Insights into the black box of multimodal meaning-making: investigating the reception of multimodality empirically

HANS-JUERGEN BUCHER
University of Trier, Germany

ABSTRACT

The intention of this article is to clarify the concept of multimodal meaning making from a recipient's perspective on the basis of empirical data. The data comes from various reception studies in which different methods like eye tracking, knowledge tests, re-narrations, guided interviews or questionnaires were combined to investigate the appropriation of multimodal artifacts such as newspapers, science videos, comics, social media postings, or commercials. The article is structured in such a way that examples and results from various reception studies on newspaper pages, commercials, comics and science videos are presented in order to introduce concepts like attention, non-linearity, intermodality, modal density and modal coherence for clarifying the process of meaning-making. Investigating meaning-making empirically should not only fill a research gap in the mostly product- and production-oriented approaches to multimodality, but also help to develop a satisfactory theory of multimodality. Eye tracking in combination with verbal data reveals that multimodal artefacts have a non-linear structure – even in linear films – that is grasped step by step by the addressees, analogous to a hypertext. The concept of non-linearity opens a path to define the relevance structure of a multimodal arrangement: the relevance of modal elements is the outcome of its mutual contextualization. With this perspective on multimodality, the central unit of analysis is neither the sign-maker, the sign itself or its materiality but the interaction between communicator and addressees, for which modal resources are used. Intermodal relations are analysed not as constellations of signs, but as complex discursive actions addressed with a special intention to an audience or a person. The theoretical result can be summarized as follows: combining action theory with cooperation theory provides the conceptual tools to analyse meaning-making as a conversational implicature of a discursive action.

KEYWORDS

action theory • attention • cooperation theory • meaning-making • multimodality • reception study

1. INTRODUCTION: THE IMPLICIT PRODUCTION BIAS IN MULTIMODALITY RESEARCH

The concept of meaning-making is one of the most central concepts in multimodality theories as it is used as a starting point for clarifying other relevant concepts like mode, social semiotics, sign, rhetorical resources, multimodal coherence or affordance. Although the term 'meaning-making' can in principle be applied both to the production of multimodal utterances and to their reception, it is used almost exclusively in a production-oriented way. 'The emphasis is on the sign-maker and their situated use of modal resources' (Jewitt, 2009: 30). In this sense, the concept is used to explain the constructivist dimension of multimodality: modes are defined as 'meaning-making resources, which people use to communicate and represent' and 'people orchestrate meaning through their selection and configuration of modes' (p. 15). The dominance of the production-centred perspective becomes even more obvious in Gunther Kress's concept of sign-making:

> Signs are *made* rather than used. The focus on sign-*making* rather than sign *use* is one of several features which distinguishes social-semiotic theory from other forms of semiotics. In a social semiotic account of meaning, individuals . . . are agentive and generative in sign-making and communication. (Kress, 2010: 54)

For critical remarks on this account of meaning and especially on the relation of arbitrariness and intentionality, see Bucher (2017: 105–107, 117–119).

The 'material turn' in multimodality research, propagated for example by John Bateman (2021), supports the bias towards a production-oriented approach to multimodality. Although he concludes that 'the characterization of materiality . . . is not that of physics but rather rests on active perceivers' embodied engagement with materials for semiotic purposes' (p. 47), his approach implicates that these active perceivers are producers who screen modal material for use and not recipients who are confronted with its multimodal outcome. As a consequence, case studies following this 'material turn' are essentially concerned with reconstructing the producer's principles of constructing multimodal messages, leaving aside the comprehension of addressees (e.g. Pflaeging and Stöckl, 2021; Wildfeuer and Bateman, 2016). Like sign-centred approaches, the material-based approaches are more concerned with an inventory of modes than with the question of how modes or 'material traces' contribute to intermodal relations which in the end constitute

the global meaning of a multimodal arrangement. Both approaches have led to a wide range of descriptive categories and thus also have contributed to raising fruitful questions for reception studies (see Castaldi, 2024). But they leave open the question of how to counter the danger of isolating mode from discourse. Accordingly, in his critique of Kress's affordance-based concept of mode, Bezemer (2024: 507) warns that a project 'to map possibilities and limitations of modes, [renders them] untenable, unless they are considered in the context of a concrete observed communicative event'. On the basis of reception studies, the article therefore argues for a program to analyse intermodal relations in the context of 'communicative events' applying a *dynamic action-theoretic and structural approach* (Fritz, 2022). The intention of the article is to integrate audience research in multimodal studies (Castaldi, 2021) by providing a toolbox for describing 'inter-modal coherence' (Stöckl and Bateman, 2022: 2) as meaningful relations between modal building blocks.

2. FROM PRODUCTION TO RECEPTION

Producer- or material-centred approaches to multimodality seem to imply that each multimodal utterance 'has' the meaning which authors intend and which all recipients understand somehow as intended by them. Of course, we know from all different forms of communication – be it a face-to-face conversation or media discourse – that this is rather implausible. It therefore makes sense to supplement the hitherto dominant perspectives with an addressee- or reception-oriented perspective, especially since both can be compatible from an epistemological point of view (Castaldi, 2021, 2024). This article puts the recipients at its centre, asking the question: How do recipients integrate the different modalities like text, picture, sound and design into a coherent meaning? This change of perspective not only fills a research gap but can also help to develop a viable theory of multimodality. To explain how and why reception studies can help deploy a theory of multimodality, one can go back to a technique long since applied in ordinary language philosophy by authors such as Austin (1956/1957), Alston (1964), or Grice (1968, 1975), and which can be labelled 'dialogical': one can reconstruct the meaning of a verbal expression by analysing the replies and reactions to that expression (Bucher, 2017).

As this article focuses on multimodality in media discourses, there are two different types of replies one can investigate: first, natural replies like letters to the editor or user comments in internet-based communication, which are investigated in 'engagement studies' (Castaldi, 2021); and, second, experimental replies elicited in a laboratory setting, which are the objects of 'experimental studies'. The latter offers the possibility of manipulating the reception process intentionally so that the special features of multimodal 'meaning-making' become visible. Besides that, the laboratory facilitates the reactions of recipients to different multimodal stimuli such as newspapers, magazines, online media, videos, video games, or scientific presentations with different

combinations of methods: eye tracking, thinking aloud during and after the reception process, re-narrations, knowledge tests, interviews, or questionnaires. These experimentally elicited reactions are normally used as data to infer media effects. In this article, these data are used methodologically as communication feedback to the different stimuli, which allow reconstructing the meaning-making process of the recipient. Hence, the documented reactions of the participants indicate how they perceive the multimodal ensemble, what they select as relevant, which modes they combine in which order, and so on. Experimental methods, so to speak, allow reconstructing the interaction between a recipient and a multimodal stimulus. They help, on the one hand, to understand the process of multimodal meaning-making and, on the other, to reconstruct the structure of multimodal discourse which is reflected in the layers and aspects of the reception results (Boy et al., 2020; Bucher and Niemann, 2012; Bucher and Schumacher, 2006). In principle, one can distinguish three types of reception data which are elicited by three types of experimental methods:

1. *Eye tracking* data, which indicate how *attention* is allocated to the different modal elements of a multimodal artefact.
2. *Knowledge tests* – recall tests such as re-narration and recognition tests such as multiple choice – in order to measure *knowledge transfer* and degrees of *comprehension.*
3. *Verbal data* from *guided interviews*, *questionnaires* or *thinking aloud* comments to investigate *ratings*, *assessments* or *emotional reactions*.

Based on the so-called eye–mind hypothesis (Just and Carpenter, 1976: 441), eye movements serve as indicators for cognitive processes and provide data beyond self-reporting methods. Thus, eye tracking data form the bridge between effects investigated with methods in points 2 and 3. By localizing attention, eye movements can indicate which multimodal aspects can be attributed to reception effects such as knowledge acquisition, rankings, evaluations or variants of understandings. Eye tracking data can be evaluated with two different methods that use various measures: a *fixation-related evaluation* according to criteria such as duration, frequency, localization, sequence, or so-called revisits of Areas of Interest (AOI), provides information about *which elements* (AOIs) were viewed for how long, how often, and when. A *process-related evaluation*, based on measures of scan paths such as their length, similarity, predictability, etc. provides information about the *sequence of AOIs* considered, the dynamics and the course of the reception (for more details, see Bucher et al., 2022: 84–94; Holmqvist et al., 2011).

The richness of the three types of data – gaze data, knowledge acquisition and verbal data – help to understand how recipients integrate the different modes and acquire a coherent understanding of the multimodal discourse. Furthermore, these data contribute to investigating the basic problem that any

theory of multimodality has to confront if it intends to clarify the meaning-making process. In regard to the *problem of compositionality*, by comparing reception data it is possible to reconstruct how the individual modes contribute to the overall meaning of a discourse and how they interact. Previous multimodality research has approached the problem of compositionality with concepts like 'intersemiosis' (O'Halloran, 2008: 470), 'intersemiotic texture' (Liu and O'Halloran, 2009: 367), 'semantic multiplication' (Fei, 2004: 239), or 'modal interrelation' (Kress, 2010: 165). Empirical research could help to clarify these terms beyond their metaphorical use and to focus on the concept of modal coherence (Bateman, 2014).

3. MULTIMODALITY IN THE LIGHT OF RECEPTION THEORIES

Applying the term 'meaning-making' to the recipient side, this key concept in multimodality research can be connected to a family of reception theories in media research that turn away from the question of media effects on a passive audience and focus on the *active acquisition* of the addressees. This family of theories is the result of a paradigm shift in media research, which Elihu Katz and David Foulkes propagated as early as in 1962. They suggest that the focus should not be on the media but on the recipients. Instead of asking 'What do the media do to people?', the question should be 'What do people do with the media?' (Katz and Foulkes, 1962: 378). The performative core of the question of what people *do with the media* has been taken up in media and communication research by social constructivist approaches (Schönbach, 2017), by Cultural Studies (Ang, 1996; Hall, 1980) and by action-theoretical reception research (Bucher et al., 2012; Renckstorf et al., 2004). What these approaches to reception research have in common is that they avoid a media-centric perspective, following the basis of classical theories of media effects, according to which media discourses are stimuli, information or content which carry their effect within themselves and determine their reception. In line with this change of perspective, Sonia Livingstone (2006: 343) concludes 'that research should investigate the activities of actual audiences in order to know how they interpret programs in everyday contexts.' The common idea behind this family of reception theories is the assumption that reception can be modelled as an *interaction* between a media stimulus and its recipients: the outcome is the result of the interplay between qualities of the media discourse on the one hand – such as its multimodal composition or its genre – and attributes of the recipients on the other hand – such as their pre-knowledge, their interests, their media literacy or the particular situation of reception. Interactive theories of reception therefore integrate bottom-up 'realistic' theories and top-down 'idealistic' theories of reception. The intention of all these approaches is to detect systematic correlations between characteristics of a media outlet and layers of reception such as knowledge

acquisition, distribution of attention or emotions (Bucher and Schumacher, 2012; Castaldi, 2021).

The concept of an interactive approach to reception fits very well with the concept of meaning-making in multimodality theories. Both agree that the meaning of a semantic unit – be it a picture, a video, an Instagram-post or a newspaper article – is not an inherent quality of these objects but the result of an active achievement. Therefore we can go one step further: using the concepts developed within the framework of interactive approaches to reception, we are able to clarify and concretize the vague notion of 'meaning-making' and its near relatives such as 'intersemiosis', 'semantic multiplication', or 'modal interrelation'. But the clarification process also works in the opposite direction. Theories of multimodality help to concretize and differentiate vague terms such as media, media content or media stimuli. This clarification is a prerequisite for designing a media-specific design of empirical reception studies that is suitable for capturing the special features of the multimodal understanding of media discourse.

In contrast to a conversation or dialogue, the term 'interaction' is to be understood here in the sense of a counterfactually assumed interaction, an idea that is at home in various approaches to reading research, audience research or in perception studies (Bucher and Schumacher, 2012). The idea was picked up in the context of interactive digital media for analysing the particular structure of these forms of communication:

> Interactivity is not a characteristic of the medium. It is a process-related construct about communication. It is the extent to which messages in a sequence relate to each other, and especially the extent to which later messages recount the relatedness of earlier messages. (Rafaeli and Sudweeks, 1997)

Within this perspective, interactivity is defined as a kind of 'as-if interactivity': users who are reading a newspaper, surfing a website or a social media platform or watching a video imply that they interact with the particular object as if it were a real partner in a face-to-face interaction. The thinking-aloud procedure, in particular, has provided a lot of evidence for this kind of 'as-if' interaction. Although the concept of interactivity can benefit from the concept of interaction – for example, in terms of the rule-based relationship between elements of a discourse – they belong to two different dimensions: the latter to the dimension of discourse and 'interactivity' to the dimension of reception (Bucher, 2004; Bucher and Schumacher, 2006).

Thus, recursivity is introduced as a criterion for interactivity, according to which every sequence of meaningful utterances and every exchange of information is to be seen in the context of previous communication. Therefore, one can conclude that meaning-making has a recursive structure in principle. In the case of multimodal objects, this means that the different modes are not

meaningful in themselves but only in contextualizing each other. Thus, for empirical reception research it is not enough to simply record and analyse the results or effects of media use, but also the process of acquisition. Although knowledge transfer, emotional reactions or evaluations allow conclusions to be drawn about certain triggers of these effects, they do not allow any conclusions to be drawn about its appropriation and processing. For investigating the recursive process of interaction with a multimedia artefact in order to identify systematic relations between features of a stimulus and effects on the part of the addressee, a method is necessary that is both process oriented and sensitive to the multimodal complexity of media discourse. Combining the minimal invasive method of eye tracking with verbal data from knowledge tests, guided interviews and questionnaires are one way to fulfill this condition. The following sections will present some empirical studies which apply these methodological designs with the intention of enlightening basic problems of multimodality research: the problem of meaning-making and the problem of intermodal relations. Therefore, the following two sections will focus on two questions. First: How does eye tracking data contribute to the recording of attention distribution and, consequently, to meaning-making? And, second: How can the contribution of a mode to the overarching meaning of a multimodal artefact be determined empirically?

4. MULTIMODALITY AND ATTENTION

4.1. Eye tracking: a window onto multimodal meaning-making

The modal elements of multimodal artefacts can be presented principally in two ways: *simultaneously*, as in the case of a combination of image and spoken language or sound, or *serially*, as in the case of a text and an associated image. In many multimodal discourses, the two types of presentation are combined – for example, in videos or animations when we hear the spoken language and simultaneously see the visuals, both elements of the discourse unfold sequentially in time. Regardless of whether multimodal orchestrations are organized spatially, temporally, or *temporally–spatially*, they require a selective reception strategy from the recipients since not all modal elements can be processed simultaneously. According to the Cognitive Load Theory (Neumann, 1987; Paas and Sweller, 2014) in all cases the process of reception demands an organization of the *allocation of attention* as the capacity of the cognitive system is not able to process all the different stimuli at the same time. Research on attention assumes that this term encompasses a variety of mechanisms of the human processing system, which have in common that 'they serve to regulate information and behavior where existing skills and routines alone cannot achieve this' and which are therefore 'closely linked to the processing of the new' (Neumann, 1996: 561, my translation). Accordingly, across the various theories, attention is ascribed two basic functions: the function of

selection, which ensures that relevant aspects of a stimulus are chosen, and the function of *integration*, which integrates the selected aspects with each other (Wolfe, 2015). Attention is a bi-directional concept, with an intentional top-down and an unintentional bottom-up direction: on the one hand, attention is attracted by salient features of the perceived object; on the other hand, it is driven by intentions or habits of the recipient (Bucher and Schumacher, 2006; Neumann, 1987). The bi-directional character of attention corresponds with an interactional theory of reception which also focuses on the dynamic interplay of stimuli-driven and intentional aspects of reception processes.

Taking this background for granted, attention could be a key concept for reconstructing the architecture of multimodality and the relevance structure of the modal elements: tracking attention processes reveal how recipients select what is relevant and integrate single modes in the process of meaning-making. Attention as well as selection processes are not directly open to observation and are therefore inferences of the observer. Only by observing several successive activities can we determine to what aspect one pays attention and to what extent, or what has been selected as relevant. Eye movements are one indicator of these activities as they depend on attentional processes (Boeriis and Holsanova, 2012; Holmqvist et al., 2011; Holsanova, 2014; Salvucci, 1999; Yarbus, 1967). They are usually not controlled consciously and can also be non-intentional, which is why we can categorize them as *behavioural indicators* (comparable to mimic and gesture) in contrast to *proactive indicators* such as actions (cursor moves, scrolling, clicking, turning over a page), utterances (e.g. comments from the thinking-aloud method), or strategies of problem-solving (back navigation, repeated reading, asking questions). Thus, data recorded by eye tracking devices can be used to obtain insight into the characteristics of the process of allocating attention – for example, its intensity, temporal structure or its selectivity. An eye-tracker collects data that allow conclusions about fixations and saccades of the eye on a given stimulus. *Fixations* are periods when the eye is relatively immobile focused on an AOI and during which conscious perception happens. They indicate the area where attention is likely to be allocated (Holmqvist et al., 2011; Rayner, 1995). *Saccades* occur when the eye jumps from one fixated area to the next (Stark and Ellis, 1981). Therefore eye movements can serve as a window to multimodal meaning-making as they indicate to which elements of a stimulus attention is allocated and which are ignored, and also in which sequential order stimuli are perceived. The *dwell time* on a particular part of the stimulus indicates the recipients' level of attention and interest, and the *navigation paths* reveal the process of integrating the single elements in an overall meaning. Eye-tracking data also reveal the quality of reception: one can deduce from the pattern of the eye movement if the visual or the textual element is read in depth or just cursorily scanned (Bucher and Schumacher, 2006; Duchowski, 2007). Aggregating data from a group of recipients allows conclusions to be drawn about the extent to which a multimodal arrangement

can direct attention, which could be helpful if it comes to its optimization. Compared to traditional retrospective methods like interviews or questionnaires, eye-tracking data are more reliable as they are obtained directly during the reception process and thus provide immediate insight into the interaction between the stimulus and the recipient. However, knowing at which object someone is looking does not mean knowing what he or she actually sees. Therefore, it is important to combine eye tracking with several verbal methods like re-narration, knowledge tests, concept mapping or interviews in order to gain additional information from the recipients. Data from this triangulation of methods provide manifold traces of visual attention and thus indicate how recipients handle the problem of compositionality, selecting what is relevant and integrating the selected elements into a coherent interpretation. Therefore, these data allow us to understand multimodality from a recipient's perspective and identify some essential requirements on a theory of multimodality, which is demonstrated with the following examples from an eye tracking study on newspaper reading.

The example in Figure 1 is a front page of the German newspaper *Die Tageszeitung* whose lead story covers Barack Obama's victory in the primaries in 2008 and his nomination as candidate for the presidency of the Democratic Party. The lead story carries the headline 'Onkel Baracks Hütte' ('Uncle Barack's Cabin') and contains a photograph of the White House in Washington, DC.

Although this front page is a rather tricky example of a multimodal arrangement, it allows some fundamental insights into multimodal meaning-making. It is obvious that the meaning of the multimodal arrangement composed of headline text, picture and lead text is more than the sum of the individual elements. Neither the picture nor the headline alone has enough semantic potential to uncover the full meaning of this front-page story. And even if all three elements are taken into account, getting the full meaning is not at all self-evident: on account of lack of prior knowledge, one may not understand the allusion to the famous novel 'Uncle Tom's Cabin' by Harriett Beecher Stowe, or identify the building as the White House in Washington, DC. Without this knowledge, which is not 'in' the multimodal arrangement of the newspaper page, one misses the point of this image–text combination.

The eye-tracking data documented in Figure 2 mirrors the strategy of making sense with the help of several modes. The different elements – headline, lead text and picture – are not perceived sequentially one after the other but *alternately* and *repeatedly*: the recipient attempts to get an overall meaning by taking each element as a context for the others. Via this process of alternately interpreting the elements, they finally achieve an overall meaning or what the recipient considers as such.

The *zigzag line* of eye movements between different semiotic elements is quite typical of interpreting multimodal stimuli and has been discovered in eye-tracking studies of all different kinds: reading newspapers (Holsanova

Figure 1. Front page of the German newspaper *die tageszeitung*, 5 June 2008: 'Onkel Baracks Hütte' ('Uncle Barack's Cabin').

et al., 2006, 2008) or comics (Bucher and Boy, 2018) as well as in surfing websites (Bucher and Schumacher, 2006; Schumacher, 2009) or watching videos (Bucher, 2012) in science communication (Boy et al., 2020) and

Figure 2. Scan path of a recipient reading the front page of the German newspaper *die tageszeitung*, 5 June 2008: 'Onkel Baracks Hütte' (see Figure 1).

even in real-world scenarios like visiting a museum or following a scientific presentation (Bucher and Niemann, 2012). So, *recursive* reception and an *interactive* acquirement of the different modal elements seem to be a characteristic pattern of understanding multimodal discourses: the interpretation

of multimodal arrangements is carried out step by step, with each step building on the previous one.

4.2. From empiricism to theory: what reception data reveal about meaning-making

The eye tracking data allow us to infer some conclusions on multimodal meaning-making:

1. Multimodality is a *dynamic phenomenon*: multimodal meaning-making is not a one-step action; understanding multimodal discourse does not happen instantaneously but step by step and recursively. Therefore, the theoretical frame of multimodal understanding should not be a theory based on media effects, but a theory based on interactive reception that brings the interpretive activities of the recipient and the characteristics of the object of reception into a systematic interrelation (Bucher and Schumacher, 2012). According to this approach, the concept of 'multiplication' as a metaphor for multimodal meaning-making is somehow misleading: multiplication implies that the elements that are multiplicated are already identified and fixed. Otherwise, we could not calculate the product. But this condition is not given in multimodal communication: with each step in the reception process, the meaning of the single element is expanded and, therefore, the context for the other elements is changed. Only due to this interpretive dynamic does the repeated fixation of the *same* element make sense. The process of appropriation documented by eye tracking data also suggests that 'semiotic resources' are not only 'a set of abstract, a-material options' (Castaldi, 2024: 5) but have a relational character: the meaning of the White House picture in the context of the newspaper title page in Figure 1 is different to the meaning it would have in a tourist guide for Washington – and, in any case, the interpretation depends on prior knowledge as demonstrated by interviews with test interviewees about this front page.
2. Multimodality inherently needs the notion of *non-linearity*: multimodal forms of communication are not just ensembles of several modes but they exhibit a non-linear structure for the recipient, be it a picture, a newspaper page, or a presentation: the recipients have to decide which elements are relevant and in which order they should be perceived. This problem of selectivity (Bucher and Schumacher, 2006) is true for films: besides their temporal structure, films are also spatially organized as 'composition of the shots' (Van Leeuwen, 2005: 181) or within 'phases' which constitute a sequence (Baldry and Thibault, 2005: 47). The fact that different viewers can see different things in the same film (as can the same viewer watching it twice) indicates that the selection of the relevant elements is an integral part of understanding and interpreting films (for more details, see the next section). Therefore, one can conclude from reception data that multimodal ensembles have a hypertextual or hypermodal structure.

3. Explaining multimodality requires a theory of *discourse and cooperation*: it is obvious that the meaning of the lead story in Figure 1 cannot be deduced by adding the meaning of the single elements – signs – because neither the picture nor the headline, nor the lead text unfold their full meaning in isolation but only in the context of each other. Therefore, the meaning of this multimodal ensemble is not *explicit* and somehow deducible directly from some salient signs of the single textual or pictorial elements, but *implicit* and only accessible inferentially. Apart from the mutual knowledge concerning the novel *Uncle Tom's Cabin* and the photograph depicting the White House, recipients have to presuppose that the newspaper journalists are following the cooperative principle and the conversational maxims, otherwise the headline 'Uncle Barack's Cabin' sounds somehow irrelevant and uninformative in the context of the coverage of the US presidential primaries. Due to the assumption that the journalist is following the cooperative principle and the conversational maxims (Grice, 1968, 1975), an inferential process is triggered by this anomalous utterance. The recipients will assume that the journalists intend to communicate with their audience, that the multimodal orchestration is not obscure but perspicuous, that the article is about something that really happened and of which the journalists have evidence (namely that Obama was elected as candidate of the Democratic Party), that each of the modal elements is relevant and that the content of the article is informative for the newspaper reader under the given temporal circumstances. Together with the relevant prior knowledge, these maxims can lead the newspaper reader to the conclusion that the article reports about Obama having won the Democratic primary and at the same time accentuates the historic dimension of this election by alluding, with a play on words, to one of the most famous anti-slavery novels. The photograph of the White House links the historical dimension with the present age and visually triggers another allusion to black-and-white symbolism.
4. Far beyond this example, it is the concept of *conversational implicature* that can explain how and why comprehension of multimodal discourse is possible, even if the meanings of iconic elements like visuals, gestures, design, or sound are not defined in a dictionary but must be worked out from the discourse itself. In contrast to Forceville's (2020: 14) assumption that 'connecting events in the film world . . . pertains to the level of "explicit meaning"', the example discussed above verifies that meanings of a multimodal artefact or of its modal elements are never explicit but even on the lowest level of interpretation always implicit and based on inferences.

5. INTERMODAL RELATIONS

5.1. Intermodal relations as linkage of communicative actions

In addition to the problem of defining 'what a mode is' (Kress, 2009), it is a methodological challenge to determine the contribution of individual modes

to the overall meaning of a multimodal artefact. Discerning the contributions of the single modes of a multimodal arrangement requires isolating them, which obviously will destroy the meaning of the whole. The following section will discuss some methodological strategies of reception research in order to investigate the meaning-making contribution of single modes provided that the context of multimodal discourse is respected. In principle, there are two empirical methods for this investigation: one can compare the reception data like eye movements, knowledge acquisition or recipient comments of several multimodal artefacts with similar content and different modal orchestrations. The differences in the reception data can then be attributed to the specific modal orchestration in each case. A second method involves switching off single modal layers of a multimodal arrangement – for example, sound, spoken language or colour – and then analysing how this changes the reception data.

Most of the reception studies deploy the first method, for example investigations of information comics (Bucher and Boy, 2018), newspaper articles (Holsanova et al., 2009), information graphics (Holsanova et al., 2008), or science videos (Boy et al., 2020). They have all shown, first, that allocation of attention depends on aspects of the multimodal arrangement and, second, that the degree of attention paid to a relevant element of a multimodal artifact is a determining variable for the degree of comprehension measured by knowledge acquisition. In information comics, for example, there are correlations between the intensity of text reading – measured by dwell time of textual elements – and the degree of comprehension – measured by recall or recognition texts. This indicates that the text mode controls the perception of the visual elements and is an essential prerequisite for understanding the entire comic (Bucher and Boy, 2018). However, modes can also influence each other negatively and thereby disrupt the meaning-making for the overall multimodal artifact (see section 5, paragraph 2). Splitting the attention between several modes increases a cognitive overload and thus decreases comprehension (Ayres and Sweller, 2014).

We can get a deeper insight into the mechanism of intermodal relations by deploying the second empirical method of manipulating the multimodal composition of communication by switching off modal layers of a stimulus – for example, the spoken language or sound of a film or video, or the personal presenter of a PowerPoint presentation or one of the layers of a Tiktok video. Comparing the reception outcome of the manipulated stimuli with the outcomes of the non-manipulated ones allows us to reconstruct the contribution of the switched-off modes to the overall meaning of the multimodal ensemble.

The example used to demonstrate this method is taken from a study which investigated the reception of a commercial with about 60 test subjects (Bucher, 2011). The commercial promotes a new mobile phone by telling the story of a person watching a soccer game on his mobile phone, which makes him so absent-minded that he forgets that he is sitting in the lounge of a hotel

and not in his living room at home. To investigate the problem of compositionality and to identify the influence of single modes on the interpretation of the video, some of the groups saw the videos *with* sound and some *without*. This experimental manipulation of switching off a modal layer was used to allow for isolating the meaning potential of the soundtrack of the film – both sounds and spoken language.

The importance of the soundtrack for understanding the commercial becomes obvious if we compare the reports of the two groups. The subjects who saw the film with sound give a *re-narration of the story* while the 'no sound' group only *describe what they saw*. The 'sound' group is able to tell a story from the perspective of the protagonist ('The protagonist is so fascinated by the TV program . . . he forgets that he is not at home'). Their stories contain relational sequences, which explains the coherence between single episodes, and they draw a conclusion on the full story in formulating the overall intention, the point of the spot ('This means: mobile TV gives you the same TV experience as watching at home'). Subjects from the 'no sound' group describe in particular what they saw ('one sees a good-looking man'; 'in a close-up his face is presented'; 'then a zoom shows the scenery'), aspects which are hardly mentioned by subjects from the 'sound' group. In addition, the verbal data of the 'no sound' group display a distinct additive structure (and then . . . and then), restricted to temporal relations between the episodes ('And then he presumably got a phone call and then he realizes that he is eating potting soil'). One of the test persons of the 'no sound' group explicitly formulates the central problem which was caused by the absence of the aural modes, particularly that she could not hear the phone ringing in the key episode of the video: 'As long as he talks with the waiter, I was somehow able to understand the plot. But from this point I did not get why he picks up his mobile and why the video ends this way' (all recipient quotations are translated from German. For more detailed information about this experiment, see Bucher, 2017).

This last comment indicates that the problem of understanding is brought about by the missing clues from the aural mode of sound. And, indeed, the relevance of the modal resources of the soundtrack for directing the recipients' attention becomes obvious when we compare the eye-tracking data of the groups with and without sound. In general, the eye movement pattern of the 'no sound' group is characterized by the fact that certain relevant areas of attention are viewed later and for longer as in the 'sound group' (Bucher, 2017). Obviously, the lack of sound prevents them from watching the video in the intended temporal dynamics. The differences between the eye-tracking data of the two groups document that the aural mode influences the allocation of attention and the perceptual processing of the visual field, which allows us to reveal the modal resources of sound and its function in audio-visual media. Obviously, a maxim for watching audio-visual messages is: 'You see what you hear'. Figure 3 verifies this function by demonstrating the influence of the ringing tone of the phone on the eye movements of a recipient.

Figure 3. How sound influences the recipients' eye movements: the ringing of a phone. Still from an eye tracking video produced with SMI mobile eye tracker 'iView X RED', showing eye movements from a scene of a TV commercial for an LG mobile phone. (own representation first published in Bucher 2011, p. 138).

The moment the cell phone rings, the attention shifts away from the waiter. This is an important observation because a salience-based bottom-up theory focusing on visual elements cannot explain why the eyes of the protagonists are moving to an area of the screen where nothing significant can be seen. This area is only relevant in the context of the story told: the recipients assume that it is the protagonist's cell phone ringing and not just any phone. In the case of the subjects who see the silent movie, the attention does not shift away from the waiter until the protagonist is completely on screen.

This short episode in the film is of high significance for getting the point of the story across. The ringing of the mobile phone makes the protagonist realize that he is not watching the football game at home but in a restaurant. This switch from the subjective reality of the protagonist to the objective reality of the narration is staged in the film by a rather complex intermodal arrangement of visual elements and sound. The ringing indicates what element in the visual mode of the video is relevant in this episode, namely the mobile phone and not the waiter. The sound directs the selection of the elements which are relevant in the visual mode – although they are not even seen. Furthermore, the ringing of the phone is an important clue for the interpretation of the protagonist's facial expression. None of these co-deployed modes is read in isolation, but each is mutually interpreted with the help of the others. Neither the ringing of the mobile phone alone nor the simultaneously presented video sequence make sense on their own, but together they

constitute an episode within the story told in the commercial. A sign-oriented approach, which for example focuses on the meaning of the ringing phone as 'someone gets a call from another person', is not suitable for explaining these complex intermodal relations. Even a 'system' which categorizes 'cohesive features' (Tseng et al., 2021) has problems explaining the complex dynamic of audio-visual meaning-making practices like this one because coherence of discourse is not constituted by cohesive links between signs or features but on the level of complex communicative actions with a hierarchical structure of sub-actions and superordinated actions.

To understand the interrelations of the different modal elements, one must integrate the actions performed with the different modes – visual and aural – into a superordinate action of communication: telling the story of a person who watches a football game on his mobile phone in a restaurant, assuming that he is at his home. And, of course, understanding the video presupposes that we integrate this narrative discourse into another higher-level structure of advertising a new cell phone. This reconstruction of the process of multimodal meaning-making with the help of eye-tracking data and verbal protocols clarifies the requirements for explaining intermodal relations. What is necessary is a theory of cooperation and inferential meaning, and a dynamic theory of communicative action which constitutes the required building blocks of a theory of multimodality. Setting out such a communicative approach as a fundamental scaffold for analysing multimodality takes the requirement seriously that 'we must always see both the modes and their contexts of use within socio-historically-situated medial practices' (Bateman et al., 2016: 71). The combination of theories proposed here provides the conceptual building blocks needed to systematically clarify what is meant by 'context of use' and 'sociohistorically situated medial practices'.

5.2 Modal density and modal coherence

Of course, comprehension of multimodal media discourse is not limited to understanding single intramodal relations nor can it be equated with the sum of such local understandings. Eye tracking data do not only contribute to the investigation of single intermodal relations but also to the overall multimodal discourse structure of media messages. Eye tracking data can be seen as reflections of the structure of multimodal arrangements, filtered by the process of allocating attention and decisions of relevance. Translated into the language of eye tracking data, intermodal relations are shifts from one AOI to another. By aggregating all these shifts of one multimodal artefact, we can calculate a transition matrix which is a full catalogue of all visual intermodal relations. If all the transitions of a group of recipients are aggregated in a transition matrix, the matrix represents the interactive profile of the corresponding multimodal artifact, which can be quantified by two measures: the *density* of transition matrix and its *entropy* (Krejtz et al., 2014). The single cells of the matrix indicate how many transitions a modal element attracts and therefore which

cognitive load it demands to be processed (Holmqvist et al., 2011: 193–196). The *density of a matrix* expresses how high the proportion of cells visited is compared to all cells. 'A low transition density is an indication of efficient and directed search, while a dense matrix would imply random search' (Holmqvist et al., 2011:341; see also Chen and Shi, 2019).

The value of *entropy* was defined mathematically by Claude E Shannon (1948) as a measure of the probability of an occurrence in an environment. For the reception of multimodal arrangements, it can be calculated on the basis of the transition matrix which captures the distribution of attention to a particular stimulus. The lower the entropy, the more predictable the gaze movements and the more homogeneous the navigation paths of the recipients. And the more homogeneous the recipients' gaze patterns are, the stronger the gaze guidance of a multimodal artefact and the higher the probability that the recipients have caught the relevant modal elements (Gwizdka, 2014; Hooge and Camps, 2013). Entropy value and transition matrix density are therefore reliable empirical measures for modal density (number of modal elements relevant for perception) and modal coherence (mutual coordination of the modes) of a multimodal ensemble as perceived by the recipients. The measures express how recipients deconstruct the density of multimodal arrangements as it is defined in picture theory in terms of a 'dense symbol system with a high degree of repleteness' (Scholz, 2000) or in multimodal interaction analysis as a combination of modal intensity and modal complexity (Norris, 2004: 79). Together, entropy value and matrix density express the potential of a multimodal artefact for gaze control and thus the scope for allocating attention and relevance.

The following screen shots of two science videos from a study on science communication in YouTube and TV (Bucher et al., 2022) document the aggregated eye movements of 18 (left) and 17 (right) test persons. The two videos, both explaining the phenomenon of *déjà vu*, evoke two quite different gaze patterns: a homogeneous one in the case of the YouTube amination film (left) and a heterogeneous one in the case of the TV presentation video (right). (see Figure 4).

One can quantify the visual differences with the eye movement measures of entropy and matrix density (see Table 1).

Table 1 documents that lower matrix density and entropy and shorter navigation paths of the YouTube video are indications of efficient and directed gaze guidance, which lead to successful knowledge transfer. The comparison of these two films reveals a dilemma of attention allocation, which is typical for presentation videos (see Wang et al., 2020): in comparison with other video types, such as animation film, narrative exploratory film or expert film, they present the worst performance in the knowledge test (Boy et al., 2020: 11). This below-average performance can be explained by a specific weakness in attentional guidance, which is expressed in the entropy and density data:

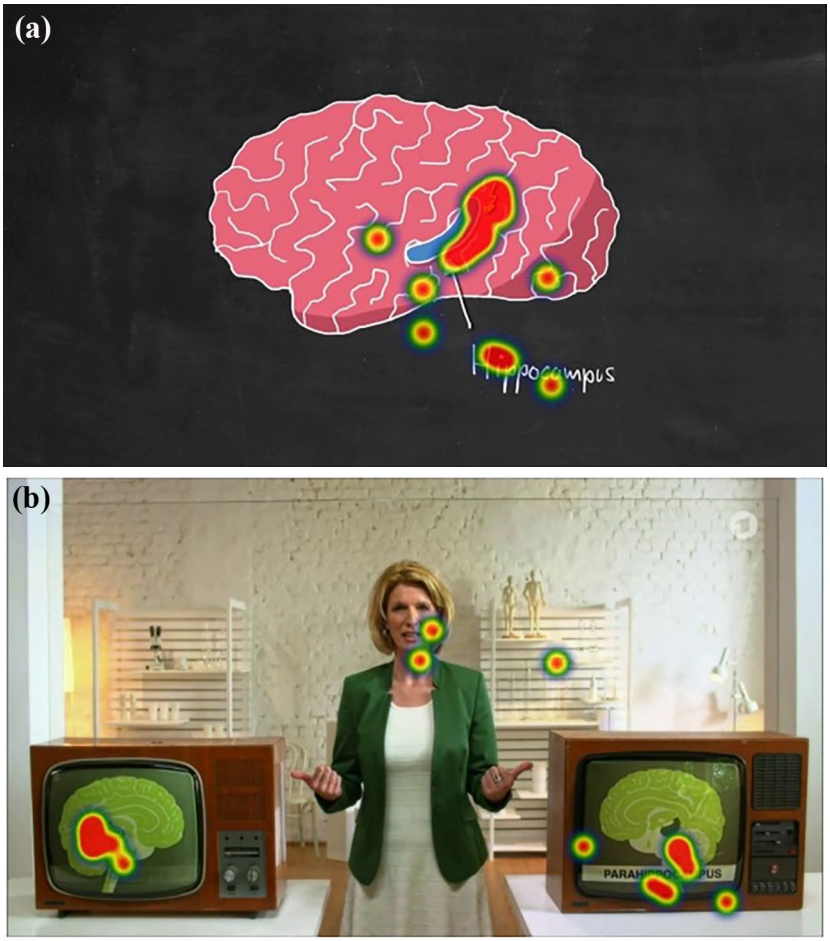

Figure 4. Heat maps of eye movements from a YouTube animation video (left, $n = 18$) and a TV presentation video (right, $n = 17$) both explaining the phenomenon of *déjà vu*. a) still from an eye-tracking video produced with SMI mobile eye tracker 'iView X RED', showing eye movements from a scene of a BYTEthinks YouTube-Video explaining *déjà vu* (https://www.youtube.com/watch?v=UObb82p-YNo&t=7s). b) still from an eye-tracking video produced with SMI mobile eye tracker 'iView X RED', showing eye movements from a scene of a YouTube video explaining *déjà vu* (Wissen vor acht: https://www.daserste.de/information/wissen-kultur/wissen-vor-acht-mensch/sendung/wissen-vor-acht-mensch-1256.html) (Own representation first published in Bucher et al. 2022, p. 153).

Table 1. Gaze guidance and knowledge transfer of a YouTube animation film and a TV presentation film explaining the issue of *déjà vu*.

	Déjà-vu video on YouTube	*Déjà-vu* video on TV
Matrix density	0.4	0.55
Entropy value	0.61	0.71
Navigation paths	short	long
Transfer of factual and structural knowledge	high	low

the simultaneous presence of a speaking person and the relevant visualizations force the recipients to split the visual reception channel into two sources: the speaker and the objects he or she is talking about. This leads to the dilemma of splitting attention between different modes and hence increases cognitive overload (Ayres and Sweller, 2014). In the case of the TV film, the dilemma is heightened further by the presenter's ambiguous pointing gesture, which refers to both screens simultaneously. The comparison of the two science videos shows that the number of modal elements has a significant influence on comprehending a multimodal artefact: if the modal density is too high, the single modes influence each other negatively and thus disturb the meaning-making for the overall multimodal artifact.

A general result of this comparison is that the lower the modal density and the higher the modal coherence of the multimodal artefact, the lower the entropy value and the matrix density. In the case of the YouTube video, both values indicate a highly homogeneous scan path and, therefore, strong gaze guidance. Based on entropy and matrix density, we can define modal density and modal coherence as two basic criteria for the communication quality of a multimodal message. The aggregation of eye tracking data is therefore a methodological means of capturing and evaluating the multimodal structure of media messages.

6. CONCLUSIONS

The intention of this article was to clarify the concept of multimodal meaning making from a recipient's perspective on the basis of empirical data. This should not only fill a research gap in the mostly product- and production-oriented approaches to multimodality but also help to develop a satisfactory theory of multimodality.

On a *methodological* level, it follows from the article that the black box of multimodal meaning-making remains closed if only either the products or the effects of multimodal artifacts are considered. Rather, a research design is needed that takes the appropriation process into account to discover systematic relations between elements of a multimodal artifact and its effects on meaning-making, attention, interpretation, knowledge and feelings. Eye tracking data, recorded during the process of reception, form the bridge between such effects and the layers of the multimodal artefact. The article demonstrates that a triangulation of eye tracking, knowledge tests, retellings, guided interviews and questionnaires provides the verbal and attentional data to make the complex of multimodal meaning-making more transparent.

A first *theoretical result* of this approach is the insight that multimodality implicates *non-linearity*: the simultaneous occurrence of several modes affords selectivity and distribution of attention. Therefore, multimodal understanding has some similarities with navigating a hypertext: each recipient has to find their path of reading, be it a newspaper page, a website, a photo, a film,

an Instagram posting or an exhibition in a museum. To find a path means, first, to select and classify the elements which are relevant and, second, to assign coherence to the selected elements, which means bringing them under a higher-level action (see Bucher, 2017: 111). Hence, understanding multimodal discourse happens step by step dynamically and is not a type of decoding. Forceville (2020: 258) concedes that 'Relevance Theory does not proffer any tools or methodologies for the analysis of specific visual and multimodal discourses.' The concept of non-linearity opens a path to define the *relevance structure* of a multimodal arrangement: the relevance of modal elements is the outcome of its mutual contextualization on the basis of constellations of communicative actions. For example, the ringing of a phone is relevant for the protagonist's facial expressions – and vice versa – in the context of a narration. This brings us to a second consequence: understanding intermodal relations implies bringing the related modal elements under a superordinate structure of action.

With this perspective on multimodality, the central unit of analysis is neither the sign-maker nor the sign itself or single modes but the *discourse* or *activity type* in which signs of different modes are used for communicative intentions. The suggestion to begin the analysis of multimodality at the level of discourse or activity type is compatible with Forceville's (2020) view that 'context and Genre-attribution are "the single most important element of the addressee's cognitive environment steering his strategy of interpreting a message, whether verbal, visual, or of any other kind"' (Forceville, 2020: 121; see also p. 241). Intermodal relations between text and image, sound and film, or design and article are analysed not as constellations of signs but as complex discursive units addressed with a special intention to an audience or a person. In line with this conception, the common ground of different modes is not to be located in semantic metafunctions but on a higher level of communication, i.e. in discourse patterns like narration, explanation, argumentation, or advertising. Considering the examples analysed, the ideational, interpersonal and textual metafunctions are either too unspecific or too restrictive to represent the communicative or functional diversity of the various media discourses. In order to make multimodality analysis even more compatible with research on genres, speech acts or forms of action, 'mode' could be defined as follows: mode is the smallest material unit with which a complete communicative act can be performed. Therefore, resources are an inherent aspect of mode: there is no mode without communicative resources and there are no communicative resources without a material mode.

As has become clear from the analysis of several examples of multimodal arrangement, the concept of *relevance* is central for a theory of multimodality. The non-linear character of a multimodal arrangement confronts the addressee with the task of selecting which modal elements are relevant for its comprehension. Thus, coherence is a sufficient condition for relevance.

One can demonstrate that, in a multimodal arrangement, a modal element coheres with another – a picture with a piece of text or a sound with a bodily reaction – by showing that they are relevant for each other. They are mutually relevant on the basis of an overarching activity type or a topic. The ringtone of the mobile phone and the facial expression are mutually relevant as discourse elements of a narration of how an absent-minded person is called back into reality by the ringtone of his or her phone. This discourse-based concept of relevance is more specific than Forceville's (2020: 40) definition following Sperber and Wilson as 'positive effect in the cognitive environment of the addressee'.

From the perspective of cooperation theory, Forceville (2014: 53) defines multimodal communication as 'ostensive-inferential'. A viable concept for explicating inference in communication is the concept of *conversational implicature*, as introduced by Grice (1975): to explain the inferential process, Grice proposes a cooperative principle and four maxims that govern all communication: the *maxim of quantity* (make your contribution as informative as is required); the *maxim of quality* (do not say what you believe to be false or for which you lack evidence); the *maxim of relation* (be relevant); and the *maxim of manner* (be perspicuous, avoid obscurity and ambiguity, and be brief and orderly). Grice (1975: 47) develops this infrastructure of cooperation for 'a maximally effective exchange of information' and himself points out that his consideration needs to be generalized to other communicative functions. In order to utilize the full potential of conversational implicature, we therefore also need a theory at the level of different forms of communication and types of discourse. A theory of communicative action can accomplish such a generalization. It provides the means to describe the structure of multimodal orchestrations as complex actions referring to the basic types of relations between actions like the '*and-then-*', the '*and-simultaneously-*', or the '*by-relation*' (Heringer, 1978: 21–44). In an action-based approach, intermodal relations are treated as relations between actions on different hierarchical levels, which represent the complex structure of a multimodal orchestration. Considered together, a theory of cooperation and inferential meaning and a theory of communicative action can be used as complementary building blocks for a theory of multimodality. Thus, the article follows the plea to 'use the full spectrum of social research methods to understand the conditions of sign making' (Bezemer, 2024: 506).

ORCID ID

Hans-Juergen Bucher https://orcid.org/0009-0001-8208-1822

DATA AVAILABILITY STATEMENT

The datasets used and analyzed for this study are available from the corresponding author on reasonable request.

REFERENCES

Alston W (1964) *Philosophy of Language*. Englewood Cliffs, NJ: Prentice Hall.

Ang I (1996) *Desperately Seeking the Audience*. London: Routledge.

Austin JL (1956/1957) A plea for excuses: The Presidential Address. *Proceedings of the Aristotelian Society*, New Series 57: 1–30.

Ayres P and Sweller J (2014) The split-attention principle in multimedia learning. In: Mayer RE (ed) *The Cambridge Handbook of Multimedia Learning*. New York, NY: Cambridge University Press, 206–226.

Baldry A and Thibault PJ (2005) *Multimodal Transcription and Text Analysis: A Multimedia Toolkit and Coursebook*. London: Equinox.

Bateman J (2014) Multimodal coherence research and its applications. In: Gruber, H and Redeker G (eds) *The pragmatics of discourse coherence*. Theories and applications, Amsterdam, John Benjamins Publishing Co: 145–177.

Bateman J Wildfeuer J and Hiippala T (2016) Multimodality: Foundations, research and analysis. *A Problem-oriented introduction*. Berlin/Boston, Walter de Gruyter.

Bateman J (2021) Dimensions of materiality towards an external language of description for empirical multimodality research. In: Pflaeging J et al. (eds) *Empirical Multimodality Research*. Berlin: De Gruyter, 35–63.

Bezemer J (2024) What modes can and cannot do: Affordance in Gunther Kress's theory of sign making. *Text & Talk* 44(4): 493–510.

Boeriis M and Holsanova J (2012) Tracking visual segmentation: Connecting semiotic and cognitive perspectives. *Visual Communication* 11(3): 259–281.

Boy B, Bucher H-J and Christ K (2020) Audiovisual science communication on TV and YouTube: How recipients understand and evaluate science videos. *Frontiers in Communication* 5, 17 December. DOI: 10.3389/fcomm.2020.608620.

Bucher H-J (2004) Online-Interaktivität: ein hybrider Begriff für eine hybride Kommunikationsform. Begriffliche Klärungen und empirische Rezeptionsbefunde (Online-Interactivity: A hybrid concept for a hybrid form of communication. Clarification of a concept and empirical reception results). In: Bieber C and Leggewie C (eds) *Interaktivität. Ein transdisziplinärer Schlüsselbegriff* (Interactivity: A transdisciplinary key concept). Frankfurt am Main: Campus Verlag, 132–167.

Bucher H-J (2011) 'Man sieht, was man hört' oder: Multimodales Verstehen als interaktionale Aneignung. Eine Blickaufzeichnungsstudie zur audiovisuellen Rezeption (One sees what one hears: Multimodal understanding as interactional appropriation). In: Schneider J and Stöckl H (eds) *Medientheorien und Multimodalität. Ein TV-Werbespot – Sieben methodische Beschreibungsansätze* (Media theory and

multimodality. One commercial – seven methodological approaches). Cologne: Herbert von Halem Verlag, 109–150.

Bucher H-J (2017) Understanding multimodal meaning making: Theories of multimodality in the light of reception studies. In: Seizov O and Wildfeuer J (eds) *New Studies in Multimodality: Conceptual and Methodological Elaborations*. London: Bloomsbury, 91–123.

Bucher H-J and Boy B (2018) How informative are information comics in science communication? Empirical results from an eye-tracking study and knowledge testing. In: Dunst A et al. (eds) *Empirical Comic Research: Digital, Multimodal and Cognitive Methods*. New York, NY: 176–196.

Bucher H-J and Niemann P (2012) Visualizing science: The reception of powerpoint presentations. *Visual Communication* 11(3): 283–306.

Bucher H-J and Schumacher P (2006) The relevance of attention for selecting news content: An eye-tracking study on attention patterns in the reception of print- and online media. *Communications: The European Journal of Communications Research* 31(3): 347–368.

Bucher H-J and Schumacher P (eds) (2012) *Interaktionale Rezeptionsforschung. Theorie und Methode der Blickaufzeichnung in der Medienforschung* (Interactional reception research. Theory and method of eye tracking in media research). Wiesbaden: Springer Verlag.

Bucher H-J, Boy B and Christ K (2022) *Audiovisuelle Wissenschaftskommunikation auf YouTube. Eine Rezeptionsstudie zur Vermittlungsleistung von Wissenschaftsvideos* (Audio-visual science communication on Youtube: A reception study on the potentials of science videos). Wiesbaden: Springer VS.

Castaldi J (2021) A multimodal analysis of the representation of the Rohingya crisis in BBC's Burma with Simon Reeve (2018): Integrating audience research. *Multimodal Critical Discourse Studies* 10(1): 55–72.

Castaldi J (2024) Refining concepts for empirical multimodal research: Defining semiotic modes and semiotic resources. *Frontiers in Communication* 9. DOI: 10.3389/fcomm.2024.1336325.

Chen Z and Shi BE (2019) Using variable dwell time to accelerate gaze-based web browsing with two-step selection. *International Journal of Human–Computer Interaction* 35(3): 240–255. DOI: 10.1080/10447318.2018.1452351.

Duchowski AT (2007) *Eye Tracking Methodology: Theory and Practice*, 2nd edn. London: Springer.

Fei V L (2004) Developing an integrative multi-semiotic model. In: O'Halloran, K L (ed) *Multimodal Discourse Analysis. Systemic Functional Perspectives*, London, New York, Continuum: 220–246.

Forceville C (2014) Relevance Theory as a model for analysing visual and multimodal communication. In: Machin D (ed.) *Visual Communication*. Berlin: De Gruyter, 51–70.

Forceville C (2020) *Visual and Multimodal Communication: Applying the Relevance Principle*. New York, NY: Oxford University Press.

Fritz G (2022) *Coherence in Discourse: A Study in Dynamic Text Theory*. Giessen: Giessen University Library Publications. DOI: 10.22029/jlupub-791.

Grice HP (1968) Utterer's meaning, sentence-meaning and word-meaning. *Foundations of Language* 4: 225–242.

Grice HP (1975) Logic and conversation. In: Cole P and Morgan JL (eds) *Syntax and Semantics*, Vol. 3. New York, NY: Speech Acts, 41–58.

Gwizdka J (2014) Characterizing relevance with eye-tracking measures. *Proceedings of the 5th Information Interaction in Context Symposium*. Regensburg: Association for Computing Machinery, 58–67.

Hall S (1980) Encoding – decoding. In: Hall S (ed.) *Culture, Media, Language*. London: Routledge, 128–138.

Heringer H-J (1978) *Practical Semantics: A Study in the Rules of Speech and Action*. The Hague: Mouton Publishers.

Holmqvist K (2011) *Eye Tracking: A Comprehensive Guide to Methods and Measures*. Oxford: Oxford University Press.

Holsanova J (2014) In the eye of the beholder: Visual communication from a recipient perspective. In: Machin D (ed.) *Visual Communication*. Berlin: De Gruyter, 331–355.

Holsanova J Holmberg N and Holmqvist K (2006) Tracing integration of text and pictures in newspaper reading. *Lund University Cognitive Studies* 125: Available at: https://www.researchgate.net/publication/304023646_Tracing_Integration_of_Text_and_Pictures_in_Newspaper_Reading (accessed 26 May 2025).

Holsanova J, Holmqvist K and Holmberg N (2008) Reading information graphics: The role of spatial contiguity and dual attentional guidance. *Applied Cognitive Psychology*. DOI: 10.1002/acp.1525.

Holsanova J, Rahm H and Holmqvist K (2006) Entry points and reading paths on newspaper spreads: Comparing a semiotic analysis with eye-tracking measurements. *Visual Communication* 5(1): 65–93.

Hooge I and Camps G (2013) Scan path entropy and arrow plots: Capturing scanning behavior of multiple observers. *Frontiers in Psychology* 4. DOI: 10.3389/fpsyg.2013.00996.

Jewitt C (ed.) (2009) *The Routledge Handbook of Multimodal Analysis*. London: Routledge.

Just MA and Carpenter PA (1976) Eye fixations and cognitive processes. *Cognitive Psychology* 8(4): 441–480.

Katz E and Foulkes (1962) On the use of mass media as 'escape': Clarification of a concept. *Public Opinion Quarterly* 26(3): S. 377– 388.

Krejtz K, Szmidt T and Krejtz I (2014) Entropy-based statistical analysis of eye movement transitions. *ETRA*, 26–28 March: 159–166.

Kress G (2009) What is mode? In: Jewitt C (ed.) *The Routledge Handbook of Multimodal Analysis*. London: Routledge, 54–67.

Kress G (2010) *Multimodality: A Social Semiotic Approach to Contemporary Communication*. London: Routledge.

Liu Y and O'Halloran KL (2009) Intersemiotic Texture: analyzing cohesive devices between language and images. *Social Semiotics* 19(4): 367–388.

Livingstone S (2006) The changing nature of audiences: From the mass audience to the interactive media user. In: Valdivia A (ed.) *The Blackwell Companion to Media Research*, 2nd edn. Oxford: Blackwell Publishing, 337–359.

Neumann O (1987) Beyond capacity: A functional view of attention. In: Heuer H et al. (eds) *Perspectives on Perception and Action*. Hillsdale, NJ: Lawrence Erlbaum Associates, 361–394.

Neumann O (1996) Theorien der Aufmerksamkeit (Theories of attention). In: Neumann O and Sanders AF (eds) *Enzyklopädie der Psychologie. Themenbereich C Theorie und Forschung*. Serie II Kognition. Band 2 Aufmerksamkeit. Göttingen: Hogrefe, 559–643.

Norris S (2004) *Analyzing Multimodal Interaction: A Methodological Framework*. New York, NY: Routledge.

O'Halloran KL (2008) Systemic functional-multimodal discourse analysis (SF-MDA): constructing ideational meaning using language and visual imagery. *Visual Communication* 7(4): 443–475.

Paas FGWC and Sweller J (2014) Implications of cognitive load theory for multimedia learning. In: Mayer RE (eds.) *The Cambridge Handbook of Multimedia Learning*, 2nd edn. New York, NY: Cambridge University Press, 27–42.

Pflaeging J and Stöckl H (2021) Tracing the shapes of multimodal rhetoric: Showing the epistemic powers of visualization. *Visual Communication* 20(3): 397–414.

Rafaeli S and Sudweeks F (1997) Networked interactivity. *Journal of Computer-Mediated Communication* 2(4). DOI: 10.1111/j.1083-6101.1997.tb00201.x.

Rayner K (1995) Eye movements and cognitive processes in reading, visual search, and scene perception. In: Findlay JM (ed.) *Eye Movement Research: Mechanisms, Processes and Applications*. Amsterdam: Elsevier, S 3–22.

Renckstorf K et al. (2004) *Action Theory and Communication Research. Recent Developments in Europe*, Vol. 3. Berlin: Mouton de Gruyter.

Salvucci DD (1999) *Mapping Eye Movements to Cognitive Processes*. Pittsburg, KS: Carnegie Mellon University.

Scholz OR (2000) A solid sense of syntax. *Erkenntnis* 52(2): 199–212.

Schönbach K (2017) Media effects: Dynamics and transactions. *The International Encyclopedia of Media Effects*. Hoboken, NY: John Wiley.

Schumacher P (2009) *Rezeption als Interaktion. Wahrnehmung und Nutzung multimodaler Darstellungsformen im Online-Journalismus* (Reception as interaction. Perception and utilization of multimodal arrangments in online-journalism). Baden-Baden: Nomos.

Shannon CE (1948) A mathematical theory of communication. *The Bell System Technical Journal* 27, July/October: 379–423, 623–656.

Stark LW and Ellis SR (1981) Scanpath revisited: Cognitive models direct active looking. In: Fisher DF (ed.) *Eyemovements: Cognition and Visual Perception*. London: Lawrence Erlbaum, S. 193–226.

Stöckl H and Bateman JA (2022) Editorial: Multimodal coherence across media and genres. In: *Frontiers in Communication* 7. DOI: 10.3389/fcomm.2022.1104128.

Tseng C- I, Laubrock J and Bateman JA (2021) The impact of multimodal cohesion on attention and interpretation in film. *Discourse, Context & Media* 44. DOI: 10.1016/j.dcm.2021.100544.

Van Leeuwen T (2005) *Introduction to Social Semiotics*. London: Routledge.

Wang J, Antonenko P and Dawson K (2020) Does visual attention to the instructor in online video affect learning and learner perceptions? An eye-tracking analysis. *Computer Education* 146: 103–779. DOI: 10.1016/j.compedu.2019.103779.

Wildfeuer J and Bateman J (2016) Lingustically oriented comic research in Germany. In: Cohn N (ed.) *The Visual Narrative Reader*. London: Bloomsbury, 19–65.

Wolfe JM (2015) Visual search. In: Fawcett JM et al. (eds) *The Handbook of Attention*. Cambridge, MA: The MIT Press, 27–56.

Yarbus AL (1967) *Eye Movements and Vision*. New York, NY: Plenum Press.

BIOGRAPHICAL NOTES

HANS-JUERGEN BUCHER is Professor Emeritus of Media Studies at the University of Trier and Research Fellow at the Karlsruhe Institute of Technology (KIT) in the Department of Science Communication. In addition to science communication and political communication, his research focuses on empirical reception research (including eye tracking), multimodality research, media discourse analysis, internet and journalism research. He is co-founder of the 'Media Language and Media Discourse' section of the German Society for Journalism and Communication Studies (DGPuK) and a reviewer for various funding agencies and advisor to the journal *Medien & Kommunikation* (M&K). In the BMBF-funded research network 'Multimodality in Crisis and Risk Communication' (MIRKKOMM), he heads the subproject 'Reception of Multimodal Crisis and Risk Communication'.

Address: The Karlsruhe Institute of Technology, Trier University, Universitaetsring, Trier 54286, Germany. [email: bucher@uni-trier.de]