frontiers | Frontiers in Artificial Intelligence

TYPE Editorial
PUBLISHED 27 June 2023
DOI 10.3389/frai.2023.1234920

OPEN ACCESS

EDITED AND REVIEWED BY
Shlomo Engelson Argamon,
Illinois Institute of Technology, United States

*CORRESPONDENCE
Alexander Mehler
✉ mehler@em.uni-frankfurt.de

RECEIVED 05 June 2023
ACCEPTED 09 June 2023
PUBLISHED 27 June 2023

CITATION
Mehler A, Lücking A and Dong T (2023) Editorial: Multimodal communication and multimodal computing.
Front. Artif. Intell. 6:1234920.
doi: 10.3389/frai.2023.1234920

Editorial: Multimodal communication and multimodal computing

Alexander Mehler[1]*, Andy Lücking[1,2] and Tiansi Dong[3]

[1]Text Technology Lab, Goethe-University Frankfurt, Frankfurt, Germany, [2]Laboratoire de Linguistique Formelle (LLF), Université Paris Cité, Paris, France, [3]Neurosymbolic Representation Learning Group, Fraunhofer IAIS, Sankt Augustin, Germany

KEYWORDS

human-object interactions (HOIs), multimodal learning and analytics, visual-linguistic interaction, text-image analysis, unified methodology

Editorial on the Research Topic
Multimodal communication and multimodal computing

After a successful and text-centered period, AI, computational linguistics, and natural language engineering need to face the "ecological niche" (Holler and Levinson, 2019) of natural language use: *face-to-face interaction*. A particular challenge of human processing in face-to-face interaction is that it is fed by information from the various sense modalities: it is *multimodal*. When talking to each other, we constantly observe and produce information on several channels, such as speech, facial expressions, hand-and-arm gestures, and head movements. To learn drive, we first learn theories about traffic rules in driving schools. After passing the examinations, we practice on the streets, accompanied by an expert sitting aside. We ask questions and follow instant instructions from this expert. These symbolic traffic rules and instant instructions shall be quickly and precisely grounded to the perceived scenes, with which the learner shall update and predict other cars behaviors quickly, then determine her/his own driving action to avoid potential dangers. As a consequence, multimodal communication needs to be *integrated* (in perception) or *distributed* (in production). This, however, characterizes multimodal computing in general (but see also Parcalabescu et al., 2021). Hence, AI, computational linguistics and natural language engineering that address multimodal communication in face-to-face interaction have to involve multimodal computing–giving rise to the next grand research challenge of those and related fields. This challenge applies to all computational areas which look beyond sentences and texts, ranging from interacting with virtual agents to the creation and exploitation of multimodal datasets for machine learning, as exemplified by the contributions in this Research Topic.

From this perspective, we face several interwoven challenges: On the one hand, AI approaches need to be informed about the principles of multimodal computing to avoid simply transferring the principles of Large Language Models to multimodal computing. On the other hand, it is important that more linguistically motivated approaches do not underestimate the computational reconstructability of multimodal representations. They might otherwise have to share experiences with parts of computational linguistics, given the success of models such as OpenAI's ChatGPT (cf. Wolfram, 2023), which confronted them with the realization that even higher-order linguistic annotations could be taken over by digital assistants and consequently render the corresponding linguistic modeling work obsolete. Again, the scientific focus on face-to-face communication seems to point to a

middle ground. This is because we are dealing with the processing of highly contextualized data whose semantics require recourse to semantic or psycholinguistic concepts such as utterance situation Schüz et al., situation models or mental models (Johnson-Laird, 2010; Ragni and Knauff, 2013; Alfred et al., 2020) or reference to concepts such as grounding (Harnad, 1990), for the automatic reconstruction of which there are not yet adequate computer-based approaches, certainly not on the basis of scenarios such as one-shot or few-shot learning, since the corresponding experiential content is not available as (annotated) mass data. The particular moment in which one finds oneself information-theoretically at this point can be formulated as follows: large domains of linguistic and multimodal interactions, if they provide a sufficient number of patterns for association learning, are well manageable with methods based on current neural networks. However, as soon as we go beyond such associative regularities and arrive at a kind of meaning constitution that includes the *about* of communicative interaction—when we are dealing, so to speak, with the alignment of immediate objects and interpretants in the sense of Peirce (1934) (cf. Gomes et al., 2007 for a reference to Peirce in AI)—we reach the limits of such models, which have by no means already been explored and which we believe we can identify once again in the area of face-to-face communication. It is obvious that AI models need to complement bottom-up approaches with top-down approaches that start from multimodal situation models grounded in face-to-face communication, or at least from the notion of discourse as put forward by Alikhani et al., an approach that finds its obvious extension in an approach more oriented to terms of social science (see, for example, Cheema et al.).

From another angle, AI applications are increasingly appearing in complex communication situations or action contexts as quasi-agentive fourth-generation interfaces (Floridi, 2014), which raises the question of their status with respect to the distinction between simulation, emulation, and realization (Pattee, 1989). Looking again at the driving example, the issue here is that AI applications are increasingly applied in real-world contexts, where their use is contextualized each time by corresponding multimodal real-world data, representing a potential grounding-relevant resource that can be re-used for fine-tuning such models or even grounding them. One could object that such an AI agent is nothing more than a simulation, which in principle cannot know anything about this its status. However, such simulations perform under real conditions in interaction with more and more humans in no longer simulatively closed systems [of agent(s) and environment(s)], and this can drive a technological development of these systems in terms of life-long learning, which can ultimately make them appear as *realizations* of *interaction partners*. But here, too, one can ask what the limits of this interaction are, even if it is multimodal. For it is something fundamentally different to process multimodally generated data than to experience it through independent production, of which the notion of telic affordance provides a vivid example, since it is based on people's habits of use, a kind of use that AI systems are mostly incapable of at present. Is it this kind of difference, such as being able to identify a telic affordance either through one's own use or merely by observing data left by uses of human agents, that constitutes one of the limitations implied above? Be that as it may, in their paper Henlein et al. explore the question of the learnability of affordances using vision-based AI models, an approach that we argue could also be interpreted as an example of measuring the implied limit(s).

The counter-scenario to agents interacting with us as artificial interactors in real-world environments is a completely virtualized scenario in which both human and artificial agents interact as avatars (see Chalmers, 2022). Here, conversely, it is the human who enters the sphere of simulation, so to speak, rather than the simulation that we encounter as a putative realization. The key research advantage of such settings is that the resulting multimodal data becomes largely amenable to direct digitization and thus automatic analysis. This concerns areas as diverse as speech data, data regarding interaction with objects, lip movement data, facial expression data, eye movement data, head movement data, manual gesture data, body movement data, and (social) space-related behavioral data, as well as (social) distance behavioral data (see Mehler et al., 2023 for a corresponding formal data model in the context of VR). Evidently, virtual worlds provide an excellent experimental environment for the study of artificial interaction. This is addressed in the work of Nunnemann et al.. It can be seen as an example of the study of grounding issues that directly affect the actors involved and thus relate to the issue of grounding interactions. This raises the broader question of how to advance semantic theories that can be experimentally falsified, as VR systems seem to fit into the paradigm of an Experimental Semiotics (Galantucci and Garrod, 2011) in exemplary way, a fit that could not have been foreseen even just a few years earlier. In other words: in VR, the research strands of face-to-face communication, dialogic communication (Galland et al.), multimodal information processing, grounding in interaction environments that may be equipped with artifacts of a wide variety of affordances, and 4th-order artificial interaction (Floridi, 2014) seem to come together in exemplary fashion, suggesting much further research in this direction in the future. The time is ripe for a fundamental expansion of the empirical base of linguistics and communication studies research that knows how to utilize the possibilities of AI-based systems experimentally for its research purposes, and conversely, for the acquisition of ever more extensive multimodal data for the situation-specific grounding of AI systems, which will ideally no longer rely solely on text windows and wordpiece or subword analogies (Song et al., 2021) (cf. the *Bag-of-Visual Words* approach of Bruni et al., 2014) to infer the putative underlying semantics from the associations shadowed in the character strings observable by means of these windows. At present, it is unclear how far this line of research is developed or to what extent other than the current greedy segmentation models or tokenizers are already emerging that can also identify multimodal ensembles as recurrent data units. Nevertheless, as in the case of transformers (Devlin et al., 2019), this line of research can point to a worthwhile direction for development.

A crucial part of the multimodal challenge is to address the question of how to assemble, let alone parse, multimodal representations. A successful multimodal system shall unify representations from different channels. The fundamental challenge is to merge the two complementary modals, namely, the neural modal and the symbolic modal, and be capable of solving problems from both perspectives (Dinsmore, 1992). Geometrical

structure is advocated as a potential cognitive representation apart from symbols or neural-networks (Gärdenfors, 2000). A recent geometric approach successfully unified large symbolic tree structures with pre-trained vector embedding precisely (Dong, 2021), and opens a new door to allow symbolic structures to have precise neural representation, and potentially remove the gap between neural modal and symbolic modal (Bechtel and Abrahamsen, 2002; Dong et al., 2022; Sun, 2023).

Multimodal representations can be compared to musical scores where the different "voices" co-occur and may (or not) be tied together by relevance (Lücking and Ginzburg, 2023) (see Mehler and Lücking, 2009 for an example and a formalization of such kinds of representations). In this respect, McNeill (1992) and Kendon (2004) have shown in seminal works that manual gesture and speech form unified messages, but without specifying systematic, computational means for analyzing multimodal utterances. Alikhani et al. argue in their contribution "*Image–text coherence and its implications for multimodal AI*" that the appropriate level for processing multimodal representations in AI is the level of *discourse*. By example of image–text pairs, they apply *coherence theory* to capture the structural, logical and purposeful relationships between images and their captions. Using a dataset of image–text coherence relations, the authors question whether simple coherence markers are accounted for in two pre-trained multimodal language models, CLIP (Radford et al., 2021) and ViLBERT (Lu et al., 2019). Alikhani et al. move on to use these results to critique and improve the architecture of machine learning models, and to develop coherence-based evaluations of multimodal AI systems.

Image–text relations are also investigated by Cheema et al.. The authors focus on the relation between images and texts in the setting of news. They propose directions for multimodal learning and analytics in social sciences. Taking a largely semiotic perspective, the authors bring together news value analysis of news media from both a production and reception perspective, and the multimodality of news articles in terms of image–text relations which go beyond (related to the coherence-driven approach by Alikhani et al.) mere captions. The framework is applied to a couple of examples and is intended to shape larger-scale machine learning applications in the context of multimodal media analysis, as exemplified by means of a number of potential uses cases.

Turning from two-dimensional pictures to objects within virtual reality, Henlein et al. present their research on Human-Object Interaction (HOI) and augment the HICO-DET dataset (Chao et al., 2018) to distinguish Gibsonian (Gibson, 1979, Chap. 8) affordances (actions to which objects "invite") and telic affordances (objects' conventionalized purposes) (Pustejovsky, 2013). They successfully train the computational model AffordanceUPT on their extended resource and show that is is able to distinguish intentional use from Gibsonian exploitation, even for new objects. Hence, Henlein et al. contribute to a better understanding of clustering of objects according to their action potentials, in particular a clustering between perceptual features and intention recognition.

(Virtual) Objects and characters are potential referents in human–human and human–computer interaction. Nunnemann et al. investigate "*The effects of referential gaze in spoken language comprehension: human speaker vs. virtual agent listener gaze*". Hence, they address multimodal computing at the interface of human and artificial communication: On the one hand, people are known to respond to virtual agent gaze (Ruhland et al., 2015). On the other hand, during referential processing eye movements to objects in joint visual scenes are closely time locked to referring words used to describes those scenes (Eberhard et al., 1995). Using eye-tracking methods Nunnemann et al. compared the influence of human speaker gaze to that of virtual agent listener gaze in sentence verification tasks. While they could replicate findings that participants draw on human speaker gaze, they do not rely on the gaze of the virtual agent. Thus, the study hints at important directions in the creation of and interaction with virtual agents, pointing out the influence of the communicative role of virtual agents (i.e., speaker vs. hearer) and potentially the need of a Theory of Mind (Krämer, 2005).

While gaze can be used for establishing reference (in particular in dangerous situations, see Hadjikhani et al., 2008), the most important linguistic devices for referring are verbal referring expressions. The form of these referring expressions is adapted to the utterance situation: Schüz et al. discuss the representation problem in the sub-field of Referring Expression Generation (REG), where expressions are depended on contexts. They provide a systematic review of a variety of visual contexts and approaches to REGs, and strongly argue for an integrated or unified perspective or methodology. The focus is on different input modalities and how they shape the information that is needed for successful reference (i.e., enable the addressee to single out the intended object), thereby complementing and going beyond established research on multimodal deictic output (e.g., Kranstedt et al., 2006; van der Sluis and Krahmer, 2007).

In conversation, interlocutors exhibit conversational strategies or styles (Tannen, 1981). Galland et al. explore communicative preferences in the context of human-computer interaction in terms of task-oriented and socially-oriented dialogue acts. By utilizing reinforcement learning, they train an artificial agent to adapt its strategy to meet the preference of a human user by combining task-oriented and socially-oriented dialog act. This is achieved by combining four components: an engagement estimator (mainly based on the user's non-verbal behavior), a topic manager (keeping track of the user's favorite topics), a conversational preferences estimator (estimates the user's task/social preference a each turn), and a dialog manager (selects the most appropriate turn according to the artificial agent's user model). Subjective experiments involving over 100 participants show a cross-modal influence: adapting to a user's preferred conversational strategy or style affects the human's perception and increases user engagement.

The Research Topic *Multimodal communication and multimodal computing* comprises six different contributions that highlight different areas and challenges of the interplay between communication and computing, as they have emerged not only due to the recent rapid development of AI methods. What unites these contributions is their common focus on multimodality, which, however, they treat from very different perspectives: be it in terms of text-image relations, the affordances detectable through images, the interaction between humans and artificial agents, or the specific status of referring expressions

in spoken language comprehension. From a methodological perspective, these approaches are interesting because they redirect the AI focus from Big Data to Small or even Tiny Data, massively emphasizing the situatedness of communication in its multiple multimodal manifestations. What we ultimately lack, however, is an approach that integrates these heterogeneous research directions and their underlying distributed data resources to ground a more comprehensive multimodal semantics in a final joint research effort by linguistics, computational linguistics, and computer science—before this will all be taken over by AI agents.

Author contributions

This Research Topic on *Multimodal communication and multimodal computing* was proposed by AM, AL, and TD. The editors worked collaboratively to decide which potential authors to invite and which papers were accepted or rejected. The single manuscripts were subject to review by the corresponding handling editor as well as peer reviewers. This editorial was drafted by AM, AL, and TD. All authors contributed to the article and approved the submitted version.

Funding

AL and AM contribution was partly supported by the German Research Foundation (DFG, project number 502018965) and TD contribution was partially supported by Federal Ministry of Education and Research of Germany as part of the Competence Center for Machine Learning ML2R (01IS18038C).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

Alfred, K. L., Connolly, A. C., Cetron, J. S., and Kraemer, D. J. M. (2020). Mental models use common neural spatial structure for spatial and abstract content. *Commun. Biol.* 3, 17. doi: 10.1038/s42003-019-0740-8

Bechtel, W., and Abrahamsen, A. (2002). *Connectionism and the Mind: Parallel Processing, Dynamics, and Evolution in Networks*. Hong Kong: Graphicraft Ltd.

Bruni, E., Tran, N.-K., and Baroni, M. (2014). Multimodal distributional semantics. *J. Artif. Intell. Res.* 49, 1–47. doi: 10.1613/jair.4135

Chalmers, D. J. (2022). *Reality+: Virtual Worlds and the Problems of Philosophy*. Allen Lane.

Chao, Y.-W., Liu, Y., Liu, X., Zeng, H., and Deng, J. (2018). "Learning to detect human-object interactions," in *2018 IEEE Winter Conference on Applications of Computer Vision, WACV* 381–389. doi: 10.1109/WACV.2018.00048

Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2019). "BERT: pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019* (Minneapolis, MN, USA) 4171–4186.

Dinsmore, J. (1992). "Thunder in the gap," in *The Symbolic and Connectionist Paradigms: Closing the Gap* (Erlbaum) 1–23.

Dong, T. (2021). *A Geometric Approach to the Unification of Symbolic Structures and Neural Networks*. Cham: Springer-Nature. doi: 10.1007/978-3-030-56275-5

Dong, T., Rettinger, A., Tang, J., Tversky, B., and van Harmelen, F. (2022). "Structure and Learning (Dagstuhl Seminar 21362)," in *Dagstuhl Reports* (Schloss Dagstuhl: Leibniz-Zentrum für Informatik) 11–34.

Eberhard, K. M., Spivey-Knowlton, M. J., Sedivy, J. C., and Tanenhaus, M. K. (1995). Eye movements as a window into real-time spoken language comprehension in natural contexts. *J. Psycholinguist. Res.* 24, 409–436. doi: 10.1007/BF02143160

Floridi, L. (2014). *The Fourth Revolution. How the Infosphere is Reshaping Human Reality*. Oxford: Oxford University Press.

Galantucci, B., and Garrod, S. (2011). Experimental semiotics: a review. *Front. Human Neurosci.* 5, 1–15. doi: 10.3389/fnhum.2011.00011

Gärdenfors, P. (2000). *Conceptual Spaces-The Geometry of Thought*. Cambridge, Massachusetts, USA: MIT Press. doi: 10.7551/mitpress/2076.001.0001

Gibson, J. J. (1979). *The Ecological Approach to Visual Perception*. Boston: Houghton Mifflin.

Gomes, A., Gudwin, R., Niño El-Hani, C., and Queiroz, J. (2007). Towards the emergence of meaning processes in computers from peircean semiotics. *Mind Soc.* 6, 173–187. doi: 10.1007/s11299-007-0031-9

Hadjikhani, N., Hoge, R., Snyder, J., and de Gelder, B. (2008). Pointing with the eyes: The role of gaze in communicating danger. *Brain Cogn.* 68, 1–8. doi: 10.1016/j.bandc.2008.01.008

Harnad, S. (1990). The symbol grounding problem. *Physica D.* 42, 335–346. doi: 10.1016/0167-2789(90)90087-6

Holler, J., and Levinson, S. C. (2019). Multimodal language processing in human communication. *Trends Cogn. Sci.* 23, 639–652. doi: 10.1016/j.tics.2019.05.006

Johnson-Laird, P. N. (2010). Mental models and human reasoning. *PNAS* 107, 18243–18250. doi: 10.1073/pnas.1012933107

Kendon, A. (2004). *Gesture: Visible Action as Utterance*. Cambridge, MA: Cambridge University Press. doi: 10.1017/CBO9780511807572

Krämer, N. C. (2005). "Theory of mind as a theoretical prerequisite to model communication with virtual humans," in *Modeling Communication with Robots and Virtual Humans*, eds. I., Wachsmuth, and G., Knoblich (Berlin and Heidelberg: Springer) 222–240. doi: 10.1007/978-3-540-79037-2_12

Kranstedt, A., Lücking, A., Pfeiffer, T., Rieser, H., and Wachsmuth, I. (2006). "Deictic object reference in task-oriented dialogue," in *Situated Communication*, eds. G., Rickheit, and I., Wachsmuth (Berlin: Mouton de Gruyter) 155–207. doi: 10.1515/9783110197747.155

Lu, J., Batra, D., Parikh, D., and Lee, S. (2019). "ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," in *Advances in Neural Information Processing Systems*, eds. H., Wallach, H., Larochelle, A., Beygelzimer, F., d'Alché-Buc, E., Fox, and R., Garnett (Red Hook, NY: Curran Associates, Inc.).

Lücking, A., and Ginzburg, J. (2023). Leading voices: Dialogue semantics, cognitive science, and the polyphonic structure of multimodal interaction. *Langu. Cogn.* 15, 148–172. doi: 10.1017/langcog.2022.30

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

Mehler, A., Bagci, M., Henlein, A., Abrami, G., Spiekermann, C., Schrottenbacher, P., et al. (2023). "A multimodal data model for simulation-based learning with Va.Si.Li-Lab," in *Proceedings of HCI International 2023, Lecture Notes in Computer Science* (Springer).

Mehler, A., and Lücking, A. (2009). "A structural model of semiotic alignment: The classification of multimodal ensembles as a novel machine learning task," in *AFRICON 2009* (IEEE) 1–6. doi: 10.1109/AFRCON.2009.5308098

Parcalabescu, L., Trost, N., and Frank, A. (2021). "What is multimodality?" in *Proceedings of the 1st Workshop on Multimodal Semantic Representations, MMSR* (Groningen, Netherlands: Association for Computational Linguistics).

Pattee, H. H. (1989). "Simulations, realizations, and theories of life," in *Artificial Life. SFI Studies in the Sciences of Complexity*, eds. C. G., Langton (Boston: Addison-Wesley) 63–77.

Peirce, C. S. (1934). *Collected Papers: Pragmatism and Pragmaticism, volume 5*. Cambridge MA: Harvard University Press.

Pustejovsky, J. (2013). "Dynamic event structure and habitat theory," in *Proceedings of the 6th International Conference on Generative Approaches to the Lexicon, GL2013* (Pisa, Italy: Association for Computational Linguistics) 1–10.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., et al. (2021). "Learning transferable visual models from natural language supervision," in *Proceedings of the 38th International Conference on Machine Learning*, M., Meila, and T., Zhang (PMLR) 8748–8763.

Ragni, M., and Knauff, M. (2013). A theory and a computational model of spatial reasoning with preferred mental models. *Psychol. Rev.* 120, 561–588. doi: 10.1037/a0032460

Ruhland, K., Peters, C. E., Andrist, S., Badler, J. B., Badler, N. I., Gleicher, M., et al. (2015). A review of eye gaze in virtual agents, social robotics and hci: Behaviour generation, user interaction and perception. *Comput. Graph. Forum* 34, 299–326. doi: 10.1111/cgf.12603

Song, X., Salcianu, A., Song, Y., Dopson, D., and Zhou, D. (2021). "Fast WordPiece tokenization," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (Punta Cana, Dominican Republic. Association for Computational Linguistics) 2089–2103. doi: 10.18653/v1/2021.emnlp-main.160

Sun, R. (2023). *The Cambridge Handbook of Computational Cognitive Sciences*. 2 edition Cambridge: Cambridge University Press. doi: 10.1017/9781108755610

Tannen, D. (1981). Indirectness in discourse: Ethnicity as conversational style. *Disc. Process.* 4, 221–238. doi: 10.1080/01638538109544517

van der Sluis, I., and Krahmer, E. (2007). Generating multimodal references. *Disc. Process.* 44, 145–174. doi: 10.1080/01638530701600755

Wolfram, S. (2023). *What Is ChatGPT Doing . . . and Why Does It Work*? Champaign, IL: Wolfram Media, Inc.