# Distant viewing and multimodality theory : prospects and challenges

Hiippala, Tuomo

Sage
2021

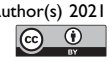
Hiippala, T 2021, 'Distant viewing and multimodality theory : prospects and challenges', Multimodality & society, vol. 1, no. 2, pp. 134-152. https://doi.org/10.1177/26349795211007094

http://hdl.handle.net/10138/336357
10.1177/26349795211007094

*Research Paper*

MULTIMODALITY & SOCIETY

Multimodality & Society
2021, Vol. 1(2) 134–152


Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/26349795211007094
journals.sagepub.com/home/mas


# Distant viewing and multimodality theory: Prospects and challenges

**Tuomo Hiippala**
*University of Helsinki, Finland*

**Abstract**
This article discusses the prospects and challenges of combining multimodality theory with distant viewing, a recent framework proposed in the field of digital humanities. This framework advocates the use of computational methods to enable large-scale analysis of visual and multimodal materials, which must be nevertheless supported by theories that explain how these materials are structured. Multimodality theory is well-positioned to support this effort by providing descriptive schemas that impose structure on the materials under analysis. The field of multimodality research can also benefit from adopting computational methods, which help to achieve the long-term goal of building large multimodal corpora for empirical research. However, despite their immense potential for multimodality research, the use of computational methods warrants caution, because they involve a number of potentially cascading risks that arise from biases inherent to the underlying data and different approaches to the phenomenon of multimodality.

**Keywords**
Multimodality, artificial intelligence, digital humanities, distant viewing, data cascades

## Introduction

The field of multimodality research has matured considerably in the past two decades, reaching a level that has allowed the field to be brought into productive dialogue with more established fields of study such as ethnography (Kress, 2011), information design

**Corresponding author:**
Tuomo Hiippala, Department of Languages, University of Helsinki, P.O. Box 24, Helsinki FI-00014, Finland.
Email: tuomo.hiippala@helsinki.fi

(Bateman, 2019) and media archeology (Thomas, 2020b), to name just a few examples. At the same time, however, most theories of multimodality remain without a robust empirical foundation and are mainly based on conjecture (cf. e.g. Bateman, 2020b). This shortcoming may be traced back to the lack of large-scale multimodal corpora, which would allow checking theories of multimodality against large volumes of real-world communicative situations and artefacts annotated for their characteristics. Such corpora remain intractable due to the time and resources needed to create them, and consequently, current multimodal corpora are more like curated collections rather than true corpora in the linguistic sense of the term (Huang, 2020; Waller, 2017).

A parallel may be drawn to the field of corpus linguistics, which reaped considerable benefits from technological advances at the time, particularly in terms of scaling up the size of linguistic corpora: optical character recognition removed the need for manual input, while natural language processing enabled annotating high volumes of texts automatically for their linguistic features. For studying the multimodality of communicative situations and artefacts, the possibilities for automatic processing remain limited, although recent advances in computer vision and natural language processing are increasingly applied to the analysis of various media in the field of digital humanities (Arnold and Tilton, 2019; Lang and Ommer, 2018; Steen et al., 2018; Wevers and Smits, 2020). Calls to explore the potential of computational methods can also be found in the field of multimodality research (see e.g. O'Halloran et al., 2018, 2021).

In this article I aim to show how theories of multimodality can inform the application of computational methods in the field of digital humanities. More specifically, I argue that theories of multimodality can help circumscribe the large-scale analysis of communicative situations and artefacts, and crucially, to relate the output of computational methods to social situations in everyday life. I begin by describing the growing interest in computational methods in the field of multimodality research, and explicate how this interest is aligned with corresponding developments in the field of digital humanities. I explore the relationship between multimodality research and digital humanities by focusing on the framework for distant viewing proposed by Arnold and Tilton (2019), who advocate the use of computational methods to enable large-scale analysis of visual and multimodal materials, but underline the need for theoretical support. I argue that this support may be provided by the general framework for multimodality proposed in Bateman et al. (2017), which can be used to derive semiotically appropriate metadata schemas for describing diverse communicative situations and artefacts. Finally, I identify certain risks involved in using computational methods, which need to be acknowledged to mitigate them.

## Background

The desire to scale up the volume of data available for analysis has surfaced recurrently across various fields and research traditions concerned with multimodal communication over the last two decades (see e.g. Allwood, 2008; Bateman, 2014b; Forceville, 1999; Huang, 2020; Kaltenbacher, 2004; Parodi, 2010). For the emerging field of multimodality research, Bateman (2014b: 238) observes:

> Discussions of single illustrative cases play an important role in the initial stages of research and theory development, but as theories mature we need to establish the degree to which they can cover and explain uses of multimodality more generally.

Moving from single illustrative cases to larger volumes of data is crucial for any research on multimodality that seeks to adopt an empirical orientation (Bateman, 2020b; Bateman and Hiippala, 2021). Without adequate support from corpora, theories of multimodality cannot be effectively exposed to communicative situations and artefacts at a scale that would capture their inherent variation to an extent that allows evaluating whether the theories and models proposed hold up in the face of data. Ideally, such corpora should enable research to proceed in a cyclic manner: validating theories and models against data, and improving them based on the results (Bateman and Hiippala 2021).

As multimodality research is slowly turning towards a more empirical direction, the need for large, systematically annotated corpora becomes more pronounced (see e.g. Pflaeging et al., 2021; Thomas, 2020a). Researchers have proposed various means of negotiating the bottleneck of time and resources needed to build multimodal corpora and to scale up their size. To exemplify, Hiippala (2016) applies computer vision algorithms to digital images of page-based media to generate annotations that conform to the schema defined in the *Genre and Multimodality* framework (Bateman, 2008), but the output nevertheless requires a considerable degree of human post-processing (see also Thomas et al., 2010). Bateman et al. (2016), in turn, show how computer vision can impose structure on filmic media by establishing visual perceptual units using shot boundary detection algorithms, and how face recognition algorithms allow tracking the appearance of characters across these shots. Similar methodologies are now developed collaboratively for audiovisual media and embodied communication as a part of the *Red Hen Lab* consortium (Steen et al., 2018).

O'Halloran et al. (2018: 12) propose to anchor computational methods more strongly to theories of multimodal discourse analysis within a 'multimodal mixed methods research framework', which has guided the subsequent development of the *Multimodal Analysis Platform* (MAS; O'Halloran et al., 2021). This general framework and its application in the MAS platform aim to provide both qualitative and quantitative insights and to support their combined analysis using visualisations, in order to enable analysts to form hypotheses in an interactive and iterative manner. As such, this framework advocates an approach that resembles the notion of *rapid probing* proposed by Kuhn (2019) for interdisciplinary collaboration between humanities and computer science under the umbrella of digital humanities.

What makes the connection between multimodality research and digital humanities particularly interesting is that combining quantitative and qualitative approaches to visual and multimodal materials is a domain where the interests of the respective fields seem to be gradually converging. Both fields wish to pursue comprehensive, large-scale analyses of multimodal communication, but digital humanities lacks a common theoretical foundation for describing the diversity of communicative situations and artefacts studied within the field. As Bateman (2017) notes, multimodality theory is well-positioned to provide the necessary theoretical support for this effort, but its conceptual

apparatus remains without empirical validation, which may be traced back to the lack of large corpora. If multimodality research seeks to adopt computational methods for building large corpora, the methodological experience of digital humanities in applying such methods may prove especially valuable. This suggests that a mutually beneficial exchange is possible, provided that the fields can identify where exactly their interests meet.

## Distant viewing as a point of contact between digital humanities and multimodality theory

One point of contact between multimodality research and digital humanities is the framework of distant viewing proposed by Arnold and Tilton (2019). This framework uses computer vision to support the large-scale analysis of visual and multimodal materials, in order to automatically identify and extract 'culturally coded elements' from photographic and filmic artefacts (Arnold and Tilton 2019: i7). The act of viewing is conceptualised as a process of interpretation – performed either by humans or algorithms – which involves making inferences about the materials. The information extracted through the process of distant viewing then serves as a foundation for further analysis.

Arnold and Tilton (2019: i5) exemplify how distant viewing can be applied to photography by drawing on the seminal work of Roland Barthes: photographs become meaningful through the act of production, that is, actively manipulating some photographic medium for some communicative purposes. This act provides the light captured on film or digital sensor with a secondary layer of connotative meanings, which are manifested in the form of 'cultural elements' realised using content, framing and composition, to name just a few examples. Arnold and Tilton (2019: i6) propose that these culturally coded elements can be made at least partially explicit by using computer vision algorithms trained – for example – to automatically detect objects in the photograph or to generate linguistic descriptions for the entire photograph or its parts. They emphasise, however, that computer vision models can never capture all culturally coded elements in a photograph, because the resulting descriptions are always reductive, that is, they foreground certain interpretations while neglecting others (Arnold and Tilton 2019: i5).

To illustrate an application of computer vision methods within the framework of distant viewing, Figure 1 uses two different computer vision algorithms to describe a single photograph. Whereas the first algorithm, *DenseCap*, generates linguistic descriptions for regions of interest detected in photographs using a language model (Johnson et al., 2016), the second algorithm, *Mask R-CNN*, detects and classifies objects in photographs and estimates their outlines (He et al., 2017). While the results are technically impressive and aptly illustrate recent advances in computer vision and natural language processing, the resulting detections, classifications and descriptions are entirely dependent on the data used to train the algorithms. *DenseCap* cannot describe objects that it has not 'seen' during training, whereas *Mask R-CNN* can only recognise objects that belong to categories present in the training data, which in this case amount to approximately 80 categories for common everyday objects, as exemplified by the objects detected in Figure 1(b) (Lin et al., 2014).


**Figure 1.** Applying two different computer vision algorithms to distant viewing of a single photograph. (a) shows how a model trained using the *DenseCap* algorithm (Johnson et al., 2016) detects regions of interest and uses a language model to generate linguistic descriptions for these regions. (b) shows how a model trained using the *Mask R-CNN* algorithm (He et al., 2017) detects, localises and classifies objects in a photograph and estimates their shape. The numbers associated with each object label reflect how confident the model is about its prediction and range from 0 to 1 (higher is better). Photograph taken and models applied by the author.

Although output from object detection algorithms such as *Mask R-CNN* can be used as a proxy for detecting human activities at carefully circumscribed locations (Väisänen et al., 2021), using these kind of reductive descriptions as the sole basis for further analysis carries considerable risks. Consider, for example, the set of unique objects detected in Figure 1(b): *bottle*, *chair*, *cup*, *dining table*, *potted plant* and *wine glass*. Although the photograph in Figure 1(b) was taken in a private home, the same set of objects could be detected in a restaurant or a furniture store – locations that host entirely different social situations. To put it simply, sets of objects become meaningful by virtue of being embedded in social interactions and contexts (cf. Nevile et al., 2014). As Arnold and Tilton (2019: i7) suggest, this warrants using multiple computational methods that complement each other: Väisänen et al. (2021), for example, complement the results of object detection using scene detection to narrow down the contexts in which particular objects appear.

However, just what kinds of descriptions can help bridge the *semantic gap* between humans and algorithms for the purposes of distant viewing remains an open question. The semantic gap is commonly conceptualised as the difference in the information that can be extracted computationally from some material and the meanings that humans can give to the same material (Smeulders et al., 2000: 1353). Humans excel in making situated discourse interpretations, draw on their embodied world knowledge and effortlessly update their inferences through abductive reasoning as additional evidence becomes available (Asher and Lascarides, 2003; Bateman and Wildfeuer, 2014b). Algorithms, in contrast, do not possess such capabilities, particularly when faced with

multimodal communication, in which meaning is distributed over various modes of communication that regularly cross sensory channels (Bateman, 2011).

Acknowledging the need to narrow down the semantic gap, Arnold and Tilton (2019: i13) observe that establishing 'code systems for specific computational interpretations through metadata extraction algorithms' should be a major focus of digital humanities, and suggest that this effort requires integrating theories of visual semiotics with state-of-the-art computational methods. Describing such code systems, however, involves several challenges, as Arnold and Tilton (2019: i4–i5) observe by pointing out how 'language' and 'images' differ in terms of their 'representational strategies' and 'code systems'. They emphasise that the 'explicit' code system of language yields itself to computational analysis, because individual words can be readily treated as minimal units of analysis for written text.

Visual materials, in turn, require developing corresponding 'code systems' or 'metadata schemas' for the materials under analysis, in order to acquire an inventory of analytical units (Arnold and Tilton, 2019: i4). The challenges of decomposing diverse visual and multimodal materials into analytical units has been discussed extensively in previous research on multimodality, as exemplified by the work on picture books (Boeriis and Holsanova, 2012), film (Bateman, 2013), comics (Bateman and Wildfeuer, 2014a), information graphics (Hiippala, 2020) and data visualisations (Aiello, 2020), to name just a few examples. Previous research has shown that assuming a distinction between 'language' and 'images' is too coarse for decomposing the materials into analytical units in a way that would respect the complexity of multimodal communication. More precisely, how some material can be decomposed into analytical units depends on the semiotic modes deployed on that material, which warrants a deeper understanding of the semiotic modes and their structure (Hiippala and Bateman, 2021). This complexity raises questions about the potential role and applicability of multimodality research within distant viewing, which I will seek to address below.

## Applying multimodality theory to distant viewing of photographic media

Bateman (2017) argues that multimodality research is well-positioned to support the computational analysis of visual and multimodal materials within digital humanities for several reasons. To begin with, what distinguishes multimodality research from semiotics is its orientation towards establishing a stronger bond between theory and data. As such, multimodality research may be characterised as a form of *applied semiotics*, which is supported by a linguistically inspired methodology that is geared towards systematic description of data at different levels of abstraction (Bateman and Hiippala, 2021). However, as the lack of large multimodal corpora shows, this kind of methodology, which is oriented towards producing detailed descriptions at multiple levels of abstraction, is difficult to apply at scale. Nevertheless, multimodality research has developed a broad array of theoretical concepts that can be used to circumscribe targets of analysis among the diverse range of materials studied in digital humanities, especially when the scope of analysis is extended beyond written language (Champion, 2017). These concepts can provide semiotically appropriate 'metadata schemas' to support distant

viewing of various materials. As we have recently shown in Hiippala and Bateman (2021), annotation schemas that describe multimodal phenomena at various levels of abstraction can be effectively used as metadata schemas for distant viewing, which can reveal underlying structural patterns in multimodal discourse.

The same applies to distant viewing of photographs, as discussed in Arnold and Tilton (2019), which can be informed by previous research on multimodal and visual communication in photographic media. In terms of theoretical concepts, Aiello (2006: 90) relates Barthes' notion of code, or 'sets of rules that are agreed upon within a given cultural system' to the concept of *semiotic resource* developed in the field of social semiotics (Kress and van Leeuwen, 2006). In contrast the fixed nature of a 'code', which assumes that meanings are simply encoded and decoded, the concept of semiotic resource postulates that formal features of photography – as exemplified by composition and framing – can receive different interpretations depending on who the producers and intended consumers are, and what is being communicated using photography. To exemplify, the semiotic resources of news photography have been explicated in great detail by Caple (2013), who shows how they are used to construe news values. Zhao and Zappavigna (2018a), in turn, describe how social media photography uses similar semiotic resources to establish intersubjectivity, whereas Mosbaek Johannessen and Boeriis (2019) show how the manipulation of a photographic medium is grounded in the embodied practices of taking photographs.

These focused descriptions of photography can also be related to more general theoretical frameworks for multimodality. The framework proposed in Bateman et al. (2017) argues for the need to clearly separate the contributions of medium, semiotic mode and genre, which are often taken as key concepts in multimodality research (Hiippala and Tseng, 2017). Within this framework, the concept of medium refers to *socially and historically situated practices that involve regularly manipulating some materiality* for communicative purposes. In the case of photography, all photographic media share the same underlying materiality, which involves capturing light on a physical substrate – such as a film or a digital sensor – or using computers to generate digital artefacts that mimic these material properties. This materiality provides photographic media with a two-dimensional, static canvas. Such a canvas can, in theory, host all semiotic modes that are compatible with a canvas that has these properties (Bateman et al., 2017: 107–109). Just which semiotic modes are actually drawn on depends on the communicative situations that unfold on these media.

As pointed out above, the definition of a medium by Bateman et al. (2017) is founded on the notion of materiality, but is equally defined by the communicative purposes that the materiality in question is regularly used for. Over time, as regular uses of materiality become sufficiently stabilised, the materiality can develop into a fully articulated medium recognised by a community of users. In other words, when members of a community observe a materiality used in particular ways 'in the wild', they can associate the materiality with particular kinds of communicative purposes. By virtue of anchoring the concept of a medium to social and historical practices, *social media photography* (Mosbaek Johannessen and Boeriis, 2019; Zhao and Zappavigna, 2018a) and *news photography* (Caple, 2013) can be characterised as distinct media, which share the same underlying materiality but serve different communicative purposes.

Furthermore, what allows drawing an additional distinction between the media of social media and news photography are the media in which these photographic media are commonly found. Whereas social media photography is generally found in user-generated content on social media platforms, news photography is embedded in various media such as news websites and printed newspapers. Alternatively, news photography is often also circulated on social media platforms. In other words, communicative situations often involve multiple media that are set in various relations to one another, as exemplified by *embedding* and *depicting* (Bateman et al., 2017: 127). Recognising these constellations helps to identify specific media and cues expectations towards the semiotic modes deployed on them.

Characterising the relationship between medium and semiotic modes is crucial for drawing further distinctions between different photographic media. Media are grounded on social and historical practices, but making and exchanging meanings can only take place using semiotic modes deployed on some medium (Bateman et al., 2017: 124). These semiotic modes (and their novel combinations) emerge and evolve on media adopted by communities of users. As such, modes are socially shaped and evolve to meet the communicative needs of a community (Kress, 2014). Bateman (2011) characterises semiotic mode as having three strata: *materiality*, which determines the kind of 'canvas' required by the mode; *expressive resources*, that is, 'regularities of form' that a community of users can associate with particular resources that belong to the semiotic mode in question; and *discourse semantics*, which are mechanisms that support the contextual interpretation of expressive resources and their combinations.

To exemplify, drawing the form of an arrow (→) requires a materiality that provides a two-dimensional canvas. This arrow can be either an illustration of a physical object – an actual arrow – or a diagrammatic element: this depends on whether this particular form is associated with expressive resources that belong to the semiotic of drawing (Riley, 2004) or the diagrammatic mode (Hiippala and Bateman, 2021; Hiippala et al., 2020). This kind of disambiguation is supported by the stratum of discourse semantics, which generates candidate interpretations for instances of expressive resources in their context of occurrence. If shape of an arrow occurs in a picture drawn by a child, the most likely interpretation involves a drawing, whereas when encountered in a diagram, the same shape is likely to be interpreted as a diagrammatic element that stands in for some process (Alikhani and Stone, 2018).

For photographic media such as social media photography and news photography, one may postulate the existence of distinct semiotic modes that are tightly aligned with these media. As pointed out above, a medium alone cannot establish the kinds of conventions that are associated with particular forms of photography, as only semiotic modes possess this capability. At the risk of confusing the reader, it is therefore equally reasonable to talk about social media and news photography as *semiotic modes* that have evolved on media of the same name. Why photographic media and semiotic modes appear conflated can be traced back to the underlying material substrate and the expressive resources that this substrate provides. Although both semiotic modes build on similar expressive resources, such as composition and framing, the choices made within these expressive resources can receive different contextual interpretations, which are guided by the stratum of discourse semantics.

To exemplify, if the photograph shown in Figure 1 were to be treated as an instance of the semiotic mode of social media photography, the formal features related to content, framing and composition could be interpreted as an implied self-portrait or 'selfie' (Zhao and Zappavigna, 2018a: 1745). In this kind of selfie, no human participants are shown, but their presence is implied through objects associated with particular social situations. Alternatively, if the same photograph were to be embedded in an interior design magazine, the example in question would likely to be interpreted as an example of editorial photography, which is intended to convey a certain kind of mood or to establish a narrative. Much of these interpretations rely on recurrent choices made within semiotic modes, which multimodality theory commonly describes using the notion of genre.

Within multimodality theory, genre is understood as staged, goal-oriented ways of achieving specific communicative goals, and plays a crucial role in narrowing down the interpretation of particular choices made within semiotic modes (Bateman, 2008). By mapping particular choices made within semiotic modes to specific communicative goals, genre invokes previous encounters with instances of the same semiotic mode and uses them to generate expectations towards the current situation at hand. Selfies, for example, can be characterised as a genre whose communicative goal is to establish and maintain intersubjectivity between photographer and audience (Zhao and Zappavigna, 2018b), which can be achieved by drawing on particular expressive resources (Zhao and Zappavigna, 2018a). Genres may also blend into one another, as exemplified by Smith (2019), who shows how social media influencers combine aspects of both selfies and landscape photography for self-branding on Instagram.

To reiterate, for many photographic media, the corresponding semiotic modes are strongly connected to the medium and the underlying material substrate. This means that the viewer must first recognise the medium to which the individual photograph belongs. With support from genre cues, recognising the medium narrows down the range of semiotic modes, which are likely to provide the appropriate discourse semantic mechanisms for interpreting the choices made within expressive resources. In many cases, these choices can be increasingly fine-grained if no other semiotic modes are brought into play, which would allow identifying the photographic medium in question. Some exceptions include aerial photography, which draws on the diagrammatic mode to support the interpretation of photographs (Bateman et al., 2017: 280), and the use of filters and graphic overlays in social media photography (Poulsen and Kvåle, 2018; Zhao and Zappavigna, 2018b).

Finally, to summarise the discussion above, I have emphasised the situated and contextual nature of photographic media, while also illustrating just how far contemporary multimodality theory has advanced in describing the production and consumption of such media. Together with general frameworks for multimodality, the previous research on photography can be used to derive multimodally informed metadata schemas for distant viewing. Without support from appropriate metadata schemas, photographs cannot be related to particular social situations and interactions, which explain why the photographs were created in the first place. Figure 2 provides an overview of the conceptual framework presented above and shows how multimodality theory provides appropriate metadata schemas that target multimodal phenomena at different levels of abstraction.

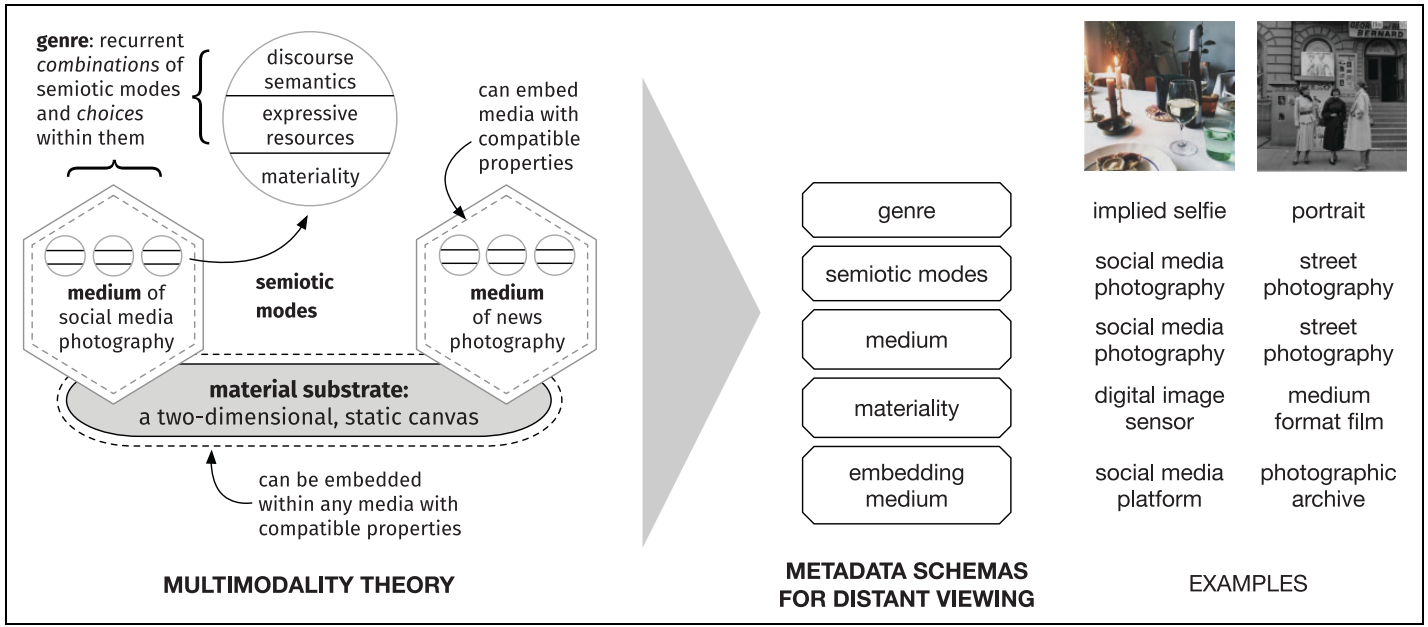



**Figure 2.** Multimodality theory as a source of metadata schemas for distant viewing. The left-hand side shows how distinct photographic media may share the same material substrate, which provides a static canvas with two dimensions. Both media afford access to semiotic modes that are compatible with this canvas. Each semiotic mode features its own set of associated expressive resources, whose contextual interpretation is guided by their discourse semantics. Photographic media can also be embedded within other media with compatible material properties. In addition, photographic media can also embed other media, such as embodied communication, which allows access to further semiotic modes. Pulling apart the media and semiotic modes provides appropriate layers of description for multimodally motivated metadata schemas, which must be then filled in through empirical research.

The categories provided by a metadata schema can serve a *conditioning* role, that is, they can provide potentially useful abstractions that characterise particular features of the data under analysis. To exemplify, the different types of selfies proposed by Zhao and Zappavigna (2018a) could be used as conditioning categories to explain particular choices made in relation to composition, angle and framing of the photographs, to name but a few examples. However, such candidate categories should never be taken for granted, and their reliability should always be evaluated by measuring agreement between analysts (Bateman and Hiippala, 2021). When a sufficiently reliable set of conditioning categories has been identified, these categories may be related to observations made in the data, a task which the framework of distant viewing allocates to computational models. Such features include, for example, the ratio between background and foregrounded objects, the presence and number of human figures, their posture and direction of gaze, and other features that can be extracted automatically using algorithms. In this way, the multimodally motivated conditioning categories may be connected to formal features in photographic media, paving the way for analysis that goes beyond mere surface features.

## Cascading risks in applying computational models

Having discussed the potential benefits of applying multimodality theory to distant viewing, I now turn towards computational models and their application in both multimodality research and distant viewing. If computational models are used to automatically produce descriptions of the visual and multimodal materials under analysis, the potential risks involved in their use must be acknowledged. To reiterate, in this context a model refers to a statistical model that has learned to perform some task – such as to detect and classify objects in photographs, as exemplified in Figure 1. These models can be trained in different ways: in supervised learning, a model learns to perform a task by observing training samples annotated by humans. In unsupervised learning, models learn from massive amounts of unannotated data through so-called proxy tasks that involve predicting, for example, a hidden word in a sequence of words, or whether a given sentence is likely to follow another. Regardless of the way the model is trained, if the training data contains human-like biases, the models trained on this data can pick them up as well (see e.g. Caliskan et al., 2017; Steed and Caliskan, 2021).

Sambasivan et al. (2021) have recently described these risks and how they accumulate using the term *data cascades*, which refers to negative downstream effects that arise from the combination of insufficient attention to training data and poor modelling decisions. Although Sambasivan et al. (2021) focus on real-world applications of artificial intelligence that can have serious and immediate consequences to human well-being, as exemplified by detecting disease or being eligible for credit, I argue that the concept is equally relevant to applying computational models in research on multimodal communication, especially if the research is intended to inform the design of interventions. Figure 3 adopts the diagram introduced by Sambasivan et al. (2021) to characterise potential risks (represented by coloured lines) that emerge from committing to particular decisions during various stages (points on the horizontal line) of developing a model. Below, I describe each stage and the risks involved in relation to multimodality research.

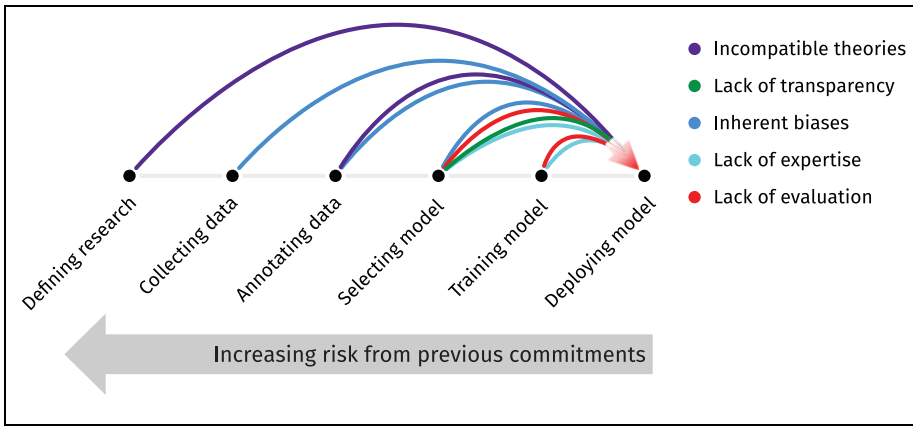



**Figure 3.** Cascading risks in applying computational models in multimodality research and distant viewing. The points positioned on the horizontal line indicate various stages of model development, whereas the coloured lines indicate risks. The combined effect of these risks increases towards the right.

To begin with risks involved when **defining the aims of the research**, it is important to understand that computer science – the field in which the models are developed – and multimodality theory – the field of application – approach the phenomenon of multimodality from different perspectives. These differences are reflected in the definitions of key theoretical concepts, their scope and depth, that is, what the concepts attempt to explain and at which level of detail. In computer science, the notion of multimodality is understood in terms of *sensory modalities*, or the ability to see, hear, touch and process natural language, whereas multimodality theory attempts to explicate how humans combine different 'forms' of communication to make and exchange meanings (Bateman et al., 2017). Computer science often assumes that meanings can be derived in a straightforward manner through visual perception, whereas multimodality theory presumes that any percepts involving communication must be related to particular semiotic modes in order to bridge perception and interpretation (Boeriis and Holsanova, 2012). By default, these approaches are not compatible, and thus computational models may lack sufficient reach when used to pursue to semiotically oriented research questions (cf. Bateman et al., 2016: 140).

Underlying assumptions about multimodality also influence the process of **collecting and annotating data**. Approaches that build on the notion of sensory modalities often assume a distinction between 'text' and 'images', which multimodality theory has found to be untenable, because semiotic modes regularly cross senses (see e.g. Bateman, 2014a). As a result of maintaining this distinction, everything that does not fall neatly into the category of 'text' is often placed under 'images', which hides or obscures the diversity of visual expressive resources across and within semiotic modes (Hiippala and Bateman, 2021). In most cases, the category of 'images' equals photographic media, which excludes drawings, diagrams and other visually driven semiotic modes that differ radically from photographs, particularly in terms of their ability to adjust the level of

abstraction for visual representation (Kress and van Leeuwen, 2006). This means that computational models trained on photographs should not be used to make predictions about non-photographic semiotic modes. The models cannot make reliable predictions, because they have not been exposed to this kind of data during training.

Another source of risk emerges from biases inherent to the training data, which can be propagated to the models trained on this data. Caliskan et al. (2017) and Steed and Caliskan (2021) have shown that both language and computer vision models can learn human-like biases from massive amounts of training data collected for unsupervised learning. To exemplify, Caliskan et al. (2017) show that compared to male names, language models are more likely to associate female names with family words than career words. Similar results have been observed for humans through implicit association tests. Steed and Caliskan (2021) find similar biases in image classifiers that have been trained in an unsupervised manner using photographs scraped from the web. To summarise, training data collected on a massive scale from the internet is not a neutral source of information, but may be expected to contain human biases. Without intervention, these biases can easily seep into models, which are then reflected in model output. This warrants caution, for example, when using language models to automatically generate descriptions of photographic media for further analysis.

**Selecting the model**(s) to be deployed in research is a crucial step for several reasons. State-of-the-art models that build on the family of algorithms known as neural networks (see e.g. LeCun et al., 2015) are often described as black boxes that obfuscate model decisions, which results in a *lack of transparency*. However, making these models more interpretable is an active area of research within the paradigm of explainable artificial intelligence (Barredo Arrieta et al., 2020). A far greater risk related to transparency emerges from applying these models through proprietary services, which do not allow direct access to the model, but simply return reductive output through an application programming interface, such as predicted labels for input data (cf. e.g. O'Halloran et al., 2021). In other words, output from models hidden behind proprietary services cannot be probed using the tools developed for explicating model decisions. A similar lack of transparency naturally extends to the data used to train the models.

Further risks related to selecting and **training models** involve lack of expertise and evaluation. To begin with, choosing and applying an appropriate model requires both technical skills and expertise in *how* models are effectively applied in humanities research and beyond (Bateman et al., 2016: 140). For example, Wevers and Smits (2020) show how image classifiers trained on large datasets can be adopted to custom categories motivated by particular research questions with considerably less data through a process known as transfer learning. However, if computer vision models are used to make predictions about the input data, e.g. to provide descriptive labels for photographs, then their performance must be evaluated against a human-annotated gold standard, because models may not necessarily generalise across domains. In other words, if a model performs well in one domain, this performance is not necessarily carried over to another domain.

To summarise, Figure 3 shows how potential sources of risk may cascade when deploying computational models in multimodality research. These risks may be mitigated by considering the source of risks at each step of model development, and taking

action as necessary, preferably already at the stage of designing the research process, or at least when interpreting the results. Without performing this kind of risk analysis for each stage, one commits to various kinds of assumptions, which may affect the results of the research. This also underlines the need for expertise from various fields, preferably spanning both humanities and computer science.

## Discussion

In this article, I have argued for the benefits of combining multimodality theory with distant viewing, while also outlining risks involved in applying computational models to research on multimodal communication. My core argument has been that multimodality theory can be used to derive semiotically appropriate metadata schemas, which can guide the process of distant viewing for a wide range of communicative situations and artefacts. In this context, semiotically appropriate refers to metadata schemas that respect particular multimodal characteristics of the modes and media under analysis (Hiippala and Bateman, 2021). Developing such schemas, however, requires moving beyond the text/image dichotomy, which should be particularly desirable for research in digital humanities, which often seeks to uncover trajectories of change in the materials under analysis (Wevers and Smits, 2020). In short, multimodality theory can offer a more fine-grained view of *what* changes and *how* by relating the observations to distinct types of multimodal phenomena, ranging from changes within a semiotic mode to combinations of modes in different media (Bateman, 2017).

However, many questions remain as to how the rich conceptual framework of multimodality theory can be most effectively applied within the framework of distant viewing (Arnold and Tilton, 2019). While multimodality theory can provide appropriate metadata schemas for distant viewing, applying these schemas to data is time-consuming manual work, which is precisely the same logjam that has so far prevented building large multimodal corpora with multiple layers of annotation. In Hiippala et al. (2020), we have recently shown that annotations created by crowd-sourced non-expert workers can be used to increase the size of multimodal corpora. However, the extent to which these annotations can support research on multimodality depends on how well the crowd-sourced annotations capture the characteristics of the modes and media under analysis. The use of crowdsourcing should not, however, be ruled out for creating annotations for multimodality research, although this would require bridging the 'discursive gap' between experts and non-experts by making theoretically motivated categories and the criteria for their application explicit to laypersons (Bateman, 2020a).

Whereas defining and applying categories to data ahead of the analysis represents a top-down approach, another alternative involves proceeding from a bottom-up direction. Thomas (2020a: 85) argues that it is easier to 'build upwards by aggregating material features' rather than attempting to pull apart the features that define a given category. As Thomas (2020a) points out, a bottom-up approach may indeed be more suitable, because computational models are often developed for narrow tasks. This places emphasis on identifying models whose tasks are aligned with the needs of multimodality research, or alternatively, applying the models in a manner that supports a bottom-up approach.

Väisänen et al. (2021), for example, use an image classification model to extract computational representations from photographs, which approximate the semantic content of photographs, rather than obtaining 'final' predictions in the form of labels for each photograph. Feeding these interim computational representations to a clustering algorithm, which groups together photographs with similar content, allowed Väisänen et al. to discover human activities in the corpus of photographs.

One must also consider the methodological implications of combining multimodality theory and distant viewing. If computational models are used to automatically process large volumes of data in order to produce annotations correspond to some semiotically appropriate metadata schema, the resulting corpus is likely to be too big and complex to be analysed manually in a productive way. As we have recently argued in Bateman and Hiippala (2021), the search for patterns among large volumes of data is best supported using statistical methods, which are inherently geared for this purpose. Statistical modelling, in particular, can be used to interrogate how different aspects of multimodal phenomena interact with each other. Moreover, statistical models can be comfortably placed against the backdrop of Peircean semiotics by treating them as iconic signs. By examining the properties of a model, we can gain information about the properties of the phenomenon described by the model (Bateman and Hiippala, 2021). Applying techniques involving statistical modelling will undoubtedly require multimodality researchers to add new skills to their toolkit or to seek suitable collaborators: interestingly, a similar development may be currently observed in the field of corpus linguistics (see e.g. Larsson et al., 2020).

Relating multimodality theory to the framework of distant viewing may also help to rethink how multimodal corpora are developed and used. Common approaches to corpus-building aim for a comprehensive representation of the artefact or situation under analysis, which inevitably leads to a situation in which some multimodal characteristics are covered in excessive detail, whereas others are described inadequately (Thomas, 2020a: 83). This imbalance naturally sets limitations to what kinds of research questions may be pursued using the corpus (cf. also Hiippala et al., 2020). Adopting techniques proposed in digital humanities, such as rapid probing (Kuhn, 2019), in combination with guidance from multimodality theory, may eventually help to rethink the role and nature of multimodal corpora. Rather than circumscribing the annotation layers needed ahead of the analysis, automatic processing and distant viewing may help to generate layers of description as required.

Finally, growing reliance on computational models warrants increased attention to the risks involved. Current research suggests that human biases are reflected in written texts and photographs used to train the models, which they pick up during training (Caliskan et al., 2017; Steed and Caliskan, 2021). Whether other sources of bias may be identified for other semiotic modes remains unclear: page layout, for instance, exhibits considerable variation across cultures, which raises the question whether models trained using data from one culture can be reliably applied to another (Bateman, 2008; Thomas, 2020a). These questions may also open new avenues of research for multimodality theory, which may be used to identify and pinpoint potential sources of bias across various modes and media.

## Conclusion

In this article, I have argued that the emerging fields of multimodality research and digital humanities have common goals, as both wish to pursue large-scale, in-depth analyses of visual and multimodal materials. Building on the shared interests may yield fruitful results, because the two fields complement each other methodologically and theoretically. Whereas digital humanities has developed new computationally driven methodologies for analysing various types of materials, multimodality theory has strived towards a general understanding of how humans make and exchange meanings. I propose that combining these perspectives may enable multimodality research to work with larger volumes of data, while providing digital humanities with a deeper understanding of the materials under analysis.

### Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

### Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

### References

Aiello G (2006) Theoretical advances in critical visual analysis: perception, ideology, mythologies, and social semiotics. *Journal of Visual Literacy* 26(2): 89–102.

Aiello G (2020) Inventorizing, situating, transforming: social semiotics and data visualization. In: Kennedy H and Engebretsen M (eds.) *Data Visualization in Society*. Amsterdam: Amsterdam University Press, pp. 49–62.

Alikhani M and Stone M (2018) Arrows are the verbs of diagrams. In: *Proceedings of the 27th International Conference on Computational Linguistics* (ed. Bender EM, Derczynski L and Isabelle P). Santa Fe, New Mexico, USA, August 2018, pp. 3552–3563.

Allwood J (2008) Multimodal corpora. In: Kytö M and Lüdeling A (eds.) *Corpus Linguistics: An International Handbook*. Berlin: Mouton de Gruyter, pp. 207–225.

Arnold T and Tilton L (2019) Distant viewing: analyzing large visual corpora. *Digital Scholarship in the Humanities* 34(suppl 1): i3–i16.

Asher N and Lascarides A (2003) *Logics of Conversation*. Cambridge: Cambridge University Press.

Barredo Arrieta A, Díaz-Rodríguez N, Del Ser J, et al (2020) Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion* 58: 82–115.

Bateman JA (2008) *Multimodality and Genre: A Foundation for the Systematic Analysis of Multimodal Documents*. London: Palgrave Macmillan.

Bateman JA (2011) The decomposability of semiotic modes. In: O'Halloran KL and Smith BA (eds.) *Multimodal Studies: Multiple Approaches and Domains*. London: Routledge, pp. 17–38.

Bateman JA (2013) Multimodal analysis of film within the GeM framework. *Ilha do Desterro* 64: 49–84.

Bateman JA (2014a) *Text and Image: A Critical Introduction to the Visual/Verbal Divide*. London and New York: Routledge.

Bateman JA (2014b) Using multimodal corpora for empirical research. In: Jewitt C (ed.) *The Routledge Handbook of Multimodal Analysis*, 2nd edn. London and New York: Routledge, pp. 238–252.

Bateman JA (2017) Multimodale Semiotik und die theoretischen Grundlagen der Digital Humanities. *Zeitschrift für Semiotik* 39(1–2): 11–50.

Bateman JA (2019) Information design and multimodality: new possibilities for engagement across theory and practice. *Information Design Journal* 25(3): 249–257.

Bateman JA (2020a) Afterword: legitimating multimodality. In: Wildfeuer J, Pflaeging J, Bateman JA, Seizov O and Tseng C (eds.) *Multimodality: Disciplinary Thoughts and the Challenge of Diversity*. Berlin, Munich, Boston: De Gruyter, pp. 297–321.

Bateman JA (2020b) Commentary: the critical role of analysis in moving from conjecture to theory. In: Stöckl H, Caple H and Pflaeging J (eds.) *Shifts Towards Image-centricity in Contemporary Multimodal Practices*. New York, NY: Routledge, pp. 86–94.

Bateman JA and Hiippala T (2021) From data to patterns: on the role of models in empirical multimodality research. In: Pflaeging J, Wildfeuer J and Bateman JA (eds.) *Empirical Multimodality Research: Methods, Applications, Implications*. Berlin and Boston: De Gruyter.

Bateman JA and Wildfeuer J (2014a) Defining units of analysis for the systematic analysis of comics: a discourse-based approach. *Studies in Comics* 5(2): 373–403.

Bateman JA and Wildfeuer J (2014b) A multimodal discourse theory of visual narrative. *Journal of Pragmatics* 74: 180–208.

Bateman JA, Tseng C, Seizov O, et al. (2016) Towards next generation visual archives: image, film and discourse. *Visual Studies* 31(2): 131–154.

Bateman JA, Wildfeuer J and Hiippala T (2017) *Multimodality: Foundations, Research and Analysis – A Problem-Oriented Introduction*. Berlin: De Gruyter Mouton.

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

Caliskan A, Bryson JJ and Narayanan A (2017) Semantics derived automatically from language corpora contain human-like biases. *Science* 356(6334): 183–186.

Caple H (2013) *Photojournalism: A Social Semiotic Approach*. London: Palgrave Macmillan.

Champion EM (2017) Digital humanities is text heavy, visualization light, and simulation poor. *Digital Scholarship in the Humanities* 32(suppl 1): i25–i32.

Forceville C (1999) Educating the eye? Kress and van Leeuwen's *Reading Images: the Grammar of Visual Design*. *Language and Literature* 8(2): 163–178.

He K, Gkioxari G, Dollár P, et al. (2017) Mask R-CNN. In: *Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV)*. Venice, Italy, 22–29 October 2017, pp. 2980–2988.

Hiippala T (2016) Semi-automated annotation of page-based documents within the genre and multimodality framework. In: *Proceedings of the 10th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities* (ed. Reiter N, Alex B and Zervanou KA). Berlin, Germany, August 2016, pp. 84–89.

Hiippala T (2020) A multimodal perspective to data visualization. In: Kennedy H and Engebretsen M (eds.) *Data Visualization in Society*. Amsterdam: Amsterdam University Press, pp. 277–293.

Hiippala T and Bateman JA (2021) Semiotically-grounded distant view of diagrams: insights from two multimodal corpora. *arXiv* 2103.04692. https://arxiv.org/abs/2103.04692.

Hiippala T and Tseng C (2017) Introduction to the special issue on media expectations and genre evolution. *Discourse, Context & Media* 20: 157–159.

Hiippala T, Alikhani M, Haverinen J, et al. (2020) AI2D-RST: a multimodal corpus of 1000 primary school science diagrams. *Language Resources and Evaluation*. DOI: 10.1007/s10579-020-09517-1.

Huang L (2020) Toward multimodal corpus pragmatics: rationale, case, and agenda. *Digital Scholarship in the Humanities*. DOI: 10.1093/llc/fqz080.

Johnson J, Karpathy A and Fei-Fei L (2016) DenseCap: fully convolutional localization networks for dense captioning. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2016)*. Las Vegas, NV, USA, 27–30 June 2016, pp. 4565–4574.

Kaltenbacher M (2004) Perspectives on multimodality: from the early beginnings to the state of the art. *Information Design Journal + Document Design* 12(3): 190–207.

Kress G (2011) Partnerships in research: multimodality and ethnography. *Qualitative Research* 11(3): 239–260.

Kress G (2014) What is mode? In: Jewitt C (ed.) *The Routledge Handbook of Multimodal Analysis*, 2nd edn. London: Routledge, pp. 60–75.

Kress G and van Leeuwen T (2006) *Reading Images: The Grammar of Visual Design*. 2nd edn. London: Routledge.

Kuhn J (2019) Computational text analysis within the humanities: How to combine working practices from the contributing fields? *Language Resources and Evaluation* 53(4): 565–602.

Lang S and Ommer B (2018) Attesting similarity: supporting the organization and study of art image collections with computer vision. *Digital Scholarship in the Humanities* 33(4): 845–856.

Larsson T, Plonsky L and Hancock GR (2020) On the benefits of structural equation modeling for corpus linguists. *Corpus Linguistics and Linguistic Theory*. Epub ahead of print. DOI: 10.1515/cllt-2020-0051.

LeCun Y, Bengio Y and Hinton G (2015) Deep learning. *Nature* 521: 436–444.

Lin TY, Maire M, Belongie S, et al. (2014) Microsoft COCO: common objects in context. In: *Proceedings of the 13th European Conference on Computer Vision (ECCV)*. Cham: Springer, pp. 740–755.

Mosbaek Johannessen C and Boeriis M (2019) Accelerating semogenesis: an ecosocial approach to photography. *Visual Communication*. DOI: 10.1177/1470357219887769.

Nevile M, Haddington P, Heinemann T and Rauniomaa M (eds.) (2014) *Interacting with Objects: Language, Materiality and Social Activity*. Amsterdam: Benjamins.

O'Halloran KL, Pal G and Jin M (2021) Multimodal approach to analysing big social and news media data. *Discourse, Context & Media* 40: 100467.

O'Halloran KL, Tan S, Pham DS, et al. (2018) A digital mixed methods research design: integrating multimodal analysis with data mining and information visualization for big data analytics. *Journal of Mixed Methods Research* 12(1): 11–30.

Parodi G (2010) Research challenges for corpus cross-linguistics and multimodal texts. *Information Design Journal* 18(1): 69–73.

Pflaeging J, Wildfeuer J and Bateman JA (eds.) (2021) *Empirical Multimodality Research: Methods, Applications, Implications*. Berlin and Boston: De Gruyter.

Poulsen SV and Kvåle G (2018) Studying social media as semiotic technology: a social semiotic multimodal framework. *Social Semiotics* 28(5): 700–717.

Riley H (2004) Perceptual modes, semiotic codes, social mores: a contribution towards a social semiotics of drawing. *Visual Communication* 3(3): 294–315.

Sambasivan N, Kapania S, Highfill H, et al. (2021) "Everyone wants to do the model work, not the data work": data cascades in high-stakes AI. In: *Proceedings of the ACM CHI Virtual Conference on Human Factors in Computing Systems*. Yokohama, Japan, 8–13 May 2021.

Smeulders AWM, Worring M, Santini S, et al. (2000) Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22(12): 1349–1380.

Smith SP (2019) Landscapes for "likes": capitalizing on travel with Instagram. *Social Semiotics*. DOI: 10.1080/10350330.2019.1664579.

Steed R and Caliskan A (2021) Image representations learned with unsupervised pre-training contain human-like biases. In: *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency*. Atlanta, GA, USA, January 2019, DOI: 10.1145/3442188.3445932.

Steen FF, Hougaard A, Joo J, et al (2018) Toward an infrastructure for data-driven multimodal communication research. *Linguistics Vanguard* 4: 20170041.

Thomas M (2020a) Making a virtue of material values: tactical and strategic benefits for scaling multimodal analysis. In: Wildfeuer J, Pflaeging J, Bateman JA, Seizov O and Tseng C (eds) *Multimodality: Disciplinary Thoughts and the Challenge of Diversity*. Berlin: De Gruyter, pp. 69–91.

Thomas M (2020b) Multimodality and media archaeology: complementary optics for looking at digital stuff? *Digital Scholarship in the Humanities*. DOI: 10.1093/llc/fqaa024.

Thomas M, Delin J and Waller RHW (2010) A framework for corpus-based analysis of the graphic signalling of discourse structure. In: *Proceedings of Multidisciplinary Approaches to Discourse (MAD 2010)*. Moissac, France, 17–20 March 2010.

Väisänen T, Heikinheimo V, Hiippala T, et al. (2021) Exploring human-nature interactions in national parks with social media photographs and computer vision. *Conservation Biology*. DOI: 10.1111/cobi.13704.

Waller RHW (2017) Practice-based perspectives on multimodal documents: corpora vs connoisseurship. *Discourse, Context & Media* 20: 175–190.

Wevers M and Smits T (2020) The visual digital turn: using neural networks to study historical images. *Digital Scholarship in the Humanities* 35(1): 194–207.

Zhao S and Zappavigna M (2018a) Beyond the self: intersubjectivity and the social semiotic interpretation of the selfie. *New Media & Society* 20(5): 1735–1754.

Zhao S and Zappavigna M (2018b) The interplay of (semiotic) technologies and genre: the case of the selfie. *Social Semiotics* 28(5): 665–682.

## Author biography

**Tuomo Hiippala** is Assistant Professor in English Language and Digital Humanities at the University of Helsinki, Finland. His current research interests include computational methods in multimodality research and multimodal corpora. His major publications include The Structure of Multimodal Documents (2015) and Multimodality: Foundations, Research and Analysis (2017, with John Bateman and Janina Wildfeuer).