# Semiotically-grounded distant viewing of diagrams : insights from two multimodal corpora

Hiippala, Tuomo

2022-05-25

Hiippala , T & Bateman , J A 2022 , ' Semiotically-grounded distant viewing of diagrams : insights from two multimodal corpora ' , Digital scholarship in the humanities , vol. 37 , no. 2 , pp. 405-425 . https://doi.org/10.1093/llc/fqab063

http://hdl.handle.net/10138/344232
https://doi.org/10.1093/llc/fqab063

# Semiotically-grounded distant viewing of diagrams: insights from two multimodal corpora


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

**John A. Bateman**
Department of English, Universität Bremen, Germany and Department of Linguistics, Universität Bremen, Germany



## Abstract

In this article, we argue for the benefits of combining large-scale analyses of visual materials currently pursued within digital humanities with insights from multimodality research, which is an emerging discipline that studies how human communication relies on appropriate combinations of expressive resources. We show that concepts developed within the field of multimodality research provide appropriate metadata schemes for various modes of expression in large corpora and datasets. We illustrate the proposed approach using a common mode of expression, diagrams, and analyse two recent multimodal diagram corpora using statistical and computational methods. Our results suggest that multimodally-motivated metadata schemes can provide a robust foundation for computational analyses of large corpora and datasets. Even if a corpus or dataset is not designed to support full-blown analyses of multimodal communication, our results imply that multimodality theory can still be used to impose tighter analytical control over a variety of visual materials.



**Correspondence:**
Tuomo Hiippala, Department of Languages, University of Helsinki, Yliopistonkatu 4, 00100 Helsinki, Finland.
**E-mail:**
tuomo.hiippala@helsinki.fi


## 1 Introduction

Whether taking place via an external medium or in face-to-face interaction, communication is naturally multimodal: that is, making and exchanging meanings involve combining multiple modes of expression in a coordinated, goal-oriented manner. There is currently growing interest in multimodal communication across various fields of research, including the digital humanities. It is then natural to consider more closely whether contemporary theories of multimodality can support the kinds of large-scale analyses commonly pursued in digital humanities and if so, to what extent. Although the field of multimodality is increasingly oriented towards empirical analysis, compiling multimodal corpora to support such analyses is still highly labour-intensive. More extensive use of computational techniques is thus a clear priority.

In this article, we consider the potential benefits of combining contemporary accounts of multimodality, computational methods, and research orientations from digital humanities with respect to one extremely common mode of expression, namely diagrams. Diagrams are found everywhere, and their structure






doi:10.1093/llc/fqab063

varies depending on the context in which they occur (Purchase, 2014). Understanding how diagrams are structured is a prerequisite for their large-scale analysis in any field. We approach the study of diagrams by drawing on two recent corpora developed for different purposes. While the first corpus was developed to support research on artificial intelligence for tasks such as automatic diagram understanding (Kembhavi *et al.*, 2016), the second corpus builds multiple layers of annotation on top of the first corpus to create a resource for studying how diagrams communicate multimodally (Hiippala *et al.*, 2020). Against this backdrop, we ask: how can modern multimodality theory inform the design and analysis of multimodal datasets and corpora? To answer this question, we critically evaluate the two datasets from the perspective of multimodality theory and the digital humanities notion of 'distant viewing' (Arnold and Tilton, 2019), and support our argument using quantitative and computational methods.

## 2 Multimodality Research and the Digital Humanities

Multimodality research is an emerging discipline that examines how communication builds on appropriate combinations of 'modes' of expression, such as natural language, illustrations, drawings, photography, gestures, layout, and many more (Wildfeuer *et al.*, 2020). Although now widely acknowledged as an inherent feature of human communication, multimodality is not always understood in the same way across the diverse fields of study where the concept has been employed. These fields include, among others, text linguistics, spoken language and gesture research, conversation analysis (Mondada, 2019), human–computer interaction (Oviatt and Cohen, 2015), and, last but not least, digital humanities (Svensson, 2010). To consolidate these perspectives, Bateman *et al.* (2017) propose a general framework for multimodality that extends beyond previous approaches by offering a common set of concepts and an explicit methodology for supporting empirical research regardless of the 'modes' and materialities involved.

Modern multimodality theory has developed a battery of interrelated theoretical constructs to support descriptive and empirical analyses of complex communicative situations and artefacts (Stöckl, 2020). Several core concepts, including in particular semiotic mode (Bateman, 2011; Kress, 2014), medium (Bateman *et al.*, 2017), and genre (Bateman, 2008; Hiippala, 2014), theorize in detail how individual forms of expression are structured and what enables them to effectively combine and co-operate with each other across a wide range of communicative contexts and situations. Despite this comprehensive theoretical apparatus, which is now mature enough to be brought into productive discussion with established fields of study such as media archaeology (Thomas, 2020b), literacy (Jewitt, 2008), ethnography (Kress, 2011), and others, the lack of large annotated corpora stands in the way of refining multimodality theory through empirical research (Thomas, 2020a).

Within linguistics, the provision of ever larger collections of authentic language use has established corpus linguistics as a major pillar of research. The corresponding treatment of multimodal data, and particularly *static* multimodal data, still lags very much behind, constituting a major bottleneck in evaluating and refining the theoretical constructs proposed. Whereas it is common for a range of automatic processing techniques to be applied to linguistic corpora, the possibilities for multimodal data remain limited. As a consequence, multimodal annotation frameworks are slow to develop and apply, and require expert annotators. Most current multimodal corpora consequently remain small and thus resemble curated collections rather than true corpora in the linguistic sense of the term. Although this problem applies in principle both to face-to-face interaction (Huang, 2021) and multimodal documents (Waller, 2017), substantial progress is now being made for some forms of audiovisual multimodal data (e.g. Steen *et al.*, 2018). The situation for multimodal documents is very different and it is this area that we focus on here.

Parallel to developments in multimodality research, within the field of digital humanities, there is a growing interest in the large-scale analysis of 'visual' communication as well (see e.g. Heftberger, 2018; Lang and Ommer, 2018; Arnold and Tilton, 2019; Burghardt *et al.*, 2020; Münster and Terras, 2020; Smits and Ros, 2020; Wevers and Smits, 2020). Arnold and Tilton (2019) propose a framework for *distant viewing*, arguing that such approaches are needed to counterbalance the strong textual orientation in digital humanities, which excludes a wealth of non-linguistic phenomena that are traditionally of interest to the humanities. Drawing on foundational work in semiotics in the tradition of Saussure and Barthes, Arnold and Tilton (2019, p. i4) observe that:

> in order to view images computationally, a representation of elements contained within the visual material—a code system in semiotics or, similarly, a metadata schema in informatics—must be constructed.

Put differently, Arnold and Tilton (2019) emphasize the need to impose analytical control over visual material, just as 'the explicit code system of written language' allows imposing structure on textual corpora by establishing units of analysis (e.g. tokens or parts of speech) and their interrelations (e.g. syntax) (Arnold and Tilton, 2019, p. i5). However, just which analytical units should be defined for modes of communication other than language and at what level of granularity remain open and hotly contested questions.

Providing such capabilities building on more traditional semiotic notions of 'code' is, however, unlikely to succeed for visual materials. As argued more extensively in Bateman (2017, p. 21–22), any model that treats communication in terms of an exchange of meanings in a process of encoding/decoding according to some fixed, static code can readily be shown to be inadequate. Early debates on this 'fixed code fallacy' focused mainly on language and excluded other modes of communication (Cobley, 2013, p. 232), which may explain why early research on multimodality already turned towards developing the concept of 'semiotic mode' as an alternative.

Semiotic modes, of which language is taken as just one alongside many others, are assumed to emerge and be shaped through social interaction within a community of users (Kress and van Leeuwen, 1996, 2001). However, as Bateman (2017, p. 22) points out, early definitions of mode were not robust enough to support empirical research, leading common conceptions of mode and code to gradually became indistinguishable as semiotics itself advanced beyond the encoding/decoding model of communication (cf. Cobley, 2013, p. 231).

Bateman (2011) addresses previous shortcomings with the definition of semiotic mode by proposing a formal account consisting of three semiotic strata, visualised in Fig. 1a. Each stratum is a prerequisite for a fully developed semiotic mode. Starting from the bottom of the inner circle, all semiotic modes must work with respect to some *materiality* that can be manipulated intentionally for communicative purposes. Such traces of manipulation must reflect formal distinctions that are pertinent for *expressive resources* available within the semiotic mode, as exemplified by differences in form between written language and line drawings, which allow us to distinguish between these expressive resources. The expressive resources are assumed to be subject to a paradigmatic organization that allows making selections among them and combining them into larger syntagmatic organizations. Third, the expressive resources and their combinations are mobilized in the service of communication by a corresponding *discourse semantics*, which supports the contextual interpretation of selections made within expressive resources, and whose operation we illustrate in a moment. This general model places no restrictions on the kinds of materiality that may be employed; for current purposes, however, we focus on static materialities with a 2D spatial extent, such as a sheet of paper or a static display presented on a screen.

Figure 1b exemplifies the application of this general model to what may be tentatively called the 'diagrammatic mode'—a semiotic mode underlying all kinds of diagrammatic representations (Hiippala and Bateman, 2020). Beginning from the bottom, diagrams always require a materiality with a 2D spatial extent. Consequently, the diagrammatic mode can theoretically draw on all expressive resources that can operate with a 2D spatial materiality, although which expressive resources are actually mobilized and the choices made within them are largely motivated by genre. The concept of genre is understood as conventionalized ways of achieving particular communicative goals (Kostelnick and Hassett, 2003; Lemke, 2005; Hiippala, 2014). In other words, the communicative goals set for a diagram shape its structure, as these conventionalized structures help invoke previous encounters with similar diagrams.

One intrinsic property of the model in Fig. 1a that multiple expressive resources such as written

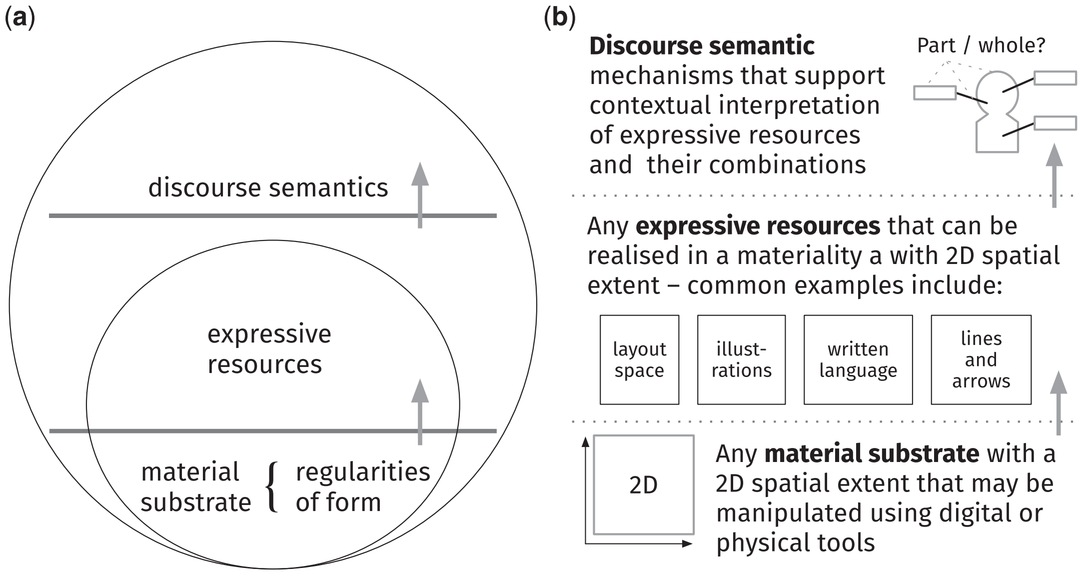


**Fig. 1** The concept of a semiotic mode and its application to diagrams. (**a**) A theoretical model of a semiotic mode. (**b**) A characterization of the diagrammatic mode.

language, illustrations, and photographs may *naturally* co-occur with each other in diagrams, thus avoiding committing to arbitrary divisions between 'verbal' and 'visual' or 'text' and 'image' (Bateman, 2014). This perspective is obviously carried over to mass media, which regularly deploy multiple semiotic modes (Bateman *et al.*, 2017, p. 124). Finally, discourse semantics guides the interpretation of expressive resources and their combinations in context. For diagrams, resolving the resulting discourse relations relies on formal cues such as spatial placement of elements or connections realized using lines and arrows in combination with world knowledge (Watanabe and Nagao, 1998; Alikhani and Stone, 2018).

This brief description illustrates the extent to which modern multimodality theory can explicate how semiotic modes operate quite generally. Describing the characteristics of each stratum of a given semiotic mode—that is, material substrate, expressive resources, and discourse semantics—is then an issue demanding empirical research. Conversely, this also shows just how much complexity is missed when operating with pretheoretical distinctions such as 'text' and 'image'. Although computational analyses of page-based media are already advancing beyond such dichotomies, as Wevers and Smits (2020) have shown by training convolutional neural networks to distinguish between instances of illustrations, photographs, and other semiotic modes in historical newspapers, we argue that advancing this effort within digital humanities will benefit still further from the input of multimodality theory.

As a form of 'applied semiotics' that seeks a close relationship between theory and data (Bateman and Hiippala, 2021), multimodality theory is well-positioned to provide a foundation for characterizing the diverse range of communicative artefacts and situations studied within digital humanities (Bateman, 2017; Hiippala, 2021). Theories of multimodality have already been used to guide the application of computational methods to both filmic (Bateman *et al.*, 2016) and page-based media (O'Halloran *et al.*, 2018), but much remains to be done in terms of applying computational methods in a way that respects the complexity of multimodal communication. While we are not suggesting that all studies that use computational methods should perform full-blown multimodal analyses for each semiotic mode encountered, we do encourage the broader application of multimodality theory to determine the analytical granularity appropriate for answering specific research questions. This allows targets of descriptions and their respective granularities to be derived systematically on the basis of developing bodies of theory. There are then both theoretical and practical reasons for adopting this approach when collecting multimodal data at scale.

As noted by Arnold and Tilton (2019, p. i4) above, it is important when constructing larger collections or corpora for computational analyses that appropriate metadata schemes are defined for organizing that data. In computer science, the definition of 'modality' (the term preferred over 'mode') is strongly aligned with the senses: our ability to see, hear, touch, and use natural language. Distinctions between sensory modalities are then built into the different research fields of computer vision, audio signal processing, and natural language processing, and these are then the sources for corresponding metadata schemes. The resulting fields are as a consequence often confined to their own 'problem spaces', even though these are increasingly converging on multimodality in tasks such as machine translation (Sulubacak *et al.*, 2020). Nevertheless, applying definitions based on sensory modalities continues. In contrast, within humanities-oriented multimodality theory, restricting approaches to sensory channels is now receiving considerable critique because defining modalities ahead of analysis solely on the basis of perceptual properties makes identification of the actual semiotic contributions being made to meaning construction more difficult. As argued extensively in Bateman *et al.* (2017), semiotic contributions regularly extend across sensory channels and their demarcations need to be teased out empirically: one cannot assume their individual characteristics in advance (Bateman, 2011, p. 17–18). It is crucial to open up a two-way communication channel between the semiotic distinctions being made and their supporting material distinctions, rather than assuming that sensory perception alone will result in appropriate segmentations.

These challenges and limitations are made fully evident by our characterization of the diagrammatic mode in Fig. 1b. First, diagrams are clearly not aligned with a single traditional modality as they cross-cut both 'vision' and 'language'. Second, what makes diagrams different from other combinations of a similar nature, such as photographs with embedded or overlaid text, is left an open question. Although assumptions of similarity concerning expressive resources *within* a single sensory modality are common in computer vision research, where objects of analysis are often reduced to mere carriers of content, this is insufficient. Haehn *et al.* (2019, p. 649), for example, report that models trained on photographs do not generalize well to diagrammatic representations without further training even though they are both clearly 'visual'. They consider this finding surprising given prior comparisons between artificial neural networks and the human visual cortex, which assume that *visual perception* suffices for reasoning about both photographs and diagrams.

From the perspective of multimodality theory, however, the differences between photographs and diagrams are rather evident: diagrams differ radically in terms of their expressive resources and discourse semantics. Diagrams are compositional, that is, they can be broken down into component parts, which may be realized using multiple expressive resources and combined into discourse structures that work towards a shared communicative goal. This allows diagrams to represent abstract concepts and phenomena that are not limited to specific slices of time and space, and so stand in strong contrast to photographs (cf. e.g. Alikhani and Stone, 2018; Greenberg, 2018). This demonstrates how it is always essential to consider material distinctions in terms of the particular semiotic modes they operate with respect to. It is precisely these semiotic modes that deliver appropriate metadata schemes for characterizing corresponding objects of analysis.

## 3 Insights from Multimodal Diagram Corpora

Having introduced the concept of a semiotic mode and how appropriate metadata schemes may be derived for individual semiotic modes through empirical research, we turn now to examine two recent diagram corpora from this perspective. These corpora originate in two different fields of research, namely artificial intelligence (Kembhavi *et al.*, 2016) and multimodality research (Hiippala *et al.*, 2020), but contain similar diagrams, as the second corpus is a subset of the first. The two corpora differ in terms of the metadata schemes used to describe the diagrams and their structure, which reflect the disciplinary interests of their respective fields. In the following sections, we unpack these metadata schemes in detail to show that their underlying assumptions significantly affect their ability to capture the multimodal structure of diagrams.

### 3.1 AI2D—a dataset for computational processing of diagrams

The first dataset is the *Allen Institute for Artificial Intelligence Diagrams* dataset (AI2D), which was developed to support research on visual question answering, automatic diagram understanding, and other computational tasks involving diagrams in the field of artificial intelligence (Kembhavi *et al.*, 2016). The AI2D dataset contains 4,903 diagrams that represent seventeen topics in elementary school natural sciences, ranging from life and carbon cycles to human physiology and food webs. The dataset models four types of diagram elements: text, arrows, arrowheads, and blobs. Whereas the first three categories are rather self-explanatory, 'blobs' is a technical term that refers to all visual expressive resources deployed in AI2D diagrams, such as line drawings, illustrations, and photographs (Tversky *et al.*, 2000). To summarize, AI2D relies on predefined categories for diagram elements, which are assumed to be known in advance.

Placing objects into predefined categories is a common strategy when crowdsourcing annotations for computer vision research (Kovashka *et al.*, 2016), but faces problems similar to those encountered when defining analytical categories ahead of actual analysis (Bateman, 2011, p. 18). To reiterate, if the expressive resources of a semiotic mode are assumed to be known in advance, then it becomes difficult to explicate just what a given semiotic mode does with its material substrate—that is, what kinds of 'regularities of form' or expressive resources are made available by the semiotic mode and what can be done with them in terms of communication. As we will see below, problems related to modelling expressive resources are then also propagated to the stratum of discourse semantics, complicating its description as well (cf. Fig. 1).

Each diagram in the AI2D dataset is provided with several types of description. All instances of text, arrows, arrowheads, and blobs were first segmented from the original diagram layout by crowdsourced workers on Amazon Mechanical Turk.[1] Diagram elements identified during *layout segmentation* provide a foundation for a *Diagram Parse Graph* (DPG), which represents the diagram elements as nodes, whereas edges define the semantic relations holding between elements. These semantic relations are described using ten relation definitions drawn from the framework proposed for diagrammatic representations by Engelhardt (2002). The following examples illustrate the application of the AI2D annotation schema using a single diagram from the dataset. The original diagram in Fig. 2 represents a rock cycle, that is, transitions between different types of rock, using a combination of an illustration (a cross-section) whose parts are described using written language. These parts set up the stages of the rock cycle, which are then related to one another using arrows and written language.

The crowdsourced workers were first requested to identify instances of diagram elements during layout segmentation. Figure 3 shows that text blocks and arrowheads were segmented using rectangular bounding boxes, whereas more complex shapes for arrows and various types of graphics were segmented using polygons. This layout segmentation illustrates well how crowdsourced annotators tend to segment diagrams to quite uneven degrees of detail. The entire cross-section of a volcano in Fig. 3 is assigned to a single blob (B0), although arguably a more accurate description would be to segment separate parts of the

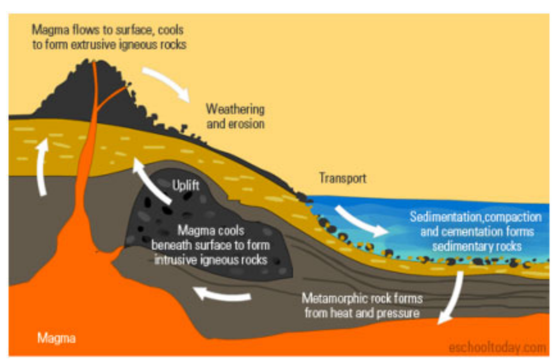


**Fig. 2** Diagram #4210 in the AI2D dataset.

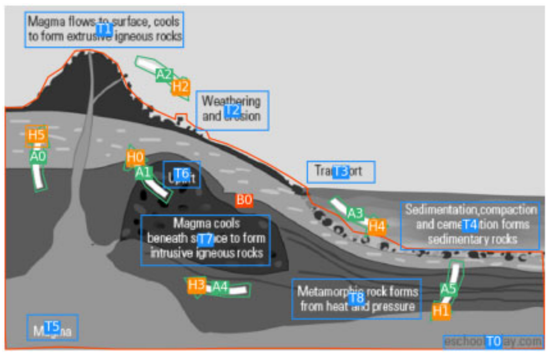


**Fig. 3** Layout segmentation. The diagram image has been converted into greyscale to highlight the diagram elements and their outlines. Each element segmented from the diagram by the crowdsourced workers is coloured according to diagram element type (blue: text; red: blob; arrow: green; arrowhead: orange) and assigned a unique identifier, laid out on top of the bounding boxes and polygons.

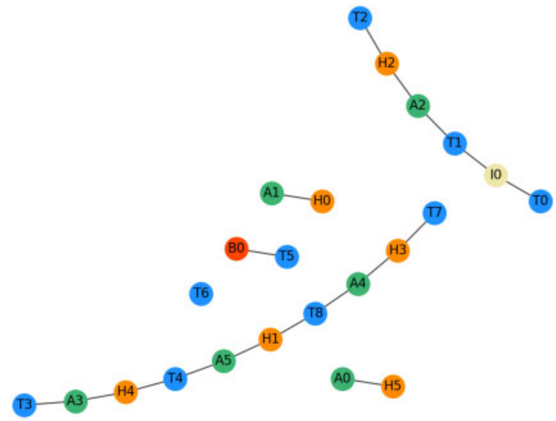


**Fig. 4** DPG. Node identifiers are carried over from the layout segmentation and refer to individual diagram elements. The element I0 stands for the entire diagram, which can be used as the 'target' node for generic elements, as exemplified by text element T0. The DPG has been visualized using the tool developed by Hiippala *et al.* (2020). Node positions are not meaningful, but determined by an algorithm in the NetworkX library (Hagberg *et al.*, 2008).

cross-section, such as magma and various layers of rock. Demarcating such meaningful regions on material substrates with a 2D spatial extent is a hallmark feature of illustrations and other expressive resources that allow visual depiction, which diagrams regularly deploy for communicative purposes in combination with written text (Richards, 2017). As we will show below, ignoring this feature places severe limitations on the description of discourse semantics possible for diagrams.

The DPG in Fig. 4 shows how elements may be connected via labelled edges to capture a variety of relations. Since for current purposes we will focus more on the presence or absence of connections, we omit the labels in the graph here. Examples of specific relations are the arrowHeadTail between arrow A2 and arrowhead H2 visible in the upper part of Fig. 3, and the interObjectLinkage relation corresponding to arrow A2, which acts in turn as a connector between text blocks T1 ('Magma flows to surface . . .') and T2 ('Weathering and erosion'). As these relations exemplify, the relations drawn from Engelhardt (2002) cover *local* relations that hold between diagram elements positioned close to each other or connected using arrows or lines (Kembhavi *et al.*, 2016, p. 239), but they neglect the relations needed to describe the *global* organization of the diagram—that is, relations between units that are made up of multiple elements (Hiippala *et al.*, 2020, p. 5–6). This is not a shortcoming of Engelhardt's (2002) framework, but rather of its limited application in the AI2D dataset.

Crowdsourcing graph-based descriptions of diagrams without a descriptive schema that covers both local and global discourse structures undoubtedly constitutes a challenging task, which may explain why isolated nodes and multiple connected components are commonly found in AI2D DPGs (see, e.g. the isolate T6 and five connected components in Fig. 4). Although the original diagram shows a rock cycle, the cyclic nature of this phenomenon is not reflected by the structure of the DPG visible in Fig. 4 at all, even though the AI2D annotation schema does in principle provide the relation definitions necessary for describing such cycles, including interObjectLinkage and intraObjectRegionLabel (Kembhavi *et al.*, 2016, p. 239).

This shortcoming is caused by insufficient detail in the layout segmentation. The crowdsourced annotators were not instructed to decompose instances of expressive resources capable of visual depiction into meaningful regions, although such resources are commonly deployed in diagrams precisely due to their capability to demarcate regions 'in the world' or described phenomena. The blob B0, which covers the entire cross-section, is not then segmented into its component parts. These parts are indicated by labels such as 'Magma' (T5) and 'Metamorphic rock forms from heat and pressure' (T8), picking out particular regions of the cross-section by visual *containment* (Engelhardt, 2002, p. 47) to set up the stages of the rock cycle. Figure 4 shows that only the label T5 ('Magma') is connected to blob B0 (via a relation of intraObjectRegionLabel). Because the cross-section (B0) constitutes a single unit, other potentially necessary relations are not then available for mapping the visually contained labels (e.g. T4, T6, T7, and T8) to the corresponding regions of the cross-section. Consequently, precisely those regions that would be needed to represent the cyclic structure are absent from the inventory of annotated diagram elements.

These challenges in decomposing diagrammatic representations relate to the well-known problem of identifying analytical 'units' in any visually invested medium. Bateman and Wildfeuer (2014) consider this issue for comics and argue for a discourse-based approach to identifying analytical units, whereby the discourse organization of some larger unit (e.g. a panel in a comic or an entire diagram) help determine which elements are to be picked up for interpretation in a given context. In other words, the stratum of discourse semantics simultaneously supports *decomposing* larger units into their component parts and *resolving* their potential interrelations.

In contrast to often criticized attempts to impose fixed units or parts on visual materials, segmentation in the discourse-based account is always pursued with the goal of maximizing *discourse coherence* via abductive reasoning (Bateman and Wildfeuer, 2014, p. 377), that is, updating discourse interpretations as additional evidence becomes available, while striving to form the most plausible explanation for a particular constellation of expressive resources. For this reason, it will often be more effective not to operate with a predefined inventory of elements (i.e. defining units bottom-up), but instead to allow the inventory of elements derived during analysis to change dynamically as interpretations are made and updated (top-down). Such mechanisms are intrinsic to the stratum of discourse semantics (Bateman, 2020).

## 3.2 AI2D-RST—adding discourse semantics to AI2D diagrams

The second corpus of diagrams, AI2D-RST, features multiple layers of expert annotations built on top of the crowd-sourced layout segmentations from the AI2D dataset (Hiippala *et al.*, 2020). The AI2D-RST corpus covers 1,000 diagrams from AI2D and attempts to provide a more comprehensive description of discourse structure in diagrams, deriving a further metadata scheme directly from the assumed discourse semantics of the diagrammatic mode. This offers a substantial bridge between the visual organization of diagrams and more conceptual–functional descriptions, and so addresses many of the issues recently raised by DeRose (2020) concerning the lack of organizational structure in more surface-oriented diagram markup.

AI2D-RST consequently employs a formalism frequently used to describe the operation of discourse semantics, that of Rhetorical Structure Theory (RST), originally developed as a theory of text organization and coherence in the 1980s (Mann and Thompson, 1988). RST attempts to describe why well-formed texts appear coherent, or why individual parts of a text appear to contribute towards a common communicative goal. The RST framework was extended for the description of multimodal documents in the early 1990s (André and Rist, 1995) and has been used as a model of discourse semantics for various modes and media, ranging from bird field guides and tourist brochures to scientific publications and product packaging (Bateman, 2008; Taboada and Habel, 2013; Thomas, 2014; Hiippala, 2015).

In order to pull apart how different expressive resources are combined into discourse structures in diagrams, AI2D-RST represents each diagram using three distinct graphs that correspond to three distinct but complementary layers of annotation: grouping, connectivity, and discourse structure. Whereas the grouping and connectivity layers attempt to capture aspects of the expressive resources, the discourse structure layer seeks to describe what kinds of interpretations diagram elements and their combinations receive in particular contexts (see Fig. 1b). Figure 5 exemplifies the graphs for all three annotation layers for the diagram introduced in Fig. 2.

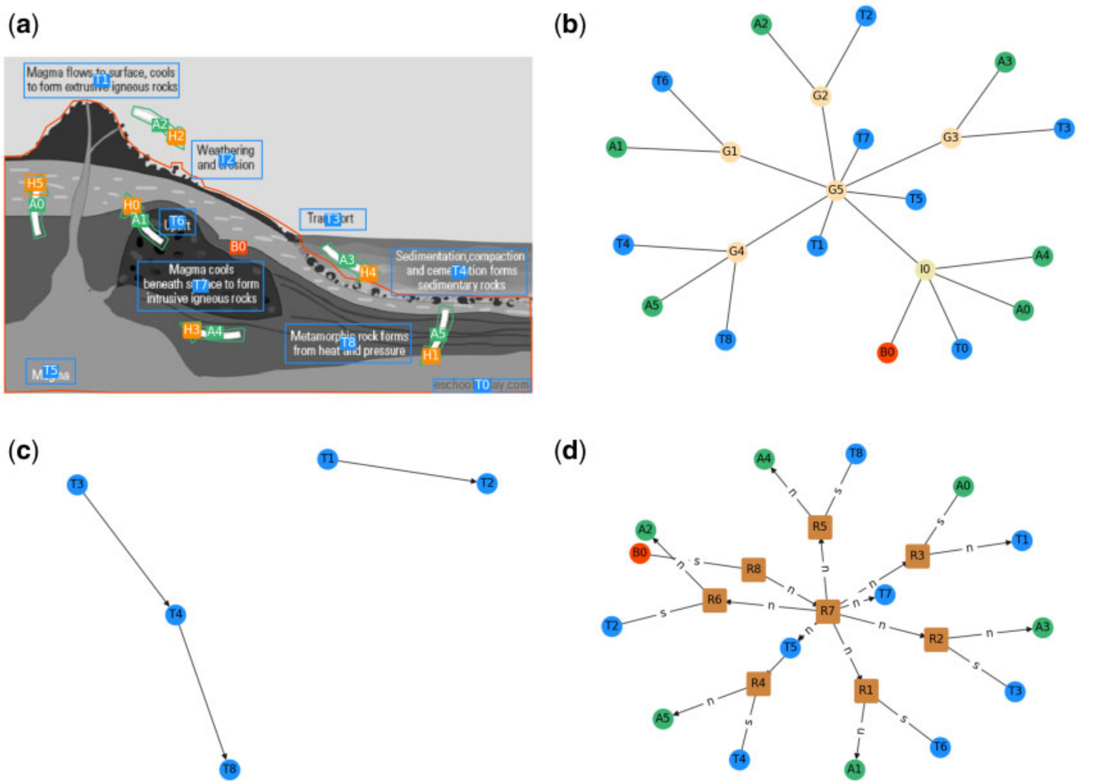


**Fig. 5** Graph-based representations of diagram structure in AI2D-RST. Note that AI2D-RST does not model arrowheads as diagram elements, as this information can be retrieved from the original AI2D annotation as required. The graphs have been created using the visualization tools developed by Hiippala *et al.* (2020). Node positions in each graph are not meaningful, but calculated by an algorithm in the NetworkX library (Hagberg *et al.*, 2008). (**a**) Layout segmentation. AI2D-RST uses the layout segmentation in AI2D to populate its inventory of analytical units. (**b**) Grouping graph. Grouping nodes with prefix G join nodes that are likely to be perceived as belonging together. I0 is used as the root node. (**c**) Connectivity graph. The edges of the graph indicate which nodes are explicitly connected to others using arrows or lines. (**d**) Discourse structure graph. Discourse relations are represented by nodes prefixed with R. Edges encode relative statuses of the participating nodes (n: nucleus; s: satellite).

To begin, Fig. 5a shows the AI2D layout segmentation which provides the inventory of diagram elements for AI2D-RST. Figure 5b then shows the grouping graph that collects diagram elements that are likely to be perceived as belonging together into visual perceptual groups loosely based on Gestalt principles of visual perception (Ware, 2012). This results in a hierarchical tree graph where grouping nodes, identified by the prefix 'G', are added to the graph as parents for nodes that are grouped together during annotation. These grouping nodes can then be picked up in subsequent annotation layers to refer to a group of diagram elements. In this way, the grouping layer provides a foundation for describing both connectivity and discourse structure layers (Hiippala *et al.*, 2020, p. 7–8).

Figure 5c shows the connectivity layer, represented using a cyclic graph whose edges represent *visually explicit* connections that are signalled using arrows and lines and which hold between diagram elements and their groups as defined in the AI2D-RST grouping layer (Hiippala *et al.*, 2020, p. 9). Because the original rock cycle diagram features several arrows without explicit sources and targets (see e.g. arrows A0, A1, and A5 in

Fig. 3), the connectivity graph in fact fails to reflect the cyclic nature of the phenomenon represented in the diagram. This goes back to the insufficient segmentation of the cross-section blob B0, which was already identified as problematic when applying the original schema proposed in AI2D to the same diagram: the arrows connect sub-regions of the cross-section (see e.g. A0 and A1), but these sub-regions are not available in the inventory of diagram elements used to populate the connectivity graph. In the case of both AI2D and AI2D-RST, the problem originates in the stratum of expressive resources and their segmentation into analytical units. Some visual expressive resources may require more fine-grained decomposition than others, but this information becomes available only when these resources are considered as a part of the discourse structure they participate in.

This brings us to the final layer for discourse structure shown in Fig. 5d, which uses RST to describe semantic relations between diagram elements as suggested above. The discourse relations defined in RST are intended to capture the communicative intentions of the designer, as judged by an analyst, and are added to the discourse structure graph as nodes prefixed with the letter 'R'; the edges of the graph describe which role an element takes in the discourse relation, namely nucleus ('n') or satellite ('s'). The notion of *nuclearity* is a key criterion in definitions of semantic relations in RST and distinguishes between units of a 'text' that are considered central to the argument unfolding and units that play only a supporting role. Following the original RST definitions, AI2D-RST represents the discourse structure layer using a tree graph (Hiippala *et al.*, 2020, p. 10–13).

The specific rhetorical relations shown in the discourse graph of Fig. 5d include identification (R1–R6), cyclic sequence (R7), and background (R8). Since AI2D-RST relies on the inventory of diagram elements provided by the original layout segmentation in AI2D—a decision motivated by the need to avoid annotating everything from scratch and to maintain compatibility with AI2D—the description of discourse semantics in AI2D-RST requires several compromises that will become evident below. Here, the annotator had concluded that most text instances serve to identify what the arrows stand for, namely stages of the rock cycle. The image showing the cross-section (B0), in turn, is placed in a background relation to the cyclic sequence relation. The definition of a background relation (Mann and Thompson, 1988) states that the satellite (B0) increases the ability to understand the nucleus (R7), which is the top-level relation assigned to the diagram's representation of the entire cycle.

The analysis in Fig. 5d remains, however, a rather crude description of the discourse structure of the diagram of Fig. 3, because the cross-section actually provides for more information than is captured by the AI2D layout segmentation, which assigns the entire cross-section to the blob B0. As the preceding discussion of the connectivity layer argued, insufficient analytical granularity results in an incomplete inventory of diagram elements. The information contained in these sub-regions is crucial for understanding what the diagram is attempting to communicate *but we cannot know that such a decomposition into sub-regions is necessary without considering the discourse organization of the entire diagram first.* This demonstrates why the decomposition of diagrams should be pursued in a top-down direction and be guided by the discourse structure (Carberry *et al.*, 2003; Bateman and Wildfeuer, 2014). This means that, even when supported by an annotation framework with sufficient local and global reach in terms of discourse structure, resulting descriptions are unlikely to be adequate unless the inventory of analytical units has been appropriately populated.

## 3.3 Layout across diagram categories

With two sizeable corpora with rich annotations at hand, we now turn to examine how decisions related to the metadata scheme in the AI2D dataset are propagated to the AI2D-RST corpus. As pointed out above, the two corpora differ in terms of the metadata schemes used to describe the diagrams, and this difference may be traced back to the research interests of artificial intelligence and multimodality research, and the assumptions these fields make about the sensory and semiotic nature of multimodality as a phenomenon (see Section 2).[2]

We begin by focusing on *layout*, a key expressive resource in the diagrammatic mode, whose importance has been underlined in both artificial intelligence and multimodality research. Research on graphic and information design has long acknowledged that document layout supports access to discourse structure, which raises the question whether layout plays a similar role in diagrams, and if so, to what extent (Waller, 2012). André and Rist (1995, p. 149) observe that 'layout has to be considered as an important carrier of meaning' in diagrams, because it generates hypotheses about their discourse structure (see also Watanabe and Nagao, 1998). For the same reason, layout should be considered a prime target for distant viewing, as the information on expressive resources and their placement in the layout is readily available in the AI2D layout segmentation. In short, layout information may provide valuable clues about the communicative goals of the diagrams.

Although both corpora share the same layout segmentation, AI2D and AI2D-RST differ in terms of how they categorize the diagrams. AI2D assigns diagrams into seventeen semantic categories that mainly correspond to the subject matter of the diagram, as exemplified by *volcano* and *rock strata*, whereas AI2D-RST defines eleven structural categories that represent abstract diagram types, such as *network*, *cycle*, and *cross-section*. These structural categories attempt to capture patterned ways of combining expressive resources made available by the diagrammatic mode that apply independently of particular subject matters. In multimodality research, such patterns are often attributed to genre, a higher-level phenomenon that generates expectations towards the content and structure of multimodal discourse by providing structural cues that evoke previous encounters with similar communicative situations and artefacts (Lemke, 2005; Bateman, 2008; Hiippala, 2014). Figure 6 shows how the 1,000 diagrams in AI2D-RST are mapped to structural categories in AI2D-RST (left) and semantic categories in AI2D (right).

For the purposes of distant viewing, the semantic and structural categories in AI2D and AI2D-RST can be treated as different metadata schemes that can impose structure on the observations made for expressive resources and their placement in the layout space. To explore differences between these schemas, we retrieve the diagrams from each semantic and structural category and calculate the spatio-visual centroid of each blob, line, and text element. Because the diagrams are of different size and orientation, we normalize the horizontal and vertical coordinates for each centroid by dividing the coordinates by the width and height of the original diagram image. We then use multivariate kernel density estimation (KDE) to estimate probabilities for the position of diagram element centroids along the horizontal and vertical axes that demarcate the layout space.

Figure 7 shows KDE for four semantic categories in the AI2D dataset: *parts of*, *life cycles*, *rock strata*, and *volcano*. The KDE plot for the category *parts of* in Fig. 7a reveals that the centroids of blob elements occur with high probability in the middle of the diagram, whereas the centroids of line and text elements are distributed evenly on the sides. In this pattern, the different elements occupy distinct areas of the layout, which suggests that the objects under description are positioned in the middle and their parts are picked out using lines and written labels positioned along the outer edges. Interestingly, *parts of* is one of the few categories in AI2D that is not strongly aligned with a specific subject matter, but represents a generic category that includes diagrams concerned with various topics.

In the category of *life cycles*, shown in Fig. 7b, the centroids for blobs occur with higher probability in areas positioned on top, bottom, left, and right of the diagram. These areas are likely to stand for particular stages of the life cycle. In contrast to the category *parts of* in Fig. 7a, blobs, lines, and text do not occupy distinct areas in the layout of *life cycles*. Conversely, the patterns for *rock strata* and *volcano* in Fig. 7c and d, respectively, bear some similarities to the pattern for *parts of*, but seem more variable in their positioning of diagram elements, as reflected by the spatial pattern and lower values for the probability density estimates.

Figure 8 shows patterns for four of the structural categories defined in AI2D-RST. The pattern for *cycle* (Fig. 8a) suggests that structural categories in AI2D-RST can bring out regularities in how diagrams use the layout space regardless of their subject matter. As the alluvial plot in Fig. 6 showed, the diagrams in the AI2D-RST category *cycle* are mapped to AI2D categories *life cycles*, *moon phases*, *rock cycle*, and *water cycle*. What is particularly striking about the layout pattern for *cycle* is that four major positions may be identified for blobs—namely top, bottom, left, and right—whereas minor positions are visible between them. Because the category *cycle* in AI2D-RST includes diagrams that use layout or arrows and lines to set up cyclic connections between entities (Alikhani and Stone, 2018), the fixedness of spatial positions for blobs is particularly interesting. In other words,

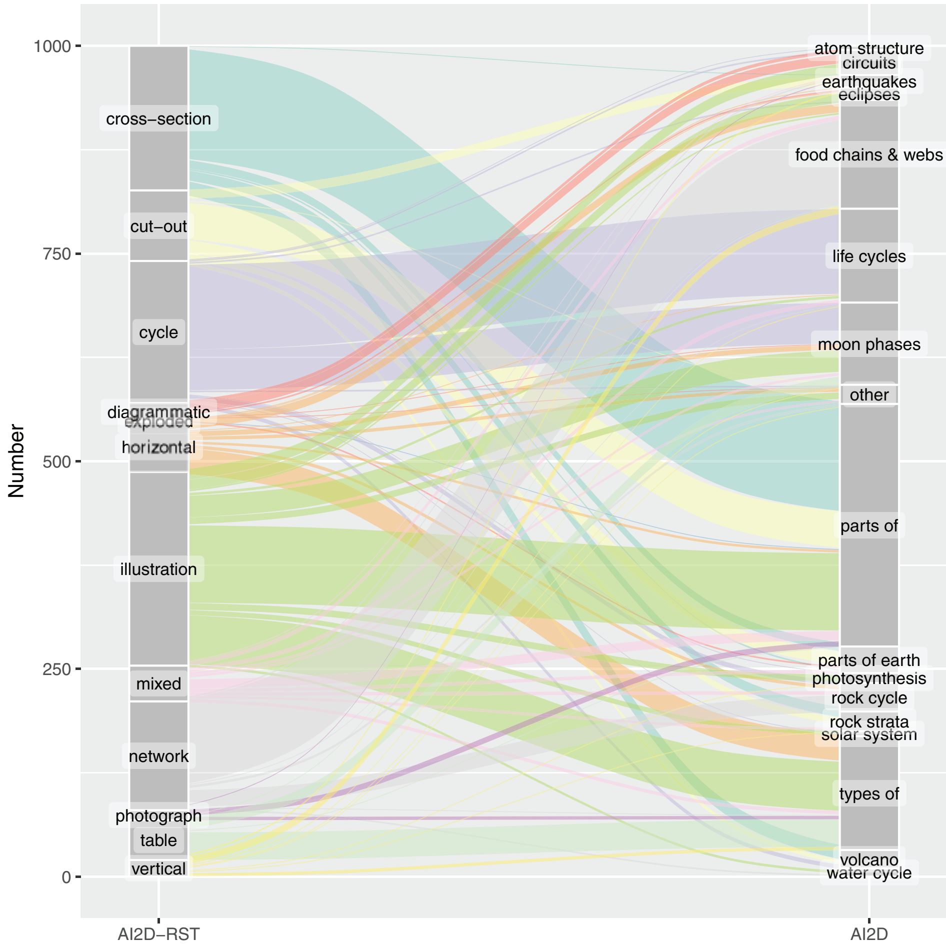


**Fig. 6** An alluvial plot mapping the structural categories in AI2D-RST to the semantic categories in AI2D. *Y*-axis shows the number of diagrams in each category. Diagrams in AI2D-RST that combine multiple categories are labelled as 'mixed'. Created using *ggalluvial* 0.12.3 for R 4.0.2.

populating these spatial positions with blobs may evoke an association with cycles, thus operating as a genre cue that encourages the viewer to consider whether such a discourse semantic interpretation holds.

The remaining examples in Fig. 8 show layout patterns for three further structural categories in AI2D-RST concerned with how diagrams represent their depicted objects (Hiippala *et al.*, 2020, p. 8–9). These include *illustration* in Fig. 8b, which covers all forms of depiction at various levels of visual detail from monochrome to colour drawings. *Illustration* is distinguished from *cross-section* in Fig. 8c and *cut-out* in Fig. 8d based on whether the internal structure of the depicted object is shown by cutting the object in half (*cross-section*) or by removing a part of the object to expose its structure (*cut-out*). As the layout patterns show, *illustrations* are more flexible in their positioning of blobs than *cross-sections* and *cut-outs*, as *illustrations* occasionally depict multiple objects in a single diagram, which causes the blob centroids to spread out. The layout patterns for *cross-sections* and *cut-outs*, in turn, cannot be distinguished from one another based on layout information alone.

This exposes certain limitations of the metadata scheme used to describe diagram elements in the

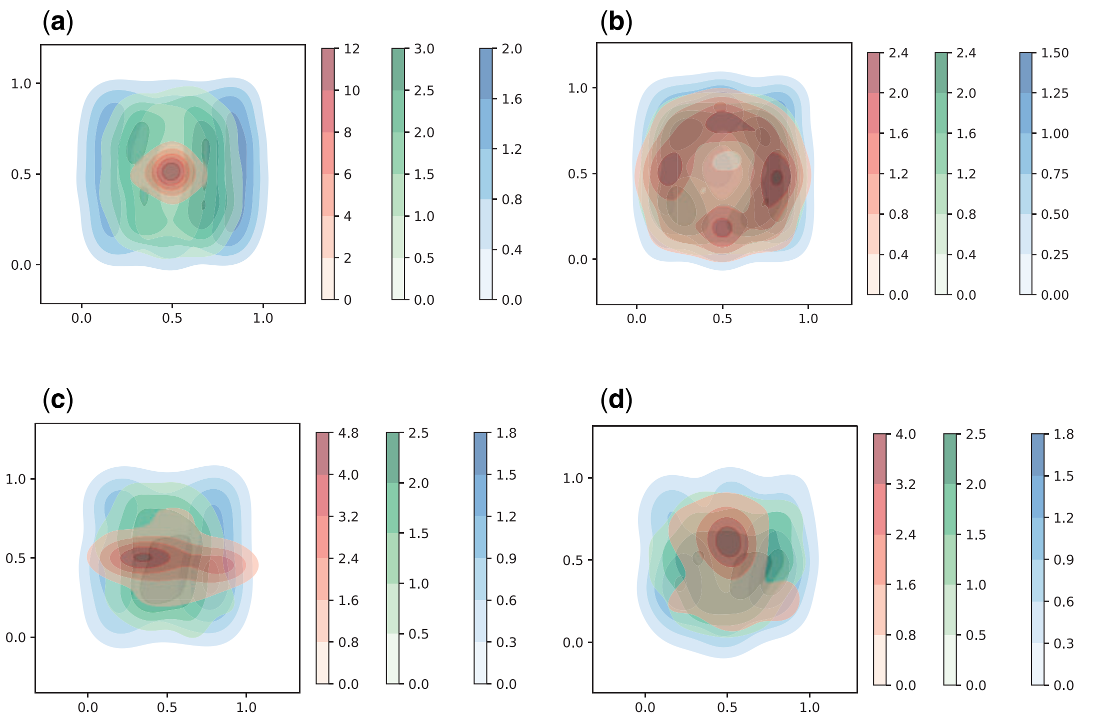


**Fig. 7** KDEs for the centroids of text (blue), arrow/line (green), and blob (red) elements for four semantic categories in the AI2D dataset. The coloured bars to the right of the graphs map the colours to the values of probability density function for each diagram element type. Created using *matplotlib* 3.3.0 (Hunter, 2007) and *seaborn* 0.10.1 for Python 3.8.5. (**a**) Parts of. (**b**) Life cycles. (**c**) Rock strata. (**d**) Volcano.

AI2D dataset, which may be traced back to defining the expressive resources *ahead* of actual analysis (see Section 3.1). Because AI2D classifies all forms of depiction as 'blobs', we cannot determine whether categories in AI2D-RST such as *illustrations*, *cycles*, and *cut-outs* prefer to draw on different expressive resources—for example coloured hand-drawn illustrations or monochrome line drawings—for depicting objects and their structure. Put differently, forms of depiction must be described more accurately to identify which expressive resources are being deployed and for which communicative purposes in each diagram category. Capturing these distinctions is crucial because the diagrammatic mode uses such expressive resources to adjust diagrams' levels of abstraction (Dimopoulos *et al.*, 2003). To exemplify, an animal may be represented by a round circle in a *network* diagram showing its role in a food web, whereas an *illustration* is more likely to use a lifelike drawing to portray the same animal. Distinguishing between these representations was not possible because in AI2D and AI2D-RST both are classified as 'blobs'. In the following section, we explore the diversity of expressive resources constituting the category of 'blobs' using computer vision methods.

## 3.4 Unpacking the expressive resources in 'blobs'

We now offer a more fine-grained description of the expressive resources collectively labelled as 'blobs' in the AI2D dataset. To explore which expressive resources are used for depiction in *illustrations*, *cycles*, and *cut-outs*, we extract all diagram elements classified as blobs from the AI2D dataset ($N$ = 20,937) and apply the method presented in Fig. 9 to characterize their visual appearance in terms of brightness and texture.

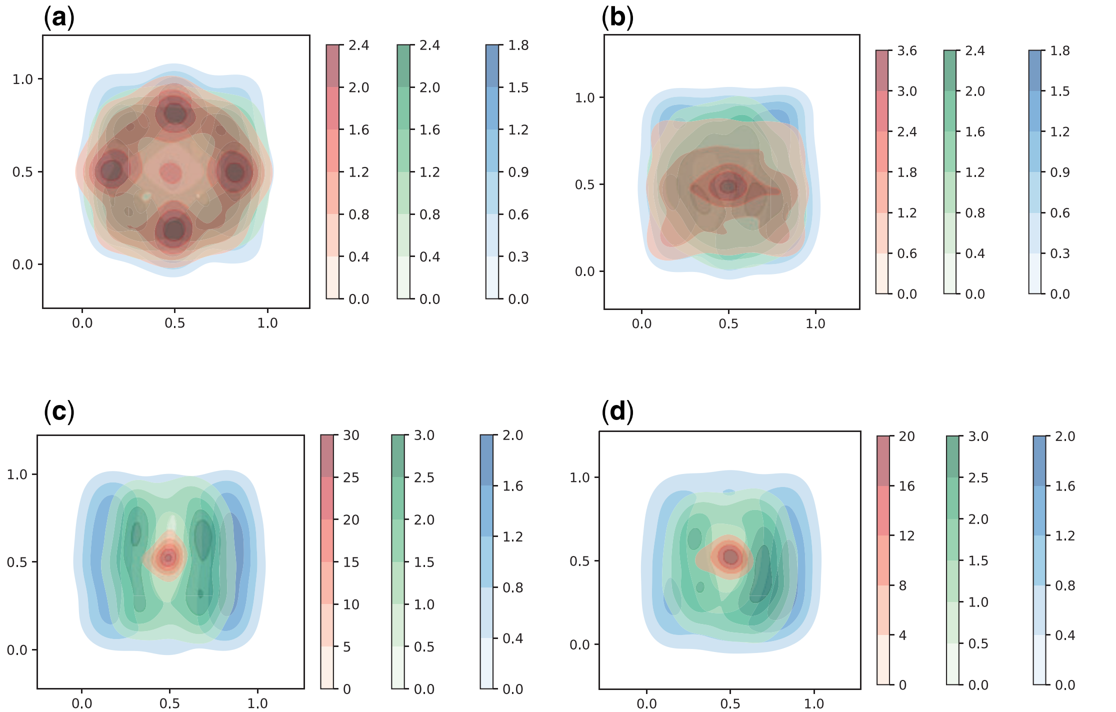

**Fig. 8** KDEs for the centroids of text (blue), arrow/line (green), and blob (red) elements for four structural categories in the AI2D-RST corpus. The coloured bars to the right of the graphs map the colours to the values of probability density function for each diagram element type. Created using *matplotlib* 3.3.0 (Hunter, 2007) and *seaborn* 0.10.1 for Python 3.8.5. (**a**) Cycle. (**b**) Illustration. (**c**) Cross-section. (**d**) Cut-out.

We first convert each blob to greyscale. Each pixel in a greyscale image is represented by a value between 0 and 255, which encodes its brightness: 0 stands for black, whereas 255 stands for white. We then describe the brightness of the entire blob by calculating a greyscale histogram with sixty-four bins, each of which covers a range of values. To exemplify, the first bin (out of sixty-four) covers values from 0 to 3. If the value of a pixel falls within this range, the value for the first bin increases by one. Distributing all the pixels across the sixty-four bins provides a sixty-four-dimensional vector that describes the brightness of the blob.

We then extract uniform Local Binary Patterns (LBPs; Ojala *et al.*, 1996) to represent the texture of each blob using the *scikit-image* library for Python (van der Walt *et al.*, 2014). Uniform LBP examines the neighbourhood of each pixel within a prespecified window and encodes information about that pixel neighbourhood using binary values: if the value of a neighbouring pixel is lower than the value of the current pixel, the neighbour receives a value of 0. Conversely, if the value is larger, the neighbouring pixel receives a value of 1. Uniform LBP collects this information into a vector of zeros and ones. This vector is then quantified by counting the number of transitions from 0 to 1 and 1 to 0. The number of transitions is aggregated into a histogram to describe the distribution of binary patterns in the image. We use LBP to examine the twenty-four neighbouring pixels positioned within a radius of three pixels from the centre pixel; this provides a twenty-six-dimensional vector for each blob. Finally, we concatenate the sixty-four-dimensional vector for brightness and the twenty-six-dimensional vector for texture into a ninety-dimensional feature vector. These features were extracted for each blob in the AI2D dataset.

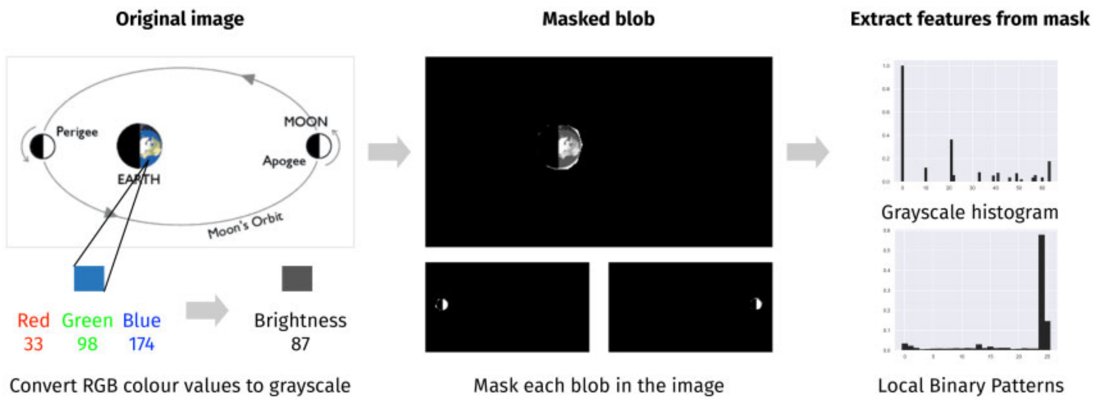


**Fig. 9** Extracting features for brightness and texture from 'blobs' in the AI2D dataset.

We then project the ninety-dimensional feature vectors into a two-dimensional space for visual exploration. To do this, we use the Uniform Manifold Approximation and Projection (UMAP) algorithm to learn a mapping between the ninety- and two-dimensional feature spaces (McInnes *et al.*, 2018). UMAP is controlled by two parameters: nearest neighbours and minimum distance. We set nearest neighbours to 200, which seeks to emphasize global over local structure when determining neighbours in the high-dimensional space, and minimum distance to 0.99, to allow loose clustering of points in the low-dimensional space preserving the broad topological structure of the high-dimensional space. Figure 10 shows the resulting visualization, which plots the two-dimensional UMAP features learned for each blob against each other along the two dimensions. Each blob is represented by its thumbnail image.

Figure 10 reveals that the category of 'blobs' covers a wide range of expressive resources. One dimension of variation may be identified along the horizontal axis, which encodes colour information: coloured drawings gradually turn into monochrome line drawings when moving from left to right. Another dimension of variation exists along the vertical axis: blobs with solid colours and texture are positioned along the outer edges, while the level of visual detail increases towards the middle. Although UMAP preserves enough information about colour and texture to capture colour differentiation (the number of colours) and modulation (the range of colour shades), Fig. 10 is not sufficient for distinguishing between modes of depiction, for example whether the blob represents an illustration, cross-section or cut-out and *which* expressive resource is mobilized for depiction.

To better understand how the expressive resources in Fig. 10 are associated with particular types of depiction captured by the structural categories in AI2D-RST (e.g. *illustration*, *cross-section*, and *cut-out*), Fig. 11 shows three hex plots that show the specific distributions of blob centroids along the horizontal (colour) and vertical (visual detail) dimensions for these three structural categories. In terms of colour, Fig. 11a reveals that blobs in *illustrations* are distributed evenly along the horizontal dimension, but peak in the region corresponding to monochrome line drawings. For *cross-sections* shown in Fig. 11b, blobs are more likely to occur in the region for drawings with colour gradients along the horizontal axis, but some can also be found in the region for black and white line drawings. Blobs from *cut-outs*, in turn, are very likely to be found in the region for coloured drawings, as shown in Fig. 11c. In terms of visual detail, blobs from all three categories peak around the region for visual detail, but especially strongly for *cut-outs*.

*Cut-outs* clearly prefer coloured drawings with the rich visual detail needed to depict three-dimensional objects and their structure. This suggests that Fig. 11 reflects functionally motivated choices related to expressive resources that enable visual depiction. In other words, the way objects are being depicted influences the choice of expressive resources. Notably, *cross-sections* draw on both coloured and

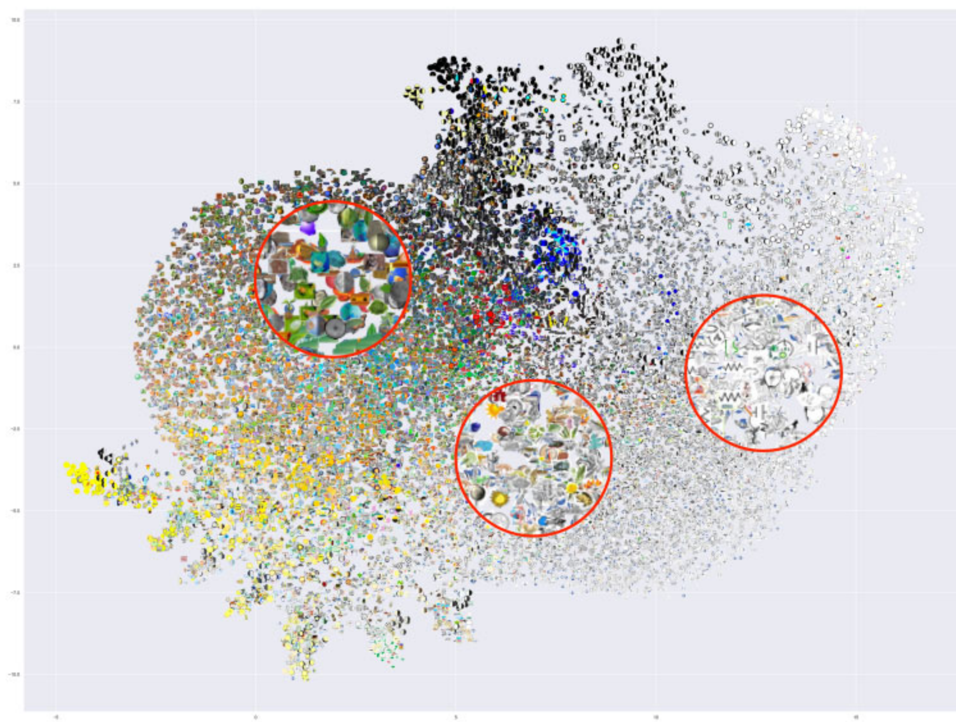

**Fig. 10** Mapping the ninety-dimensional feature space for grey histogram and LBPs to two dimensions for plotting using the UMAP dimensionality reduction algorithm. Each blob is represented by its thumbnail. The three loupes demarcated in red zoom into different regions of the plot. Created using the *matplotlib* 3.3.0 (Hunter, 2007) and *seaborn* 0.11.0 libraries for Python 3.8.5.

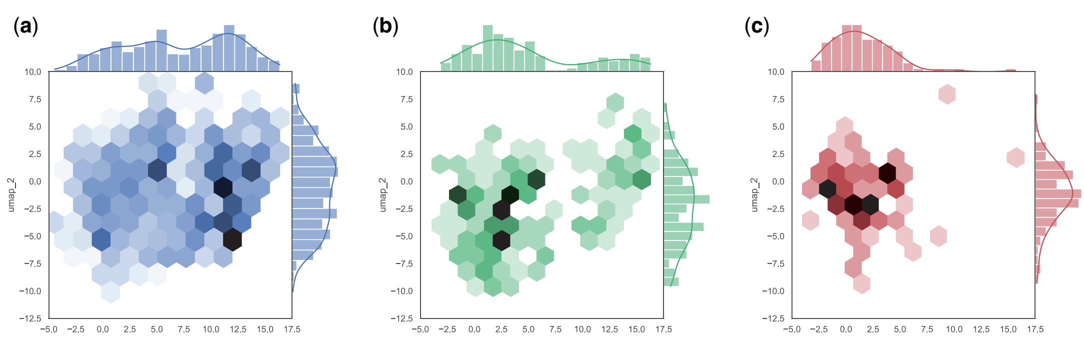


**Fig. 11** The distribution of blob centroids across three structural categories in AI2D-RST. The marginal plots show histograms with 20 bins and a KDE for each UMAP dimension. Created using the *matplotlib* 3.3.0 (Hunter, 2007) and *seaborn* 0.11.0 libraries for Python 3.8.5. (**a**) Illustration. (**b**) Cross-section. (**c**) Cut-out.

monochrome drawings, which are both suitable for depicting the internal structure of objects from a side view. *Illustrations*, in turn, are more flexible in terms of the choice of expressive resources that they may draw on for depiction. These results are generally aligned with the findings of Dimopoulos *et al.* (2003,

p. 200), who manually annotated over 2,800 primary school science diagrams for four features that affect the degree of abstraction: (1) the presence of geometrical shapes and symbols, (2) the variety of colours used, (3) their range of shades, and (4) contextualization, that is, whether the depiction uses an illustrated or plain colour background. Given that our method produces similar insights without using annotations for these particular features suggests that the generic metadata schema indeed captures key characteristics of the diagrammatic mode in this domain, such as the selection of expressive resources, which may then be linked back to particular communicative needs.

## 4 Discussion

The results of the analysis of diagram layouts in Section 3.3 emphasize the flexibility of the diagrammatic mode: the same structural configuration can be used to realize diagrams that deal with different topics, as long as the phenomena represented share certain common features (cf. Fig. 6). We characterized these structural configurations in terms of genre, which refers to high-level patterns of expressive resources (cf. Bateman, 2008). Our results show that these genre-based structural categories appear to correspond to high-level layout patterns, as shown in Fig. 8. However, Hiippala *et al.* (2020, p. 21–22) observe that genre patterns are fluid and a single diagram may incorporate high-level patterns associated with multiple genres. Therefore, the layout patterns in Fig. 8 should not be seen as representing fixed templates for diagram design, but reflect instead conventional ways of organizing expressive resources in the 2D layout space which are motivated by communicative functions. The degree to which these patterns are conventionalized varies from one genre to another.

In terms of the expressive resources analysed in Section 3.4, our analysis has revealed just how much of the variation is hidden when an under-differentiating label such as 'blob', 'image', or 'visual' is used. The results also underline the difference between photographic images and other means of visual depiction, which have radically different capabilities for representation. Diagrams regularly draw on expressive resources capable of visual depiction that allow their representational 'accuracy' to be adjusted to match their communicative needs (cf. Dimopoulos *et al.*, 2003; Greenberg, 2018). Moreover, in terms of methodology, it is worth noting that extracting features from blobs using a convolutional neural network pretrained on *photographs* did not produce meaningful UMAP clusters. Wevers and Smits (2020, p. 200) report on similar experiences when applying pretrained neural networks to historical newspapers, where multiple semiotic modes capable of visual depiction are used alongside photographs. This suggests that many computer vision techniques that work well on photographs may not be directly transferrable to other forms of depiction.

Generally, our results underline the importance of paying attention to modelling particular expressive resources when building corpora for specific semiotic modes. Without describing the expressive resources at a sufficient degree of accuracy and granularity, the analysis will not be able to capture what the semiotic mode does with its material substrate, nor to provide an inventory of discourse units needed to represent its discourse semantics appropriately. Just which expressive resources are made available by a semiotic mode must be answered through empirical research. As Thomas (2014) notes, interpretation—in this case, concerning how to best describe the expressive resources within a semiotic mode—should be motivated by observations made in the data and delayed until unavoidable. The method proposed in Section 3.4 can be used to support the process of interpretation by providing a bird's-eye view to expressive resources in a corpus.

These results demonstrate that the concept of a semiotic mode can be applied productively to creating corpora and training data for digital humanities and artificial intelligence research because it adds specifically significant structures over the measured data. There is also much potential in using the stratum of discourse semantics as the basis for defining crowdsourcing tasks alongside common strategies adopted in computer vision research (cf. Kovashka *et al.*, 2016). Rather than requesting the workers to identify and place elements into predefined categories, the crowdsourcing tasks could be used to tease out discourse interpretations related to the data and convert them into descriptions of discourse structures. To exemplify, if a diagram features an element that consists of written text, it would be natural to ask whether the text refers to a particular entity in the diagram. This would allow aspects of the inherently dynamic nature of discourse interpretations to be injected into the annotation tasks.

Finally, we argue that our analysis shows that pre-theoretical distinctions between 'text' and 'image' rarely hold in multimodal communication, and they are seriously under-differentiating for the digital humanities or any other field concerned with multimodality (Bateman, 2014; Bateman *et al.*, 2017). While we fully agree with Arnold and Tilton (2019) on the need to develop methods that enable the large-scale analysis of various media, we have demonstrated here that this effort must be supported by a solid theoretical foundation that reveals rather than hides the complexity of multimodal communication (Bateman, 2017; Hiippala, 2021). This foundation is needed for tackling issues that are traditionally of concern to the humanities, such as trajectories of change over time, whose computational analysis is still largely limited to linguistic material.

## 5 Conclusion

In this article, we argued that contemporary theories of multimodality can inform computational approaches to studying multimodal communication in the field of digital humanities and beyond. By analysing two multimodal corpora consisting of primary school science diagrams, we showed how multimodality theory can reveal descriptive shortcomings that lead to analytical blind spots, which become increasingly pronounced when using computational methods as advocated by distant viewing (Arnold and Tilton, 2019). However, Arnold and Tilton (2019, p. i13) acknowledge that the framework of distant viewing cannot specify what kinds of metadata schemes are needed to describe the data in a way that support the optimal use of computational methods, but argue that developing such metadata schemes should constitute a major area of research in digital humanities.

We propose that any effort to define metadata schemes for visual and multimodal materials can be informed by multimodality theory, which allows for a semiotically appropriate treatment of diverse media and the semiotic modes they deploy. This calls for increased attention to the communicative goals set for the artefact or situation under analysis, as these determine the extent to which individual semiotic modes must be decomposed into analytical units to achieve a sufficiently coherent description of multimodal discourse. Producing such descriptions for a wide range of historical and contemporary media will require a large-scale effort, but at the same time, grounding the analysis in semiotics offers a far stronger basis for addressing research questions that are traditionally of concern to humanities, while continuing to leverage the power of computational methods for finding patterns in large volumes of data.

## Notes

1. https://www.mturk.com.
2. The code used for analysis is available at https://doi.org/10.5281/zenodo.4761066.

## References

**Alikhani, M. and Stone, M.** (2018). Arrows are the verbs of diagrams. In *Proceedings of the 27th International Conference on Computational Linguistics*, Santa Fe, New Mexico, USA, pp. 3552–63.

**André, E. and Rist, T.** (1995). Generating coherent presentations employing textual and visual material. *Artificial Intelligence Review*, **9**: 147–65.

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

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

**Bateman, J. A.** (2011). The decomposability of semiotic modes. In O'Halloran, K. L. and Smith, B. A. (eds), *Multimodal Studies: Multiple Approaches and Domains*. London: Routledge, pp. 17–38.

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

**Bateman, J. A.** (2017). Multimodale semiotik und die theoretischen grundlagen der digital humanities. *Zeitschrift für Semiotik*, **39**(1–2): 11–50.

**Bateman, J. A.** (2020). The foundational role of discourse semantics beyond language. In Zappavigna, M. and

Dreyfus, S. (eds), *Discourses of Hope and Reconciliation. On J. R. Martin's Contribution to Systemic Functional Linguistics. Routledge Studies in Multimodality*. London: Bloomsbury, pp. 39–55.

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

**Bateman, J. A., Tseng, C., Seizov, O.**, et al. (2016). Towards next generation visual archives: Image, film and discourse. *Visual Studies*, **31**(2): 131–54.

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

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

**Burghardt, M., Heftberger, A., Pause, J., Walkowski, N.-O. and Zeppelzauer, M.** (2020). Film and video analysis in the digital humanities—an interdisciplinary dialog. *Digital Humanities Quarterly*, **14**(4).

**Carberry, S., Elzer, S., Green, N., McCoy, K. and Chester, D.** (2003). Understanding information graphics: A discourse-level problem. In *Proceedings of the Fourth SIGdial Workshop of Discourse and Dialogue*. Available from: https://www.aclweb.org/anthology/W03-2101.

**Cobley, P.** (2013). Semiotic models of communication. In Cobley, P. and Schulz, P. J. (eds), *Theories and Models of Communication*. Berlin and Boston: De Gruyter Mouton, pp. 223–40.

**DeRose, S.** (2020). What is a diagram, really? In *Proceedings of Balisage: The Markup Conference 2020', Vol. 25 of Balisage Series on Markup Technologies*. Washington, DC, July 27–31.

**Dimopoulos, K., Koulaidis, V. and Sklaveniti, S.** (2003). Towards an analysis of visual images in school science textbooks and press articles about science and technology. *Research in Science Education*, **33**: 189–216.

**Engelhardt, Y.** (2002). *The Language of Graphics: A Framework for the Analysis of Syntax and Meaning in Maps, Charts and Diagrams*, PhD thesis, Institute for Logic, Language and Computation. Netherland: University of Amsterdam.

**Greenberg, G.** (2018). Content and target in pictorial representation. *Ergo*, **5**(33): 865–98.

**Haehn, D., Tompkin, J. and Pfister, H.** (2019). Evaluating 'graphical perception' with CNNs. *IEEE Transactions on Visualization and Computer Graphics*, **25**(1): 641–50.

**Hagberg, A., Swart, P. and Schult, D.** (2008). Exploring network structure, dynamics, and function using NetworkX. In *Proceedings of the 7th Python in Science Conference*. pp. 11–15.

**Heftberger, A.** (2018). *Digital Humanities and Film Studies: Visualising Dziga Vertov's Work*. Cham: Springer.

**Hiippala, T.** (2014). Multimodal genre analysis. In Norris, S. and Maier, C. D. (eds), *Interactions, Images and Texts: A Reader in Multimodality*. Berlin and New York: De Gruyter Mouton, pp. 111–123.

**Hiippala, T.** (2015). *The Structure of Multimodal Documents: An Empirical Approach*. New York: Routledge.

**Hiippala, T.** (2021). Distant viewing and multimodality research: Prospects and challenges. *Multimodality & Society*, **1**(2).

**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).

**Hiippala, T. and Bateman, J. A.** (2020). Introducing the diagrammatic mode. *arXiv* 2001.11224. Available from: https://arxiv.org/abs/2001.11224.

**Huang, L.** (2021). Toward multimodal corpus pragmatics: Rationale, case, and agenda. *Digital Scholarship in the Humanities*, **36**(1): 101–14.

**Hunter, J.** (2007). matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, **9**(3): 90–5.

**Jewitt, C.** (2008). Multimodality and literacy in school classrooms. *Review of Research in Education*, **32**(1): 241–67.

**Kembhavi, A., Salvato, M., Kolve, E., Seo, M. J., Hajishirzi, H. and Farhadi, A.** (2016). A diagram is worth a dozen images. In *Proceedings of the 14th European Conference on Computer Vision (ECCV 2016)*, Cham: Springer, pp. 235–51.

**Kostelnick, C. and Hassett, M.** (2003). *Shaping Information: The Rhetoric of Visual Conventions*. Carbondale: Southern Illinois University Press.

**Kovashka, A., Russakovsky, O., Fei-Fei, L. and Grauman, K.** (2016). Crowdsourcing in computer vision. *Foundations and Trends in Computer Graphics and Vision*, **10**(3): 177–243.

**Kress, G.** (2011). Partnerships in research: multimodality and ethnography. *Qualitative Research*, **11**(3): 239–60.

**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.** (1996). *Reading Images: The Grammar of Visual Design*. London: Routledge.

**Kress, G. and van Leeuwen, T.** (2001). *Multimodal Discourse: The Modes and Media of Contemporary Communication*. London: Arnold.

**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): 845–56.

**Lemke, J. L.** (2005). Multimedia genres and traversals. *Folia Linguistica*, **39**(1–2): 45–56.

**Mann, W. C. and Thompson, S. A.** (1988). Rhetorical structure theory: Toward a functional theory of text organization. *Text*, **8**(3): 243–81.

**McInnes, L., Healy, J., Saul, N. and Grossberger, L.** (2018). UMAP: Uniform manifold approximation and projection. *The Journal of Open Source Software*, **3**(29): 861.

**Mondada, L.** (2019). Contemporary issues in conversation analysis: Embodiment and materiality, multimodality and multisensoriality in social interaction. *Journal of Pragmatics*, **145**: 47–62.

**Münster, S. and Terras, M.** (2020). The visual side of digital humanities: A survey on topics, researchers, and epistemic cultures. *Digital Scholarship in the Humanities*, **35**(2): 366–89.

**O'Halloran, K. L., Tan, S., Pham, D.-S., Bateman, J. A. and Vande Moere, A.** (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.

**Ojala, T., Pietikäinen, M. and Harwood, D.** (1996). A comparative study of texture measures with classification based on featured distributions. *Pattern Recognition*, **29**(1): 51–9.

**Oviatt, S. and Cohen, P. R.** (2015). *The Paradigm Shift to Multimodality in Contemporary Computer Interfaces*. San Rafael (CA): Morgan & Claypool.

**Purchase, H. C.** (2014). Twelve year of diagrams research. *Journal of Visual Languages and Computing*, **25**(2): 57–75.

**Richards, C.** (2017). Technical and scientific illustration: picturing the invisible. In Black, A., Luna, P., Lund, O. and Walker, S. (eds), *Information Design: Research and Practice*. London: Routledge. pp. 85–106.

**Smits, T. and Ros, R.** (2020). Quantifying iconicity in 940k online circulations of 26 iconic photographs. In *Proceedings of the Workshop on Computational Humanities Research (CHR 2020)*. pp. 375–384.

**Steen, F. F., Hougaard, A., Joo, J.**, et al. (2018). Toward an infrastructure for data-driven multimodal communication research. *Linguistics Vanguard*, **1**: 20170041.

**Stöckl, H.** (2020). Linguistic multimodality—multimodal linguistics: a state-of-the-art sketch. In Wildfeuer, J., Pflaeging, J., Bateman, J. A., Seizov, O. and Tseng, C. (eds), *Multimodality: Disciplinary Thoughts and the Challenge of Diversity*. Berlin: De Gruyter, pp. 41–68.

**Sulubacak, U., Caglayan, O., Grönroos, S.-A.**, et al. (2020). Multimodal machine translation through visuals and speech. *Machine Translation*, **34**(2–3): 97–147.

**Svensson, P.** (2010). The landscape of digital humanities. *Digital Humanities Quarterly*, **4**(1).

**Taboada, M. and Habel, C.** (2013). Rhetorical relations in multimodal documents. *Discourse Studies*, **15**(1): 65–89.

**Thomas, M.** (2014). Evidence and circularity in multimodal discourse analysis. *Visual Communication*, **13**(2): 163–89.

**Thomas, M.** (2020a). Making a virtue of material values: tactical and strategic benefits for scaling multimodal analysis. In Wildfeuer, J., Pflaeging, J., Bateman, J. A., 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).

**Tversky, B., Zacks, J., Lee, P. and Heiser, J.** (2000). *Diagrams 2000: theory and application of diagrams*. In: *Lines, Blobs, Crosses and Arrows: Diagrammatic Communication with Schematic Figures*. Berlin: Springer, pp. 221–230.

**van der Walt, S., Schönberger, J. L., Nunez-Iglesias, J.**, et al. (2014). scikit-image: image processing in Python. *PeerJ*, **2**(e453).

**Waller, R. H. W.** (2012). Graphic literacies for a digital age: The survival of layout. *The Information Society*, **28**(4): 236–52.

**Waller, R. H. W.** (2017). Practice-based perspectives on multimodal documents: Corpora vs connoisseurship. *Discourse, Context & Media*, **20**: 175–90.

**Ware, C.** (2012). *Information Visualization: Perception for Design*. 3rd edn. Amsterdam: Elsevier.

**Watanabe, Y. and Nagao, M.** (1998). Diagram understanding using integration of layout information and textual information. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (ACL'98/COLING'98)*. Association for Computational Linguistics, Montreal, Quebec, Canada, pp. 1374–80.

**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.

**Wildfeuer, J., Pflaeging, J., Bateman, J. A., Seizov, O. and Tseng, C. (eds)** (2020). *Multimodality: Disciplinary Thoughts and the Challenge of Diversity*. Berlin: De Gruyter.