge for AI in communication research, given that linguistic context and expertise are required to reach this mapping.

3.6 Multimodal feedback in interaction

We have a relatively good understanding of the linguistic structures used for anticipating and resolving trouble [177][19][63], but most of this research was based on observations, written or audio language, narrations and elicitations.

Only recent analyses have drawn attention to the importance of using naturalistic conversational data for this research [44]. The role and the usage of the multimodal cues as a part of trouble management resource the interlocutors employ in face-to-face interactions is largely under-researched and we have no understanding of which and how multimodal cues are combined with vocal and manual (non-)lexical signals in everyday conversation. Apart from that, very little is known about the strategies used for trouble management in sign languages. Understanding the role of the various cues in the trouble management resources in face-to-face interaction in signed and spoken languages has important implications for the conception of successful communication.

We define feedback as an interactional behavior that displays interlocutors' perception or understanding of the course of the conversation.

A central topic of multimodal modeling concerns the variability in non-verbal and non-manual conversational units and how they help shape human communication. One large source of variability exists in facial and head movement to signal feedback in face-to-face interaction. Movements seem to have specific meanings, but no clear picture exists about these mappings and how variable they are. The goal of research in this area is to understand how the various facial and head movements are linked to feedback functions to better understand conversational phenomena, such as turn-taking, feedback, and trouble management. This topic is multimodal, as it involves the use of visual (manual signs, gestures and non-manual cues, e.g. facial expressions, eyebrow, head, torso or shoulder movements) and vocal signals (such as lexical items and non-lexical vocalizations). To study feedback we use the available online corpora and collect naturalistic data in a variety of spoken and signed languages: German Sign Language (DGS)[109], Russian Sign Language (RSL) [17] and Ukranian Sign Language [18], spoken Russian [17], spoken German or spoken Polish [115]. For the transcription of the spoken language data, we try to use automatic speech recognition (ASR) but we encounter various challenges, recently documented by Liesenfeld and colleagues [124], such as the deletion of hesitations and non-lexical vocalizations like laughter (see also [71]). Languages other than English already pose a challenge for many ASR systems. While English seems to contain the lowest error rate, languages like Polish, German and Russian, perform significantly poorer when using most speech-to-text systems.

Applying ASR to natural interaction poses a much bigger challenge. Human interaction typically features a rapid back-and-forth between participants, with a normal distribution of turn transition times centered around 0–200 ms [121], with many turns occurring in slight overlap. Tested ASR systems record substantially fewer speaker transitions and no overlapping annotations, and they lose linguistically relevant chunks of language [124] (see above).

We use annotation software ELAN [216] to manually annotate further feedback signals (e.g., head movement, eyebrow-raising, mouth movement, torso orientation, shoulder shrugging, and other movements). Then machine learning is used to determine which multimodal cues are used in overlap and which are not. This involves investigating how facial movements are used in different languages, and how they are used in sign languages compared to spoken languages. Data are manually annotated for about 70 different categories of facial movements using ELAN. These include, for example: "head tilt to the left", "eyebrows raised", or "nose wrinkled".

The annotation team has incrementally developed a set of annotation guidelines. The annotations are double-checked by deaf and hearing researchers. At each time step, the annotator defines whether there is a feedback element, which features are moving, and a start and end boundary. In this way, the data is segmented from a continuous stream of video data into a discrete list of multimodal feedback cues. Timestamps are used to keep the data aligned with the video data. The timestamps also allow the duration of the feedback to be measured, as well as linking to potentially relevant pre- and post-feedback items. Annotation relies on many human judgments. It is important to note that the process is incremental, as it is not easy to define all the criteria and rules in advance.

The project under consideration is naturally divided into an annotation part and a statistical modeling part. For the annotation part, formal modeling is used to represent the criteria and logic for annotation and extraction. Whether or not this annotation process can be fully and formally modeled is still an open question. The project has not yet aimed to do so, and it is an open question whether this is possible. Parts of the annotation process are formally modeled using the annotation guidelines mentioned above. These are verbal descriptions that are formal because they are written down in a precise way. The annotations are extracted using a script that contains rules about which features to extract, which features to combine, and when certain features belong together. These rules are kept very simple and do not depend much on the context. For the data analysis part, several standard statistical and machine learning models are used to analyze the resulting data set. The main goal is to find out which features are clustered together to signal feedback in interaction and whether these features are similar between signed and spoken languages under investigation. The features ideally reflect a faithful representation of what is important for successful communication. The challenges of this problem are related to typical analysis decisions, such as sample size, data preprocessing, feature selection, and model comparison. Other decisions, e.g. in clustering analyses, are related to distance and linkage methods. Overall, these modeling techniques are very standard and well-understood: these are statistical models and therefore do not have an understanding of the nuances of the data that only humans can understand.

The question then is which of the modeling steps described so far can or cannot be automated. Both the annotation and statistical modeling steps can benefit to some extent from NLP / ML tools. The annotation part could benefit the most from automation. This process is currently very time-consuming and expensive with a tendency towards low inter-annotator agreement [151] since it relies heavily on human judgment. Computer vision tools such as MediaPipe can be used to automatically extract initial sets of relevant features. MediaPipe provides several tools for facial feature extraction, and it is possible to train new models for new features. For example, MediaPipe predicts 52 facial "blendshape" scores, which are movements such as "brow down left", "cheek squint left", "mouth roll down", and so on. These scores largely overlap with the features that are currently coded manually. OpenFace generally detects more details than MediaPipe. For example, OpenFace recognizes facial action units according to the Facial Action Coding System.[20] However, not all non-manual movements are likely to be currently detected in the video data by a Computer Vision tool. In a study by Paggio and colleagues [160], head movement recognition software trained on spoken language conversational video data annotated frame-wise with visual and acoustic features was found to predict head movement only with 0.75% accuracy. Especially subtle head nods signalling feedback are likely to be missed in the data. As the automatic recognition of non-manual elements (especially head movements) in video-recorded face-to-face dyadic conversations does not seem to be accurate and reliable enough at the current stage, a number of recent studies pursue a combined method: manual annotation in the video-recorded data and partly automatized extraction of particular measurements with a CV tool. A reliably working combination of such methods is, however, needed to carry out systematic, data-driven research in multimodal, signed and spoken language use. It seems possible to train ML models for this as well. Improved, specialized tools would mitigate at least some of the manual workload, enabling to process a larger amount of data.

Going from a set of features to a set of feedback labels is a more undefined problem so far. Typical feedback categories discussed in the literature are e.g., *continuer*, *newsmarker*, *assessment*, *open request*, *acknowledgment*. Many of these concepts are based on spoken language, which might look differently in sign language and other language settings such as in online communication. Therefore, the project is currently set up to postpone the use of any of these labels prior to the analysis of multimodal behaviour. Data analysis such as clustering is used to determine whether it makes sense to distinguish concepts such as that of a "continuer".

The approach just outlined uses statistics and ML methods such as k-nn clustering. The idea is to explore the mapping between these labels and the features in an unsupervised way. Multimodal LLMs could be useful to automate this process. Existing LLMs are not yet able to handle multimodal data. However, multimodal LLMs are under development (see above). The goals of such multimodal LLMs are not necessarily the same as what we or other cognitive scientists are interested in, but it is possible to reuse models and learn from them in both ways.

Ideally, one would like to find out what categories of feedback can be produced by movements of the face, head or torso. This requires a high level of understanding of a conversation. It is not clear that this is fully possible with automation tools such as LLMs. In this sense, it seems to be a limit for the application of current generative AI systems to learn how to process a conversational context to detect, qualify, and ultimately "understand" human behavior such as facial or head movements. We see this as an ongoing challenge for AI.

[20] https://en.wikipedia.org/wiki/Facial_Action_Coding_System

3.7 Dyadic Social Behavior – Basic Research and Clinical Applications

Investigating dyadic social behavior is a vital aspect of visual communication research, underscoring the intricate interconnectedness inherent in interpersonal interactions. This area of research also holds significant implications for addressing psychiatric and neurodevelopmental conditions often associated with challenges in interpersonal reactivity, such as Autism Spectrum Disorder (ASD)[165]. In this context, multimodality refers, on the one hand, to different techniques for recording social-communicative behavior such as facial and body movement analysis, physiological response monitoring, neuroimaging, and eye tracking, with video recordings, specialized lab equipment, or wearable devices [156,186]. On the other hand, the notion of multimodality can be extended to the integration of several "semantic" layers of social interaction, including emotions, attention, conversational themes, social settings, and cultural contexts. Such "higher-level modalities" manifest throughout a variety of channels, including physiology, voice, face, body posture, and gaze patterns. Typically, they can not be accurately interpreted from a single type of raw data but can be conceived as coherent ensembles of multimodal signals that together convey social and emotional meaning [84]. These may entail easily discernible events such as a "smile", a "gaze towards an object", or a "pointing gesture", but also specific, more complex sequences of such events.

In clinical contexts, another layer of complexity emerges: part of psychiatric diagnosis involves using multimodal aspects of behavior to characterize individuals in relation to"typical" social functioning norms, with "atypical" behavior across various contexts regarded as "symptoms" indicative of a clinical condition [217,7]. Autism Spectrum Disorder (ASD) serves as a prime example of this diagnostic approach: Direct observation of social behavior via standardized tools is crucial for diagnosis. Specialized instruments (e.g. Autism Diagnostic Observation Schedule, ADOS) [130] have been developed to encapsulate both the subjective clinical impression and quantifiable aspects of autism-related behavioral patterns [185]. This method underscores the essential role of multimodal observation in identifying and understanding the nuanced spectrum of ASD.

Data Collection and Processing Techniques: Recent technical developments allow to employ advanced multimodal data collection methods. These include using tools like MediaPipe and OpenPose for detailed body pose and facial configuration analysis, mobile eye-tracking glasses, microphones for verbal communication capture, wristbands for physiological monitoring, and portable neuroimaging devices. Manual annotation and evaluation by human experts plays a crucial role in categorizing communicative behaviors, especially for higher order multimodal social-communicative states (e.g. joint attention, emotional/ communicative states etc.) and clinical symptomatology [84,185]. Processing this data involves building on available tools and processing pipelines that use already available and established ML/AI methods [140]: analyzing facial expressions and categorizing basic emotion types and "action units" (according to Ekman and Friesen [52]) and intensity (based on video or single images), analyzing body poses for detection of body and head gestures, and creating 3D skeletal models, and processing eye tracking data for saccades, fixations, and matching to a respective world video. Transcription of speech can be greatly sped up with the help of automatic tools and allows for sophisticated annotation and analysis of conversational elements, emotional tone, and specific linguistic properties [108,176].

Solvable Challenges and Immediate Future Directions in Multimodal Data Modeling: With available tools, it is well possible today to model data channels separately with high sophistication, using modality-specific pipelines. For such single-modality analyses (e.g. gesture detection, facial expression categorization, gaze mapping, and speech processing) automated pipelines will likely continue to evolve and will be able to capture more and more subtle nuances in the near future. For example, refinements and improvements are possible by integrating temporal dynamics into gesture and facial expression models and increasing the range of classified behavioral elements (e.g. types of gestures, types of facial emotions; [171,39]). For several challenges, there are no ready-to-use tools available, yet, but these will likely be available in the near future. These include fusing multiple video recordings into a full detailed 3D representation of the whole scenery, and simultaneous tracking of body configurations of multiple persons with just a limited number of video cameras [93] and/or sparse arrays of body sensors, or mapping eye gaze data onto a scenery in 3D gaze coordinates [79]. Thus, in principle, currently available uni-modal ML/AI methods and analyses and their future refinements might be sufficient to solve many problems related to the classification of modality-specific interactional entities with high precision (e.g. gesture classification based on pose data, or expressed emotion based on facial video data). However, vast amount of (annotated) data will be needed and may exceed what is feasible to produce in academic research.

Automation and Future Aspirations in Multimodal Research: Complex *multimodal* integration is currently neglected, but is pivotal to fully grasp the semantic meaning of social situations: For example, in a conversation between two individuals, if one person nods slightly while listening, this gesture can typically be interpreted as an acknowledgment or a sign of agreement, encouraging the speaker to continue. However, this nod, when paired with a brief, yet pointed, glance towards a wristwatch before returning to the speaker, transforms the message entirely. It would then be more appropriate for the speaker to recognize this as a subtle sign of impatience and a reminder of time constraints, and to stop talking instead of continuing. The core issue here is the necessity of integrating "world-knowledge" and attributing mental states to individuals for accurate interpretation of the situation. A single-modality cue, like a brief nod, can convey vastly different messages depending on context. Noticing the additional glance at an object, such as a watch, only clarifies meaning when combined with an understanding of internal states (e.g., impatience), social conventions (e.g., politeness), and the object's significance (a watch indicating time). Thus, the combination of multi-modal ambiguous cues and the breadth of contextual aspects may exponentially increase the range of possible interpretations. As a consequence, the endeavor to automate the analysis of social behavior for discerning nuanced meanings or identifying clinical symptoms could be reframed as striving toward the creation of a highly capable perceptual agent with a nuanced internal world model. This agent would need to possess advanced, universal capabilities for navigating social and communicative scenarios, equipped with perceptual precision and intuition comparable to humans. The question of whether this is actually theoretically possible or even desirable is beyond the scope here, similar to the discussion around Artificial General Intelligence (AGI) as a general solution for automating complex cognitive tasks.

A more feasible and realistic approach might be to develop generative multimodal models that are able to simulate human behavior for specific interactive scenarios, similar to how GPT-based LLMs mimic human linguistic output. The burgeoning field of virtual and augmented reality (VR/AR), particularly within platforms like the "Metaverse", may present a future opportunity to capture and analyze nuanced recordings of dyadic interactions on a massive scale. Such datasets could be analyzed with similar methodologies as used in large language models (LLMs) that create powerful generative models from large text corpora. As in the case of advanced models like GPT-4, given enough data in combination with further human reinforcement-learning based refinement, surprisingly capable models could potentially emerge.

The key to this approach would be not to operate on the "raw data" signals but to distill a lexicon of "tokens" or behavioral "subwords" that represent the smallest units of meaningful interaction. For instance, consider the nuanced interplay of gestures and expressions in a social interaction: a *nod* accompanied by a *smile*, further contextualized by direct *eye contact*, could be identified as the basic behavioral tokens. These tokens, when observed in a sequence, such as a smile followed by a nod, and then sustained eye contact, might collectively signify an "acknowledgment" or convey "friendliness". The temporal configuration of these tokens, their order, and duration, may convey further subtleties of the interaction: For example, a quick nod with a fleeting smile might denote a polite, yet perfunctory acknowledgment typical among acquaintances. In contrast, a prolonged smile combined with a nod and extended eye contact could be interpreted as a warm, genuine greeting, indicative of a deeper rapport between friends. Employing a granular, token-based strategy as a mediator between raw data and nuanced social interpretation offers numerous benefits: it simplifies complexity by decomposing interactions into fundamental units, enhances processing efficiency and reduces data requirements, broadens model applicability across varied scenarios, supports gradual learning, and increases the model's interpretability. This perspective of a stratified bottom-up methodology, emphasizes the need for a nuanced understanding of the semantic and social building blocks of behavior. Identifying these "tokens" or "subwords" is not just about capturing gestures or expressions but understanding their significance within a rich tapestry of human interaction. This approach could unlock new perspectives in social behavior research and clinical diagnostics, offering a granular view of interpersonal dynamics without the constraints of formal pre-defined theoretical models. At the same time, this strategy allows for the thoughtful integration of established theoretical concepts by defining appropriate tokens. By adopting a layered modeling strategy, the analysis remains data-driven yet can be enhanced and guided by pertinent theories from psychology, linguistics, and medicine.

The current state of the field seems to be optimally suited for such an approach: With the increasing availability of high-precision modality-specific analysis pipelines, the intermediate "tokens" could be assessed automatically, and their multimodal integration could then further be analyzed on this more abstract, semantically meaningful level. Ensembles of these tokens, their time course, and their complexities could be further stacked in increasing layers of abstraction, i.e. from basic building blocks of interaction, to transient states of communication, up to categorizations such as typical and atypical behavior and diagnostic symptoms [185]. However, the ambition to automate the analysis of complex social behaviors for diagnostics in clinical settings should only be pursued with caution. The intricate variability of human behavior, shaped by diverse psychological and neurodevelopmental backgrounds, may not be fully captured. Furthermore, there are critical ethical concerns with respect to potential oversimplification, privacy issues and individual rights in medical care settings. Despite these challenges, the potential of automated analyses to augment clinical decision-making remains compelling. By providing a detailed and quantitatively rich portrait of behavior, AI-supported tools could offer clinicians a deeper understanding of patients' behavior and symptoms, enhancing diagnostic precision and personalized care strategies. However, the deployment of these technologies demands a careful balance, ensuring they serve as adjuncts to, rather than replacements for, the nuanced judgment of healthcare professionals. Respecting the ethical boundaries and the multifaceted nature of human behavior is crucial in realizing the benefits of AI in clinical applications, advancing patient care while safeguarding individual dignity and privacy.

3.8 Diagrams and LLMs

Diagrams represent a distinct research area within communication studies, yet they share overlaps and similarities with many other fields. While LLMs are now making remarkable progress in recognizing images and classifying objects within them, they often struggle to distinguish diagrams from other pictures. This challenge is not surprising, given that humans also encounter difficulties in this area. In most cultures, there is a similar intuitive understanding of what constitutes a diagram and how it differs from a picture. However, in research, there lacks a precise criterion for identifying diagrams as such and distinguishing them from language or other forms of representation [10,21].

The absence of a standardized criterion for diagrams is partly because diagrams can be utilized in various ways, and their design function seems almost limitless. For instance, while the traditional diagram is visual, there have been uses of 'audio diagrams' since the 19th century [53], and the 20th and 21st centuries saw the invention of various 'haptic diagrams' [69,154]. These can be objects that are both tangible and visible or visual diagrams that have been

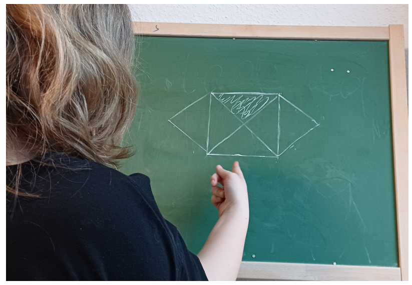


Fig. 7. Children pointing to a specific area of a diagram

translated into Braille [201]. From the Middle Ages to early modern times, diagrams were also expressed through gestures [170]. For example, many children learn from an early kindergarten age to communicate diagrams by assembling different objects, indicating the medium for conveying a diagram can be highly versatile [70] (see Figure 7).

Focusing on visual diagrams can be utilized either unimodally or multimodally, depending on the requirements. Visual programming language (VPL) in human-machine interaction, for instance, aims to be as unimodal, iconic, and rule-based as possible, yet intuitive [21]. In contrast, diagrams in human-human interaction, such as in mathematics education [205], are typically multimodal, iconic, spontaneous, and intuitive. Nonetheless, most diagrams are multimodal in that they are often accompanied by words, facial expressions, pointing gestures and gestures in general from the person drawing the diagram in a communicative situation. These situations form a kind of 'multimodal ensemble' (see Section 3.4) which frequently consists of words, facial expressions, gestures, and the diagram itself. When the communication situation is set aside, most diagrams are multimodal in that they incorporate heterogeneous elements, including geometric elements like lines, circles, parallelograms, etc., arranged spatially, alongside words or symbols. While the geometric elements are usually perceived as inherently diagrammatic, the words or symbols are often seen as non-diagrammatic. Hence, most diagrams that feature this multimodal ensemble are described as 'heterogeneous diagrams' [16].

Currently, diagrams are viewed from three perspectives [145]: (1) the 'suspicious view', which regards diagrams at best as heuristic tools; (2) a 'practical view', recognizing diagrams as capable of representing information or solving problems in specific contexts; and (3) a 'formal view', which considers diagrams as a formal language. The feasibility of the latter perspective was first demonstrated in the 1990s by distinguishing the syntax and semantics of diagrams, enabling metamathematical proofs concerning diagrams [191]. This significant advancement achieved by a philosopher was subsequently applied to mathematics and artificial intelligence [65,98]. Presently, psychology and cognitive research

[21] Seehttps://snap.berkeley.edu/project?username=ten_6044&projectname=Match%21.

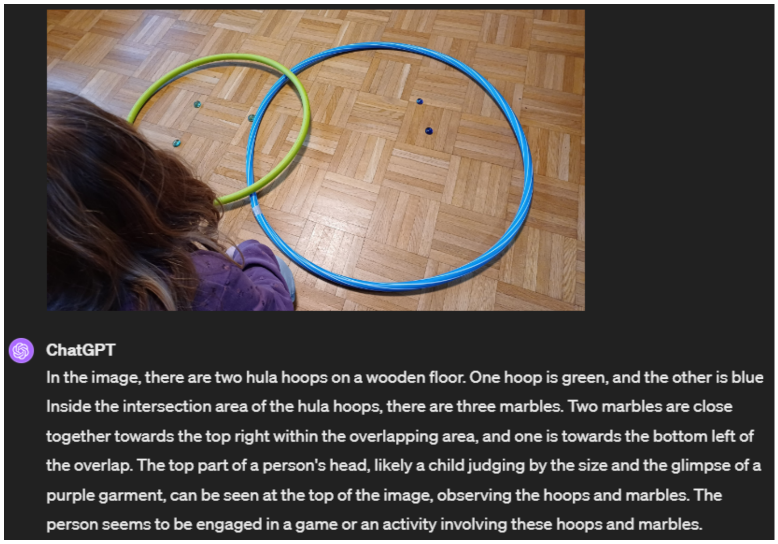


Fig. 8. GPT 4.0 (February 2024) not only fail to recognize the diagram created by the hoops but also mistakes the position of the marbles in the picture.

are particularly focused on exploring how the intuitive advantages of diagrams (e.g., free rides, observational advantages, etc.) [189,190,194] can be merged with formal diagram languages [24,25,78,92], while also addressing the long-recognized disadvantages of diagrams, such as ambiguity or uncertainty issues [64,100].

Diagrams have been utilized in artificial intelligence for years in a variety of ways. They are frequently used in the development of artificial intelligence [153], in the field of explainable AI, in the solution of certain problems and as a medium for certain applications. LLMs can be used either to obtain written information via diagrams, to analyze diagrams or to generate diagrams. The difficulty currently lies in recognizing a diagram as such and distinguishing it from an ordinary image. Language AIs are perfectly capable of doing this thanks to image recognition and different outputs. However, most LLMs already fail to recognize a diagram in a picture in which, for example, children use a diagram consisting of two partially overlapping hoops to classify marbles according to color (see Figure 8).

If we stick with this example, the next step after the identification of a diagram would be to correctly classify it. However, further problems arise here, which are due to the insufficient research on diagrams to date. The majority of diagram use takes place outside of academic research, namely in professions such as graphic designers, information designers, communication designers, etc. Numerous guides, books and, above all, websites have emerged from the practical

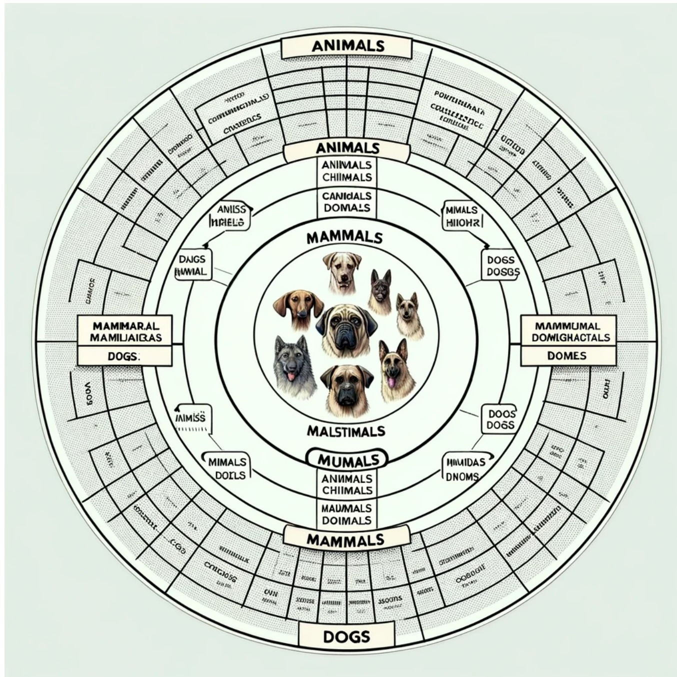
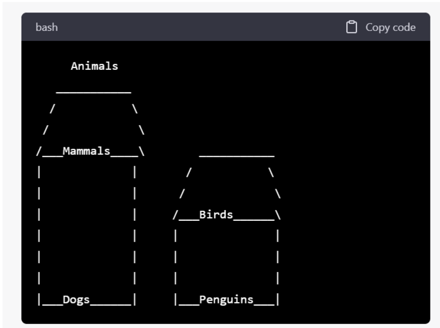


Fig. 9. Screenshot taken from GPT 4.0 in February 2024, which has now resorted to Dall-E for the same prompt as Figure 10(left screenshot).

environment of these industries, which present lists with examples of the different types of diagrams.[22]

However, these lists are not standardized and differ concerning the names of diagram types or the classification of diagrams under one type. However, there already exist approaches to creating metataxonomies [54]. As there was almost no diagram research at all between around 1880 and 1990 with the onset of the so-called ‘crisis in intuition’ [76,101], LLMs can neither draw on a broad data set nor on a uniform classification. Even if there is already a great deal of research on the above example of the two hoops or circles, the definition is still not clear. This is because the unimodal or multimodal context also plays a role in diagrams. These contexts must respond to which diagrams a corresponding diagram is associated with, or which words, gestures etc. a diagram is associated with in context. The two partially overlapping hoops or circles can be interpreted as a set diagram, an Euler diagram, a Venn diagram or something else [58]. It is the context that is crucial here, and the amount of information about these contexts that an LLM can access. Diagrams may be intuitive to humans at first glance, but the closer they are examined, the more difficulties they reveal that machines cannot solve.

[22] See e.g. https://flowingdata.com/chart-types/, https://datavizcatalogue.com/, https://datavizproject.com/



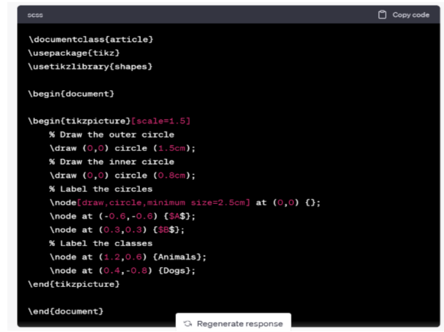


Fig. 10. The left screenshot is from GPT 3.5 from Spring 2023. The prompt asked for generating a Venn diagram including dogs and mammals as an example. The result was the Euler-type diagram shown. GPT 4.0 (without plugins) also delivers a similar result in February 2024. However, in 4.0 the containment relation is better represented by nested squares in ASCII.
The right screenshot is taken from GPT 3.5 in spring 2023, which issued the tikz code for the same prompt as mentioned in the left one.

If images with such geometric shapes are analyzed in LLMs without further information, the programs not only fail to differentiate between picture and diagram, but also fail to recognize the type of diagram correctly or hallucinate. These are some of the problems with diagram recognition that need to be resolved in the future.

However, LLMs do not yet perform well in the generation and creation of diagrams either. Image-generating AIs such as Midjourney or Dall-E can create countless diagram types, but these often do not correspond syntactically to the conventions or rules and are usually semantically meaningless (see Figure 9). Language-based AIs such as ChatGPT can now access these image generations, but then hallucinate results that do not correspond to the logical principles of the respective diagrams. Instead of a Venn diagram, for example, GPT generates an Euler diagram. LLMs work better, for example, when they generate diagrams in ASCII code or in tikZ (see the right screenshot in Figure 10), HTML (see the left screenshot in Figure 10) etc. Simple Euler or Venn diagrams are often syntactically correct, even if there can be confusion between the diagram types in the explanation given by the LLM.

Diagrammatic theorem provers or programmes in the field of visual computing work much more successfully than language-based AIs. However, if these are not themselves based on AI, but on sound and complete algorithms, they can already be successfully integrated into LLMs as plugins (see Figure 11). The results for certain tasks are correct if the right plugins are selected. A lot therefore depends on the prompt, the respective architecture and the choice of plugin. This in turn determines the explanation of the produced diagram also provided by the respective LLM.

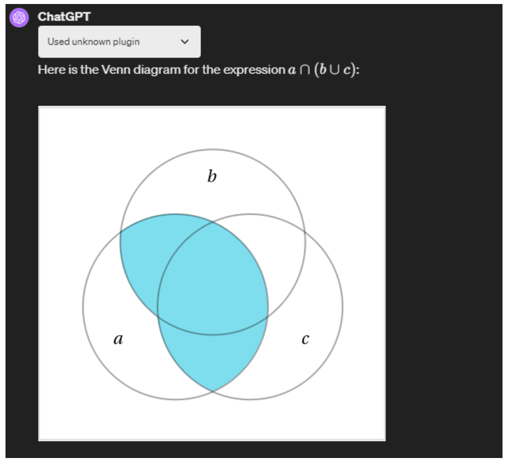


Fig. 11. Screenshot taken from GPT 4.0 in February 2024, which successfully completed a task using a Venn Diagram. The Wolfram Alpha plugin was utilized to assist.

It is therefore important for diagram recognition and diagram creation that manually set up databases in the relevant specialist areas to classify the diagram types correctly. A model for such research is Leonardi.DB, for example, in which the diagram type 'Aristotelian diagram' or 'Square of Opposition' is precisely specified and fed with many data sets [42].

3.9 Multimodal Perception of Emotional Expression in Voices and Faces by Cochlear Implant Users

In emotion perception research, the ongoing development of more sophisticated AI holds considerable promise for enhancing its ability to learn complex patterns in emotional expression. AI can analyze emotional expressions in voices and faces through techniques like machine-learning-supported acoustic analysis (e.g., [45]), natural language processing (NLP), and computer vision. As AI systems continue to advance, they are anticipated to improve in their ability to integrate information from multiple modalities (including voice tone, facial expressions, and body language). Moreover, there is a prospect for AI systems to even become better at understanding and interpreting emotions in context, considering situational factors that influence emotional expression. Advancements in processing speed and hardware capabilities may further empower AI systems to perform real-time emotion analysis - a development which, for instance, would be an important advancement in applications, such as human-computer interaction, virtual assistants, and emotion-aware technologies. Notably, a new empirical approach utilizes AI to investigate so-called neural decoding. This approach uses multivariate analyses of brain recordings to permit emotion classification (e.g., for vocal, facial, or video stimuli) on the basis of a perceiver's brain activity (e.g., [126]).

At the same time, ongoing research in AI may lead to improved cochlear implants (CIs) - sensory prostheses designed to provide auditory sensations to individuals with severe-to-profound hearing loss by directly stimulating the auditory nerve. This is achieved through (1) recording sound via a microphone, then (2) transforming the sound into electrical signals via a sound processor, (3) transmitting these magnetically through the scull to an intracranial receiver, and (4) feeding them into a limited number of stimulation electrodes (typically, 6-22) which are surgically implanted into the cochlea. In general, improved AI might result in more sophisticated signal processing algorithms that enhance both speech intelligibility and sound quality for CI users. Moreover, AI might play a role in tailoring CI settings to individual users based on their unique auditory profiles, leading to improved outcomes and user satisfaction. AI can also be used to develop personalized and adaptive programming strategies for CIs based on individual CI user's responses and preferences. In fact, similar approaches may also be pursued in the context of vision impairment [117].

However, while AI has made significant strides in understanding emotional expressions in both voices and faces, the ability to fully comprehend how CI users perceive these emotional cues remains a complex challenge. This is because understanding the subjective experience of CI users involves several factors:

1. Sensory Perception: Hearing with a CI is electronic, not biological, and can sound very different than "normal" biological hearing. Because a CI only enables rudimentary hearing, the human brain must learn to process the input via cortical plasticity – which in this case refers to adaptations of the brain to deprivation-induced altered sensory input in one but not other sensory modalities. Thus, understanding how CI users perceive vocal emotions, and how vocal and facial emotions are integrated into their perceptual systems [50], requires considering how their brain interprets the signals provided by the device.
2. Individual Variability: Perception of emotional cues can vary greatly among CI users, presumably due to a multitude of factors that include age of implantation, duration of deafness, auditory rehabilitation, and CI hardware and software. However, these huge interindividual differences and their influencing factors are not fully understood and remain part of ongoing CI research [51]. In this context, we anticipate that emerging efforts to establish central CI registries (e.g., [195]) will be instrumental to promote the (currently lacking) multi-center studies with large numbers of participants that are essential to identify the relative contributions of multiple factors for rehabilitation outcome.

Whereas AI can help analyze large datasets of CI users' responses in computer experiments that test their perceptual abilities to recognize vocal and facial emotions (both with unimodal and multimodal stimuli), understanding the subjective experience of emotion perception requires more than just data analysis. It involves interdisciplinary research combining insights from neuroscience, psychology, linguistics, and AI. In the absence of appropriate contextualization, AI technology may well produce misleading results.

Researchers are working on developing AI systems that can better understand and adapt to individual differences in sensory perception, including those of CI users. However, a complete understanding of how CI users perceive emotional voices and dynamic moving faces remains an ongoing challenge for research in this area.

3.10 Multimodal Modeling from a Dynamic Perspective

AI can be put to several uses, in general as well as in the context of multimodality (cf. Section 3.3):

- AI as tool – here the focus is on automatized task solving like, in the context of multimodal data, pre-processing and/or annotation.
- AI as companion – here the focus is on AI for social or information-seeking interaction. It is within this focal area that the behavior of AI can be compared to human behavior (Turing test). One challenge posed by multimodality as studied within ViCom (see Section 1.2) is that it requires to proceed from purely verbal T2 systems to sensorimotor, robotic T3 systems, that are able to actually produce and perceive multimodal behavior in the first place [82].
- AI as cognitive modeling – here the focus is on AI as an implemented model of cognition, that is, reverse engineering of neuro-cognitive capabilities.

Given the predominantly data analytic perspective spelled out in the previous section, we are mainly concerned with the first two foci, AI as a tool or as a companion. We touch on the second and third areas again in Section 4.

In light of the analyses in Section 3.1–Section 3.4, the following picture emerges regarding the dynamics of multimodal modeling in the context of ever-improving AI: A central aspect of the automation gap described by the projects concerns the mapping of multimodal data to linguistic models to explicate the underlying (e.g., syntactic, semantic, or pragmatic) structures. This mapping, like any modeling, is characterized by a number of informational uncertainties that take on a special character in the context of the underlying linguistic theories. We are dealing, so to speak, with a line that leads from the respective (1) primary data via their representation by means of (2) secondary or tertiary data and various intermediate representations to the (3) theoretical concepts of an already developed or still to be developed theory. In this way, several points of reference for the uncertainties described so far can be identified. The reason for this is that this chain of terms (data-driven, inter-representational, theory-driven) is linked to a series of decisions whose uncertainties determine the relationship between theory and data. Without suggesting a preferred direction, we choose the bottom-up variant in our enumeration of uncertainties, starting with the data:

1. *Primary data:* Instead of assuming that the data streams relevant for multimodal computing are predetermined per se, we can assume a wider range of possibilities beyond the established paths in linguistics. This range beyond the usual suspects (such as eye-tracking data and audio or video recordings)

concerns possibilities that arise not only from the use of interfaces with ever-increasing resolution. Rather, it concerns the possibilities of advanced data collection devices such as motion capture suits or full body trackers for VR, biometric sensors, EEG headbands (electroencephalography), EMG devices (electromyography), skin conductance sensors, or gait analysis technologies. The point is that ever new data along ever new data streams promotes bottom-up approaches, which make data exploration preferable to theory-driven top-down approaches. However, this could lead to a gradual devaluation of theoretical approaches along a technological process that guides data selection without theoretical embedding: it is then a technology-driven view of multimodal data rather than a theory-driven one. Thus, we face a *data-related decision space* that concerns the data types to be analyzed for multimodal modeling and the streams that instantiate them (in terms of subtypes and their resolution). In a positive sense, this data space increases the opportunities for purely exploratory, data-driven approaches as it continues to expand. In a negative sense, this creates pressure for theory development, on the basis of which, for example, we should be able to substantiate the formation of multimodal ensembles that can be found in the corresponding data streams.

2. *Intermediary representation:* For the success of AI technologies for modeling processes of multimodal fusion and fission, it is crucial that data from each modality are mapped into the same representation space based on the same (usually vectorial, numerical) representational terms. The more modalities are to be mapped, the more heterogeneous the provenance of such representations becomes, which, due to their representational homogeneity, can be directly related, linked or amalgamated. In the modeling chain described above, we mentioned the special role of representations between primary data and theoretical terms. The problem that arises at this point is that by relying on established representation models from computer science, the linguistic modeler adopts (even if only implicitly) a modeling language that does not necessarily meet his representational requirements (e.g. with regard to closeness to theory, theoretical grounding or motivability and explicitness). Once again, the modeler is confronted with a decision space in which actors from outside his discipline become active and, due to their disciplinary background, decide on representation issues that might be incomprehensible to him. From an exploratory point of view, this may be tempting (taking away the modeling decisions "to see what the computer does with the data"). However, the uncertainties involved increase with the success of such approaches, because the connection between data and theory is designed and controlled independently of the latter – possibly without any attempt to connect to a theory. LLMs, for example, are based on so-called subwords, which are selected from a large corpus according to roughly two criteria: they should occur as often as possible and make up the character stream of the corpus as completely as possible. Transformer models trained on such vocabularies are known to be very successful, even if the resulting subwords have nothing or very little in common with linguistically motivated

word forms or affixes. A representation gap occurs when the modeler refers to word forms as theoretical terms, even though they no longer play a role in the modeling used. This gap is all the greater in areas such as multimodal computing, where there is already theoretical uncertainty about the relevant "multimodal subwords". Thus, exploratory approaches based on heuristics around "multimodal subwords" and the modeling of their similarity and contiguity associations using neural networks should be strictly accompanied by theoretical approaches that take a closer look at the representational terms of multimodal computing.

3. *Theory formation:* The third level concerns the theoretical terms, for which we are primarily interested, in the light of the project descriptions in Section 3.1–Section 3.4, in the aspect of their openness or change in the course of their confrontation with the data. More precisely, this is about situations in which the annotation of multimodal data makes the modeler aware of the inadequacy of his theoretical terms, which makes the modification of these terms indispensable. The crucial question is the openness of the theory to such perturbations, or the degree of its changeability and adaptability, or the flexibility of the theoretical terms. We can ask, for example, whether such adaptations are merely a matter of rearranging the terms, adding additional labels for classes or relations, or whether the required changes are more extensive, calling into question the validity of the terms as a whole, which could ultimately lead to their abandonment, alternatively following a data-driven bottom-up approach. Thus, while under (1) and (2) we are dealing with questions of informational uncertain one-to-many mappings between data types, data streams, and data resolutions on the one hand, and their secondary or n-ary representation on the other, (3) is about the temporal dynamics of the underlying theories as a result of multimodal modeling, which may lead to re-entering the data-representation-theory circle again and again.

The project descriptions all referred to a special term that we called *multimodal ensemble* in Section 3.4. In light of the three-part chain described so far, this term itself can be located at three levels: (1) At the data level, we can segment a (vertically ordered) time slice of our (horizontally ordered) input data streams to qualify it as a manifestation of a particular ensemble (using some annotation software). (2) At the representation level (assuming a unified vector space model), one can ask which vector operations, starting from the vector representations of its elementary monomodal representations, generate which representation of the latter ensemble as a unit that can be related to which other units in the same representation space (this is ideally done completely automatically). (3) On the theoretical level, questions about the status of such ensembles are addressed in theoretical terms. For example, how the composition of an ensemble allows the suspension of certain (optional) modalities, while other modalities are obligatory. (4) As a unifying question for all these levels, one might ask to what extent repeated concomitances of multimodal data streams condition the constitution of a distinguishable multimodal ensemble, which as a whole becomes (syntagmatically) combinable or (paradigmatically) interchangeable with other ensembles as if they begin to take on properties of gestalts or even signs.

According to the variety of decisions that have to be made to be able to address multimodal ensembles as observables of a theory of multimodal communication, the statistical connections of the first three levels, as addressed by level (4), are characterized by uncertainties to which linguistic research is expected to provide answers, *without expecting that AI will replace this research.* Thus, while a primarily bottom-up approach runs the risk of following positivism that over-interprets the data streams that, for whatever reason, are technically available as direct access to multimodal ensembles, a primarily top-down approach runs the risk of developing concepts without an empirical foundation that are likely to require modifications and adjustments all too quickly in the course of confrontation with the data.

The detection of feedback behavior Section 3.6 can be taken as an example here: From a linguistic point of view, the recognition of such behavior and its semantic interpretation requires recourse to the respective conversational context, possibly even to the underlying conversational history. Assuming the context model of transformers, these examine context windows based on the underlying subword vocabulary, with the respective conversation appearing as a stream of subwords (and their vector representations) that is traversed window by window. Despite the efficiency of this approach, it is clear that with the window width and the focus on subwords, we are using a context model that undermines the flexibility of human context interpretation. For example, the interpreter may discover ad hoc ensembles in the conversation that make it possible to classify a behavior as feedback in the first place. Such an ensemble would manifest a kind of firstness, of which it cannot be assumed that a generative AI has seen enough training examples to identify the function associated with it, especially if the data streams to which it has access undermine the streams that a human perceives. The AI then does not "see" the ensemble in question (e.g. as a set or sequence of multimodal subwords). Conversely, however, we will follow a human interpretation only if our theory provides concepts that allow us to model such ensembles and thus create the conditions for their empirical and systematic observation. Human interpretation may be flexible and open, but for our theoretical purposes, we need a controllable approach that ensures the intersubjectivity of interpretation. At this point, which is characterized by the tension between contextual fixity and the lack of flexibility of automated methods on the one hand, and contextual openness with little formal commitment on the other, an approach comes into play that is able to integrate these two perspectives. At this point, we are thinking of an approach from the field of human computation, more precisely from the spectrum of evolutionary approaches [110], in which man and machine take on the functions of innovation and selection, thus each acting in a dual role [142], but in such a way that, as this interaction evolves, theory formation is simultaneously driven primarily by humans, while the machine focuses on optimizing the exploration of data streams and their representations, thereby bridging between data and theory.

4 Long-Term Obstacles

A fundamental prerequisite for contemporary AI systems are the need for training with relevant data before they can effectively process a problem (cf. Section 2). If a system has not seen the relevant data in any form, it cannot process it. To illustrate this, consider a model that is trained only on text data and therefore cannot be extended to image data or data from other modalities. Such models can therefore not solve application problems for which no measurable data is available, even in the long term. This concerns data as diverse as those from neuronal voxel measurements, inner monologues, or past events that have never been recorded and cannot be retrospectively constructed. In other words, it is unlikely that a generative AI will be able to read minds by being shown videos of non-speaking faces; nor is it likely that prompts will be contextualized by past events for which there are no recordings. And monitoring neural processes and structures currently seem to be limited to cellular resolution [215,161].

To think beyond such usual suspects for examples of unattainability by generative AI, one can think of several other candidates. Whatever comes to mind in this context, one has to face the situation that the corresponding research will generate significant amounts of documents that will essentially serve as a training base for generative AI to do both: link the new data with the data it already has access to (i.e., embed it in its representation space(s)), and play its statistical game to produce meaningful, well-interpretable texts in response to prompts for the supposedly untapped field of application. This consideration makes it difficult to identify problems that current generative AI will definitely not be able to overcome, even in the very long term.

To make this point very clear: language is a tool that is used to exchange information, which can relate in particular to things that are not present in the immediate, perceptible environment. Even if AI has no access to some primary data, it becomes able to recognize and respond to topics corresponding to inaccessible sources *if* there are written testimonials describing such data. Think of historical works narrating the past, diaries which reveal the author's inner episodes, or science fiction novels describing otherworldly scenarios[23]. With respect to multimodal research this means that the more papers published, the more text data for AI. If this is true, then there is just one way to constrain AI systems: Don't feed the AI! In fact, there are a couple of conceivable circumstances that might lead to this effect.

One positive development is that AI systems are suddenly solving tasks that humans were previously unable to solve. But in combination with these systems, people then develop a new understanding of precisely these tasks and continue to develop (e.g. Chess [179], Go [192]). However, it also has the opposite effect, in that these systems destroy the data basis on which they were trained. One of

[23] This despite the AI being caught in a symbolic cycle, lacking the possibility to break out of the "web of words" by means of grounding symbols in experience [81].

the best-known examples is the impact of ChatGPT on Stack Overflow, where Stack Overflow traffic dropped significantly after the release of ChatGPT.[24]

Another direction of this development could be that the internet will be flooded with more and more automatically generated data, which will become less and less distinguishable from human-generated data as these systems improve [178]. This means that future models will be trained more and more on self-generated data, which could also lead to an expected decline in model performance.

In a similar manner, communication spaces might come into existence, that exclude AI chatbots. This can happen in response to social media platforms that are more and more overtaken by dialogical AI systems while human user want to interact with other human users. Communication going on in the new, AI-free channels will then be out of the recognitional reach of the AI systems.

In addition to these social effects, political effects are also to be expected, for instance, that certain data is no longer legally accessible, or may no longer be used for the training of LLMs. In the long run, such effects may be much more influential and relevant to the development and improvement of AI models than the actual development of these systems.

Author contribution

	Sections													
	1	2	3	3.1	3.2	3.3	3.4	3.5	3.6	3.7	3.8	3.9	3.10	4
AH	■	■	□											■
AB		□							■					
RB											■			■
AĆ		□			■									
AG								■						
FK								■						
JL											■			■
AL	□	□		■		■	■						□	■
AM	■	□	■										■	■
PP								■						
PSR								■						
JS									■					
MSR										■				
SRS												■		
CvE												■		

Legend: ■ main work; □ minor contribution.

[24] https://www.similarweb.com/blog/insights/ai-news/stack-overflow-chatgpt/

References

1. Aguirre-Celis, N., Miikkulainen, R.: Understanding the semantic space: How word meanings dynamically adapt in the context of a sentence. In: Proceedings of the 2021 Workshop on Semantic Spaces at the Intersection of NLP, Physics, and Cognitive Science (SemSpace). pp. 1–11 (2021)
2. Aiyappa, R., An, J., Kwak, H., Ahn, Y.Y.: Can we trust the evaluation on chatgpt? arXiv preprint arXiv:2303.12767 (2023)
3. Akbari, H., Yuan, L., Qian, R., Chuang, W.H., Chang, S.F., Cui, Y., Gong, B.: Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text. Advances in Neural Information Processing Systems **34**, 24206–24221 (2021)
4. Alkaissi, H., McFarlane, S.I.: Artificial hallucinations in chatgpt: implications in scientific writing. Cureus **15**(2) (2023)
5. Almazrouei, E., Alobeidli, H., Alshamsi, A., Cappelli, A., Cojocaru, R., Debbah, M., Goffinet, E., Heslow, D., Launay, J., Malartic, Q., Noune, B., Pannier, B., Penedo, G.: Falcon-40B: an open large language model with state-of-the-art performance. Findings of the Association for Computational Linguistics: ACL **2023**, 10755–10773 (2023)
6. Ambrazaitis, G., House, D.: The multimodal nature of prominence: some directions for the study of the relation between gestures and pitch accents. In: Proceedings of the 13th International Conference of Nordic Prosody. pp. 262–273 (2023). https://doi.org/10.2478/9788366675728-024
7. American Psychiatric Association: Diagnostic and statistical manual of mental disorders, 5th ed. Tech. rep., American Psychiatric Association, Arlington (2013)
8. Amici, F., Liebal, K.: Testing Hypotheses for the Emergence of Gestural Communication in Great and Small Apes (Pan troglodytes, Pongo abelii, Symphalangus syndactylus). International Journal of Primatology (Dec 2022). https://doi.org/10.1007/s10764-022-00342-7, https://doi.org/10.1007/s10764-022-00342-7
9. Andonova, E., Taylor, H.A.: Nodding in dis/agreement: a tale of two cultures. Cognitive Processing **13**(S1), 79–82 (Aug 2012). https://doi.org/10.1007/s10339-012-0472-x, citation Key: Andonova2012
10. Anger, C., Berwe, T., Olszok, A., Reichenberger, A., Lemanski, J.: Five dogmas of logic diagrams and how to escape them. Language & Communication **87**, 258–270 (Nov 2022). https://doi.org/10.1016/j.langcom.2022.09.001, https://www.sciencedirect.com/science/article/pii/S0271530922000775
11. Anil, R., Dai, A.M., Firat, O., Johnson, M., Lepikhin, D., Passos, A., Shakeri, S., Taropa, E., Bailey, P., Chen, Z., Chu, E., Clark, J.H., Shafey, L.E., Huang, Y., Meier-Hellstern, K., Mishra, G., Moreira, E., Omernick, M., Robinson, K., Ruder, S., Tay, Y., Xiao, K., Xu, Y., Zhang, Y., Abrego, G.H., Ahn, J., Austin, J., Barham, P., Botha, J., Bradbury, J., Brahma, S., Brooks, K., Catasta, M., Cheng, Y., Cherry, C., Choquette-Choo, C.A., Chowdhery, A., Crepy, C., Dave, S., Dehghani, M., Dev, S., Devlin, J., Díaz, M., Du, N., Dyer, E., Feinberg, V., Feng, F., Fienber, V., Freitag, M., Garcia, X., Gehrmann, S., Gonzalez, L., Gur-Ari, G., Hand, S., Hashemi, H., Hou, L., Howland, J., Hu, A., Hui, J., Hurwitz, J., Isard, M., Ittycheriah, A., Jagielski, M., Jia, W., Kenealy, K., Krikun, M., Kudugunta, S., Lan, C., Lee, K., Lee, B., Li, E., Li, M., Li, W., Li, Y., Li, J., Lim, H., Lin, H., Liu, Z., Liu, F., Maggioni, M., Mahendru, A., Maynez, J., Misra, V., Moussalem, M., Nado, Z., Nham, J., Ni, E., Nystrom, A., Parrish, A., Pellat, M., Polacek, M., Polozov, A., Pope, R., Qiao, S., Reif, E., Richter, B., Riley, P.,

Ros, A.C., Roy, A., Saeta, B., Samuel, R., Shelby, R., Slone, A., Smilkov, D., So, D.R., Sohn, D., Tokumine, S., Valter, D., Vasudevan, V., Vodrahalli, K., Wang, X., Wang, P., Wang, Z., Wang, T., Wieting, J., Wu, Y., Xu, K., Xu, Y., Xue, L., Yin, P., Yu, J., Zhang, Q., Zheng, S., Zheng, C., Zhou, W., Zhou, D., Petrov, S., Wu, Y.: PaLM 2 technical report. arXiv 2305.10403 (2023)
12. Archer, D.: Unspoken diversity: Cultural differences in gestures. Qualitative sociology **20**, 79–105 (1997)
13. Aruin, A.S., Latash, M.L.: Directional specificity of postural muscles in feed-forward postural reactions during fast voluntary arm movements. Experimental Brain Research **103**(2), 323–332 (Mar 1995). https://doi.org/10.1007/BF00231718, https://doi.org/10.1007/BF00231718
14. Bangerter, A., Oppenheimer, D.M.: Accuracy in detecting referents of pointing gestures unaccompanied by language. Gesture **6**(1), 85–102 (2006)
15. Baroni, M.: Grounding distributional semantics in the visual world. Language and Linguistics Compass **10**(1), 3–13 (2016)
16. Barwise, J., Etchemendy, J.: Chapter VIII Heterogeneous Logic. Logical reasoning with diagrams pp. 179–200 (1996)
17. Bauer, A.: Russian multimodal conversational data (May 2023). https://doi.org/10.18716/DCH/A.00000016, https://dch.phil-fak.uni-koeln.de/bestaende/datensicherung/russian-multimodal-conversational-data
18. Bauer, A., Poryadin, R.: Russian sign language conversations. https://dch.phil-fak.uni-koeln.de/bestaende/datensicherung/russian-sign-language-conversations (2023). https://doi.org/10.18716/DCH/A.00000028
19. Bavelas, J.B., Coates, L., Johnson, T.: Listeners as co-narrators. Journal of Personality and Social Psychology **79**(6), 941–952 (2000). https://doi.org/10.1037/0022-3514.79.6.941
20. Beckman, M.E., Ayers-Elam, G.: Guidelines for ToBI Labelling: Version 3. Ohio State University. http://www.ling.ohio-state.edu/~tobi/ame_tobi/labelling_guide_v3.pdf (1997)
21. Bellucci, F., Pietarinen, A.V.: Two dogmas of diagrammatic reasoning: a view from existential graphs. In: Peirce on Perception and Reasoning: From icons to logic, pp. 174–195. Routledge (2017)
22. Bertsch, A., Alon, U., Neubig, G., Gormley, M.R.: Unlimiformer: Long-range transformers with unlimited length input. arXiv preprint arXiv:2305.01625 (2023)
23. Betker, J., Goh, G., Jing, L., Brooks, T., Wang, J., Li, L., Ouyang, L., Zhuang, J., Lee, J., Guo, Y., et al.: Improving image generation with better captions. Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf **2**(3), 8 (2023)
24. Bhattacharjee, R., Chakraborty, M.K., Choudhury, L.: Venn$_{i_{o_1}}$: A diagram system for universe without boundary. Logica Universalis **13**(3), 289–346 (Sep 2019). https://doi.org/10.1007/s11787-019-00227-z, https://doi.org/10.1007/s11787-019-00227-z
25. Bhattacharjee, R., Moktefi, A.: Revisiting peirce's rules of transformation for euler-venn diagrams. In: Basu, A., Stapleton, G., Linker, S., Legg, C., Manalo, E., Viana, P. (eds.) Diagrammatic Representation and Inference. pp. 166–182. Lecture Notes in Computer Science, Springer International Publishing, Cham (2021). https://doi.org/10.1007/978-3-030-86062-2_14
26. Boersma, P., Weenink, D.: Praat: Doing phonetics by computer, www.praat.org/
27. Bolly, C.: CorpAGEst Annotation Manual (II. Speech Annotation Guidelines) (2016)

28. Bolt, R.A.: “put-that-there”: Voice and gesture at the graphics interface. SIGGRAPH Comput. Graph. **14**, 262–270 (1980). https://doi.org/10.1145/965105.807503
29. Borodo, M.: Multimodality, translation and comics. Perspectives **23**(1), 22–41 (2015)
30. Braunschweiler, N.: The Prosodizer – automatic prosodic annotations of speech synthesis databases. In: Proceedings of Speech Prosody. vol. 2006 (2006)
31. Bressem, J., Ladewig, S.H., Müller, C.: Linguistic annotation system for gestures. In: Müller, C., Cienki, A., Fricke, E., Ladewig, S., McNeill, D., Teßendorf, S. (eds.) Body – Language – Communication. An International Handbook on Multimodality in Human Interaction, Handbücher zur Sprach- und Kommunikationswissenschaft / Handbooks of Linguistics and Communication Science (HSK) 38/1, vol. 1, chap. 71, pp. 1098–1124. De Gruyter Mouton, Berlin and Boston (2013). https://doi.org/10.1515/9783110261318.1098
32. Brown, L., Prieto, P.: Gesture and prosody in multimodal communication. In: Haugh, M., Kádár, D.Z., Terkourafi, M. (eds.) The Cambridge Handbook of Sociopragmatics, chap. 21, pp. 430–453. Cambridge University Press, Cambridge UK (2021)
33. Bulat, L., Clark, S., Shutova, E.: Speaking, seeing, understanding: Correlating semantic models with conceptual representation in the brain. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. pp. 1081–1091 (2017)
34. Cao, Y., Li, S., Liu, Y., Yan, Z., Dai, Y., Yu, P.S., Sun, L.: A comprehensive survey of ai-generated content (aigc): A history of generative ai from gan to chatgpt. arXiv preprint arXiv:2303.04226 (2023)
35. 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
36. Chen, J., Ho, C.M.: Mm-vit: Multi-modal video transformer for compressed video action recognition. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 1910–1921 (January 2022)
37. Chu, J., Liu, Y., Yang, Z., Shen, X., Backes, M., Zhang, Y.: Comprehensive assessment of jailbreak attacks against llms. arXiv preprint arXiv:2402.05668 (2024)
38. Clark, H.H.: Using Language. Cambridge University Press, Cambridge (1996)
39. Cowen, A.S., Keltner, D.: What the face displays: Mapping 28 emotions conveyed by naturalistic expression. American Psychologist **75**(3), 349–364 (Apr 2020). https://doi.org/10.1037/amp0000488, http://dx.doi.org/10.1037/amp0000488
40. Davidson, D.: Three varieties of knowledge. Royal Institute of Philosophy Supplement **30**, 153–166 (1991). https://doi.org/10.1017/S1358246100007748
41. De Jong, N.H., Wempe, T.: Praat script to detect syllable nuclei and measure speech rate automatically. Behavior research methods **41**(2), 385–390 (2009)
42. Demey, L., Smessaert, H.: A database of aristotelian diagrams: Empirical foundations for logical geometry. In: International Conference on Theory and Application of Diagrams. pp. 123–131. Springer (2022)
43. Diessel, H.: Demonstratives, joint attention, and the emergence of grammar. Cognitive Linguistics **17**(4), 463–489 (2006). https://doi.org/10.1515/COG.2006.015
44. Dingemanse, M., Enfield, N.J.: Other-initiated repair across languages: towards a typology of conversational structures. Open Linguistics **1**(1) (Jan 2015). https://doi.org/10.2478/opli-2014-0007, https://www.degruyter.com/doi/10.2478/opli-2014-0007

45. 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), 7561 (2022)
46. Doherty, E., Davila Ross, M., Clay, Z.: Multimodal communication development in semi-wild chimpanzees. Animal Behaviour (Feb 2023)
47. Dong, L., Xu, S., Xu, B.: Speech-transformer: a no-recurrence sequence-to-sequence model for speech recognition. In: 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP). pp. 5884–5888. IEEE (2018)
48. Douglas, P.H., Moscovice, L.R.: Pointing and pantomime in wild apes? Female bonobos use referential and iconic gestures to request genito-genital rubbing. Scientific Reports **5**(1) (Sep 2015). https://doi.org/10.1038/srep13999
49. Düking, P., Sperlich, B., Voigt, L., Van Hooren, B., Zanini, M., Zinner, C.: ChatGPT generated training plans for runners are not rated optimal by coaching experts, but increase in quality with additional input information. Journal of Sports Science and Medicine **23**, 56–72 (2024). https://doi.org/10.52082/jssm.2024.56
50. 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) (2022)
51. von Eiff, C.I., Skuk, V.G., Zäske, R., Nussbaum, C., Frühholz, S., Feuer, U., Guntinas-Lichius, O., Schweinberger, S.R.: Parameter-specific morphing reveals contributions of timbre to the perception of vocal emotions in cochlear implant users. Ear and Hearing **43**(4), 1178 (2022)
52. Ekman, P., Friesen, W.V.: The repertoire of nonverbal behavior: Categories, origins, usage, and coding. Semiotica **1**(1), 49–98 (1969)
53. Engelen, J., Bernareggi, C.: Ascience: a thematic network on access to scientific university courses by visually impaired students. In: CHALLENGES FOR ASSISTIVE TECHNOLOGY. vol. 20, pp. 304–309. IOS Press (Jan 2007), https://lirias.kuleuven.be/56044
54. Engelhardt, Y., Richards, C.: A framework for analyzing and designing diagrams and graphics. In: Diagrams (2018), https://api.semanticscholar.org/CorpusID:49189675
55. 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)
56. Esteve-Gibert, N., Guellaï, B.: Prosody in the Auditory and Visual Domains: A Developmental Perspective. Frontiers in Psychology **9** (2018), https://www.frontiersin.org/articles/10.3389/fpsyg.2018.00338
57. Feldman, J.: The neural binding problem(s). Cognitive neurodynamics **7**(1), 1–11 (2013). https://doi.org/10.1007/s11571-012-9219-8
58. Fish, A., Stapleton, G.: Defining euler diagrams: simple or what? In: International Conference on Theory and Application of Diagrams. pp. 109–111. Springer (2006)
59. Frege, G.: Der Gedanke. Beiträge zur Philosophie des deutschen Idealismus **1**(2), 58–77 (1918)
60. Frieder, S., Pinchetti, L., Griffiths, R.R., Salvatori, T., Lukasiewicz, T., Petersen, P.C., Chevalier, A., Berner, J.: Mathematical capabilities of chatgpt. arXiv preprint arXiv:2301.13867 (2023)
61. Fuchs, S., Kadavá, Š., Pouw, W., Ćwiek, A., Walker, B., Fay, N., Winter, B.: Exploring the sound structure of novel vocalizations. In: Proceedings of EVOLANG 2024. Madison, Wisconsin, USA (2024)

62. Galaz García, C., Bagstad, K.J., Brun, J., Chaplin-Kramer, R., Dhu, T., Murray, N.J., Nolan, C.J., Ricketts, T.H., Sosik, H.M., Sousa, D., et al.: The future of ecosystem assessments is automation, collaboration, and artificial intelligence. Environmental Research Letters **18** (2023)
63. Gardner, R.: When Listeners Talk: Response tokens and listener stance, Pragmatics & Beyond New Series, vol. 92. John Benjamins Publishing Company, Amsterdam (Jan 2001). https://doi.org/10.1075/pbns.92, http://www.jbe-platform.com/content/books/9789027297426
64. Giaquinto, M.: Crossing Curves: A Limit to the Use of Diagrams in Proofs†. Philosophia Mathematica **19**(3), 281–307 (08 2011). https://doi.org/10.1093/philmat/nkr023, https://doi.org/10.1093/philmat/nkr023
65. Giardino, V.: Diagrammatic proofs in mathematics: (Almost) 20 years of research. In: Sriraman, B. (ed.) Handbook of the History and Philosophy of Mathematical Practice, pp. 1–23. Springer International Publishing, Cham (2020). https://doi.org/10.1007/978-3-030-19071-2_46-1, https://doi.org/10.1007/978-3-030-19071-2_46-1
66. Ginzburg, J., Cooper, R., Hough, J., Schlangen, D.: Incrementality and HPSG: Why not? In: Abeillé, A., Bonami, O. (eds.) Constraint-Based Syntax and Semantics: Papers in Honor of Danièle Godard. CSLI Publications, Stanford, CA (2020)
67. Ginzburg, J., Lücking, A.: I thought pointing is rude: A dialogue-semantic analysis of pointing at the addressee. In: Grosz, P., Martí, L., Pearson, H., Sudo, Y., Zobel, S. (eds.) Proceedings of *Sinn und Bedeutung 25*. pp. 276–291. SuB 25 (2021). https://doi.org/10.18148/sub/2021.v25i0.937, https://ojs.ub.uni-konstanz.de/sub/index.php/sub/article/view/937, special Session: Gestures and Natural Language Semantics
68. Ginzburg, J., Mazzocconi, C., Tian, Y.: Laughter as language. Glossa **5**(1), 104 (2020). https://doi.org/10.5334/gjgl.1152
69. Goldstein, L.: Chapter 15 Teaching syllogistic to the blind. In: Gorayska, B., Mey, J.L. (eds.) Advances in Psychology, Cognitive Technology, vol. 113, pp. 243–255. North-Holland (Jan 1996). https://doi.org/10.1016/S0166-4115(96)80035-5, https://www.sciencedirect.com/science/article/pii/S0166411596800355
70. Gonitsioti, H., Christidou, V., Hatzinikita, V.: Enhancing scientific visual literacy in kindergarten: young children 'read' and produce representations of classification. The International Journal of Science, Mathematics and Technology Learning **20**(1), 1–15 (2013). https://doi.org/10.18848/2327-7971/CGP/v20i01/48996, https://cgscholar.com/bookstore/works/enhancing-scientific-visual-literacy-in-kindergarten
71. Gregori, A., Amici, F., Brilmayer, I., Ćwiek, A., Fritzsche, L., Fuchs, S., Henlein, A., Herbort, O., Kügler, F., Lemanski, J., Liebal, K., Lücking, A., Mehler, A., Nguyen, K.T., Pouw, W., Prieto, P., Rohrer, P.L., Sánchez-Ramón, P.G., Schulte-Rüther, M., Schumacher, P.B., Schweinberger, S.R., Struckmeier, V., Trettenbrein, P.C., von Eiff, C.I.: A roadmap for technological innovation in multimodal communication research. In: Duffy, V.G. (ed.) Digital Human Modeling and Applications in Health, Safety, Ergonomics and Risk Management. pp. 402–438. Springer Nature Switzerland, Cham (2023)
72. Gregoromichelaki, E., Kempson, R., Purver, M., Mills, G.J., Cann R., R., Meyer-Viol, W., Healey, P.G.: Incrementality and intention-recognition in utterance processing. Dialogue and Discourse **2**(1), 199–233 (2011). https://doi.org/10.5087/dad.2011.109

73. Grice, M., Baumann, S., Benzmüller, R.: German intonation in autosegmental-metrical phonology. In: Jun, S.A. (ed.) Prosodic Typology: The Phonology of Intonation and Phrasing, pp. 55–83. Oxford University Press, Oxford (2005)
74. Grice, M., Reyelt, M., Benzmüller, R., Mayer, J., Batliner, A.: Consistency in Transcription and Labelling of German Intonation with GToBI. In: Proceedings of the Fourth International Conference on Spoken Language Processing. vol. 3, pp. 1716–1719. Philadelphia and USA (1996)
75. Hagoort, P., van Berkum, J.: Beyond the sentence given. Philosophical Transactions of the Royal Society B: Biological Sciences **362**(1481), 801–811 (2007). https://doi.org/10.1098/rstb.2007.2089
76. Hahn, H.: The crisis in intuition. In: Hahn, H., McGuinness, B. (eds.) Empiricism, Logic and Mathematics: Philosophical Papers, pp. 73–102. Vienna Circle Collection, Springer Netherlands, Dordrecht (1980). https://doi.org/10.1007/978-94-009-8982-5_7, https://doi.org/10.1007/978-94-009-8982-5_7
77. Halevy, A., Norvig, P., Pereira, F.: The unreasonable effectiveness of data. IEEE intelligent systems **24**(2), 8–12 (2009)
78. Hammer, E.M.: Logic and visual information. CSLI Publications, Stanford (1995)
79. Han, E.: Integrating mobile eye-tracking and vslam for recording spatial gaze in works of art and architecture. Technology|Architecture + Design **5**(2), 177–187 (Jul 2021). https://doi.org/10.1080/24751448.2021.1967058, http://dx.doi.org/10.1080/24751448.2021.1967058
80. Han, K., Xiao, A., Wu, E., Guo, J., XU, C., Wang, Y.: Transformer in transformer. In: Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., Vaughan, J.W. (eds.) Advances in Neural Information Processing Systems. vol. 34, pp. 15908–15919. Curran Associates, Inc. (2021), https://proceedings.neurips.cc/paper_files/paper/2021/file/854d9fca60b4bd07f9bb215d59ef5561-Paper.pdf
81. Harnad, S.: The symbol grounding problem. Physica D: Nonlinear Phenomena **42**(1-3), 335–346 (1990). https://doi.org/10.1016/0167-2789(90)90087-6
82. Harnad, S.: Minds, machines and Turing. In: Moor, J.H. (ed.) The Turing Test: The Elusive Standard of Artificial Intelligence, pp. 253–273. Springer Netherlands, Dordrecht (2003). https://doi.org/10.1007/978-94-010-0105-2_14
83. Harnad, S.: Language writ large: Llms, chatgpt, grounding, meaning and understanding. arXiv 2402.02243 (2024)
84. 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**, 673982 (Aug 2021). https://doi.org/10.3389/fpsyg.2021.673982
85. Hassani, H., Silva, E.S.: The role of chatgpt in data science: how ai-assisted conversational interfaces are revolutionizing the field. Big data and cognitive computing **7**(2), 62 (2023)
86. Herbort, O., Krause, L.M., Kunde, W.: Perspective determines the production and interpretation of pointing gestures. Psychonomic Bulletin & Review **28**, 641–648 (2021). https://doi.org/10.3758/s13423-020-01823-7
87. Herbort, O., Kunde, W.: Spatial (mis-)interpretation of pointing gestures to distal referents. Journal of Experimental Psychology: Human Perception and Performance (2015). https://doi.org/10.1037/xhp0000126, online First Publication
88. Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., Casas, D.d.L., Hendricks, L.A., Welbl, J., Clark, A., et al.: Training compute-optimal large language models. arXiv preprint arXiv:2203.15556 (2022)
89. Hohwy, J.: The predictive processing hypothesis. The Oxford Handbook of 4E cognition pp. 129–145 (2018)

90. Holler, J.: Speakers' use of interactive gestures as markers of common ground. In: Kopp, S., Wachsmuth, I. (eds.) Proceedings of Gesture Workshop 2009, pp. 11–22. No. 5934 in Lecture Notes in Artificial Intelligence, Springer, Berlin and Heidelberg (2010)
91. 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, opinion
92. Howse, J., Molina, F., Taylor, J., Kent, S., Gil, J.: Spider diagrams: A diagrammatic reasoning system. Journal of Visual Languages and Computing **12**(3), 299–324 (2001)
93. Huang, B., Shu, Y., Zhang, T., Wang, Y.: Dynamic multi-person mesh recovery from uncalibrated multi-view cameras. In: 3DV (2021)
94. Huang, J., Tan, M.: The role of chatgpt in scientific communication: writing better scientific review articles. American Journal of Cancer Research **13**(4), 1148 (2023)
95. Ienaga, N., Cravotta, A., Terayama, K., Scotney, B.W., Saito, H., Busa, M.G.: Semi-automation of gesture annotation by machine learning and human collaboration. Language Resources and Evaluation **56**(3), 673–700 (2022). https://doi.org/10.1007/s10579-022-09586-4
96. Jack, R.E., Blais, C., Scheepers, C., Schyns, P.G., Caldara, R.: Cultural confusions show that facial expressions are not universal. Current biology **19**(18), 1543–1548 (2009)
97. Jackendoff, R.: Foundations of Language. Oxford University Press, Oxford, UK (2002)
98. Jamnik, M.: Mathematical reasoning with diagrams. Lecture Notes, Center for the Study of Language and Information (2001), https://press.uchicago.edu/ucp/books/book/distributed/M/bo3614100.html
99. Jang, J., Ye, S., Seo, M.: Can large language models truly understand prompts? a case study with negated prompts. In: Transfer Learning for Natural Language Processing Workshop. pp. 52–62. PMLR (2023)
100. Johansen, M.W.: What's in a diagram? on the classification of symbols, figures and diagrams. In: Model-Based Reasoning in Science and Technology: Theoretical and Cognitive Issues, pp. 89–108. Springer (2013)
101. Johansen, M.W., Pallavicini, J.L.: Entering the valley of formalism: trends and changes in mathematicians' publication practice—1885 to 2015. Synthese **200**(3), 239 (2022)
102. Kadavá, Š., Ćwiek, A., Stoltmann, K., Fuchs, S., Pouw, W.: Is gesture-speech physics at work in rhythmic pointing? Evidence from Polish counting-out rhymes. In: Proceedings of the 20th International Congress of Phonetic Sciences. Prague, Czech Republic (Apr 2023). https://doi.org/10.31219/osf.io/67fzc, https://osf.io/67fzc
103. Kendon, A.: Gesture: Visible Action as Utterance. Cambridge University Press, Cambridge, MA (2004)
104. Khan, S., Naseer, M., Hayat, M., Zamir, S.W., Khan, F.S., Shah, M.: Transformers in vision: A survey. ACM computing surveys (CSUR) **54**(10s), 1–41 (2022)
105. Kiela, D., Bulat, L., Clark, S.: Grounding semantics in olfactory perception. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers). pp. 231–236 (2015)
106. Kiela, D., Clark, S.: Learning neural audio embeddings for grounding semantics in auditory perception. Journal of Artificial Intelligence Research **60**, 1003–1030 (2017)

107. Kisler, T., Schiel, F., Sloetjes, H.: Signal processing via web services: the use case WebMAUS. In: Digital Humanities Conference 2012. p. 5 (2012)
108. Klin, A.: Attributing social meaning to ambiguous visual stimuli in higher-functioning autism and Asperger syndrome: The Social Attribution Task. Journal of Child Psychology and Psychiatry, and allied displines **41**(7), 831–46 (Oct 2000)
109. Konrad, R., Hanke, T., Langer, G., Blanck, D., Bleicken, J., Hofmann, I., Jeziorski, O., König, L., König, S., Nishio, R., Regen, A., Salden, U., Wagner, S., Worseck, S., Schulder, M.: My dgs – annotated. public corpus of german sign language, 3rd release (2020). https://doi.org/10.25592/dgs.corpus-3.0, https://doi.org/10.25592/dgs.corpus-3.0, citation Key: Konrad2020
110. Kosorukoff, A.: Human based genetic algorithm. In: IEEE International Conference on Systems, Man, and Cybernetics. vol. 5, pp. 3464–3469 vol.5 (2001), http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=972056
111. Krahmer, E., Swerts, M.: The effects of visual beats on prosodic prominence: Acoustic analyses, auditory perception and visual perception. Journal of Memory and Language **57**(3), 396–414 (Oct 2007). https://doi.org/10.1016/j.jml.2007.06.005, http://www.sciencedirect.com/science/article/pii/S0749596X07000708
112. Kranstedt, A.: Situierte Generierung deiktischer Objektreferenz in der multimodalen Mensch-Maschine-Interaktion. No. 313 in Diski, Aka, Berlin (2008), phD Thesis, Bielefeld University
113. Kranstedt, A., Lücking, A., Pfeiffer, T., Rieser, H., Wachsmuth, I.: Deictic object reference in task-oriented dialogue. In: Rickheit, G., Wachsmuth, I. (eds.) Situated Communication, pp. 155–207. Mouton de Gruyter, Berlin (2006)
114. Krivokapić, J.: Gestural coordination at prosodic boundaries and its role for prosodic structure and speech planning processes. Philosophical Transactions of the Royal Society B: Biological Sciences **369**(1658), 20130397 (Dec 2014). https://doi.org/10.1098/rstb.2013.0397, https://royalsocietypublishing.org/doi/10.1098/rstb.2013.0397
115. Kuder, A., Bauer, A.: Polish multimodal conversational data (May 2023). https://doi.org/10.18716/DCH/A.00000017, https://dch.phil-fak.uni-koeln.de/bestaende/datensicherung/polish-multimodal-conversational-data
116. Ladd, D.R.: Intonational Phonology. Cambridge University Press, Cambridge, 2 edn. (2008)
117. Lane, J., Rohan, E.M.F., Sabeti, F., Essex, R.W., Maddess, T., Barnes, N., He, X., Robbins, R.A., Gradden, T., McKone, E.: Improving face identity perception in age-related macular degeneration via caricaturing. Scientific Reports **8**, 15205 (2018)
118. Lawson-Adams, J., Dickinson, D.K.: Building lexical representations with non-verbal supports. Reading Research Quarterly **56**(3), 603–622 (2021)
119. Lee, U., Jung, H., Jeon, Y., Sohn, Y., Hwang, W., Moon, J., Kim, H.: Few-shot is enough: exploring ChatGPT prompt engineering method for automatic question generation in english education. Education and Information Technologies pp. 1–33 (2023)
120. Levinson, S.C.: Deixis. In: Horn, L.R., Ward, G. (eds.) The Handbook of Pragmatics, chap. 5, pp. 97–121. Blackwell (2008)
121. Levinson, S.C., Torreira, F.: Timing in turn-taking and its implications for processing models of language. Frontiers in Psychology **6**(731) (2015). https://doi.org/10.3389/fpsyg.2015.00731
122. Li, C., Gan, Z., Yang, Z., Yang, J., Li, L., Wang, L., Gao, J.: Multimodal foundation models: From specialists to general-purpose assistants. arXiv preprint arXiv:2309.10020 **1**(2), 2 (2023)

123. Liebal, K., Slocombe, K.E., Waller, B.M.: The language void 10 years on: multimodal primate communication research is still uncommon. Ethology Ecology & Evolution pp. 1–14 (Jan 2022). https://doi.org/10.1080/03949370.2021.2015453, https://www.tandfonline.com/doi/full/10.1080/03949370.2021.2015453
124. Liesenfeld, A., Lopez, A., Dingemanse, M.: The timing bottleneck: Why timing and overlap are mission-critical for conversational user interfaces, speech recognition and dialogue systems. In: Proceedings of the 24th Meeting of the Special Interest Group on Discourse and Dialogue. p. 482–495. Association for Computational Linguistics, Prague, Czechia (2023). https://doi.org/10.18653/v1/2023.sigdial-1.45, https://aclanthology.org/2023.sigdial-1.45
125. Litwin, P., Miłkowski, M.: Unification by fiat: Arrested development of predictive processing. Cognitive Science **44**, e12867 (2020). https://doi.org/10.1111/cogs.12867
126. Liu, C., Mao, Z., Zhang, T., Liu, A.A., Wang, B., Zhang, Y.: Focus your attention: A focal attention for multimodal learning. IEEE Transactions on Multimedia **24**, 103–115 (2020)
127. Liu, Y., Deng, G., Xu, Z., Li, Y., Zheng, Y., Zhang, Y., Zhao, L., Zhang, T., Liu, Y.: Jailbreaking chatgpt via prompt engineering: An empirical study. arXiv preprint arXiv:2305.13860 (2023)
128. Liu, Y., Han, T., Ma, S., Zhang, J., Yang, Y., Tian, J., He, H., Li, A., He, M., Liu, Z., et al.: Summary of ChatGPT-related research and perspective towards the future of large language models. Meta-Radiology p. 100017 (2023)
129. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11976–11986 (2022)
130. Lord, C., Rutter, M., DiLavore, P., Risi, S., Gotham, K., Bishop, S., et al.: Autism diagnostic observation schedule–2nd edition (ADOS-2). Los Angeles, CA: Western Psychological Corporation **284** (2012)
131. Lücking, A.: Modeling co-verbal gesture perception in type theory with records. In: Ganzha, M., Maciaszek, L., Paprzycki, M. (eds.) Proceedings of the 2016 Federated Conference on Computer Science and Information Systems. Annals of Computer Science and Information Systems, vol. 8, pp. 383–392. IEEE (09 2016). https://doi.org/10.15439/2016F83
132. Lücking, A.: Witness-loaded and witness-free demonstratives. In: Coniglio, M., Murphy, A., Schlachter, E., Veenstra, T. (eds.) Atypical Demonstratives. Syntax, Semantics and Pragmatics, pp. 255–284. No. 568 in Linguistische Arbeiten, De Gruyter, Berlin and Boston (2018)
133. Lücking, A., Bergmann, K., Hahn, F., Kopp, S., Rieser, H.: The Bielefeld speech and gesture alignment corpus (SaGA). In: Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality. pp. 92–98. LREC 2010, 7th International Conference for Language Resources and Evaluation, Malta (2010). https://doi.org/10.13140/2.1.4216.1922
134. Lücking, A., Ginzburg, J.: Leading voices: Dialogue semantics, cognitive science, and the polyphonic structure of multimodal interaction. Language and Cognition **15**(1), 148–172 (2023). https://doi.org/10.1017/langcog.2022.30
135. 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)

136. Lücking, A., Pfeiffer, T., Rieser, H.: Pointing and reference reconsidered. Journal of Pragmatics **77**, 56–79 (2015). https://doi.org/10.1016/j.pragma.2014.12.013
137. Lugaresi, C., Tang, J., Nash, H., McClanahan, C., Uboweja, E., Hays, M., Zhang, F., Chang, C.L., Guang Yong, M., Lee, J., Chang, W.T., Hua, W., Georg, M., Grundmann, M.: Mediapipe: A framework for building perception pipelines (2019)
138. Maerten, A.S., Soydaner, D.: From paintbrush to pixel: A review of deep neural networks in ai-generated art. arXiv 2302.10913 (2023)
139. Marcus, G., Southen, R.: Generative AI has a visual plagiarism problem. Experiments with midjourney and dall-e 3 show a copyright minefield. IEEE Spectrum, 06 Jan 2024 (2024)
140. 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 science and clinical practice. iScience **26**(4), 106348 (Apr 2023). https://doi.org/10.1016/j.isci.2023.106348, http://dx.doi.org/10.1016/j.isci.2023.106348
141. McNeill, D.: Hand and Mind – What Gestures Reveal about Thought. Chicago University Press, Chicago (1992)
142. Mehler, A., Hemati, W., Gleim, R., Baumartz, D.: VienNA: Auf dem Weg zu einer Infrastruktur für die verteilte interaktive evolutionäre Verarbeitung natürlicher Sprache. In: Lobin, H., Schneider, R., Witt, A. (eds.) Forschungsinfrastrukturen und digitale Informationssysteme in der germanistischen Sprachwissenschaft. vol. 6, pp. 149–176. De Gruyter, Berlin (2018)
143. 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)
144. Meskó, B.: The impact of multimodal large language models on health care's future. Journal of Medical Internet Research **25**, e52865 (2023)
145. Moktefi, A.: Diagrams as scientific instruments. in Andras Benedek & Agnes Veszelszki (eds.), Visual, Virtual, Veridical, series Visual Learning, vol. 7 (Nov 2017), https://www.academia.edu/33378854/Diagrams_as_scientific_instruments
146. 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
147. Moreno, R., Mayer, R.: Interactive multimodal learning environments: Special issue on interactive learning environments: Contemporary issues and trends. Educational psychology review **19**, 309–326 (2007)
148. Morris, M.R., Sohl-dickstein, J., Fiedel, N., Warkentin, T., Dafoe, A., Faust, A., Farabet, C., Legg, S.: Levels of AGI: Operationalizing progress on the path to AGI. arXiv 2311.02462 (2023)
149. Muhammad, G., Alshehri, F., Karray, F., El Saddik, A., Alsulaiman, M., Falk, T.H.: A comprehensive survey on multimodal medical signals fusion for smart healthcare systems. Information Fusion **76**, 355–375 (2021)
150. Mundy, P., Newell, L.: Attention, joint attention, and social cognition. Current directions in psychological science **16**(5), 269–274 (2007). https://doi.org/10.1111/j.1467-8721.2007.00518.x
151. Naert, L., Reverdy, C., Larboulette, C., Gibet, S.: Per channel automatic annotation of sign language motion capture data. In: Proceedings of the LREC2018 8th

Workshop on the Representation and Processing of Sign Languages: Involving the Language Community. p. 139–146. European Language Resources Association (ELRA), Miyazaki, Japan (2018), https://www.sign-lang.uni-hamburg.de/lrec/pub/18014.pdf, citation Key: Naert2018
152. Nagrani, A., Yang, S., Arnab, A., Jansen, A., Schmid, C., Sun, C.: Attention bottlenecks for multimodal fusion. Advances in Neural Information Processing Systems **34**, 14200–14213 (2021)
153. Nakatsu, R.T.: Diagrammatic Reasoning in AI: Decision-making and Problem-solving With Diagrams. John Wiley & Sons (Dec 2009)
154. Nilsson, J.F.: A cube of opposition for predicate logic. Logica Universalis **14**(1), 103–114 (Mar 2020). https://doi.org/10.1007/s11787-020-00244-3, https://doi.org/10.1007/s11787-020-00244-3
155. Nota, N., Trujillo, J.P., Holler, J.: Facial signals and social actions in multimodal face-to-face interaction. Brain Sciences **11**(8), 1017 (2021). https://doi.org/10.3390/brainsci11081017, https://www.mdpi.com/2076-3425/11/8/1017
156. Oberwelland, E., Schilbach, L., Barisic, I., Krall, S., Vogeley, K., Fink, G., Herpertz-Dahlmann, B., Konrad, K., **Schulte-Rüther**, t.: Look into my eyes: Investigating joint attention using interactive eye-tracking and fMRI in a developmental sample. NeuroImage **130**, 248–260 (2016). https://doi.org/10.1016/j.neuroimage.2016.02.026
157. OpenAI: ChatGPT (feb 06 version) [large language model]. https://chat.openai.com/chat (2023)
158. OpenAI: Gpt-4 technical report. arXiv 2303.08774 (2023)
159. Ostendorf, M., Ross, K.: A Multi-level Model for Recognition of Intonation Labels. In: Sagisaka, Y., Campbell, N., Higuchi, N. (eds.) Computing Prosody, pp. 291–308. Springer US, New York, NY (1997)
160. Paggio, P., Jongejan, B., Agirrezabal, M., Navarretta, C.: Detecting head movements in video-recorded dyadic conversations. In: Proceedings of the 20th International Conference on Multimodal Interaction: Adjunct. ICMI '18, Association for Computing Machinery (2018). https://doi.org/10.1145/3281151.3281152
161. Paulk, A.C., Kfir, Y., Khanna, A.R., Mustroph, M.L., Trautmann, E.M., Soper, D.J., Stavisky, S.D., Welkenhuysen, M., Dutta, B., Shenoy, K.V., Hochberg, L.R., Richardson, R.M., Williams, Z.M., Cash, S.S.: Large-scale neural recordings with single neuron resolution using neuropixels probes in human cortex. Nature Neuroscience **25**, 252–263 (2022). https://doi.org/10.1038/s41593-021-00997-0
162. Peng, R.D.: Reproducible research in computational science. Science **334**(6060), 1226–1227 (2011)
163. Perlman, M.: Debunking two myths against vocal origins of language. Interaction Studies **18**(3), 376–401 (Dec 2017). https://doi.org/10.1075/is.18.3.05per
164. Perniss, P.: Why we should study multimodal language. Frontiers in Psychology **9**, 1109 (2018). https://doi.org/10.3389/fpsyg.2018.01109
165. Poustka, L., Schulte-Rüther, M.: Autismus-Spektrum-Störungen bei Kindern und Jugendlichen. In: Fegert, J., Resch, F., Plener, P., Kaess, M., Döpfner, M., Konrad, K., Legenbauer, T. (eds.) Psychiatrie und Psychotherapie des Kindes- und Jugendalters, pp. 1–23. Springer Berlin Heidelberg, Berlin, Heidelberg (2022). https://doi.org/10.1007/978-3-662-49289-5_123-1
166. Pouw, W., Dixon, J.A.: Entrainment and Modulation of Gesture–Speech Synchrony Under Delayed Auditory Feedback. Cognitive Science **43**(3), e12721 (2019). https://doi.org/10.1111/cogs.12721, https://onlinelibrary.wiley.com/doi/abs/10.1111/cogs.12721

167. Pouw, W., Fuchs, S.: Origins of vocal-entangled gesture. Neuroscience & Biobehavioral Reviews **141**, 104836 (Oct 2022). https://doi.org/10.1016/j.neubiorev.2022.104836, https://www.sciencedirect.com/science/article/pii/S0149763422003256
168. 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)
169. Ray, P.P.: Chatgpt: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. Internet of Things and Cyber-Physical Systems (2023)
170. Reichenberger, A., Lemanski, J., Bhattacharjee, R.: The role of gestures in logic. Visual Communication (Upcoming)
171. 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 (Jan 2020). https://doi.org/10.3758/s13428-020-01350-2, http://dx.doi.org/10.3758/s13428-020-01350-2
172. Rohrer, P.L.: A temporal and pragmatic analysis of gesture-speech association. A corpus-based approach using the novel MultiModal MultiDimensional (M3D) labeling system. Ph.D. thesis, Nantes Université (2022)
173. Rohrer, P.L., Tütüncübasi, U., Vilà-Giménez, I., Florit-Pons, J., Esteve-Gibert, N., Ren, P., Shattuck-Hufnagel, S., Prieto, P.: The MultiModal MultiDimensional (M3D) labeling system. https://doi.org/10.17605/osf.io/ankdx (2023)
174. Rosenberg, A.: Classification of prosodic events using quantized contour modeling. In: Proceedings of HLT-NAACL. pp. 721–724 (2010)
175. Rosenberg, A., Hasegawa-Johnson, M.: Automatic Prosody Labelling and Assessment. In: Gussenhoven, C., Chen, A. (eds.) The Oxford Handbook of Language Prosody, pp. 645–656. Oxford University Press, Oxford (2020). https://doi.org/10.1093/oxfordhb/9780198832232.013.43
176. Rumpf, A.L., Kamp-Becker, I., Becker, K., Kauschke, C.: Narrative competence and internal state language of children with Asperger Syndrome and ADHD. Research in Developmental Disabilities **33**(5), 1395–1407 (Sep 2012). https://doi.org/10.1016/j.ridd.2012.03.007
177. Sacks, H., Schegloff, E.A., Jefferson, G.: A simplest systematics for the organization of turn-taking for conversation. Language **50**(4), 696 (Dec 1974). https://doi.org/10.2307/412243
178. Sadasivan, V.S., Kumar, A., Balasubramanian, S., Wang, W., Feizi, S.: Can ai-generated text be reliably detected? (2023)
179. Sadler, M., Regan, N.: Game Changer. New in Chess (2019)
180. Sallam, M.: Chatgpt utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. In: Healthcare. vol. 11, p. 887. MDPI (2023)
181. Salvagno, M., Taccone, F.S., Gerli, A.G., et al.: Can artificial intelligence help for scientific writing? Critical care **27**(1), 1–5 (2023)
182. Sankey, M., Birch, D., Gardiner, M.: The impact of multiple representations of content using multimedia on learning outcomes across learning styles and modal preferences. International Journal of Education and Development using ICT **7**(3), 18–35 (2011)
183. Saravia, E.: Prompt Engineering Guide. https://github.com/dair-ai/Prompt-Engineering-Guide (12 2022)

184. Schepens, J., Marx, N., Gagl, B.: Can we utilize large language models (llms) to generate useful linguistic corpora? a case study of the word frequency effect in young german readers (2023)
185. Schulte-Rüther, 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 (Jan 2023). https://doi.org/10.1111/jcpp.13650
186. Schulte-Rüther, M., Otte, E., Adigüzel, 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 (Feb 2017). https://doi.org/10.1002/aur.1654
187. Schweitzer, A.: Production and Perception of Prosodic Events-Evidence from Corpus-based Experiments. Ph.D. thesis, Universität Stuttgart, Stuttgart (2010), http://elib.uni-stuttgart.de/opus/volltexte/2011/6031/pdf/Dissertation_Schweitzer.pdf
188. Schweitzer, A., Möbius, B.: Experiments on automatic prosodic labeling. In: Proceedings of the 10th International Conference on Speech Communication and Technology. pp. 2515–2518. Brighton (2009). https://doi.org/10.21437/Interspeech.2009-663
189. Shimojima, A.: Operational constraints in diagrammatic reasoning. In: Allwein, G., Barwise, J. (eds.) Logical Reasoning with Diagrams. Oxford University Press (1996)
190. Shimojima, A.: Semantic properties of diagrams and their cognitive potentials. CSLI Publications, Stanford, California (2015)
191. Shin, S.J.: The logical status of diagrams. Cambridge University Press, Cambridge (1995). https://doi.org/10.1017/CBO9780511574696, https://www.cambridge.org/core/books/logical-status-of-diagrams/27130C396E0899C90BC632B4C7617E2B
192. Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., Chen, Y., Lillicrap, T., Hui, F., Sifre, L., van den Driessche, G., Graepel, T., Hassabis, D.: Mastering the game of Go without human knowledge. Nature **550**, 354–359 (2017). https://doi.org/10.1038/nature24270
193. van der Sluis, I.: Multimodal Reference. Studies in Automatic Generation of Multimodal Referring Expressions. Uitgevershuis BuG, Groningen, NL (2005), phD thesis, Univ. van Tilburg
194. Stapleton, G., Jamnik, M., Shimojima, A.: What makes an effective representation of information: a formal account of observational advantages. Journal of Logic, Language and Information **26**(2), 143–177 (2017). https://doi.org/10.1007/s10849-017-9250-6
195. Stöver, T., Plontke, S.K., Guntinas-Lichius, O., Welkoborsky, H.J., Zahnert, T., Delank, K.W., Deitmer, T., Esser, D., Dietz, A., Wienke, A., Loth, A., Dazert, S.: Structure and establishment of the German Cochlear Implant Registry (DCIR). HNO **71**(Suppl 1), 82–92 (2023)
196. Sun, C., Shrivastava, A., Singh, S., Gupta, A.: Revisiting unreasonable effectiveness of data in deep learning era. In: Proceedings of the IEEE international conference on computer vision. pp. 843–852 (2017)
197. Syrdal, A.K., McGory, J.: Inter-transcriber reliability of ToBI prosodic labeling. In: 6th International Conference on Spoken Language Processing (ICSLP 2000).

vol. vol.3, pp. 235–238 (2000), https://www.isca-speech.org/archive/icslp_2000/i00_3235.html
198. Taori, R., Gulrajani, I., Zhang, T., Dubois, Y., Li, X., Guestrin, C., Liang, P., Hashimoto, T.B.: Alpaca: A strong, replicable instruction-following model. Stanford Center for Research on Foundation Models. https://crfm. stanford. edu/2023/03/13/alpaca. html **3**(6), 7 (2023)
199. Team, G., Anil, R., Borgeaud, S., Wu, Y., Alayrac, J.B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A.M., Hauth, A., et al.: Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805 (2023)
200. Törnberg, P.: ChatGPT-4 outperforms experts and crowd workers in annotating political twitter messages with zero-shot learning. arXiv preprint arXiv:2304.06588 (2023)
201. Torres, M.J.R., Barwaldt, R.: Approaches for diagrams accessibility for blind people: a systematic review. In: 2019 IEEE Frontiers in Education Conference (FIE). pp. 1–7 (Oct 2019). https://doi.org/10.1109/FIE43999.2019.9028522, https://ieeexplore.ieee.org/document/9028522, iSSN: 2377-634X
202. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Ferrer, C.C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., Fuller, B., Gao, C., Goswami, V., Goyal, N., Hartshorn, A., Hosseini, S., Hou, R., Inan, H., Kardas, M., Kerkez, V., Khabsa, M., Kloumann, I., Korenev, A., Koura, P.S., Lachaux, M.A., Lavril, T., Lee, J., Liskovich, D., Lu, Y., Mao, Y., Martinet, X., Mihaylov, T., Mishra, P., Molybog, I., Nie, Y., Poulton, A., Reizenstein, J., Rungta, R., Saladi, K., Schelten, A., Silva, R., Smith, E.M., Subramanian, R., Tan, X.E., Tang, B., Taylor, R., Williams, A., Kuan, J.X., Xu, P., Yan, Z., Zarov, I., Zhang, Y., Fan, A., Kambadur, M., Narang, S., Rodriguez, A., Stojnic, R., Edunov, S., Scialom, T.: Llama 2: Open foundation and fine-tuned chat models. arXiv 2307.09288 (2023)
203. Trujillo, J.P., Holler, J.: Interactionally embedded gestalt principles of multimodal human communication. Perspectives on Psychological Science (2023). https://doi.org/10.1177/17456916221141422, online first
204. Tutton, M.: When and why the lexical Ground is a gestural Figure. Gesture **12**(3), 361–386 (2012). https://doi.org/10.1075/gest.12.3.04tut
205. Uesaka, Y., Manalo, E., Ichikawa, S.: What kinds of perceptions and daily learning behaviors promote students' use of diagrams in mathematics problem solving? Learning and Instruction **17**(3), 322–335 (2007)
206. Ungerer, F., Schmid, H.J.: An introduction to cognitive linguistics. Pearson, Harlow, 2 edn. (2006)
207. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
208. Wagner, P., Malisz, Z., Kopp, S.: Gesture and speech in interaction: An overview. Speech Communication **57**, 209–232 (Feb 2014). https://doi.org/10.1016/j.specom.2013.09.008, http://www.sciencedirect.com/science/article/pii/S0167639313001295
209. Wagner, P., Ćwiek, A., Samlowski, B.: Exploiting the speech-gesture link to capture fine-grained prosodic prominence impressions and listening strategies. Journal of Phonetics **76**, 100911 (Sep 2019). https://doi.org/10.1016/j.wocn.2019.07.001, http://www.sciencedirect.com/science/article/pii/S009544701830038X

210. Wang, D.Q., Feng, L.Y., Ye, J.G., Zou, J.G., Zheng, Y.F.: Accelerating the integration of chatgpt and other large-scale ai models into biomedical research and healthcare. MedComm–Future Medicine **2**(2), e43 (2023)
211. Watkins, R.: Guidance for researchers and peer-reviewers on the ethical use of large language models (llms) in scientific research workflows. AI and Ethics pp. 1–6 (2023)
212. Wei, X., Cui, X., Cheng, N., Wang, X., Zhang, X., Huang, S., Xie, P., Xu, J., Chen, Y., Zhang, M., et al.: Zero-shot information extraction via chatting with chatgpt. arXiv preprint arXiv:2302.10205 (2023)
213. Wicke, P.: Probing language models' gesture understanding for enhanced human-ai interaction. arXiv 2401.17858 (2024)
214. Wightman, C.W., Ostendorf, M.: Automatic labeling of prosodic patterns. IEEE Transactions on Speech and Audio Processing **2**(4), 469–481 (1994). https://doi.org/10.1109/89.326607
215. Winding, M., Pedigo, B.D., Barnes, C.L., Patsolic, H.G., Park, Y., Kazimiers, T., Fushiki, A., Andrade, I.V., Khandelwal, A., Valdes-Aleman, J., Li, F., Randel, N., Barsotti, E., Correia, A., Fetter, R.D., Hartenstein, V., Priebe, C.E., Vogelstein, J.T., Cardona, A., Zlatic, M.: The connectome of an insect brain. Science **379**(6636), eadd9330 (2023). https://doi.org/10.1126/science.add9330
216. Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., Sloetjes, H.: ELAN: A professional framework for multimodality research. In: In Proceedings of the 5th International Conference on Language Resources and Evaluation. pp. 1556–1559. LREC 2006 (2006)
217. World Health Organization (WHO): International classification of diseases, eleventh revision (icd-11). https://icd.who.int/browse11 (2019/2021)
218. Wu, S., Fei, H., Qu, L., Ji, W., Chua, T.S.: Next-gpt: Any-to-any multimodal llm. CoRR **abs/2309.05519** (2023)
219. Xu, K., Zhong, G., Deng, Z., Zhang, K., Huang, K.: Self-supervised generative learning for sequential data prediction. Applied Intelligence **53**, 20675–20689 (2023). https://doi.org/10.1007/s10489-023-04578-5
220. Yadlowsky, S., Doshi, L., Tripuraneni, N.: Pretraining data mixtures enable narrow model selection capabilities in transformer models. arXiv preprint arXiv:2311.00871 (2023)
221. Zhang, C., Bengio, S., Hardt, M., Recht, B., Vinyals, O.: Understanding deep learning (still) requires rethinking generalization. Communications of the ACM **64**(3), 107–115 (2021)
222. Zhang, Y., Gong, K., Zhang, K., Li, H., Qiao, Y., Ouyang, W., Yue, X.: Meta-transformer: A unified framework for multimodal learning. arXiv preprint arXiv:2307.10802 (2023)
223. Zhou, K., Zhu, Y., Chen, Z., Chen, W., Zhao, W.X., Chen, X., Lin, Y., Wen, J.R., Han, J.: Don't make your LLM an evaluation benchmark cheater. arXiv preprint arXiv:2311.01964 (2023)