Lang Resources & Evaluation
https://doi.org/10.1007/s10579-020-09517-1

ORIGINAL PAPER

AI2D-RST: a multimodal corpus of 1000 primary school science diagrams

Tuomo Hiippala[1] · Malihe Alikhani[2] · Jonas Haverinen[1] · Timo Kalliokoski[1] · Evanfiya Logacheva[1] · Serafina Orekhova[1] · Aino Tuomainen[1] · Matthew Stone[3] · John A. Bateman[4]

Accepted: 28 October 2020

Abstract This article introduces AI2D-RST, a multimodal corpus of 1000 English-language diagrams that represent topics in primary school natural sciences, such as food webs, life cycles, moon phases and human physiology. The corpus is based on the Allen Institute for Artificial Intelligence Diagrams (AI2D) dataset, a collection of diagrams with crowdsourced descriptions, which was originally developed to support research on automatic diagram understanding and visual question answering. Building on the segmentation of diagram layouts in AI2D, the AI2D-RST corpus presents a new multi-layer annotation schema that provides a rich description of their multimodal structure. Annotated by trained experts, the layers describe (1) the grouping of diagram elements into perceptual units, (2) the connections set up by diagrammatic elements such as arrows and lines, and (3) the discourse relations between diagram elements, which are described using Rhetorical Structure Theory (RST). Each annotation layer in AI2D-RST is represented using a graph. The corpus is freely available for research and teaching.

✉ Tuomo Hiippala
tuomo.hiippala@helsinki.fi

Malihe Alikhani
malihe@pitt.edu

Matthew Stone
mdstone@cs.rutgers.edu

John A. Bateman
bateman@uni-bremen.de

[1] Department of Languages, University of Helsinki, P.O. Box 24, 00014 Helsinki, Finland

[2] School of Computing and Information, University of Pittsburgh, Pittsburgh, USA

[3] Department of Computer Science, Rutgers University, New Brunswick, USA

[4] Faculty 10: Linguistics and Literary Studies, Bremen University, Bremen, Germany

Published online: 05 December 2020

Keywords Diagrams · Multimodality · Rhetorical Structure Theory · Graphs

1 Introduction

Diagrams are a common feature of many everyday media: they can be found everywhere from scientific publications and instruction manuals to newspapers and school textbooks. Barbara Tversky, a cognitive psychologist who has made pioneering contributions to the study of diagrams, observes that their generic purpose is “to structure information to enable comprehension, inference and discovery” (Tversky 2017, p. 350). Due to their widespread use, diagrams have been studied from various perspectives. Previous research has examined their visual perception (Hegarty and Just 1993; Ware 2012), structure and functions (Engelhardt 2002; Purchase 2014; Engelhardt and Richards 2018) and their role as a tool for supporting thinking and reasoning (Tversky 2015) and use in education and instruction (Tippett 2016), to name but a few examples.

In this article, we make a novel contribution to the study of diagrams by presenting AI2D-RST, a corpus of 1000 English-language diagrams that represent topics in primary school natural sciences. The diagrams are described using a new multi-layer annotation schema that seeks to capture their multimodal structure. Our approach to multimodality is linguistically-inspired and semiotically-oriented, that is, we seek to systematically describe how expressive resources such as natural language, illustrations, line art, photographs, lines, arrows and layout are combined in diagrams to make and exchange meanings. To do so, we build on the general framework for multimodal communication proposed in Bateman et al. (2017) and its application to diagrams as set out in Hiippala and Bateman (2020).

The current work is situated within the emerging field of multimodality research, which studies how appropriate combinations of expressive resources emerge in communicative situations (see e.g. Wildfeuer et al. 2020). Despite their growing influence in various fields of study broadly concerned with human communication, many approaches to multimodality remain without adequate empirical support. Although building multimodal corpora is often presented as a solution to this shortcoming due to the success of corpus-based methods in linguistics, developing and applying complex multimodal annotation frameworks requires ample time and resources, and consequently the resulting corpora remain small (Waller 2017; Huang 2020).

AI2D-RST seeks to reduce the need for time and resources and to scale up the volume of data by building multimodally-informed expert annotations on top of pre-existing crowdsourced annotations from the Allen Institute for Artificial Intelligence Diagrams (AI2D) dataset (Kembhavi et al. 2016). The second part of the name, RST, refers to Rhetorical Structure Theory, a theory of discourse structure which we use to describe how diagrams combine multiple expressive resources to fulfil their communicative goals (Mann and Thompson 1988; Taboada and Mann 2006; Hiippala and Orekhova 2018). Overall, the AI2D-RST corpus is intended to serve a dual purpose: to support empirical research on the multimodality of diagrams and their computational processing.

2 Developing multimodal resources for diagrams research

There is a long-standing interest in the computational processing and generation of diagrammatic representations (André and Rist 1995; Watanabe and Nagao 1998; Bateman et al. 2001; Carberry et al. 2003; Bateman and Henschel 2007), which is now resurfacing as recent advances in computer vision and natural language processing are brought to bear on diagrammatic representations (Seo et al. 2015; Sachan et al. 2018, 2019; Choi et al. 2018; Kim et al. 2019; Haehn et al. 2019). Much of this work is driven by research on well-defined tasks such as information retrieval and question answering, whose scope is increasingly extended beyond natural language to cover other modes of expression as well.

Just how these other modes of expression *and* their combinations should be described in order to create multimodal resources that can support further research on multimodality remains an open question. This requires an empirical approach, as creating multimodal resources for modes of expression beyond natural language raises questions about fundamental issues such as segmentation: how to decompose modes of expression such as diagrams into their constituent parts? We have recently argued in Hiippala and Bateman (2020) that any attempt at a systematic description of diagrams must acknowledge the specific characteristics of the *diagrammatic mode*—an abstract system capable of instantiating various types of diagrams appropriate for their context of occurrence (cf. e.g. Bateman and Henschel 2007).

Previous research points at two key characteristics of the diagrammatic mode that need to be accounted for: the use of layout space (Watanabe and Nagao 1998) and their multimodal discourse structure (Carberry et al. 2003), which are often strongly intertwined in multimodal artefacts with a 2D spatial extent, such as entire page-based documents (Hiippala 2013). Firstly, diagrams have a *spatial* organisation in the form of a layout, which can be used to set up discourse relations between instances of expressive resources, including natural language, arrows, lines, illustrations, photographs, line drawings and potentially any resource that may be realised in 2D space (Watanabe and Nagao 1998). How these expressive resources are organised in the layout space can also serve as a strong signal about the purpose and structure of the diagram by generating expectations towards its discourse structure (Holsanova et al. 2009).

This brings us to the second point: diagrams combine expressive resources into *discourse structures*, which must be resolved to make sense of what the diagram in question attempts to communicate. For this reason, Carberry et al. (2003) argue that understanding diagrams should be framed a discourse-level problem, a view that has found support in our recent work on the diagrammatic mode (Hiippala and Bateman 2020). This, however, raises another issue related to segmentation: many theories of discourse assume that discourse segments are identified before determining their interrelations (Grosz and Sidner 1986; Mann and Thompson 1988).

Establishing an inventory of discourse segments for diagrams is a particularly challenging task, as the level of detail needed for segmentation varies from one diagram to another, depending on the combination of expressive resources present and the discourse structures they participate in. To exemplify, a 2D cross-section of

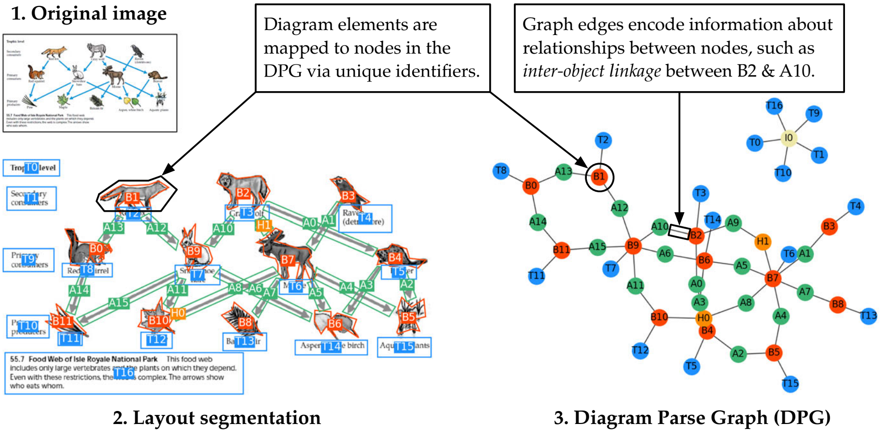

1. Original image
Diagram elements are mapped to nodes in the DPG via unique identifiers.
Graph edges encode information about relationships between nodes, such as *inter-object linkage* between B2 & A10.
2. Layout segmentation
3. Diagram Parse Graph (DPG)
55.7 Food Web of Isle Royale National Park This food web includes only large vertebrates and the plants on which they depend. Even with these restrictions, the web is complex. The arrows show who eats whom.
B1
B2
A10
I0

Fig. 1 (1) A thumbnail of the original diagram image scraped from the web, (2) its crowdsourced layout segmentation (converted into greyscale to bring out the annotation) and (3) a Diagram Parse Graph (DPG) for diagram #274 in AI2D. Diagram element types are coded using same colours in both layout segmentation and DPG: text blocks (blue), blobs (red), arrows (green), arrowheads (orange) and image constant (Navajo white)

an object, whose structure is picked out and described using textual labels, must be decomposed into analytical units to provide a sufficiently accurate description of its multimodal structure, whereas an illustration of an entire object does not need to be decomposed to the same extent (for a detailed discussion of challenges related to diagram segmentation, see Hiippala and Bateman 2020).

Keeping the role of layout and discourse structure in mind, the following sections explicate how we built a new, multimodally-informed annotation schema with multiple layers of description on top of existing crowdsourced annotations for expressive resources and their placement in the diagram layout. To do so, we start by introducing the AI2D dataset, which provided the crowdsourced annotations. We then address certain issues with the AI2D annotation schema before motivating our decision to adopt Rhetorical Structure Theory for describing the discourse structure of diagrams in AI2D-RST.

3 The Allen Institute for Artificial Intelligence Diagrams (AI2D) dataset

The AI2D dataset (Kembhavi et al. 2016)[1] was developed to support research on computational tasks such as automatic diagram understanding and visual question answering (see e.g. Kim et al. 2018). The dataset contains 4903 English-language diagrams that represent topics in primary school natural sciences, such as life cycles, food webs and circuits. Each diagram is assigned to one of 17 semantic categories that correspond to topics in this domain.

[1] The AI2D dataset is publicly available from the Allen Institute for Artificial Intelligence at https://allenai.org/plato/diagram-understanding/ (Accessed September 3, 2020).

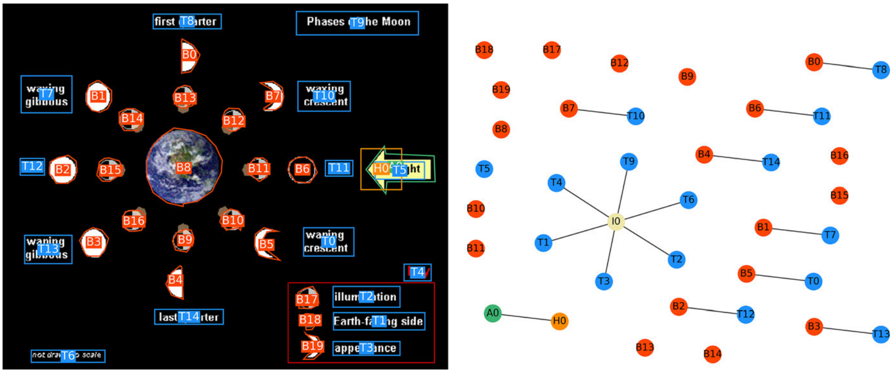


Fig. 2 Layout segmentation (left) and Diagram Parse Graph (DPG, right) for diagram #2728. The numerous disconnections in the DPG result from the lack of relation definitions for describing how *groups* of diagram elements, such as those formed by illustrations of moon phases and their verbal descriptions (e.g. illustration B3 and the text 'waning gibbous' in T13), relate to each other as a part of the global discourse structure of the diagram

Building on Engelhardt's (2002) framework for describing diagrammatic representations, Kembhavi et al. (2016, p. 239) model four types of diagram elements: 'blobs' (e.g. illustrations, line art, photographs and other visual expressive resources), written text, arrows and arrowheads. In addition, Kembhavi et al. (2016) define ten potential relationships that can hold between individual diagram elements, which are also drawn from the framework proposed by Engelhardt (2002). These include, among others, relations such as INTRA-OBJECT LABEL, INTRA-OBJECT LINKAGE and ARROW DESCRIPTOR, which seek to capture how diagram elements relate to each other (for a full list of relations, see Kembhavi et al. 2016, p. 239) AI2D represents diagram structure using a *Diagram Parse Graph* (DPG), in which the nodes stand for diagram elements whereas the edges encode information about the relations that hold between them. For computational tasks, the node features can be populated using word embeddings or visual features extracted using object detectors, depending on the diagram element type in question.

Figure 1 shows the crowdsourced layout segmentation and DPG for diagram 274 in the AI2D dataset. The diagrams were scraped from Google Image Search by using chapter titles in primary school science textbooks (for ages 6–11) as search terms. The annotations were crowdsourced using Amazon Mechanical Turk by breaking down the process of segmenting the layout and constructing a DPG into piecemeal annotation tasks. These tasks involved identifying diagram elements, categorising them and defining their interrelations (Kembhavi et al. 2016, p. 243). Kembhavi et al. (2016, p. 242) report that the 4903 diagrams in AI2D contain approximately 118,000 diagram elements and 53,000 relationships.

Previous research using the AI2D dataset has shown that inferring the meaning of arrows and lines is context-dependent, and the viewers consistently map the arrows to real-world processes (Alikhani and Stone 2018). Hiippala and Orekhova (2018), in turn, consider the AI2D annotation schema from the perspective of multimodality research and argue that DPGs conflate the description of various multimodal structures, such as the visual grouping of diagram elements and connections expressed using arrows and lines. Pulling these structures apart could help understand how diagrams operate multimodally, e.g. whether discourse relations are typically signalled explicitly using arrows and lines or implicitly using the layout space (Watanabe and Nagao 1998; Carberry et al. 2003).

Moreover, the relation definitions drawn from Engelhardt (2002) cover mainly *local* relations between diagram elements, as exemplified by relations such as INTRA-OBJECT LABEL, which is used to describe instances in which one diagram element acts as a label for another. The focus on such local relations between individual diagram elements causes the AI2D annotation schema to fall short in describing the *global* organisation of a diagram, or how larger units formed by multiple diagram elements relate to each other (see Fig. 2).

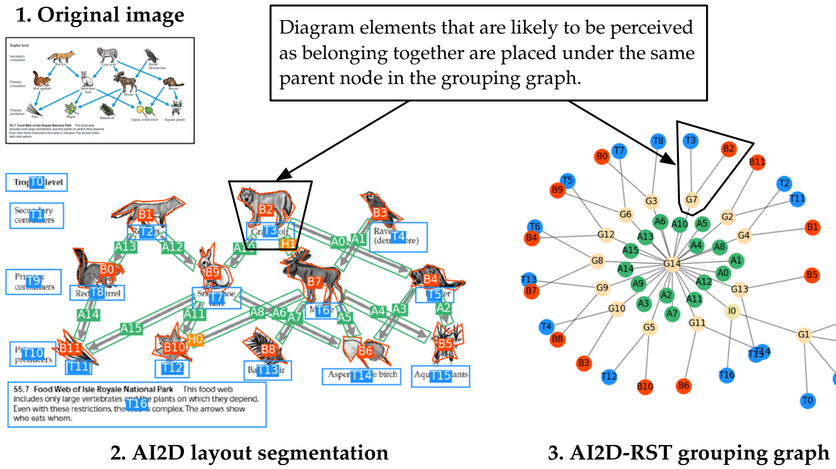
To summarise, the motivation for developing AI2D-RST can be traced back to two observations. First, the limited scope of relation definitions drawn from Engelhardt (2002) in AI2D led us to consider Rhetorical Structure Theory (RST) as an alternative for describing discourse relations in diagrams, given its previous successful applications to multimodal discourse (see e.g. Taboada and Habel 2013; Thomas 2014; Hiippala 2015). However, during the exploratory work reported in Hiippala and Orekhova (2018), it became evident that a direct conversion to RST was not feasible, but required introducing additional annotation layers to establish the units of analysis, as proposed in Bateman (2008).

Second, combining a theory of discourse structure with local and global reach, such as RST, with a multi-layer annotation schema that captures the combinations of expressive resources and their spatial organisation could be used to study whether diagrams signal discourse relations explicitly e.g. using arrows and lines, or whether they are implicit and require the viewers to draw on world knowledge (see also Hiippala and Bateman 2020). Furthermore, access to crowdsourced layout segmentations allows scaling up corpus size. With these two observations in mind, we now turn to describe the AI2D-RST annotation schema and its application to the AI2D diagrams.

4 Developing the AI2D-RST corpus

4.1 The AI2D-RST annotation schema

The AI2D-RST annotation schema describes the multimodal structure of diagrams using four annotation layers. These layers, named *grouping*, *macro-grouping*, *connectivity* and *discourse structure*, are introduced in the following sections. The annotation layers are represented using graphs, which are populated using diagram elements from the AI2D layout segmentation (see Fig. 1). The unique identifiers for diagram elements are also carried over from the AI2D layout segmentation to the AI2D-RST graphs, in order to enable cross-references across annotation layers. This kind of stand-off approach to annotation separates the description of different multimodal structures, but allows combining them as necessary using the unique identifiers, which are shared across annotation layers.



1. Original image

Diagram elements that are likely to be perceived as belonging together are placed under the same parent node in the grouping graph.

2. AI2D layout segmentation

3. AI2D-RST grouping graph

Fig. 3 (1) A thumbnail of the original diagram image scraped from the web, (2) its crowdsourced layout segmentation (converted into greyscale to bring out the annotation) and (3) the AI2D-RST grouping graph for diagram #274. The grouping graph organises diagram elements that are likely to be perceived as belonging together into groups. These groups are added to the grouping graph as parent nodes for the diagram elements that belong together. For an example, see the illustration of a wolf (*B2*) and the text 'Gray wolf' (*T3*) in the layout segmentation and their corresponding nodes in the AI2D-RST grouping graph). Both *B2* and *T3* are children of the grouping node *G7*, which can be used to refer to both diagram elements in the annotation layers for connectivity and discourse structure

4.1.1 Grouping

The grouping layer describes which diagram elements form visual groups, that is, which elements are likely to be perceived as belonging together. The principles behind the grouping layer correspond loosely to Gestalt principles of perception, which often guide the design of diagrams and other visualisations (Ware 2012, p. 179). To exemplify, the principle of *proximity* states that elements close to each other are considered to belong together. A brief introduction to Gestalt principles of pattern perception and how they influence the process of interpretation is provided in Bateman (2008, pp. 58–61).

In AI2D-RST, the grouping annotation is represented using an undirected, acyclic tree graph, such as the one shown on the right-hand side in Fig. 3. In Fig. 3, the root node of the graph is the image constant *I0*, which stands for the entire diagram. In contrast to AI2D, the AI2D-RST grouping layer includes nodes for only three types of diagram elements, namely blobs, text and arrows, but introduces another node type: groups. Diagram elements that form a visual unit in the layout are placed under the same parent node in the grouping graph. These nodes have the prefix G in their identifier, which stands for a group.

Conversely, besides grouping elements together, the grouping graph also represents which elements are considered independent, or in other words, do not belong to any visual groups. In Fig. 3, such independent units include the arrows *A0–15* that set up the network of connections between the groups of illustrations and

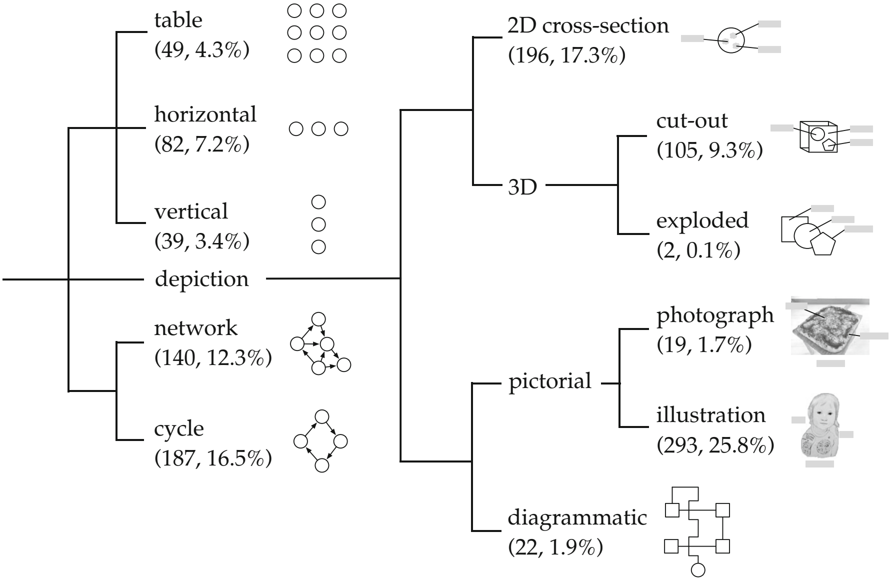


Fig. 4 A typology of macro-groups. The numbers in parentheses give the raw count for each macro-group and their proportion of the AI2D-RST corpus ($N = 1134$)

their labels *G2–13*. These connections are described in the connectivity layer in order to avoid making arbitrary decisions about whether arrows should be grouped with their sources or targets (see Sect. 4.1.3).

To summarise, the grouping graph provides a foundation for the subsequent annotation layers, namely macro-grouping, connectivity and discourse structure *by providing the necessary units of analysis*. In practice, the grouping graph allows diagram elements that form visual groups to be picked up for description in other annotation layers by referring to the identifiers of their grouping nodes.

4.1.2 Macro-grouping

Macro-grouping captures the generic principles that govern diagram structure above the level of visual groups identified in the grouping layer, in order to describe why such visual groupings of expressive resources exist in the first place. To draw on an example, the grouping graph shown in Fig. 3 consists of the groups *G2–G13*, which combine an illustration and a written label, and the arrows *A0–15*. Both groups and arrows form a single visual group, *G14*, which may be appropriately characterised as a network. We term such constellations of visual groups *macro-groups*, because they combine multiple visual groups and diagram elements into larger structures.

Due to its close relation to the grouping layer, macro-grouping annotation is incorporated into the grouping graph. If the diagram consists of a single macro-group, macro-grouping information is assigned to the root node of the grouping graph (the image constant *I0*), but if the diagram features multiple macro-groups, this information may be assigned to grouping nodes as well. Figure 3 exemplifies a diagram with multiple macro-groups. The food web under grouping node *G14* is assigned the macro-group *network*, whereas the categories on the left under the grouping node *G1* form a *vertical* organisation, whose function is to provide labels for visual groups that constitute the network.

Figure 4 shows a typology of macro-groups developed on the basis of our initial analysis of diagram types in the AI2D-RST corpus. As such, the scope of the typology is not intended to cover the space of possibilities within the entire diagrammatic mode, but is limited to the domain represented by the diagrams in AI2D-RST. In addition to describing the larger organisations of visual groups, macro-groups are intended to provide a set of structural categories that correspond to different *diagram types*, in contrast to the semantic categories in AI2D, which are based on the subject matter of the diagram. In this way, the macro-groups can also be used as target labels for training classifiers.

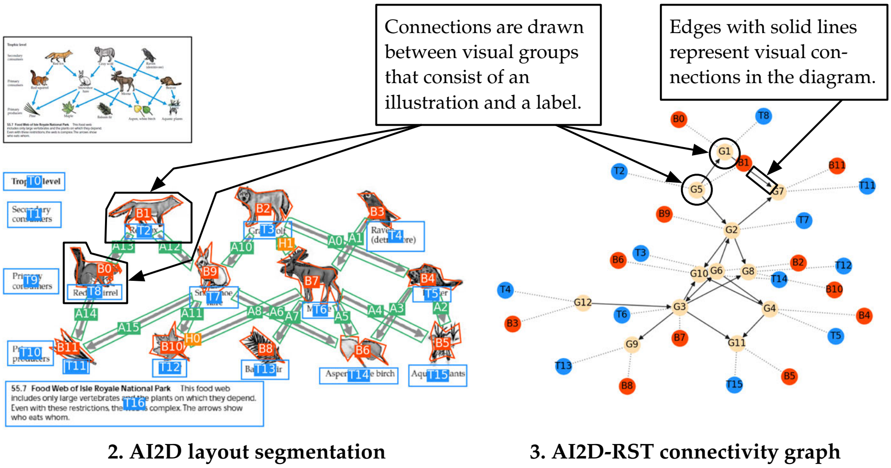
4.1.3 Connectivity

The connectivity layer describes connections between diagram elements and their groups, which are signalled visually using arrows, lines and other diagrammatic elements capable of expressing connectivity (Tversky et al. 2000). In AI2D-RST, the connectivity annotation covers visually explicit connections between diagram elements only, that is, the arrows and lines must have a clear source and a target, in order to allow the connections to be represented using graphs (cf. Alikhani and Stone 2018, p. 3554). The AI2D-RST annotation schema defines three types of connections based on directionality: undirected, directed and bidirectional.

The connectivity annotation is represented using a cyclic mixed graph, which means that the graph may feature both undirected and directed edges. Figure 5 exemplifies a connectivity graph, whose visualization has been enhanced with edges from the grouping graph (for the original grouping graph, see Fig. 3), because the connections in Fig. 5 are likely to be perceived to hold between *visual groups* of elements, rather than individual elements, such as labels or illustrations. Annotating connectivity according to visually explicit connections between individual elements, which originate and terminate in both labels and illustrations, as exemplified by the directed connection between the text block *T3* (‘Gray wolf’) and the illustration of a hare in *B9*, results in an incomplete representation of connectivity. This shows why visual groups are needed as basic units of analysis for a graph-based representation of connectivity, which also illustrates how the grouping layer supports other annotation layers by providing the necessary units of analysis.

4.1.4 Discourse structure

Whereas the grouping and connectivity layers seek to capture diagrammatic structures that are *explicitly* available for visual inspection, the discourse structure



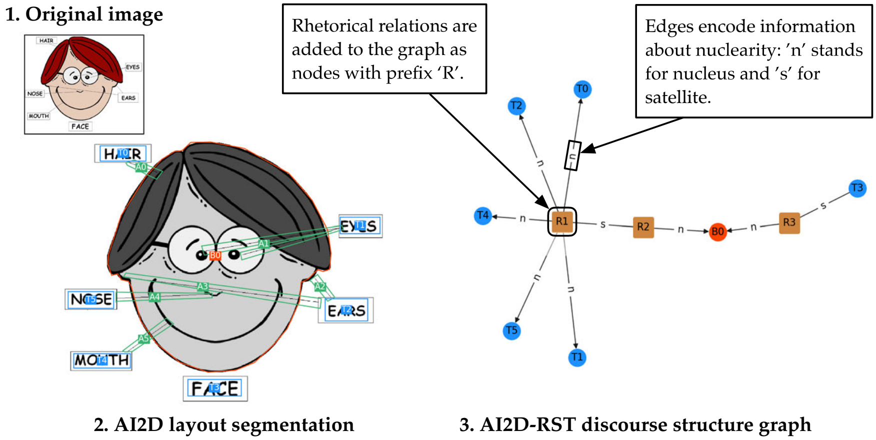
Connections are drawn between visual groups that consist of an illustration and a label.

Edges with solid lines represent visual connections in the diagram.

2. AI2D layout segmentation

3. AI2D-RST connectivity graph

Fig. 5 (1) A thumbnail of the original diagram image scraped from the web, (2) its crowdsourced layout segmentation (converted into greyscale to bring out the annotation) and (3) the AI2D-RST connectivity graph for diagram #274. In the connectivity graph, the edges with solid lines correspond to arrows in the layout segmentation, whereas edges with dashed lines represent edges in the *grouping* graph, which join diagram elements into visual groups



1. Original image

Rhetorical relations are added to the graph as nodes with prefix 'R'.

Edges encode information about nuclearity: 'n' stands for nucleus and 's' for satellite.

2. AI2D layout segmentation

3. AI2D-RST discourse structure graph

Fig. 6 (1) A thumbnail of the original diagram image scraped from the web, (2) its crowdsourced layout segmentation (converted into greyscale to bring out the annotation) and (3) the AI2D-RST discourse structure graph for diagram #0. The multinuclear JOINT relation *R1* joins together the labels *T0–2* and *T4–5*, which serve a similar communicative purpose in the diagram, that is, pick out parts of the illustration *B0* for description. Part-whole relations are described using the ELABORATION relation *R2*, in which the JOINT relation *R1* acts as a satellite and the illustration *B0* as the nucleus. Another relation on the highest level of the hierarchy is drawn between the illustration *B0* and the text *T3* ('FACE') that describes the entire diagram, which is annotated as PREPARATION (*R3*). The edge labels 'n' and 's' stand for nucleus and satellite, respectively

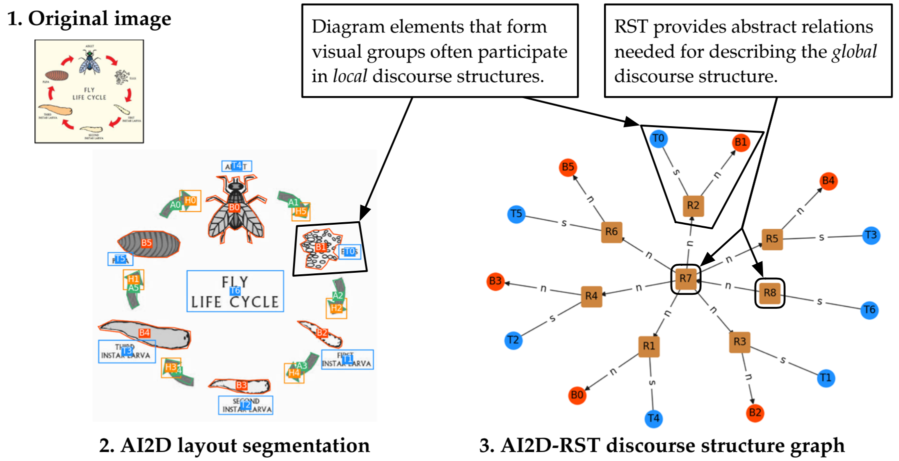

1. Original image

Diagram elements that form visual groups often participate in *local* discourse structures.

RST provides abstract relations needed for describing the *global* discourse structure.

2. AI2D layout segmentation

3. AI2D-RST discourse structure graph

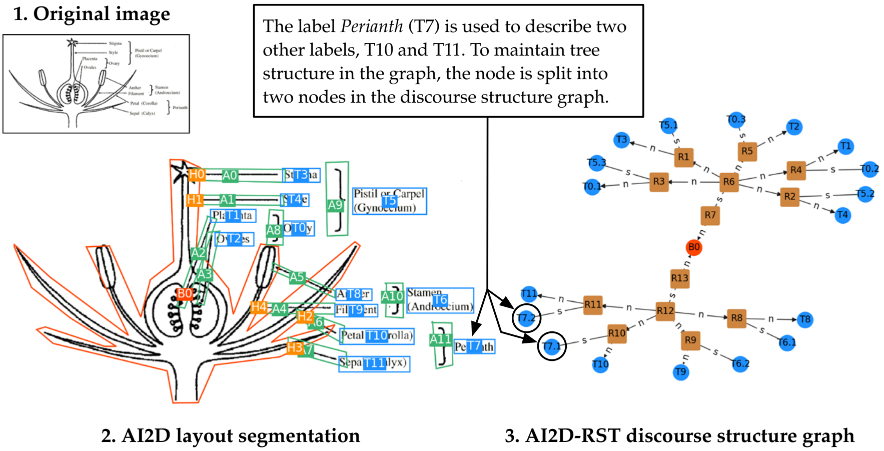
Fig. 7 (1) A thumbnail of the original diagram image scraped from the web, (2) its crowdsourced layout segmentation (converted into greyscale to bring out the annotation) and (3) the AI2D-RST discourse structure graph for diagram #2185. The diagram features three distinct types of rhetorical relations. Whereas the IDENTIFICATION relations (*R1–6*) are mainly local in the sense that the participating diagram elements form visual groups, the CYCLIC SEQUENCE (*R7*) and PREPARATION (*R8*) describe the global discourse organisation of the diagram, or how larger formations of discourse units relate to each other

layer attempts to describe the *implicit* discourse relations that hold between diagram elements and their groups, which viewers may recover from the diagram structure. As such, the discourse structure layer provides the crucial link between multimodal structure and communicative intentions in the AI2D-RST corpus.

For describing the discourse structure of diagrams, AI2D-RST uses Rhetorical Structure Theory (RST; see e.g. Mann and Thompson 1988; Taboada and Mann 2006), a theory of textual organisation and coherence which has been previously extended to diagrams in natural language generation (André and Rist 1995; Bateman et al. 2001; Bateman and Henschel 2007) and for describing discourse relations in research on multimodal documents and other artefacts (Bateman 2008; Thomas 2009; Taboada and Habel 2013; Hiippala 2015). This extension of RST, which may be described as *multimodal RST*, provides the foundation for discourse structure annotation in AI2D-RST, as exemplified in Fig. 6.

Both ‘classical’ and multimodal RST provide a set of discourse relations with criteria for their application (Mann and Thompson 1988; Bateman 2008). For annotating discourse relations in the AI2D-RST corpus, we used the relation definitions presented in Hiippala (2015, pp. 221–223) which combines the classical RST relations from Mann and Thompson (1988) with the multimodal extension proposed in Bateman (2008). We also introduced an additional relation, CYCLIC SEQUENCE, which is used to describe repeating sequences (see the example in Fig. 7). Our application of RST relations is described in great detail in the annotation guide that accompanies the AI2D-RST corpus (see Sect. 4.4).

We drew on these relation definitions to describe how elementary discourse units—which in AI2D-RST correspond to diagram elements or their groups—relate


1. Original image

The label *Perianth* (T7) is used to describe two other labels, T10 and T11. To maintain tree structure in the graph, the node is split into two nodes in the discourse structure graph.

2. AI2D layout segmentation

3. AI2D-RST discourse structure graph

Fig. 8 (1) A thumbnail of the original diagram image scraped from the web, (2) its crowdsourced layout segmentation and (3) the AI2D-RST discourse structure graph for diagram #3194. The diagram features three distinct types of rhetorical relations: IDENTIFICATION (*R1–5*, *R8–11*), ELABORATION (*R7*, *R13*) and JOINT (*R6*, *R12*). To preserve the tree structure of the graph, several diagram elements are represented by multiple nodes in the discourse structure graph, as these elements participate in multiple rhetorical relations. To exemplify, the label *Perianth* (*T7*) describes two other labels, *Petal (Corolla)* (*T10*) and *Sepal (Calyx)* (*T11*). We describe this relation as IDENTIFICATION, as the label *T7* identifies that the labels *T10* and *T11* collectively form a part named *Perianth*. In the discourse structure graph, the IDENTIFICATION relations (*R10* and *R11*) both feature a copy of *T7* as a satellite to preserve the tree structure

to each other. Depending on the relation, one discourse unit may be considered nuclear, or more important, whereas other units act as satellites that play a secondary role. RST terms such relations *asymmetric*. *Symmetric* relations, in turn, may have multiple nuclei, indicating equal status among discourse units. The example in Fig. 6 exemplifies both symmetric (*R2*, *R3*) and asymmetric (*R1*) relations and illustrates how RST relations are represented in the discourse structure graph. Relations are added to the graph as nodes whose identifier has the prefix R, whereas the edges between these nodes carry information on nuclearity, that is, whether the participating diagram elements act as nuclei or satellites.

Figure 7 shows a more complex example, which illustrates the benefit of adopting RST for describing the discourse structure of diagrams. As pointed out above in Sect. 3, the relation definitions in the AI2D annotation schema are largely constrained to local relations between adjacent elements. RST, in turn, provides abstract relations that can handle the description of global discourse organisation as well, or how larger constellations of diagram elements relate to each other.

RST analyses are commonly represented using recursive tree diagrams, although this is not a requirement set by the theory (Taboada and Mann 2006, p. 435). Wolf and Gibson (2005) have argued that tree structures are too constrained for an accurate representation of discourse structure, because a single discourse unit may be picked up as a part of multiple discourse relations. They propose using graphs as an alternative data structure, which would allow discourse units to participate in multiple relations and abolish the hierarchical tree structure.

The discourse structure layer, however, preserves the hierarchical structure and uses a directed acyclic tree graph to represent RST analyses. This decision is motivated by the use of layout space in diagrams, which is regularly used to set up discourse relations between diagram elements (Waller 2012; Watanabe and Nagao 1998). The inherently spatial organisation of diagrams makes constraining the application of discourse relations difficult, particularly in terms of spatial adjacency, that is, limiting relations to elements that are positioned close to each other (cf. Bateman 2008, p. 158). We argue that preserving the tree structure imposes additional control on the application of RST relations.

We do, however, acknowledge that like multimodal documents, diagrams can 're-use' discourse units in different rhetorical relations (Bateman 2008, p. 159). To account for diagram elements that take on the role of satellites or nuclei in multiple rhetorical relations, we split the diagram elements to preserve the hierarchical structure, as shown in Fig. 8. This involves creating copies of a node in the graph, which are identified using a decimal in the node name, such as *T7.1* or *T7.2*. Each copy of the node may be then picked up in the RST analysis while preserving the tree structure. Because the original identifiers are preserved as attributes of the split nodes in the discourse structure graph, the acyclic tree graphs can be easily converted into cyclic graphs favoured by Wolf and Gibson (2005), if necessary.

4.2 Annotators and training

The AI2D-RST diagrams were annotated by five students pursuing BA or MA degrees in English, who received approximately 10 h of initial training in the form of introductory sessions covering each annotation layer. They also received detailed feedback on their initial work and could pose questions about the application of the annotation schema using an online tool for team collaboration. The annotators were also supported by a document that provided guidelines and preferred solutions to common annotation problems, which is available in the repository associated with this article. We return to discuss the impact that the collaborative annotation process may have had on the reproducibility of the annotation framework at the end of Sect. 5. Annotating the corpus took approximately 6 months and cost 50,000€.

4.3 The annotation tool

We developed an in-house tool to annotate the diagrams. The tool provides a command line interface for building graphs, which are initially populated by nodes from the original AI2D layout segmentation. The tool is written in Python 3.6 and makes extensive use of the *matplotlib* (Hunter 2007), *NetworkX* (Hagberg et al. 2008), *OpenCV* (Bradski and Kaehler 2013) and *pandas* (McKinney 2010) libraries. The tool and its source code are available with an open license at https://doi.org/10.5281/zenodo.3384751.

4.4 Acquiring the corpus

The AI2D-RST corpus is available for download as JSON files in the Language Bank of Finland: http://urn.fi/urn:nbn:fi:lb-2020060101. Python functions for loading and processing the corpus are provided separately at https://doi.org/10.5281/zenodo.3384751.

5 Measuring the reliability of the annotation

We measured inter-annotator agreement when 355 diagrams had been annotated. At this stage, the annotators were assumed to have familiarised themselves with the annotation schema. Because the data was annotated by five annotators, we used Fleiss' kappa (κ) as implemented in the *statsmodels* Python library (Seabold and Perktold 2010) as the metric for measuring inter-annotator agreement. We report both the original κ statistic, as proposed by Fleiss (1971), which is calculated using marginal probabilities for each category, and the free-marginal κ proposed by Randolph (2005), which assumes a uniform distribution over all categories. We refer to Fleiss' original definition as marginal κ and Randolph's alternative as uniform κ. In addition, we used the *irr* library (Gamer et al. 2019) for the R programming language (R Core Team 2019) to calculate class-wise marginal κ scores for grouping, macro-grouping, connectivity and discourse structure annotations. The results are reported in Sects. 5.1, 5.2, 5.3 and 5.4. Finally, in Sect. 5.5 we model annotator reliability using MACE (Hovy et al. 2013). The raw annotations are provided as CSV files at https://doi.org/10.5281/zenodo.3384751.

5.1 Grouping

To evaluate the reliability of grouping layer annotation introduced in Sect. 4.1.1, we sampled the 355 diagrams without replacement for 10% of visual groups composed of diagram elements only, excluding groups whose child nodes included other grouping nodes. This amounted to 256 groups, whose elements were highlighted in the AI2D layout segmentation and presented to the annotators. The annotators were then asked whether the elements form a visual group, as defined in the AI2D-RST annotation schema. If the annotators considered the grouping valid, a follow-up question requested the annotators to name Gestalt principle or annotation guideline that justified their choice. If multiple principles or guidelines were applicable, the annotators were asked to choose the most prominent one. For inter-annotator agreement between five annotators and 256 groups, the marginal κ was 0.836, while the uniform κ was 0.894.

Table 1 shows class-wise agreement for Gestalt principles and annotation guidelines, which are sorted in descending order based on their marginal κ values. The results suggest that the annotation guide supported the consistent description of the data. Most cases in the guideline category consisted of label—line combinations, such as those shown in Fig. 6. In principle, such combinations could be grouped together based on several Gestalt principles, such as proximity, continuity and

Table 1 Class-wise marginal κ scores for Gestalt principles and annotation guidelines

Category	Description	κ	z-score	p-value
Guideline	The AI2D-RST guidelines state that the elements are grouped together	0.929	47.008	< 0.001
Proximity	The diagram elements are placed close to each other in the layout space	0.851	43.046	< 0.001
Closure	The element encloses the other element	0.776	39.243	< 0.001
Similarity	The elements are similar in terms of their visual appearance.	0.622	31.453	< 0.001
No-group	The elements do not form a valid group according to the AI2D-RST schema	0.410	20.766	< 0.001
Continuity	The elements form a continuous unit	0.210	10.623	< 0.001
Connectedness	The elements are connected to each other	− 0.003	− 0.159	0.874
Symmetry	The elements form a symmetrical shape	− 0.002	− 0.079	0.937

Table 2 Class-wise marginal κ scores for macro-groups

Macro-group	κ	z-score	p-value	Frequency in corpus
Network	0.884	30.480	<0.001	0.123
Cycle	0.876	30.204	<0.001	0.165
Cut-out	0.849	29.271	<0.001	0.093
Slice	0.754	25.996	<0.001	0.173
Horizontal	0.726	25.031	<0.001	0.072
Diagrammatic	0.718	24.785	<0.001	0.019
Illustration	0.709	24.458	<0.001	0.258
Vertical	0.702	24.228	<0.001	0.034
Table	0.247	8.537	<0.001	0.043
Photograph	0.162	5.604	<0.001	0.017

connectedness, but explicating annotation patterns for common diagrammatic structures such as labels and their connecting lines seems to make the decisions less arbitrary. In addition, common spatial- and attribute-based relations that build on Gestalt principles such as proximity, closure and similarity (Engelhardt 2002, p. 30), are annotated consistently in the AI2D-RST corpus.

5.2 Macro-grouping

For measuring inter-annotator agreement on the macro-groups introduced in Sect. 4.1.2, we sampled the 355 diagrams without replacement for 33% of macro-groups, which amounted to 119 macro-groups. The annotators were presented with the AI2D layout segmentation and the AI2D-RST grouping graph, which highlighted the node that had been assigned with macro-grouping information. The annotators were then asked which macro-group they would assign to the node in question. For inter-annotator agreement on macro-groups, the marginal κ was 0.784 and the uniform κ was 0.800.

Table 2 gives class-wise marginal κ values for macro-groups in descending order. Agreement is particularly high for visually distinctive macro-groups such as networks, cycles and cut-outs, which occur frequently in the AI2D-RST corpus (see also Fig. 4). The values are considerably lower for less common macro-groups such as tables and photographs. Photographs, in particular, are rarely preferred as the main visual expressive resource in the AI2D-RST corpus, as diagrams in the corpus favour illustrations, cut-outs and cross-sections for depiction. For these prominent macro-groups, agreement remains substantial.

5.3 Connectivity

For connectivity annotation (see Sect. 4.1.3), we sampled the 355 diagrams without replacement for 10% of connections holding between diagram elements or their groups, which resulted in 239 connections. The source and target of each connection were highlighted in the AI2D layout segmentation and presented to the annotators, who were then asked to place the connection into one of four categories: directed,

Table 3 Class-wise marginal κ scores for connectivity

Connection	κ	z-score	p-value	Frequency in corpus
Directed	0.910	44.512	<0.001	0.511
Bidirectional	0.908	44.402	<0.001	0.004
Undirected	0.900	44.003	<0.001	0.485
No connection	0.192	9.392	<0.001	N/A

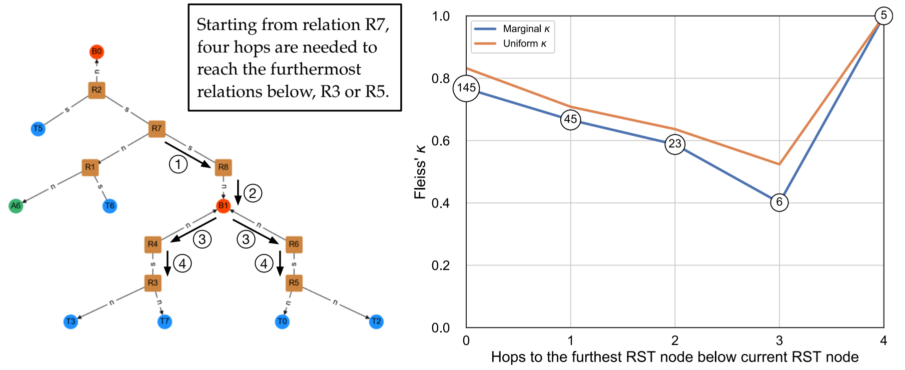
Table 4 Class-wise marginal κ scores for discourse relations

Discourse relation	κ	z-score	p-value	Frequency in corpus
CYCLIC SEQUENCE	0.924	44.029	< 0.001	0.033
PREPARATION	0.870	41.471	< 0.001	0.054
PROPERTY-ASCRIPTION	0.870	41.468	< 0.001	0.070
JOINT	0.827	39.419	< 0.001	0.109
IDENTIFICATION	0.798	37.998	< 0.001	0.439
CONNECTED	0.766	36.492	< 0.001	0.030
SEQUENCE	0.689	32.844	< 0.001	0.015
ELABORATION	0.620	29.540	< 0.001	0.134
CIRCUMSTANCE	0.449	21.388	< 0.001	0.029
CONTRAST	0.308	14.656	< 0.001	0.024
CLASS-ASCRIPTION	0.266	12.680	< 0.001	0.028
CONJUNCTION	0.249	11.848	< 0.001	0.003
DISJUNCTION	0.249	11.848	< 0.001	0.003
LIST	0.182	8.659	< 0.001	0.007
NONVOLITIONAL CAUSE	0.138	6.553	< 0.001	0.004
NONVOLITIONAL RESULT	0.078	3.738	< 0.001	0.006
MEANS	0.066	3.129	0.002	0.003
CONDITION	− 0.001	− 0.042	0.966	0.001
PURPOSE	− 0.001	− 0.042	0.966	N/A
RESTATEMENT	− 0.003	− 0.126	0.900	0.004

undirected, bidirectional or no connection. Measuring inter-annotator agreement returned a marginal κ of 0.878 and uniform κ of 0.916. Table 3 gives class-wise marginal κ values for each connection type. Apart from no connection, agreement is high across all types of connectivity, as might be expected with a low number of categories, which are also visually distinctive and whose structural features are relatively easy to formalise (see Alikhani and Stone 2018, p. 3554).

5.4 Discourse structure

For evaluating inter-annotator agreement on the discourse structure layer introduced in Sect. 4.1.4, we sampled the 355 diagrams without replacement for 10% of the relations, amounting to a total of 227 RST relations. The AI2D layout segmentation



Starting from relation R7, four hops are needed to reach the furthermost relations below, R3 or R5.

Fig. 9 Fleiss' marginal and uniform κ for RST relations at different depths of the RST tree. We measured the position of RST nodes in the tree by calculating the number of hops needed to reach the furthermost RST node in the subtree below, as illustrated on the left-hand side. On the right-hand side, the balloons give the number of samples observed for each hop. The X-axis gives the number of hops: a value of zero indicates that the RST relation is close to the edge of the tree

and the AI2D-RST discourse structure graph were presented side-by-side to the annotators, highlighting the RST relation node to be annotated in the discourse structure graph. Measuring overall agreement on the RST relations returned a marginal κ of 0.733 and a uniform κ of 0.783.

Table 4 provides class-wise marginal κ scores for RST relations that the annotators used during the inter-annotator agreement experiment in a descending order. The results show that annotators consistently agree on common RST relations such as CYCLIC SEQUENCE, which is used to annotate recurring cycles formed by diagram elements, and PREPARATION, which is used to describe the relationship between a title and an entire diagram. These RST relations are associated with visually distinctive macro-groups (cycles) and relatively fixed diagram elements (titles), which is likely to increase agreement. The same applies to frequently occurring relations defined between a label and an object or its part, such as PROPERTY-ASCRIPTION, IDENTIFICATION and ELABORATION, whose specific use cases were defined in the annotation guide. In short, the development of an annotation guide seemed to support the consistent annotation of RST relations. Compared to previous studies of inter-annotator agreement using multimodal RST (e.g. Taboada and Habel 2013), the κ scores for the AI2D-RST discourse structure layer are promising, as relations with a $\kappa > 0.62$ cover 88.4% of RST relations in the corpus.

Figure 9 provides an alternative view to the reliability of the discourse structure annotation by measuring inter-annotator agreement at different depths of the RST tree graph. Not surprisingly, agreement is highest at the leaves of the tree graph (hop 0) with a marginal k of 0.767 and a uniform k of 0.832. These consistently annotated relations mainly cover local discourse structures illustrated in Fig. 7, as exemplified by IDENTIFICATION ($N = 81$), JOINT ($N = 21$) and PROPERTY-ASCRIPTION ($N = 21$). As the κ values for hops 1–3 show, agreement decreases for relations that are positioned up the tree, which represent the more abstract relations that hold between larger discourse units. Surprisingly, annotators consistently agree on how the

Table 5 MACE reliability estimates for annotators and specific tasks

Task	Ann. 1	Ann. 2	Ann. 3	Ann. 4	Ann. 5
Grouping	0.9133	0.9378	0.9040	0.9601	0.9430
Macro-grouping	0.8851	0.8052	0.9351	0.8574	0.8954
Connectivity	0.9478	0.9382	0.9531	0.9364	0.9631
Discourse structure	0.8452	0.8698	0.8912	0.8021	0.9249

relations closest to the root (hop 4) should be annotated. It should be noted, however, that sample sizes are very small for hops 3 and 4 and therefore warrant caution.

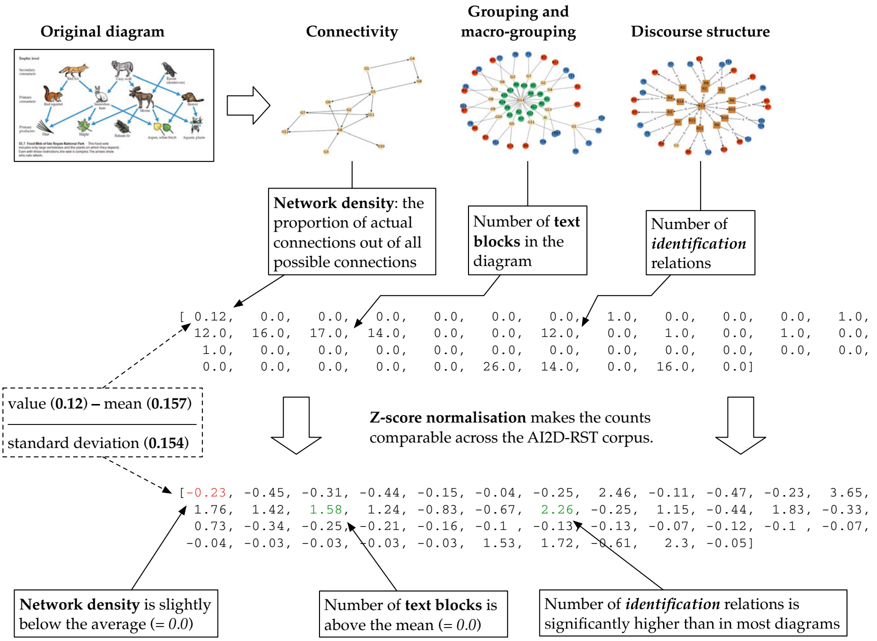
5.5 Modelling annotator reliability

In addition to measuring inter-annotator agreement, we estimated annotator reliability using MACE (Hovy et al. 2013). MACE, which stands for Multi-Annotator Competence Estimation, models the annotation process by treating the labels as latent variables and uses unsupervised learning to estimate the model parameters. The model seeks to predict whether the annotator is answering dutifully or choosing the answers at random. Hovy et al. (2013, p. 1124) show that MACE reliability estimates correlate strongly with annotator proficiency. Table 5 shows MACE reliability estimates using default settings, which suggests dutiful annotation with slightly varying competences between annotators.

5.6 On the reliability and reproducibility of the AI2D-RST annotation schema

Overall, the inter-annotator agreement measures suggest that the AI2D-RST annotation is applied consistently to the diagrams. The results are particularly promising given that inter-annotator agreement was measured between five annotators. However, it is important to acknowledge that measuring inter-annotator agreement using metrics such as Fleiss' κ often involve compromises. In the case of RST, for instance, measuring agreement over a single relation in a given context is very different from constructing entire RST trees and comparing them between annotators. To improve the evaluation of annotation reliability, future studies applying multimodal RST should follow up on recent developments in research on the automatic comparison of RST trees (see e.g. Wan et al. 2019). Alternatively, the approach illustrated in Fig. 9 could be used sample relations along the depth of the RST tree in a balanced manner, in order to ensure that agreement is evaluated for both local and global discourse structures.

In terms of the annotation schema, it should be noted that the expert annotators helped to develop the AI2D-RST annotation schema by discussing specific examples with each other, which were then documented in the annotation guide. This violates several principles of reproducibility set out for *content analysis* in Krippendorff (2013). However, as Artstein and Poesio (2008, p. 575) point out, content analysis treats the annotation process as an *experiment* about whether some



Network density: the proportion of actual connections out of all possible connections

Number of **text blocks** in the diagram

Number of ***identification*** relations

```
[ 0.12,   0.0,   0.0,   0.0,   0.0,   0.0,   0.0,   1.0,   0.0,   0.0,   0.0,   1.0,
  12.0,  16.0,  17.0,  14.0,   0.0,   0.0,  12.0,   0.0,   1.0,   0.0,   1.0,   0.0,
   1.0,   0.0,   0.0,   0.0,   0.0,   0.0,   0.0,   0.0,   0.0,   0.0,   0.0,   0.0,
   0.0,   0.0,   0.0,   0.0,   0.0,  26.0,  14.0,   0.0,  16.0,   0.0]
```

value (**0.12**) – mean (**0.157**)
standard deviation (**0.154**)

Z-score normalisation makes the counts comparable across the AI2D-RST corpus.

```
[-0.23, -0.45, -0.31, -0.44, -0.15, -0.04, -0.25,  2.46, -0.11, -0.47, -0.23,  3.65,
  1.76,  1.42,  1.58,  1.24, -0.83, -0.67,  2.26, -0.25,  1.15, -0.44,  1.83, -0.33,
  0.73, -0.34, -0.25, -0.21, -0.16, -0.1 , -0.13, -0.13, -0.07, -0.12, -0.1 , -0.07,
 -0.04, -0.03, -0.03, -0.03, -0.03,  1.53,  1.72, -0.61,   2.3, -0.05]
```

Network density is slightly below the average (= *0.0*)

Number of **text blocks** is above the mean (= *0.0*)

Number of ***identification*** relations is significantly higher than in most diagrams

Fig. 10 Extracting simple features from diagrams in the AI2D-RST corpus by counting the instances of different features across the annotation layers. The features are then normalised to make them comparable

properties may be consistently detected in a text, whose success is determined by reproducibility of the annotation. In computational linguistics, annotation serves different purposes, such as creating resources for training and evaluating algorithms, which differs from the goals set for content analysis (Reidsma and Carletta 2007).

Riezler (2014, p. 240), however, also calls for attention to the consequences of violating the requirement of independence, that is, allowing the annotators to discuss annotation tasks. This is likely to generate implicit knowledge among the annotators, which increases agreement among annotators but hinders reproducibility. This kind of implicit knowledge gives rise to circularity in annotation, which has been acknowledged as a problem in multimodality research (Thomas 2014). Given the collaborative annotation procedure, it is likely that the AI2D-RST annotations exhibit a degree of circularity.

To evaluate and improve the reproducibility of the AI2D-RST framework, future work should employ naive annotators, who are assigned tasks that do not build on concepts introduced in the annotation framework (see e.g. Asheghi et al. 2016). This kind of non-theoretical grounding (Riezler 2014) could help to break circularity by evaluating, for instance, whether naive annotators perceive diagram elements to form visual groups (grouping) or whether arrows and lines are considered to signal connections between individual diagram elements or visual groups (connectivity). For discourse structure annotation, Yung et al. (2019) introduce a multi-step procedure for sourcing descriptions of discourse relations

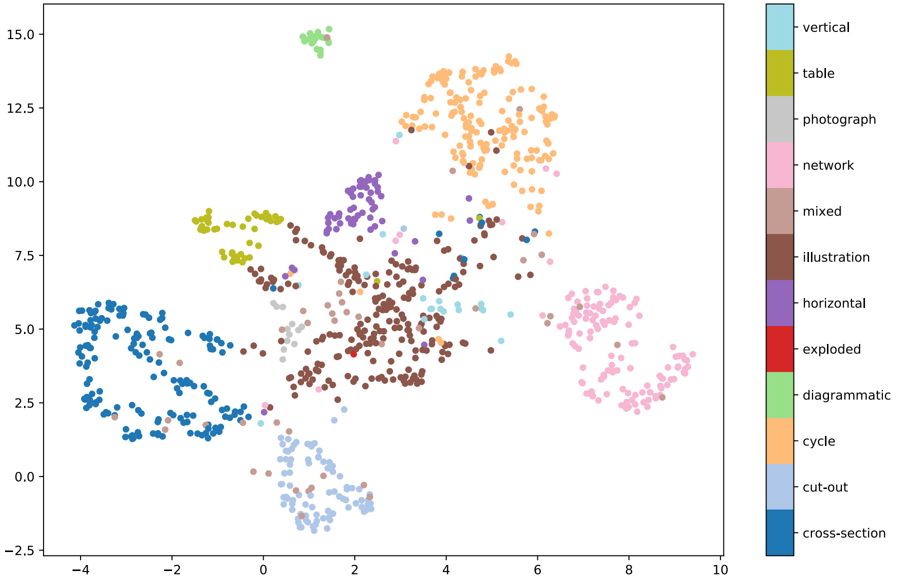


Fig. 11 A visualization showing 2-dimensional UMAP embeddings learned from the 46-dimensional feature vectors extracted using the technique in Fig. 10. Each point corresponds to a single diagram in the AI2D corpus, which are coloured according to their macro-groups

from naive annotators. Adopting this approach in multimodal RST, however, would require additional efforts to accommodate the presence of multiple expressive resources.

6 Exploring the AI2D-RST corpus

In this section, we present a brief exploratory analysis of the AI2D-RST corpus. We begin with a rather straightforward approach illustrated in Fig. 10, which makes minimal use of the graph-based representations by simply counting instances of diagram elements, macro-groups, rhetorical relations, nuclei and satellites, and types of connections in each diagram. Finally, we also calculate network density for the connectivity graph, which measures the proportion of actual edges present in the graph out of all possible edges. We concatenate these values into a 46-dimensional feature vector and use z-score normalization to scale the values of each dimension to have a mean of 0 and a standard deviation of 1. This provides each diagram in the AI2D-RST corpus with a normalised 46-dimensional feature vector that represents its multimodal structure.

Figure 11 shows a visualisation that uses the Uniform Manifold Approximation and Projection algorithm (UMAP; see McInnes et al. 2018) to reduce the 46-dimensional feature vectors to two dimensions for a visual exploration of the AI2D-RST corpus. When mapping points between high- and low-dimensional spaces, UMAP seeks to preserve both local and global structure of the points across the two spaces. In other words, points that are close to each other in the 46-dimensional space should be close to each other in the two-dimensional space, whereas points that are distant from each other in the 46-dimensional space should remain distant in the two-dimensional space as well.

The UMAP embeddings in Fig. 11 show distinct clusters that correspond to specific macro-groups, such as cycles, cross-sections, cut-outs and networks, which illustrate the space of structural variation among the AI2D-RST diagrams. It should be noted, however, that the macro-grouping annotation is explicitly encoded into the 46-dimensional feature vector. This information is thus directly available to UMAP for learning the 2-dimensional embeddings, which the algorithm leverages when clustering points in the low-dimensional space.

Nevertheless, the visualisation in Fig. 11 can yield valuable insights into the structural variation among the AI2D-RST diagrams. Firstly, diagrams that feature several macro-groups (see Sect. 4.1.2) can be found within all major clusters, which suggests that even simple count-based features can capture structural distinctions in diagrams. The diagrams labelled as ‘mixed’ are particularly interesting, as they may yield information on which macro-groups are readily combined with each other in the AI2D-RST corpus. The clusters for individual macro-groups, in turn, appear to capture variation within macro-groups, as exemplified by the clusters for networks and cross-sections, which seem to form two parts. Whether such formations within clusters reflect alternative structural configurations of expressive resources within specific macro-groups warrants further analysis.

Secondly, diagrams that feature rigid layouts, such as tabular, horizontal and vertical macro-groups, are not only positioned close to each other, but also form a continuation of the cluster for illustrations. This is not surprising, as tabular, vertical and horizontal macro-groups are typically used to organise *multiple instances* of visual depictions and their verbal descriptions for presentation, in which the local discourse structures are similar to *individual* illustrations (for examples of local discourse structures, see Fig. 7). The clusters for cut-outs and cross-sections, in turn, are distinct from illustrations, which may be traced back to differences in their discourse structure. Whereas cut-outs and cross-sections typically use labels to pick out parts or regions of a visual depiction, illustrations use labels to identify the entire object. This distinction is captured by their discourse structure annotation.

Thirdly, the diagrammatic macro-group forms a tight cluster, which is clearly separate from other macro-groups. Although the sample size for this macro-group is fairly small ($N = 22$), this is an interesting observation as the UMAP embeddings seem to capture a fundamental difference between the diagrammatic macro-group and other macro-groups in the corpus, which may be traced back to their discourse structure. The diagrammatic macro-group features schematic diagrams such as circuit diagrams, whose elements have *fixed* meanings, as exemplified by standardised symbols for switches, connections, circuit breakers and the like.

Because their diagram elements have fixed meanings that do not need to be recovered discursively from their context of occurrence, schematic diagrams resist RST analysis. Put differently, there is no need for the viewer to resolve discourse relations between diagram elements, as all the information needed for making sense of the diagram is communicated using arrows and lines that signal connections between diagram elements with fixed meanings. Although these connections are captured by the AI2D-RST connectivity layer, this raises questions about the need to revise the AI2D-RST annotation schema, if it were to be extended to domains featuring many types of schematic diagrams, in order to draw out their differences.

This brief exploratory analysis has illustrated how the AI2D-RST corpus can be used to support empirical research on the multimodality of diagrams. As pointed out above, the features extracted from the corpus made minimal use of the properties of the graph-based representations (see Fig. 10). The properties of graphs could be exploited to a much larger extent using algorithms such as graph neural networks, which learn representations of graph-structured data by passing and receiving features between neighbouring nodes (see e.g. Wu et al. 2019). Such methods could be particularly useful for learning representations of discourse structure in diagrams, allowing their computational representation to encode interactions between diagram elements. However, learning these representations directly from the data can be complicated by the relatively small number diagrams in AI2D-RST.

7 Discussion

Developing the AI2D-RST corpus showed that exploiting readily-available annotations can be used to increase the size of richly-annotated multimodal corpora, but this comes at a cost, particularly for annotating their discourse structure. As explicated in Hiippala and Bateman (2020), identifying the elementary discourse units required by RST and other discourse annotation frameworks is particularly complicated for diagrams, because the extent to which diagrams need to be decomposed to achieve a sufficient inventory of elementary discourse units varies from one diagram to another. In short, the level of detail needed for decomposition depends on the *combination* of expressive resources and the discourse relations they participate in (see Sect. 2).

Because the AI2D layout segmentation does not provide this kind of discourse-driven decomposition at various levels of detail, the AI2D-RST annotation schema had to make compromises in the description of discourse structure. The example in Fig. 6 illustrates this issue aptly: the written labels are used to pick out parts of the illustration, and to achieve a maximally accurate RST analysis of the diagram, the illustration should be decomposed into its component parts. However, as the crowdsourced annotators were not instructed to decompose visual expressive resources during layout segmentation, the elementary discourse units needed for a maximally coherent representation of discourse structure within RST are not available (for a discussion of similar problems in annotating comics, see Bateman and Wildfeuer 2014).

This shortcoming also carries implications for crowdsourcing annotations for the diagrammatic mode in any domain. Because the discourse structure determines to what extent the diagram must be decomposed, defining crowdsourcing tasks developed for the annotation of photographic images is unlikely to work for identifying the ‘building blocks’ of diagrams (cf. Kovashka et al. 2016). Instead of defining semantic object classes (i.e. what the diagram element represents), these building blocks should correspond to expressive resources available to the diagrammatic mode, such as written language, arrows, lines and other diagrammatic elements. Crucially, these expressive resources must be complemented by sufficiently fine-grained descriptions of graphic expressive resources, such as line drawings, coloured illustrations, cut-outs, cross-sections and exploded views, and photographs, to name just a few examples. In short, pre-theoretical notions such as 'language' and 'image' are not sufficiently fine-grained to capture the motivated use of distinctive graphic expressive resources in diagrams (cf. Bateman 2014).

Although the development of AI2D-RST revealed various challenges discussed above, we argue that the corpus is still a valuable resource for studying how the diagrammatic mode is used in the domain of primary school natural sciences and beyond. In the study of multimodal discourse, the corpus could be used for investigating whether discourse relations between diagram elements are signalled visually using arrows and lines or spatially using layout (cf. Watanabe and Nagao 1998), thus complementing the linguistic research on signalling of discourse relations by Das and Taboada (2018). Such empirically-backed insights could be particularly valuable to educational research on the visual perception of diagrammatic representations, and their role in constructing mental models and learning more generally (Tippett 2016; Menendez et al. 2020). Another avenue of further research involves the automatic annotation of diagram corpora. The AI2D-RST corpus covers just over 20% of the AI2D dataset, which raises the question whether the 1000 diagrams in AI2D-RST are sufficient for teaching algorithms to generate AI2D-RST annotations for the remaining 3900 diagrams in AI2D.

8 Concluding remarks

In this article we introduced AI2D-RST, a new multimodal corpus of 1000 English-language primary school science diagrams, which combines crowdsourced and expert annotations to provide a rich description of their multimodal structure. The multi-layered, stand-off annotation schema developed for AI2D-RST accounts for (1) the visual grouping of diagram elements, (2) how their connections are signalled using arrows and lines, and (3) the discourse relations between diagram elements using Rhetorical Structure Theory. We measured agreement between five annotators: the results suggest that the annotation schema may be reliably applied to describe diagrams in the AI2D-RST corpus.

As our brief exploratory analysis of the AI2D-RST corpus showed, the combination of multiple annotation layers and graph-based representations can yield valuable insights into the multimodal structure of diagrams. As such, the corpus can support empirical research on diagrams as a mode of expression and their computational processing. In terms of methodology, developing the AI2D-RST corpus illustrated how crowdsourcing low-level annotations and building expert descriptions on top of them can be used to increase the size of corpora in the field of multimodality research. Insights from linguistically-inspired multimodality research, in turn, can also inform the creation of resources for research on the computational processing and generation of diagrams.

Funding Open access funding provided by University of Helsinki including Helsinki University Central Hospital.

References

Alikhani, M., & 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. 3552–3563.

André, E., & Rist, T. (1995). Generating coherent presentations employing textual and visual material. *Artificial Intelligence Review*, *9*, 147–165.

Artstein, R., & Poesio, M. (2008). Inter-coder agreement for computational linguistics. *Computational Linguistics*, *34*(4), 555–596.

Asheghi, N. R., Sharoff, S., & Markert, K. (2016). Crowdsourcing for web genre annotation. *Language Resources & Evaluation*, *50*(3), 603–641.

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

Bateman, J. A. (2014). *Text and image: A critical introduction to the visual/verbal divide*. London: Routledge.

Bateman, J. A. & Henschel, R. (2007). Generating text, layout and diagrams appropriately for genre, *in* I. van der Sluis, M. Theune, E. Reiter and E. Krahmer, eds, 'Proceedings of the Workshop on Multimodal Output Generation MOG 2007', Centre for Telematics and Information Technology (CTIT), University of Twente, pp. 29–40.

Bateman, J. A., Kamps, T., Reichenberger, K., & Kleinz, J. (2001). Towards constructive text, diagram, and layout generation for information presentation. *Computational Linguistics*, *27*(3), 409–449.

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

Bradski, G., & Kaehler, A. (2013). *Learning OpenCV: Computer vision in C++ with the OpenCV library* (2nd ed.). Sebastopol: O'Reilly.

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

Choi, J., Krishnamurthy, J., Kembhavi, A., & Farhadi, A. (2018). Structured set matching networks for one-shot part labeling. *In* 'Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition', pp. 3627–3636.

Das, D., & Taboada, M. (2018). RST Signalling Corpus: A corpus of signals of coherence relations. *Language Resources and Evaluation*, *52*(1), 149–184.

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

Engelhardt, Y., & Richards, C. (2018). A framework for analyzing and designing diagrams and graphics, *in* P. Chapman, A. Moktefi, S. Perez-Kriz and F. Bellucci, eds, 'Diagrams 2018: Diagrammatic Representation and Inference', Vol. 10871 of *Lecture Notes in Computer Science*, Springer, pp. 201–209.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, *76*(5), 378–382.

Gamer, M., Lemon, J., Fellows, I., & Singh, P. (2019). irr: Various coefficients of interrater reliability and agreement. R package version 0.84.1. https://CRAN.R-project.org/package=irr

Grosz, B. J., & Sidner, C. L. (1986). Attention, intentions, and the structure of discourse. *Computational Linguistics*, *12*(3), 175–204.

Haehn, D., Tompkin, J., & Pfister, H. (2019). Evaluating 'graphical perception' with CNNs. *IEEE Transactions on Visualization and Computer Graphics*, *25*(1), 641–650.

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

Hegarty, M., & Just, M. A. (1993). Constructing mental models of machines from text and diagrams. *Journal of Memory and Language*, *32*(6), 717–742.

Hiippala, T. (2013). The interface between rhetoric and layout in multimodal artefacts. *Literary and Linguistic Computing*, *28*(3), 461–471.

Hiippala, T. (2015). *The structure of multimodal documents: An empirical approach*. New York: Routledge.

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

Hiippala, T. and Orekhova, S. (2018). Enhancing the AI2 diagrams dataset using rhetorical structure theory. *in* Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), European Language Resources Association (ELRA), Paris, pp. 1925–1931.

Holsanova, J., Holmberg, N., & Holmqvist, K. (2009). Reading information graphics: The role of spatial contiguity and dual attentional guidance. *Applied Cognitive Psychology*, *23*, 1215–1226.

Hovy, D., Berg-Kirkpatrick, T., Vaswani, A., & Hovy, E. (2013). Learning whom to trust with MACE. *In* Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Atlanta, Georgia, pp. 1120–1130.

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

Hunter, J. (2007). matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, *9*(3), 90–95.

Kembhavi, A., Salvato, M., Kolve, E., Seo, M. J., Hajishirzi, H., & Farhadi, A. (2016). A diagram is worth a dozen images. *In* Proceedings of the 14th European Conference on Computer Vision (ECCV 2016), Springer, Cham, pp. 235–251.

Kim, D., Kim, S., & Kwak, N. (2019) , Textbook question answering with multi-modal context graph understanding and self-supervised open-set comprehension. *In* Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019), Association for Computational Linguistics, Florence, Italy, pp. 3568–3584.

Kim, D., Yoo, Y., Kim, J., Lee, S. & Kwak, N. (2018). Dynamic graph generation network: Generating relational knowledge from diagrams. *In* Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2018), pp. 4167–4175.

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

Krippendorff, K. (2013). *Content Analysis: An Introduction to its Methodology* (3rd ed.). Thousand Oaks, CA: Sage.

Mann, W. C., & Thompson, S. A. (1988). Rhetorical Structure Theory: Toward a functional theory of text organization. *Text*, *8*(3), 243–281.

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

McKinney, W. (2010). Data structures for statistical computing in Python. *In* S. van der Walt and J. Millman, eds, Proceedings of the 9th Python in Science Conference, pp. 51–56.

Menendez, D., Rosengren, K. S., & Alibali, M. W. (2020). Do details bug you? Effects of perceptual richness in learning about biological change. *Applied Cognitive Psychology*, *34*(5), 1101–1117.

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

R Core Team. (2019). *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Randolph, J. J. (2005). Free-marginal multirater kappa (multirater κ-free): An alternative to Fleiss' fixed-marginal multirater kappa. *In*: Proceedings of the Joensuu Learning and Instruction Symposium.

Reidsma, D., & Carletta, J. (2007). Reliability measurement without limits. *Computational Linguistics*, *34*(3), 319–326.

Riezler, S. (2014). On the problem of theoretical terms in empirical computational linguistics. *Computational Linguistics*, *40*(1), 235–245.

Sachan, M., Dubey, A., Hovy, E. H., Mitchell, T. M., Roth, D., & Xing, E. P. (2019). Discourse in multimedia: A case study in extracting geometry knowledge from textbooks. *Computational Linguistics*, *45*(4), 627–665.

Sachan, M., Dubey, K. A., Mitchell, T. M., Roth, D., & Xing, E. P. (2018). Learning pipelines with limited data and domain knowledge: A study in parsing physics problems. *In* Proceedings of the 32nd Conference on Neural Information Processing Systems (NeurIPS 2018).

Seabold, S., & Perktold, J. (2010). Statsmodels: Econometric and statistical modeling with Python. *In* 9th Python in Science Conference. pp. 57–61.

Seo, M., Hajishirzi, H., Farhadi, A., Etzioni, O., & Malcolm, C. (2015). Solving geometry problems: Combining text and diagram interpretation. *In* Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP 2015)', Association for Computational Linguistics, Lisbon, Portugal, pp. 1466–1476.

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

Taboada, M., & Mann, W. C. (2006). Rhetorical structure theory: Looking back and moving ahead. *Discourse Studies*, *8*(3), 423–459.

Thomas, M. (2009). Localizing pack messages: A framework for corpus-based cross-cultural multimodal analysis, PhD thesis, University of Leeds.

Thomas, M. (2014). Evidence and circularity in multimodal discourse analysis. *Visual Communication*, *13*(2), 163–189.

Tippett, C. D. (2016). What recent research on diagrams suggests about learning with rather than learning from visual representations in science. *International Journal of Science Education*, *38*(5), 725–746.

Tversky, B. (2015). The cognitive design of tools of thought. *Review of Philosophy and Psychology*, *6*(1), 99–116.

Tversky, B. (2017). Diagrams: Cognitive foundations for design. In A. Black, P. Luna, O. Lund, & S. Walker (Eds.), *Information design: Research and practice* (pp. 349–360). London: Routledge.

Tversky, B., Zacks, J., Lee, P., & Heiser, J. (2000). Lines, blobs, crosses and arrows: Diagrammatic communication with schematic figures. *Diagrams 2000: Theory and Application of Diagrams* (pp. 221–230). Berlin: Springer.

Waller, R. H. W. (2012). Graphic literacies for a digital age: The survival of layout. *The Information Society*, *28*(4), 236–252.

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

Wan, S., Kutschbach, T., Lüdeling, A., & Stede, M. (2019). RST-Tace: A tool for automatic comparison and evaluation of RST trees. *In* Proceedings of the Workshop on Discourse Relation Parsing and Treebanking 2019, Association for Computational Linguistics, Minneapolis, MN, pp. 88–96.

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

Watanabe, Y., & 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. 1374–1380.

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

Wolf, F., & Gibson, E. (2005). Representing discourse coherence: A corpus-based study. *Computational Linguistics*, *31*(2), 249–288.

Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., & Yu, P. S. (2019). A comprehensive survey on graph neural networks. http://arxiv.org/abs/1901.00596

Yung, F., Demberg, V., & Scholman, M. (2019). Crowdsourcing discourse relation annotations by a two-step connective insertion task. *In:* Proceedings of the 13th Linguistic Annotation Workshop, Association for Computational Linguistics, Florence, Italy, pp. 16–25.