Written labels in elementary school science diagrams: linguistic patterns and discourse relations

Jonas Haverinen

Master's thesis

Master's Programme in English Studies

Faculty of Arts

University of Helsinki

January 2022

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

Tiivistelmä

Tiedekunta: Humanistinen tiedekunta

Koulutusohjelma: Englannin kielen ja kirjallisuuden maisteriohjelma

Opintosuunta: Yleinen

Tekijä: Jonas Christian Haverinen

Työn nimi: Written labels in elementary school science diagrams: linguistic patterns and discourse relations

Työn laji: Maisterintutkielma

Kuukausi ja vuosi: Tammikuu 2022

Sivumäärä: 56; 3 liitesivua

Avainsanat: Multimodaalisuus, diagrammit, kieliteknologia, RST

Ohjaaja tai ohjaajat: Tuomo Hiippala

Säilytyspaikka: Helsingin yliopiston kirjasto

Muita tietoja:

Tiivistelmä:

Communication, by nature, is multimodal: it uses various forms (modes) of communication, such as spoken or written language, illustrations, and many others to create meaning. Multimodality research is the study of communicative situations that rely on such various modes and their combinations. One form of multimodality very commonly seen in everyday life comes in diagrams, which can convey complex concepts by combining visual expressive resources (such as illustrations or photographs), written language, and diagrammatic elements such as lines and arrows.

The primary aim of my thesis is to establish whether the linguistic structures of written labels – that is, textual elements – in diagrams can inform the decomposition of visual expressive resources. Put simply, I seek to find if said visual elements can more accurately be divided into further, more granular units in accordance with linguistic patterns in their accompanying textual elements. To answer my main research question, I posit three sub-questions. First, whether certain diagram types (macro-structures), such as tables, cycles, or cross-sections co-occur with specific linguistic patterns; second, if different rhetorical functions found in diagrams employ different linguistic structures; and third, if these functions are signaled by other means in tandem with written language. Answering these questions can help in designing future multimodal corpora and their annotation schemata, increasing annotation accuracy and possibilities for their processing.

I approach diagrams from the perspective of multimodality, highlighting them as discursive artefacts. This is enabled by the diagrammatic mode, which establishes how discourse semantics can function in the context of diagrams and how their interpretation is dynamic; each element or combination of multiple elements can in turn contextualize or be a part of others on a different scale. I discuss the concepts of coherence and cohesion as they relate to multimodal artefacts: different elements, even if not linguistic, can combine to create semantically meaningful connections between constituents in such an artefact. To exemplify this, I also apply Rhetorical Structure Theory (RST), which formalizes how units of discourse are interconnected and form a communicative whole. RST employs rhetorical relations such as ELABORATION and IDENTIFICATION to describe how units and their combinations relate to other parts of a discursive whole.

The data I use consists of two interrelated and complementary multimodal corpora: AI2D and AI2D-RST. AI2D is a collection of primary-school textbook science diagrams, annotated for blobs (visual expressive resources), labels, and diagrammatic elements, created for question-answering purposes. It also contains the linguistic data in each of the corpus's diagrams. AI2D-RST contains a subset of the diagrams in AI2D, expanding them with additional annotation layers for information on macro-structures, visual connectivity, and RST, describing each element's rhetorical relation in the diagram.

I computationally find each rhetorical relation containing a label in AI2D-RST, noting its type, the type of the diagram it appears in, and fetching the labels' linguistic content from AI2D. I process each label's contents with spaCy, a library for natural language processing, for linguistic elements such as phrase types, part-of-speech patterns, and average word counts.

There are indeed some differences in how distinct rhetorical relations and macro-groups use language: for example, cycles contain the most verb phrases and highest word count, indicating the use of written language to explicate certain processes. As linguistic patterns differ across these classes and are contextualized by surrounding elements, approaching diagrams from a discursive standpoint may be beneficial for future empirical multimodality research as well as designing more intuitive and precise annotation schemata. With larger datasets and further research, sets of rules containing linguistic structures and layout information may be developed to increase accuracy in the computational analysis of diagrams.

Table of Contents

1 Introduction 2
1.1 Multimodality and diagrams 2
1.2 Research questions 5
2 Theoretical framework and key concepts 6
2.1 Multimodality research 6
2.2 The diagrammatic mode 11
2.3 Cohesion and coherence 17
2.4 Rhetorical Structure Theory (RST) 19
2.5 Signaling coherence relations 20
3 Data 21
3.1 AI2D 21
3.2 AI2D-RST 23
3.2.1 Annotation schema 24
4 Methods 27
5 Analysis 31
5.1 Most common part-of-speech patterns 32
5.2 Unique POS patterns across diagram types 36
5.3 Average word counts 38
5.4 Phrase classes 40
6 Discussion 42
7 Conclusion 48
Bibliography 50
Appendices 57
Appendix A 57
Appendix B 58
Appendix C 59

1 Introduction

1.1 Multimodality and diagrams

Communication is inherently multimodal – that is, it relies on various intentional combinations of spoken and written language, illustrations, photographs, and other "modes" to make and exchange meanings. As interest in the study of these modes of communication is increasing, multimodality is slowly becoming an emerging discipline (see e.g. Bateman et al., 2017; Wildfeuer et al., 2020). Bateman et al. (2017, p. 7) summarize multimodality broadly as "characterising communicative situations … which rely upon combinations of different 'forms' of communication to be effective." These communicative situations encompass an extremely board range, from audiovisual material and face-to-face interaction to newspaper websites and school textbooks, to name a few. Just which of these "forms of communication" are used depends on the communicative situation.

Diagrams are one ubiquitous form of communication in daily life. They are commonly used to convey information of various topics in many different ways and can be found everywhere from instruction manuals to textbooks and news programs (see e.g. Purchase, 2014; Kembhavi et al., 2016; Hiippala et al., 2021). Because diagrams are deployed for communicative purposes in so many contexts, it is hardly surprising that they have become a topic of interest in various fields such as psychology, artificial intelligence, cognitive and social sciences, human-computer interaction, as well as applied linguistics (Purchase, 2014; Wildfeuer et al., 2020; Hiippala and Bateman, 2021). For example, in the field of applied cognitive psychology, diagrams have been used to study the guidance of attention allocation via eye-tracking, highlighting the importance of layout for processing and integrating information (Holsanova, Holmberg, and Holmqvist, 2009); they have also recently been used to research how knowledge is learned and transferred based on different diagrams depicting animal life cycles (Menendez, Rosengren, and Alibali, 2020). In order to support empirical research more efficiently in the disciplines in which diagrams are used and studied, further attention needs to be directed to the multimodal discourse structure of diagrams.

Watanabe and Nagao (1998) suggest that natural language is vital for understanding diagrams, as natural language and layout patterns can together communicate information more efficiently than either could by itself. They show that different linguistic structures can serve different semantic purposes in diagrams: one structure may name a pictured plant, while another might indicate an individual part thereof. It seems, then, that written language guides viewers to interpret diagrams in certain ways, and therefore cannot be overlooked when discussing the communicative potential of diagrams. Watanabe and Nagao (1998) also indicate a need for larger, computationally accessible corpora of diagrams for further computational studies. Such corpora (see Section 3) were unavailable at the time but can now be found annotated with both linguistic and layout data, among others (Kembhavi et al., 2016; Hiippala and Orekhova, 2018; Hiippala et al., 2021). The challenge then becomes how to efficiently support the computational analysis of corpora that are too large for human analysts.

To analyze linguistic structures found in diagrams, natural language processing (NLP) can be used to computationally extract and process written language for numerous attributes. Because written language can be quite explicit in guiding the viewer in its content, structure, and layout, it can be a major constituent of coherence in diagrams, and certainly worth researching in a discourse-oriented approach to them. In fact, many of the key concepts used in this study originate in linguistics, but can be used in multimodal “texts” as well.

As noted by Hiippala et al. (2021), more recent advances in computer vision and natural language processing may assist in the generation and processing of diagrams. Computer vision research is driven by photographic media, whereas diagrams and other forms of visual and pictorial expression have received less attention. Crucially, diagrams are radically different from photographs due to properties such as *compositionality* – that is, diagrams combine multiple forms of communication into discourse organizations. A single diagram may include illustrations, text, and connecting lines all in one, which photographic material seldom contains; including diagrammatic and linguistic representation in addition to pictorial (see e.g. Greenberg, 2021) then requires approaches not limited to purely photographic material. Kembhavi et al., 2016 (p. 325) also point out that understanding of such “rich” visuals is scarce.

There are exceptions as well, however, such as Haehn, Tompkin, and Pfister (2019), who encourage further research into machine graphical perception. Because prior research is scarce, the structure and semantic functions of possible written language may not have received enough attention in the computational processing of rich visual material, either – not to mention how written language and other modes are combined in effect. Multimodality theory offers a promising framework to approach this problem with, as the discourse relations between different modes of expression are a common topic in the emerging field.

As diagrams can provide numerous ways to communicate a given concept, their designers must choose which ones to use and how if they wish to efficiently convey information. A diagram's discourse structure then reflects its communicative intentions, which highlights the potential of the written language found in diagrams as a tool for explicitly guiding viewers to infer and further decompose visual elements. Written *labels*, which are linguistic elements in diagrams, often function to describe entire elements or parts thereof; as such, they are particularly useful for communicative purposes. Labels can be identified by their proximity and relative placement to the element they describe, or by connectivity, in which the label is connected to the described element via a diagrammatic element (such as a line or an arrow). Figure 1 shows an example of labels describing entire objects, while Figure 2 demonstrates how labels identify individual parts of a depicted object.

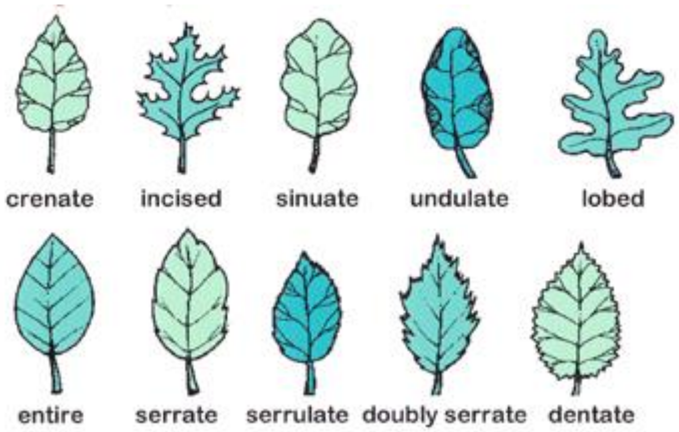

Figure 1. An example of labels classifying individual elements (types of tree leaves). The labels are located below each element, exemplifying consistent proximity. Diagram #4405 from AI2D (Kembhavi et al., 2016; see Section 3).

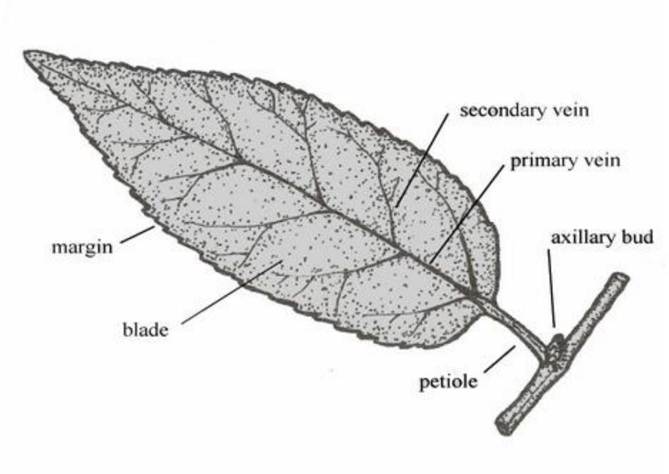

Figure 2. An example of labels describing parts of the element (an illustration of a tree leaf). The labels are connected to the described parts of the element via lines. Diagram #3149 from AI2D (Kembhavi et al., 2016).

As can be seen from Figures 1 and 2, linguistic and diagrammatic elements serve as guides for interpreting illustrations, drawings, graphic shapes, and other forms of visual representation in diagrams (Tversky et al., 2000). Lines, written language, and their placements clarify the information the diagram is attempting to convey and guide viewers in interpreting it. Whether viewers interpret labels as describing complete elements or only parts of them may be affected by the presence of these attributes.

1.2 Research questions

My main research question, then, is as follows: "Can the linguistic structure of written labels inform the decomposition of visual expressive resources in diagrams?" More specifically, I seek to answer the question above by answering the following questions:

1. Do different types of diagrams generally use different linguistic structures in their labels?
2. Do certain linguistic patterns co-occur with specific rhetorical relations between visual expressive resources and their labels?
3. How are the discursive functions of labels with different linguistic structures signaled for viewers to interpret?

Finding answers to the presented questions would assist in understanding diagrams as a form of communication and aid various diagram research efforts. The specific inherent discursive characteristics of diagrams should be considered when creating annotation schemas for multimodal corpora (see Hiippala and Bateman, 2021). Such data could be found useful in both crowdsourced and automatic interpretation of diagrams.

To answer the questions introduced above, I combine theories and approaches from the studies of multimodality, discourse structure, diagrams research, and computational linguistics to answer some of these questions. I first establish essential theories and background literature; next, I provide an overview of the data; I then go over the methods used to process said data; then, I present my analysis; and finally, discuss its further changes and implications for discourse-oriented diagram research.

2 Theoretical framework and key concepts

In this chapter, I will establish the theories and approaches I follow in this thesis. I begin by going over multimodality research and facets thereof pertaining to my topic; I then introduce the diagrammatic mode, which guides my approach; I follow this by defining the concepts of cohesion and coherence; next, I present Rhetorical Structure Theory (RST), which has been applied to multimodal material in previous studies; finally, I discuss how different coherence relations that hold between elements are signaled.

2.1 Multimodality research

Since multimodality is a ubiquitous phenomenon, it has unsurprisingly been studied in diverse fields. Multimodality research is the study of multimodal artefacts and situations, as well as the communicative phenomena found therein. The emerging field is expanding rapidly due to the increasing interest in multimodal communication in various disciplines and contexts. Because of the diversity of different approaches to multimodality, there is no single universal theory of multimodal communication; the field is as heterogenous and broad as the disciplines it is studied within (see e.g. O'Halloran & Smith, 2011; Bateman et al., 2017; Wildfeuer et al., 2020). Nonetheless, as O'Halloran and Smith (2011, p. 2) note:

> There has been a clear movement towards the development of generalisations applicable beyond the particular concerns of those studying within particular domains of reference or with particular academic backgrounds and with application to the study of multimodal phenomena in general.

There are, then, approaches to multimodality that are applicable across disciplines and beneficial to a diverse range of studies on the topic. As multimodal phenomena are researched in an increasing capacity, theories and approaches to their analysis are necessary. Authors such as O'Halloran and Smith (2011), Jewitt et al. (2016), Bateman et al. (2017), and Wildfeuer et al. (2020) provide possible frameworks for the systematic analysis of communicative situations with combined expressive resources, indicating the emergence of multimodality studies as a field.

Stöckl (2020, p. 41) proposes that contemporary linguistic and semiotic approaches to multimodality are defined by two central tenets, described as "the semiotic dictum that communication relies on a whole host of different signing modes and their combination, and the linguistic concerns evident since the advent of pragmatics and text linguistics with a gradual extension of context." The first of these tenets then enables different signing modes (such as language, illustrations, and music) to be analyzed from the perspectives of pragmatics and semiotics. The second establishes that multimodality is characterized by an extension of existing topics in pragmatics and text linguistics, such as discourse and related phenomena; these phenomena, such as cohesion and coherence (see Section 2.3), are then applied to forms of communication other than language. This in turn enables the co-contextualization of such forms of communication, wherein each provides additional context for the others. The issue that follows, then, is what exactly constitutes such a form of communication – that is, a *mode* – and how they could be distinguished.

In order to draw systematic and accurate distinctions between different communicative situations, it is essential to establish their material components: such situations can unfold in different ways using different *materialities*. For example, a comic strip and a face-to-face discussion are drastically different forms of communication tracing back to the different materialities they employ, and as such, a mode is partially defined by its material aspects. Bateman et al. (2017) and Bateman (2021, pp. 40–42) classify four dimensions of materiality: temporality (static or dynamic), space (two-dimensional or three-dimensional), role (observer or participant), and transience (permanent or fleeting). As these dimensions vary, so do the possible communicative situations within them. A comic, for example, would be static, two-dimensional, and permanent, which an in-person, face-to-face chat is certainly not; the latter also includes the possibility of participation. Because of these vast differences in communicative possibilities, materiality is a necessary part of multimodal analysis that can be built on.

A mode then has a material stratum: a kind of "canvas" (Bateman et al., 2017, pp. 86–87) that can be modified. The canvas is not necessarily literal, but instead describes how multimodal artefacts can be perceived and how the materiality can be manipulated for communication; materiality, as it relates to multimodality, can then be anything from a digital screen to physical actions transpiring over a period of time. A two-dimensional, non-aural materiality such a piece of paper can hardly support the same modes of expression as face-to-face communication, which relies on body language and sound; films differ noticeably from literature in how they can present information due to the properties and restrictions of their corresponding materialities, as a page is completely static. Furthermore, as noted by Bateman (2021), it is not uncommon for communicative situations to "exhibit highly complex material structures with distinct levels of embedding, each of which then requires its own material classification" (p. 42).

Regarding the further distinction of modes, Kress et al. (2001, p. 43) provide an example:

> [T]he question of whether X is a mode or not is a question specific to a particular community. As laypersons we may regard visual image to be a mode, while a professional photographer will say that photography has rules and practices, elements and materiality quite distinct from that of painting, and that the two are distinct modes.

What may constitute a mode in one community or context might therefore not be recognized similarly in another. This also shows that materiality should ideally be described through both how it can be manipulated as well has how it is perceived: even in a given context of perception, there may be multiple materialities at play, and a single substrate can carry various modes. In the context of diagrams research, cut-outs and photographs may serve distinctly different purposes, for example.

A materiality can then be manipulated intentionally for communication. Such manipulation must consistently follow formal distinctions associated with a given semiotic mode. Any such "regularities of form" on a materiality can then develop into *expressive resources* associated with that semiotic mode. As noted by Hiippala and Bateman (2021, p. 3) specifically regarding diagrams, this is “exemplified by differences in form between written language and line drawings, which allow us to distinguish between these resources.” That is, written language and line drawings have different principles of organization that result in distinct forms and patterns thereof. A mode can then make use of a number of different expressive resources depending on how its materiality can be manipulated for communicative purposes and what properties the materiality possesses – it can be spatial, temporal, two-dimensional, three-dimensional, static, or dynamic, among others, and all of these properties and their different combinations allow for different kinds of communication. It is important to note, however, that the attributes of materiality do not necessarily restrict what can be *represented* by it. As Bateman et al. (2017, p. 102) note:

> [W]hile we cannot derive from a particular canvas any statements about what can be represented, we can say much about how it might be represented. A particular canvas will only make certain material distinctions available and not others, and this is what is available for meaning-making.

For example, a static materiality such as a piece of paper may still represent movement through a sequence of illustrations; these are the kinds of limitations defined by the underlying materiality of the semiotic mode.

Ultimately, the contextual interpretation of different combinations and selections of expressive resources is facilitated by *discourse semantics*, which guide viewers’ interpretation. Hiippala and Bateman (2021) give an example from the viewpoint of diagrams by stating that “resolving the resulting discourse relations relies on formal cues such as spatial placement of elements or connections realised using lines and arrows in combination with world knowledge” (p. 4). The above aspects in unison constitute a semiotic mode, as defined by Bateman (2011; see also Bateman et al., 2017; Hiippala & Bateman, 2021).

Another central aspect of multimodal analysis is the concept of *medium*. Bateman et al. (2017, p. 123) summarize the concept broadly as “a historically stabilised site for the deployment and distribution of some selection of semiotic modes for the achievement of varied communicative purposes.” Hence, the medium (such as a book or a film) determines what kinds of modes can be used therein – although it should be noted that semiotic modes themselves are abstractions and not intrinsically bound to any particular medium, but instead can be mobilized within a given medium if it provides a compatible materiality. This results in certain properties of multimodal communication being medium-specific: books and films employ distinct combinations of semiotic modes in different ways and said combinations can be anticipated from an artefact of the corresponding medium. One would hardly expect encountering moving images or musical tracks in a book, although they could be represented via other modes, as discussed above.

The notion of *genre* is also of importance in multimodal research. Genre ascribes classes to certain recognizable patterns of conventions used to achieve an intended communicative purpose (see e.g. Bateman et al. 2017 pp. 128–131). Genres enable those who participate in communication within them to set certain expectations for them, which also help guide their interpretation; examples of genre could then be a crime film, a science textbook diagram, or perhaps a master's thesis. Any multimodal *text* – that is, a unit resulting in the combination of semiotic modes afforded by a medium (Bateman et al., 2017, pp. 131–133) – participates in a genre. As Stöckl (2020, p. 65) summarizes:

> Placing multimodal ensembles firmly in the context of a rhetorical situation means to look at them as comprising a rhetor's goal, rhetorical strategies as deliberate choices and combinations of semiotic resources, and a recipient's task-based communicative engagement in the resulting multimodal structure. Such a consistently rhetorical approach allocates genre a central role in shaping and constraining multimodal discourse interpretation. Any multimodal artifact would then first of all be an exemplar instantiating or realizing a genre's underlying functional, logical, structural, and stylistic regime.

This illustrates how the importance of genre as a guide for the rhetorical interpretation of multimodal artefacts cannot be understated (see also Bateman, 2008).

The notion of genre helps in the analysis of diagrams more specifically, as well. Different types of diagram structures, such as cycles, tables, cross-sections, or exploded views effectively function as genres: these various types of diagrams guide viewers towards certain interpretations. For example, viewers may infer the cyclical sequence between some of a cycle's constituents at first glance, while this might not be the case with a table. There are then certain conventions in place that illustrate functional differences and set viewer expectations for diagrams' contents.

In summary, due to the complexity of multimodality, pursuing multimodal analyses requires well-defined theoretical concepts set in appropriate relations with each other. Modes and their combinations co-textualize and contextualize one another. This allows for various methods of communication, the expressive resources of which are formed by manipulating the materiality they are set upon and guided by their corresponding discourse semantics. The medium determines the kinds of materialities available for a text of multimodal communication. Any combination of semiotic modes for achieving communicative goals is bound to exhibit patterns determined by genre, which provides additional context for its construction and interpretation. These concepts enable the systematic analysis of multimodal artefacts.

2.2 The diagrammatic mode

As modes of expression and their different organizations and combinations form discursive wholes and discourse semantics can take place on various levels in multimodal artefacts, exactly how far those modes can be decomposed may affect the successful interpretation thereof. Discourse semantics thus establishes *discourse units*, which are compositional in nature: while a multimodal artefact can act as a page-level discourse unit, further decomposing such artefacts requires viewers to re-apply the appropriate discourse semantics to outline units on a more granular level. Just as other forms of multimodal discourse, diagrams organize instances of expressive resources into discourse structures, the interpretation of which is supported by the stratum of discourse semantics. Diagrams use various elements such as arrows, text and images to communicate different things; they can be divided into their component parts, each of which functions as a part of the whole diagram. Because proper decomposition can affect the ability with which a diagram's message is interpreted, the question then becomes how diagrams might be decomposed more accurately in order to fully comprehend the messages they are trying to communicate.

Discussing sufficient decomposition of diagrams, particularly in the corpora used for this thesis (see the Section 3), Hiippala et al. (2021, pp. 663–664) note:

> 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 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.

The dataset contains multiple diagrams whose visual expressive resources have not been segmented at the same level of detail as the labels. This begs the question if the granularity of labels corresponds with the needed level of detail in visual segmentation. Hiippala et al. (2021) follow this by stating that “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” (p. 683) and that “these expressive resources must be complemented by sufficiently fine-grained descriptions of graphic expressive resources” (p. 684). Because effectively uncovering the discursive structure of diagrams necessitates increasingly granular identification and analysis of discourse units, it can be argued that the annotation schema used to describe the diagrams within the corpora does not capture the structure of diagrams as multimodal artefacts appropriately.

To address this problem, Hiippala and Bateman (2020; 2021) introduce the *diagrammatic mode*, the purpose of which is to view and analyze diagrams as a semiotic mode (see also Bateman et al., 2017, pp. 279–294). The diagrammatic mode is built on a theoretical model of a semiotic mode introduced by Bateman (2011). Bateman (2011) proposes a new model of a semiotic mode, which consists of three strata: the *materiality*, which acts as a basis to and defines the kinds of resources that can be used; the formally distinct *expressive resources* created by manipulating the materiality in different ways, which can also be selected and combined to create broader organizations; and the *discourse semantics*, which guide readers’ interpretations of those expressive resources. Figure 3 shows a visual representation of the strata in said theoretical model of a semiotic mode and the diagrammatic mode.

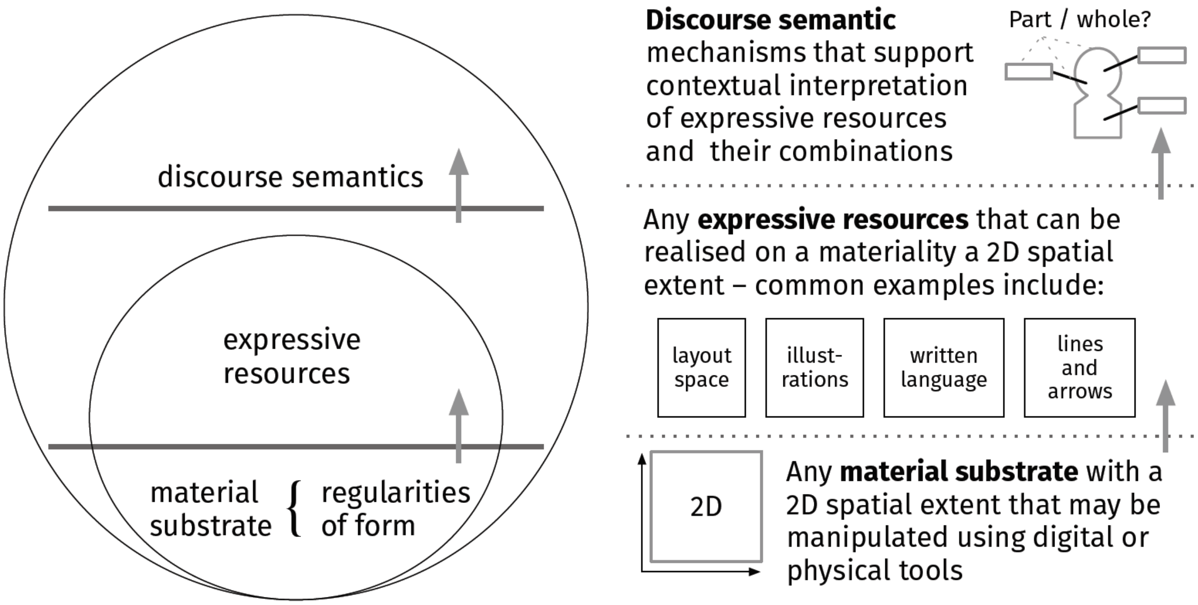
discourse semantics
expressive resources
material substrate
regularities of form
Discourse semantic mechanisms that support contextual interpretation of expressive resources and their combinations
Part / whole?
Any expressive resources that can be realised on a materiality a 2D spatial extent – common examples include:
layout space
illust- rations
written language
lines and arrows
2D
Any material substrate with a 2D spatial extent that may be manipulated using digital or physical tools

Figure 3. A theoretical model of a semiotic mode (left) and the diagrammatic mode (right) with their corresponding strata, as illustrated by Hiippala and Bateman (2021, p. 3). Used with permission.

Based on the model by Bateman (2011), the diagrammatic mode is accordingly divided into three semiotic strata that correspond to materiality, expressive resources, and discourse semantics of a semiotic mode, realized via multimodal aspects. The strata in the diagrammatic mode are, respectively, a materiality with two-dimensional extent to support diagrams; semiotic resources that require this material (such as layout space and written language); and the mechanics guiding the interpretation of said semiotic resources.

As any fully developed semiotic mode, the diagrammatic mode provides discourse semantics, the purpose of which "is to identify candidate interpretations which are then resolved *dynamically* against the context in which the expressive resources appear" (Hiippala and Bateman, 2021, p. 3, italics in original). Expressive resources can then allow viewers of a diagram (who may have less information or *world knowledge* of the depicted phenomenon than the author) to infer meaning from it. As Bateman (2011, p. 22) states: "discourse semantic rules control when and how world knowledge is considered in the interpretation process." For example, without written labels, arrows in a food network would have to be interpreted by viewers as replacements for processes such as "eats" or "is eaten by" using their world knowledge (see Alikhani and Stone, 2018), as exemplified by Figure 4. World knowledge is thus used by viewers to fill any gaps left in the diagram's discourse structure.

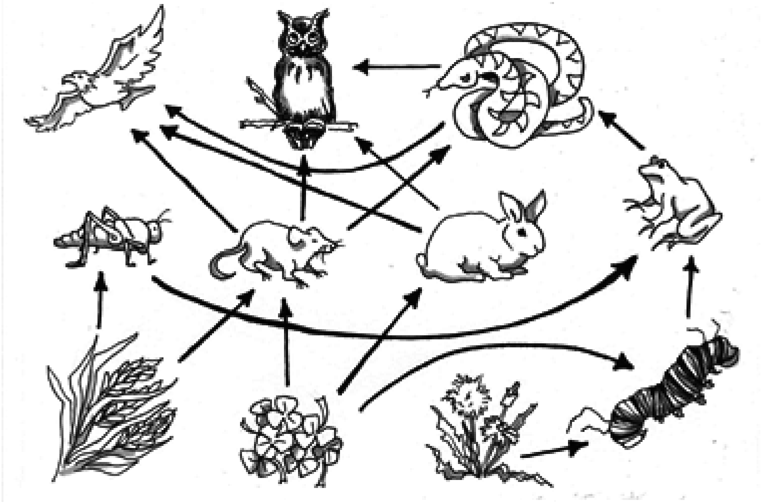

Figure 4. An example of a food web in which world knowledge is required to interpret the meaning of the diagram's arrows. Diagram #450 from AI2D (Kembhavi et al., 2016; see Section 3).

As a fully-articulated semiotic mode, the diagrammatic mode provides mechanics that guide interpretation (as well as their combinations) on the level of discourse semantics; it can be argued that the amount of world knowledge the receiver needs is dependent on the discourse semantics used in the diagram. Hiippala and Bateman (2020) establish that

> [t]he contribution of discourse semantics is also not limited to guiding the interpretation of *local* discourse relations that hold between two or more diagram elements, because such local interpretations are also always evaluated within the context provided by the *global* discourse organisation, which may as a consequence already nudge a viewer towards particular candidate interpretations rather than others (p. 4, italics in original).

In practice, along with the notion of dynamic interpretation, this means that instead of only following a diagram's structure from each small unit up towards the global representation, viewers may – and perhaps should – interpret each element in light of everything else presented visibly in the text.

The importance of entire diagrams' rhetorical structure is especially apparent in the example shown in Figure 5: although illustrated as a cross-section with numerous regions, the image has been annotated in the data as a single element.

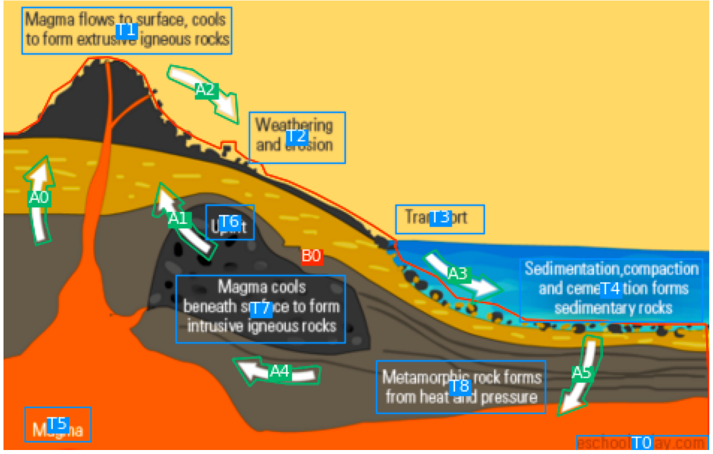
Magma flows to surface, cools to form extrusive igneous rocks
T1
A2
Weathering and erosion
T2
A0
A1
Uplift
T6
B0
Transport
T3
A3
Sedimentation,compaction and cemeT4tion forms sedimentary rocks
Magma cools beneath surface to form intrusive igneous rocks
T7
A4
Metamorphic rock forms from heat and pressure
T8
A5
Magma
T5
eschoolT0ay.com

Figure 5. The annotation of diagram #4210 from AI2D (Kembhavi et al., 2016). The cross-section of the volcano has not been decomposed and instead has been annotated as a single blob.

Seeing as this annotation is insufficient in its decomposition of presented information, Hiippala and Bateman (2021) argue that "[t]his information is crucial for understanding what the diagram is attempting to communicate *but we cannot know that such a decomposition is necessary without considering the rhetorical discourse organisation of the diagram as a whole*" (p. 10, italics in original). The example diagram contains a cycle without it being explicitly signaled using arrows. Instead, only some arrows occur, and the rest seems importantly implied by the written language present that describes the cycle using entire phrases such as "Metamorphic rocks form from heat and pressure". Hiippala and Bateman (2021) conclude that discourse-oriented segmentation of the corpus could be beneficial in terms of further research, which informs my thesis regarding the importance of linguistic structure in labels.

Previous research in the broad field of artificial intelligence and natural language processing can offer additional insight into the matter. Watanabe and Nagao (1998) find that "[p]attern information and natural language information used together can complement and reinforce each other to enable more effective communication than can either medium alone" (p. 1374) and that inferring the meaning of diagrams can be noticeably more difficult without linguistic guidance (see Figure 4). The study uses diagrams of flora, with each diagram consisting of illustrations, written language and possibly diagrammatic elements such as lines. Figure 6 includes a similar example from the AI2D dataset (Kembhavi et al., 2016) used in this thesis (see also Figures 1 and 2).

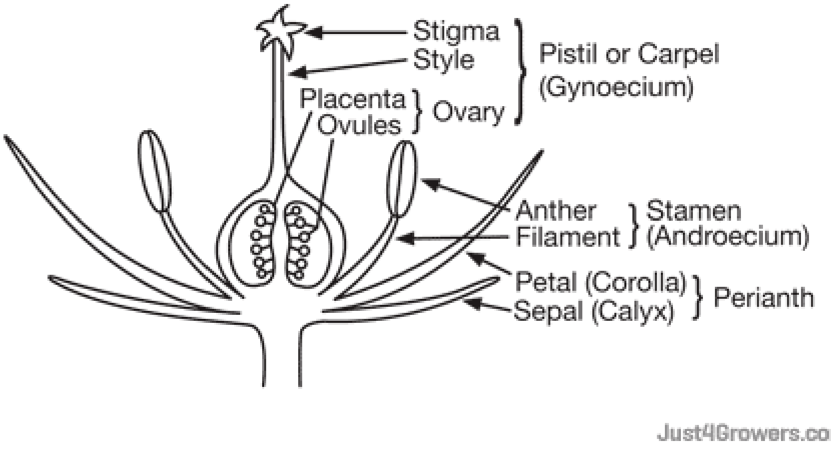

Figure 6. An example of a diagram pointing out parts of a flora similar to those used by Watanabe and Nagao (1998). Diagram #3118 from AI2D (Kembhavi et al., 2016).

Watanabe and Nagao (1998) use two kinds of data to determine how semantic interpretation is guided in the data: layout and natural language of labels (that is, occurrences of written language). For layout, Watanabe and Nagao (1998) examine adjacency and connection: a label is connected if it is connected to its corresponding element via a line, and adjacent if it is next to its corresponding element, but not connected. In terms of linguistic patterns, the study finds five specific phrasal patterns in the data, some of which occur precisely in examples of *ascribing properties to* the flora but not when *identifying parts* of them (and vice versa). Synthesizing layout and language patterns, the authors establish a step-by-step ruleset that accurately helps come to a conclusion of the various semantic relations in the data, but lament the lack of computational resources in the process: being able to computationally extract and analyze large quantities of diagrams would help establish these rulesets more concretely with a larger sample size and save the effort of manually encoding linguistic and diagrammatic elements for similar studies. At the time of writing, there were also no large, widely available, and annotated multimodal corpora of diagrams fit for the purpose; however, this is now afforded by AI2D and AI2D-RST (see Section 3). Given that the data used by Watanabe and Nagao (1988) is very similar to some of the diagrams found in the corpora used in this thesis, and the fact that the corpora contain annotated information on semantic relations, written language and layout of the diagrams' elements, a similar approach may help identify linguistic patterns that guide viewers towards precise segmentation and interpretation.

It therefore seems plausible that written language and the structure thereof can be of value to the study, interpretation, and generation of diagrams. Exploring how diagrams are structured discursively using expressive resources and the relations that hold between them with the aid of NLP and multimodality might provide insight into how much of the information is given to viewers and what is left for them to interpret – and whether present written language can accordingly guide them to more precise decomposition.

2.3 Cohesion and coherence

Discussing diagrams as a semiotic mode allows for approaching them as discursive, multimodal artefacts, which in turn enables the application of other concepts developed in multimodality theories to them. Being communicative in nature, multimodal artefacts can then be approached from the perspectives of cohesion and coherence, which in the context of multimodality aim to establish how different modes fit and function together in the multimodal artefact, (co-)contextualized by each other and the genres they exhibit.

Cohesion, coined as a concept by Halliday and Hasan (1976), originates in linguistics. Cohesion refers to how the same object is referred to in different ways throughout a text and how different elements fit together. As a phenomenon, cohesion is *non-structural* in nature; that is, instead of depending on the structural arrangement of a text, it can be applied within individual units or across them, even on the scale of an entire text. In other words, cohesion is not dependent on the presence of specific discourse units; instead, it is mostly concerned with semantics and subject matter. Cohesive ties can be formed via repetition or referring to the same entity or concept in different ways (see e.g. Hiippala, 2015, p. 18). Being non-structural, cohesion is readily adaptable in the context of multimodality. To further describe cohesion, specifically between text (in the sense of written language) and image, Bateman (2014a) finds:

> This reliance on dependency in interpretation rather than structural configurations requiring particular types of diagrammatical elements has made it relatively natural to consider the possibility that similar relations might hold *even when the elements in a cohesive tie are not linguistic elements at all.* (p. 165, italics in original)

This suggests that cohesive ties can hold between two multimodal elements, linguistic or not. Using specifically captions as an example, Bateman (2014a) states that if the image and text are “*designed* to operate together” (p. 165, italics in original), it is appropriate to speak of cohesion between the two elements, and so also beneficial to view the combination of written language and image as a *textual unit*. Semiotic ties between written language and image in multimodal discourse have been established to hold just as well as purely linguistic data (see also Liu and O’Halloran, 2009). Bateman (2014b) posits that cohesion is one approach that can help in interpreting multimodal coherence but notes that “[a]lthough the ‘cohesive’, non-structural approach to analysis certainly allows many connections to be made explicit within any multimodal artefact, it is considerably less effective as a tool for engaging critically with [them]” (p. 165).

Coherence, in turn, is a structural phenomenon. It is created by the segments of a text – or in the context of multimodality, instances of different expressive resources or semiotic modes – functioning in different *coherence relations* to one another and forming structures, be they informational, genre-specific, or thematic (see e.g. Gruber and Redeker, 2014). Establishing coherence is a central task in discourse semantics, and a coherent text creates plausible explanations for the relations that hold between a text’s constituents, thus supporting its interpretation. According to Gruber and Redeker (2014, p. 2), coherence relations “establish semantic or pragmatic relations between units that express (simple or complex) propositions or illocutions” and “describe how parts of a discourse combine recursively to form larger chunks and eventually the whole structure.” The units then can range in scope from a single element in a text to the entire discourse. This idea of recursion is also relevant for the analysis of diagrams, as they may contain such “chunks” in a multitude of different organizations. In the following section, I introduce one framework for systematically describing coherence relations: Rhetorical Structure Theory (see Section 2.4) first established by Mann and Thompson (1988). Coherence relations can also be signaled in various ways; in purely linguistic material, a common method for this is the use of connectives, but other markers are able to signal these relations as well (see Section 2.5).

2.4 Rhetorical Structure Theory (RST)

Seeing as discursive ties hold between different modes, a model for the analysis of multimodal discourse can be applied. *Rhetorical Structure Theory* (Mann and Thompson, 1988), henceforth referred to as *RST*, is one prominent framework for analyzing discourse structures. Although initially developed in the field of linguistics, RST has been successfully applied to multimodal material across varying scales, from individual modes to entire documents (see e.g. André and Rist, 1995; Bateman, 2008; Bateman 2014; Bateman et al., 2017; Hiippala and Orekhova 2018; Hiippala et al. 2021; Taboada and Habel, 2013).

RST is concerned with discourse structure; it seeks to explicate how discourse units are related to each other and how they work towards a shared communicative goal. Smaller-scale units may participate in coherence relations, which can then participate in further, higher-level relations. For example, there may be a relation between two units, which then together as a discourse unit participate in a relation with another unit. Examples of such relations are ELABORATION, BACKGROUND, ENABLEMENT, SEQUENCE, and RESTATEMENT; the distribution of these relations varies by application. Depending on the context in which RST is applied, further relations may arise, such as IDENTIFICATION and PROPERTY-ASCRIPTION (Bateman 2008) or CYCLIC SEQUENCE (Hiippala et al., 2021) in multimodal analysis. RST defines two types of discourse relations, which Bateman and Delin (2006, p. 591) summarize as follows:

> [T]here are two kinds of rhetorical relations: asymmetric relations, where one of the related rhetorical units is singled out as the rhetorical head, or nucleus, and symmetric relations, also termed multinuclear, where all of the related units are of equal status.

The nucleus of an asymmetric rhetorical relation is considered more central for the text, which its satellites support and enhance by providing, for example, elaboration or enablement. In symmetric relations, each segment is equally needed for coherence. In multimodal artefacts, the same discourse unit may also simultaneously participate in multiple relations, which offer different perspectives to said unit (see Bateman, 2014, pp. 219–220; Hiippala et al., 2021).

Crucially, Mann et al. (1989, p. 15) posit that RST relations are to be considered *plausibility judgments* instead of certainty – in other words, they are judgments made by the analyst based on the text about what the author or designer of a text plausibly intended to communicate to its intended audience. Such inferences about communicative intent are made based on different aspects in the text that imply certain discursive functions, the scrutiny of which discourse semantics facilitates.

2.5 Signaling coherence relations

Different types of coherence relations can be signaled by different markers: often in linguistic material, *connectives* (that is, connective expressions) such as conjunctions can signal coherence. Certain cues can be more explicit in signaling specific relations, although there is no universally applicable model of correspondence between cues and relations. The frequency and types of connectives also vary by genre and level of discourse from an individual element (such as phrase) to the entire text (see e.g. Gruber and Redeker, 2014).

Das and Taboada (2019) approach signaling via discourse markers and relation semantics (that is, connections between parts of discourse). The study uses RST and shows that while some relations can be inferred through explicit discourse markers, they can be ambiguous enough that combined signals are needed to glean meaning in many cases. In essence, multiple signals working simultaneously can strengthen coherence. Although not venturing beyond written language, the study demonstrates the value in scrutinizing multiple signals in unison to infer coherence relations.

Multimodal artefacts include numerous co-operating signals and can be approached from the viewpoint of coherence relations. Therefore, an approach regarding multimodal signaling of coherence relations in the context of diagrams and their various more specific genres can be beneficial for approaching them as a form of discourse. Alikhani and Stone (2018) find that arrows function discursively in diagrams very much like verbs in natural language, while Watanabe and Nagao (1998) demonstrate how layout information can be used to assess the discursive purpose of written labels more accurately. These studies show that diagrammatic elements and layout can contribute to coherence, and thus serve as signals for coherence relations in diagrams just as much as written language and illustrations.

This concludes the discussion of contemporary theories of multimodality and their application to diagrams. In what follows, I discuss the data used in this thesis, how the relevant corpora were compiled, how they are structured, and how they relate to the theories and approaches shown here. I also further examine their annotation schemata.

3 Data

In this section, I introduce two recent corpora of diagrams: AI2D and AI2D-RST. The corpora are interrelated, as the diagrams present in AI2D-RST are a subset of AI2D. The two corpora have been developed in different fields of research, and as such, for different purposes: AI2D is intended for training algorithms to answer questions about diagrams and their structure in the field of artificial intelligence, while AI2D-RST was developed for researching multimodality and discourse coherence. I first explain the content, annotation schema, and annotation process of AI2D, after which I do the same for AI2D-RST, illustrating their uses and differences.

3.1 AI2D

The first corpus is the Allen Institute for Artificial Intelligence Diagrams dataset (henceforth referred to as AI2D) compiled by Kembhavi et al. (2016) to support research on automatic diagram understanding and question answering. The dataset consists of some 5,000 grade-school natural science diagrams that cover topics such as food webs, rock cycles, and human anatomy.

AI2D's annotation schema is partially built on Engelhardt's (2002) theory of a "visual grammar" for diagrams, etc. and explores structural and semiotic aspects of graphic representations. The theory proposes a visual syntax, in which a composite graphic object consists of a graphic space, other graphic objects, and a set of graphic relations. This means that a graphic object can be a compound, consisting of smaller graphic objects, and that this can apply recursively. The most elementary graphic objects Engelhardt (2002, p. 24) compares to linguistic morphemes. Building on this, Engelhardt (2007) claims "that all graphics are based on the possibility of combining graphic constituents (graphic objects) of different syntactic categories" (p. 27) and asks what level of detail these elementary, "basic signs" (p. 29) can be found on. In light of this, Kembhavi et al. (2016) describe diagrams as "a composite graphic that consists of a graphic space, a set of constituents, and a set of relationships involving these constituents" (p. 238).

Kembhavi et al. (2016) then establish the different relationships and constituents that appear in the data. Ten relation definitions are provided; these relations are mostly local in scale, as opposed to assessing the discourse on the level of the entire diagram, and graphic in that they describe the visual constituents and the connections between them – this is identified as an issue by Hiippala and Orekhova (2018). Said connections can be spatial or based on attributes such as color or linkage, for example. The established relations include (but are not limited to) Intra-Object Label, in which a text box names an entire object; Intra-Object Region Label, which refers to a region within an object; Intra-Object Linkage, which denotes a label referring to a region within a visual object using an arrow; Inter-Object Linkage, in which two objects are connected via an arrow; as well as Arrow Head Assignment, which is an arrowhead connected to its tail. Engelhardt's (2002) theory as described by Kembhavi et al. (2016, p. 4) classifies the different constituent elements as illustrative (such as drawings), textual, diagrammatic (arrows and lines), informative (i.e. legends), and decorative. A vital distinction lies in the fact that the theory adopted by Kembhavi et al. (2016) is based on structure and not discourse semantics.

The diagrams in AI2D were collected via web-scraping Google Image Search using terms extracted from grade-school science textbook chapter titles. A total of over 5,000 diagrams were collected and then annotated via Amazon Mechanical Turk (AMT), a crowdsourcing platform. Crowdsourcing, as comprehensively defined by Pedersen et al. (2013, p. 7), is "[a] collaboration model enabled by people-centric web technologies to solve individual, organizational, and societal problems using a dynamically formed crowd of interested people who respond to an open call for participation." Crowdsourcing may be used to access collective intelligence and lower costs (Pedersen et al., 2013, p. 1). As discussed in Section 2, multimodality is a complex phenomenon, and so crowdsourcing descriptions of multimodal communication is challenging; thus, to ensure accessibility for non-expert workers, the crowdsourcing tasks need to be simple. To make the annotation process feasible and to maintain agreement between AMT annotators, Kembhavi et al. (2016) split the task into a six-step sequence.

The six-step sequence consisted of distinct phases, beginning with identifying basic constituents such as images, text, and arrows. This very simplistic distinction however neglects the possible different communicative aspects afforded by multimodal artefacts (see Section 2). The sequence then continued with categorizing the constituents and finally forming multiple-choice questions and answers for the diagram in its entirety (Kembhavi et al., 2016, p. 243). This division already presupposes separate functions for different semiotic modes. The annotation as such splits each diagram into its constituents by type – blob, text box, arrow, or arrowhead. The annotation schema therefore does not regard diagrams as discursive wholes with more granular discourse units and combinations of different modes of expression, but as a collection of visual components. A crucial point also lies in that annotators were not instructed to decompose the constituent type of *image* into further component parts, even though this is a common occurrence in diagrams (see Section 2). This seems to have resulted in the insufficient composition of images, as the corpus contains numerous blobs outlined in full instead of granular parts that may otherwise be interpreted as analytical units (see Figure 5). It should also be noted that the schema does not separate blobs further into illustrations, photographs et cetera, which may differ in their semantics (see Section 2; see also Greenberg, 2021).

The annotation process generated approximately 150,000 annotations and 15,000 multiple-choice questions. Kembhavi et al. (2016) introduce Diagram Parse Graphs (DPG) to encode the annotated data, representing the diagrams' different elements and the connections between them. In DPGs, annotated objects are represented by nodes, while the edges correlate to the relations between them.

AI2D has successfully been used in other projects to examine semantics in diagrams, such as Alikhani and Stone (2018) to interpret the function of arrows from the perspectives of linguistics, crowdsourcing, and machine learning. The study finds that in certain contexts, arrows serve largely the same purpose as verbs in the dataset, although it also notes potential limitations in automatic semantic parsing of the corpus; in particular, the lack of stroke order and sometimes ambiguous general purpose of diagrammatic elements (Alikhani and Stone, 2018, p. 3559).

3.2 AI2D-RST

To assist in the empirical research of diagrams from a multimodal perspective as well as their computational processing, Hiippala et al. (2021) present the AI2D-RST corpus of diagrams, stating that "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 crowd-sourced annotations" (p. 662) from AI2D. This is to say that the inventory of analytical units in AI2D-RST was populated by the crowdsourced annotations; elements are precisely as annotated for AI2D.

Whereas AI2D used crowdsourcing for its annotation process, AI2D-RST was annotated by five experts with backgrounds in English, trained in the annotation process and familiar with RST. As a result, the annotation quality is notably consistent – however, as the annotators were permitted to discuss individual cases, questions of potential circularity between annotators and reproducibility arise (Hiippala et al., 2021, pp. 679–681). The lack of visual decomposition in the original AI2D annotation hinders AI2D-RST somewhat, as more detailed annotation would in turn enable more precise RST analysis. Ideally, if multimodal cohesion can be used to guide future annotation attempts, these issues could be minimized via naive annotators and consistent, sufficient decomposition.

Hiippala and Orekhova (2018) propose adopting Rhetorical Structure Theory (RST; Mann and Thompson, 1988) as a model of discourse structure for diagrammatic representations, as RST has previously been applied to multimodal artefacts successfully (Hiippala and Orekhova, 2018, p. 1925; see Section 2.2). Since diagrams are multimodal in nature, multimodal analysis can contribute significantly to the dataset, aiding in examining diagrams' communicative properties. Hiippala and Orekhova (2018) note that local discourse relations are not sufficient for capturing the discourse semantics of diagrams. According to Hiippala and Orekhova (2018), a more comprehensive theory is needed, and RST provides one such theory. Moreover, the study finds exemplary rhetorical relations in the corpus by using RST. The application of RST enables detailed and descriptive representation of the relations that elements and their combinations participate in.

3.2.1 Annotation schema

Hiippala et al. (2021) apply the ideas proposed in Hiippala and Orekhova (2018) to create a corpus of 1,000 AI2D diagrams with multiple layers of annotation. The annotation schema introduced by Hiippala et al. (2021) has four distinct layers of information for each diagram: grouping, macro-grouping, connectivity, and RST. Each annotation layer is represented by a graph in which nodes represent elements (blobs, text, and diagrammatic). On the grouping layer, nodes can also stand for groups of said elements, with edges connecting the elements and groups – this can apply recursively. The RST layer, on the other hand, includes nodes for rhetorical relations, with edges connecting it to the nucleus (or nuclei) and possible satellites, also potentially recursively. The connectivity layer contains directed edges to represent directed connectivity: if node B0 is the source of an arrow and B1 is its target, the node is directed from B0 to B1. Edges can also be bidirectional. In the case of lines, which do not imply direction, the edge is undirected.

As its name implies, the grouping layer is intended to group elements of the diagram together. Elements are annotated as a group on this layer if they are likely to be interpreted together. The choices were justified using Gestalt principles of perception (see e.g. Ware, 2012, pp. 181–187); for example, elements which were similar (such as in color or shape) or spatially close to each other were annotated as a group. These can be examples of cohesion in which an image and text are designed to operate together (see Section 2.2): a common occurrence of grouping in AI2D-RST is to group elements hierarchically, such as an illustration with its label in cycles, as seen in Figure 7.

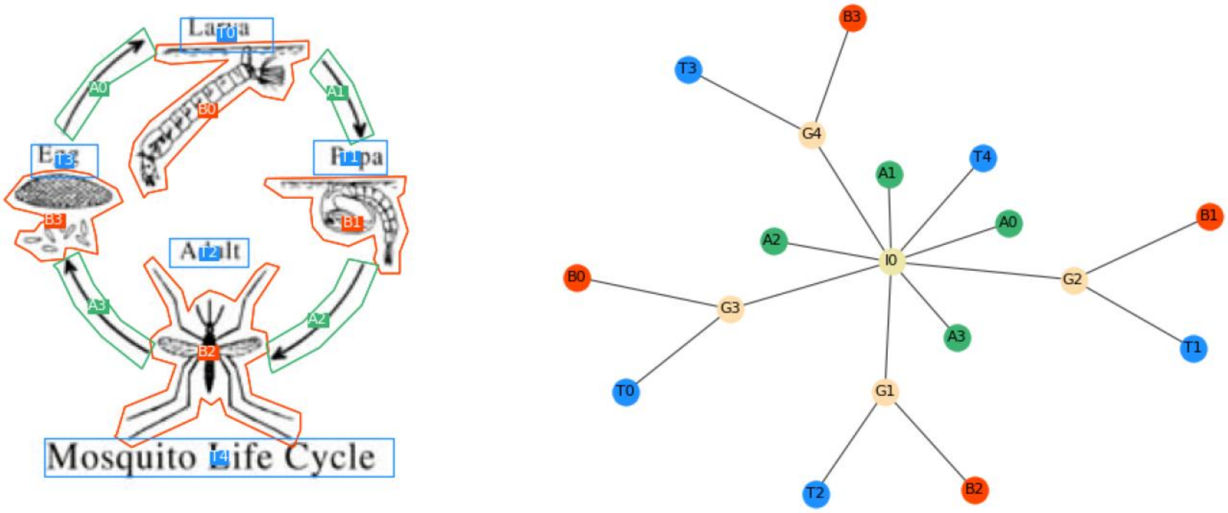

Figure 7. An example of hierarchical grouping in AI2D-RST. Each label has been grouped together with its corresponding illustration, such as T0 (“Larva”) with B0, illustrating the larval stage of a mosquito, to create the group node G3. Each of the 4 groups is then considered a stage of the cycle. Diagram #840.

As per Hiippala et al. (2021), “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*" (p. 668, italics in original) allowing for the annotation and labeling of entire groups of elements.

Macro-grouping refers to a diagram's structure and can describe the type of diagram in question (Hiippala et al., 2021, pp. 668–669), essentially functioning as a genre in how it guides the construction and interpretation of its different semiotic modes in its context (see Section 2.1), as an arrow in a cycle may represent a very different phenomenon from one in a network. Possible macro-groups include networks (such as food chains), cycles (rock cycles and life cycles), tables, and cross-sections (anatomy), among others. It should be noted that a diagram may have multiple groups with different types on the macro-grouping layer. For example, a diagram can consist of a cycle with more information elements in a vertical structure next to it; or a diagram may be a table of cross-sections.

The connectivity layer represents explicit visual connections between elements, realized via diagrammatic elements such as arrows and lines. Hiippala et al. (2021, p. 669) specify that these connections must have clear sources and targets, and that they can be undirected, directed, or bidirectional.

Lastly, the RST layer concerns the *discourse structure* of a diagram and applies Rhetorical Structure Theory to its constituents, which can be any annotated element by itself or a group thereof present on the grouping layer. As Hiippala et al. (2021) state:

> Whereas the grouping and connectivity layers seek to capture the diagram structure that is *explicitly* available for visual inspection, the discourse structure 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. (pp. 669–671, italics in original)

As such, rhetorical relations require inference from viewers' part. These relations are heavily informed by the prior layers, which establish if certain expressive resources belong together, if they follow a hierarchy, and if they are connected visually. This information can indicate what a diagram is attempting to communicate, and as such, applying RST allows users to uncover the abstract rhetorical meaning expressed by the explicit elements and groups in the diagrams. In the RST layer's graph, a rhetorical relation is represented by a node (marked with the appropriate RST relation category), the children of which in turn represent the relation's constituents. It is also vital to note that a relation may function as a nucleus or satellite of another relation: for example, a CYCLIC SEQUENCE depicting an insect's life cycle may be constructed of various IDENTIFICATION relations in which a written label names the stage shown as an illustration.

AI2D and AI2D-RST contain various types of data – element placement, connectivity, visual macrostructure, linguistic content, and rhetorical relations, just to name some – of a large number of diagrams. This information can be extracted, combined, and examined in various ways. The corpora contain different data on the diagrams, but are complementary: the linguistic content of labels and their spatial coordinates, for example, can be retrieved from AI2D, while the rhetorical relations they function in are found in AI2D-RST. Both corpora exist in JSON format, which simplifies their simultaneous processing. The corpora are interrelated, as AI2D-RST is a subset of AI2D. This enables pursuing analysis using both in my thesis, and in fact is necessary for the analysis of co-occurrence between linguistic patterns and rhetorical structure in the diagrams present in them.

4 Methods

Large multimodal corpora are only just emerging. As a result of these newly available and expanding volumes of data, new methods for their analysis are required (Steen et al., 2018; Huang, 2021), which poses a considerable challenge. In the total of 1,000 diagrams in AI2D-RST, there are 8,647 labels that participate in rhetorical relations; as such, manually annotating or processing the linguistic data alone would hardly be practical. To cope with the volume of data and multiple cross-referenced annotation layers, I use computational methods to study the linguistic structure of text elements in the corpus's diagrams. I process the data in Python 3.9 using the spaCy NLP library (Honnibal et al., 2020).[1]

[1] The code used for this thesis can be found on GitHub at https://github.com/Havoq/process_AI2D-RST and under the DOI https://doi.org/10.5281/zenodo.5834586 on Zenodo.

spaCy is an open-source natural language processing library for Python, which is capable of performing a wide range of basic natural language processing tasks, such as tokenizing texts, tagging them for their parts-of-speech, and parsing their syntactic structure. The library applies a statistical language model to the input text and makes predictions pertaining to its linguistic attributes. To examine the linguistic content in the corpus, I extract each label participating in a rhetorical relation and parse it via spaCy.

To prepare the data for analysis, I take the following steps:

1. I iterate over the entire AI2D-RST corpus (the diagrams of which, again, are a subset of AI2D).
2. Each diagram can be represented as a graph, which can in turn be parsed and processed with the NetworkX library (Hagberg et al., 2008) to find edges, nodes, and their various attributes such as neighbors and predecessors. By iteration, the child nodes of macro-groups and rhetorical relations can be found and analyzed.
3. Each rhetorical relation in a diagram is then examined: the RST layer of AI2D-RST enables the extraction of labels that participate in a relation, as the relation nodes are connected to their nuclei and satellites. How labels participate in rhetorical relations is vital for this study, and so I find the labels through relations instead of simply listing them.
4. I fetch the label’s value as a string from the diagram’s corresponding AI2D JSON file using the IDs of the diagram and its annotated elements; this enables the linguistic analysis of the label’s content.
5. I iterate over the graph to find nodes of the “group” type to establish potential macro-structures the labels may be a part of. The function for this is recursive: for each group, if it has a macro-structure field, that field is set as the predominant macro-structure, after which each of its child nodes is further checked for a macro-structure or added to the currently predominant macro-group. If a label does not belong to a smaller macro-structure, it is classified as belonging to the macro-structure associated with the entire diagram. Assigning macro-groups to labels at this point enables finding co-occurrences between different macro-structures and linguistic structures.

These steps are needed to gather and organize the data for analysis. There is also an additional step that is required in order to sufficiently access and recognize the rhetorical function of every label in the corpus, resulting from an artefact of the annotation schema used in AI2D-RST.

The AI2D-RST corpus uses the JOINT relation as a shorthand to group together multiple elements. This relation is not present in the original schema for RST, but is used in AI2D-RST for elements with a shared rhetorical purpose: if a number of elements serve the same purpose and have similarities on the grouping and connectivity layers, they are annotated as the nuclei of a JOINT relation instead of each being an individual relation. Figure 8 shows an example of the relation's use.

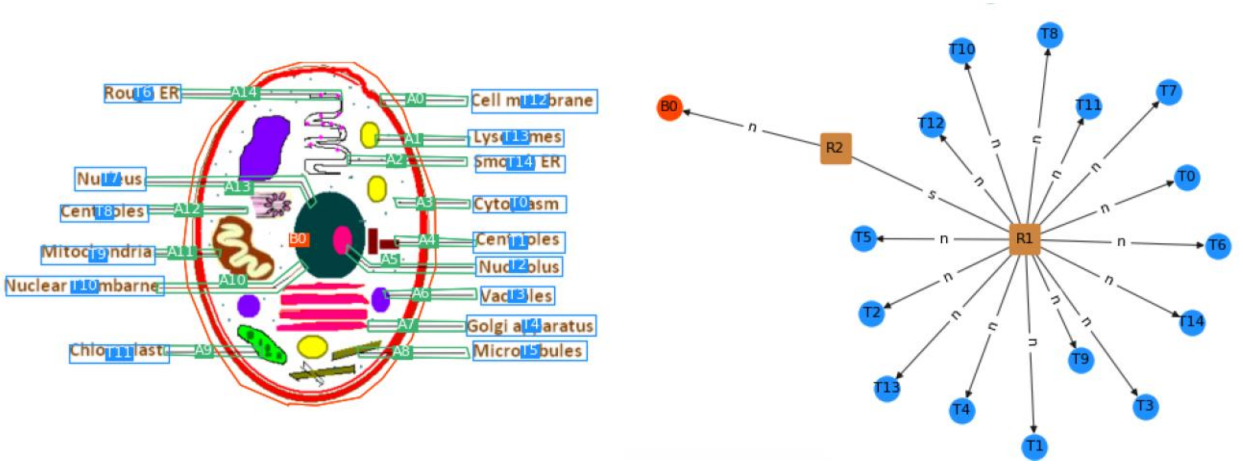

Figure 8. The function of a JOINT relation in the AI2D-RST corpus. The relation node R1 represents a JOINT, which then acts as a satellite to the ELABORATION relation (used to denotate part-whole relations) for the blob B0. Instead of having to annotate each text element as the satellite of an individual ELABORATION relation, JOINT is used as a shorthand to group the text elements together.

Since JOINT does not function as a true rhetorical relation but rather as a schema, each instance of JOINT needs to be parsed accordingly. The script therefore substitutes the JOINT for each of its nuclei, so that the nuclei function in the corresponding manner within the relation whose satellite JOINT acts as. For example, each of the labels in Figure 2 is interpreted as functioning as the satellite of the ELABORATION relation instead.

Resulting from this process, each label then has a set of data attached to it: the ID of the diagram of origin, the ID and type of the relation in which the label appears, the macro-structure in which the label appears, the label's ID, the label's linguistic content, and whether the label functions as a nucleus or satellite in its relation of origin. The labels are then added as rows to a DataFrame – structurally equivalent to a table – from the *pandas* Python library (the pandas development team, 2021) for further processing. If a label participates in more than one relation (for example, if the same text functions in both a PROPERTY-ASCRIPTION and a CONTRAST relation), it is added to the DataFrame the same number of times so that it may be accounted for in each context. Table 1 demonstrates the DataFrame's structure.

index	diagram_id	relation_id	relation_type	macro_group	label_id	content	role	nucleus_type
5423	4759	6XABRE	property-ascription	table	T11	Spiny with sharp stiff points	sat	blobs
2609	1065	9P6MI9	elaboration	cut-out	T4	Vacuole	sat	blobs
3578	490	OPXJUX	connected	network	T11	oceanic fishes	nuc	
8527	2061	445PCG	connected	network	T9	Decomposers	nuc	
5645	3078	WH8R4G	elaboration	cross-section	T1	LEUKAEMIC CELLS	sat	blobs

Table 1. A sample of five randomly selected rows of the DataFrame, exemplifying its structure. The columns include the ID of the diagram of origin, the ID of the relation from AI2D-RST, the relation type and macro-group as annotated in AI2D-RST, the label's ID, linguistic content and rhetorical role – either nucleus (nuc) or satellite (sat). The final column contains the relation's nucleus (text, blobs, arrows, or group) where applicable, provided it is not the label itself.

Each row in the resulting DataFrame is then processed. First, I use spaCy to tag the label for its part-of-speech class (POS), which produces a string of POS tags, representing a pattern. An example of this is NOUN VERB for the label value "rain falls". All proper nouns are converted into nouns for this analysis, as they serve the same purpose structurally, and to streamline the analysis. Punctuation is completely removed from this POS pattern; although it may show further complexity in sentences, simplifying patterns will make the data more uniform for the purpose of this study.

Each label is also parsed for phrase classes, as I compare the percentages of noun and verb phrases in each rhetorical relation and macro-group. spaCy's dependency parsing model, which largely follows the Universal Dependencies formalism (see de Marneffe et al., 2021), allows identifying the head of a phrase, whose POS tag may then be retrieved to determine the phrase type. This then enables the label (or a part thereof) to be classified as a verb or noun phrase, for example. If a label consists of more than one phrase (this occurs, albeit rarely), all phrases are processed and counted separately. Furthermore, the number of unique POS patterns is counted for each RST relation and macro-group to see how much variation there is in the linguistic structure of said relations and structures. In addition to POS patterns, I calculate the average word count for labels in each category. This may provide additional insight into the linguistic complexity of labels in certain categories.

5 Analysis

In this chapter, I present my analysis of AI2D-RST's labels and how their linguistic structure differs by the rhetorical relation and macrostructure they participate in. To scrutinize the results, I construct heatmaps and tables produced by processing the data using the methods described above. I first show the number of labels in each category for an overview of the data; then I present heatmaps of the most common part-of-speech (POS) patterns in each rhetorical relation and macrostructure to display the differences in linguistic structures participating in them; after this, I discuss the number of unique POS patterns in each category to see how much linguistic structures may vary by the contexts in which they are employed. Finally, I compare the number of noun phrases against the number of verb phrases in the corpus to explore how they are used along with different types of diagrams to signal coherence or limiting the amount of necessary world knowledge.

Table 2 shows the total number of labels in AI2D-RST for each rhetorical relation and macro-structure type.[2]

[2] There are very rare instances in the data in which a label may not belong to any group due to only individual blobs having been annotated as belonging to a macro-group (such as illustration) without including its surrounding labels in the same macro-structure; alternatively, individual sections of a diagram may have been annotated as representing separate macro-groups without the diagram as a whole having a macro-group associated with it. Such labels have been left out of the macro-structure categories, but retained in the relations they appear in.

Relation	Labels
elaboration	3633
identification	3065
connected	591
property-ascription	440
preparation	336
class-ascription	208
cyclic sequence	93
circumstance	72
restatement	34
sequence	30
means	21
background	18
nonvolitional-result	18
disjunction	8
nonvolitional-cause	8
contrast	4
enablement	2
list	2
condition	2
conjunction	2
volitional-result	1

Macro-group	Labels
cross-section	1916
illustration	1881
network	1306
cycle	1263
cut-out	768
table	467
horizontal	432
vertical	199
photograph	117
diagrammatic	95
exploded	11

Table 2. The total number of labels under each category.

The tables demonstrate the most common macro-structural and rhetorical categories within which written language is used in the corpora. ELABORATION and IDENTIFICATION far surpass other rhetorical relations in this category, as they are the most prominent discourse relations in AI2D-RST. The next most common relation for written language is CONNECTED, which denotes networks in the corpora. Networks often contain textual elements (or units of textual and pictorial elements) interconnected by diagrammatic elements (see Figure 4); these elements have been annotated as CONNECTED with all the network's lower-level units as the nuclei.

5.1 Most common part-of-speech patterns

To find answers to my research questions of whether different types of diagrams contain different linguistic structures in their labels and if certain patterns occur in specific rhetorical relations, I produce heatmaps via the *seaborn* library (Waskom, 2021; see also Hunter, 2007), which I examine for occurrences of each of the five most common POS patterns for each rhetorical relation (Figure 9) and macro-group (Figure 10). Visualization provides a practical overview of large volumes of data, and heatmaps afford high readability. Normalizing the counts is necessary to make the observations comparable across the data; notably, these heatmaps have been normalized across rows and not columns. This means that the brightest point in each row is in relation to that category's occurrences, and the highest ratio on one row may not be numerically equivalent to that on another. The unmodified tables for the occurrences can be found in Appendices A (for rhetorical relations) and B (for macro-structures).

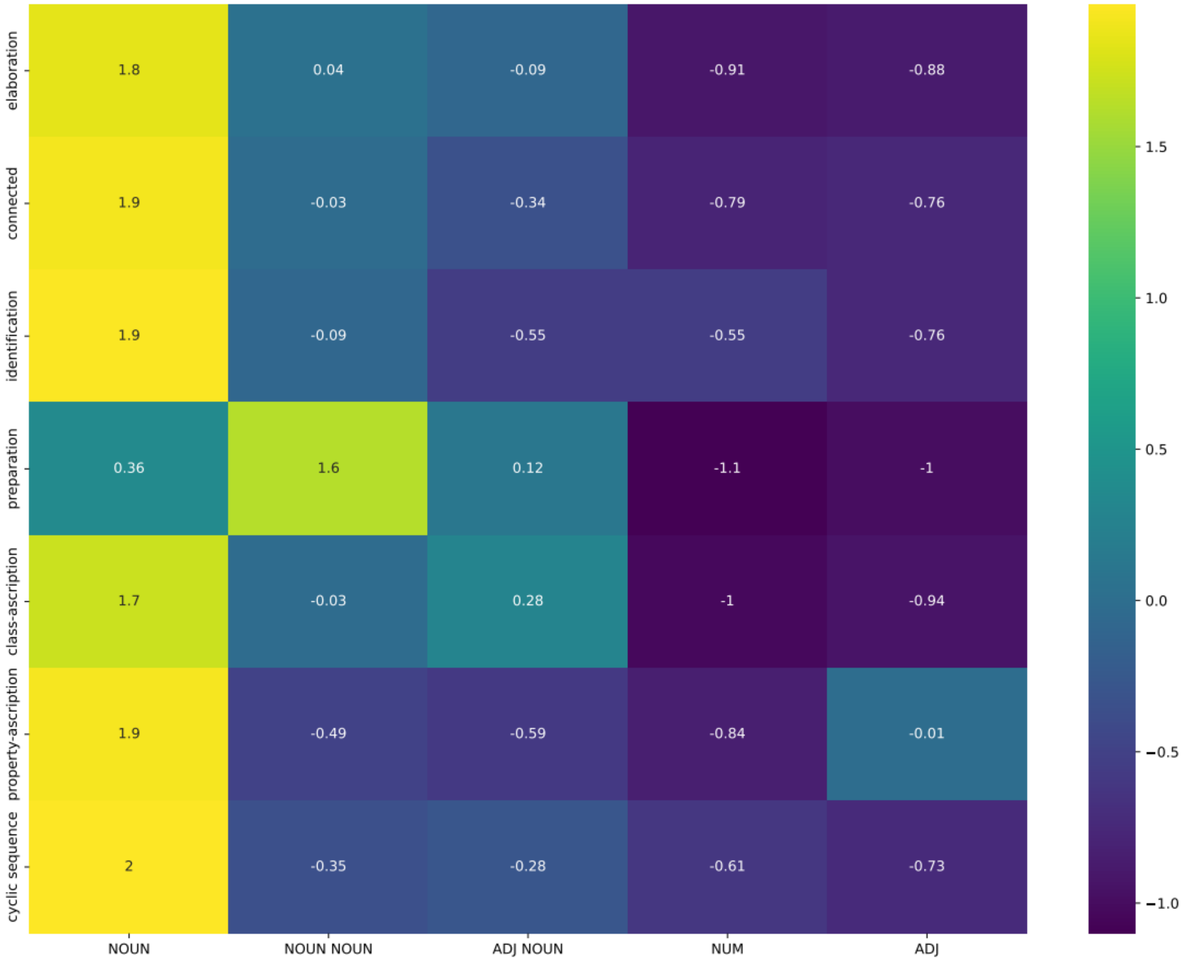

Figure 9. The spread of the five most common POS patterns per rhetorical relation. The heatmap has been normalized across rows.

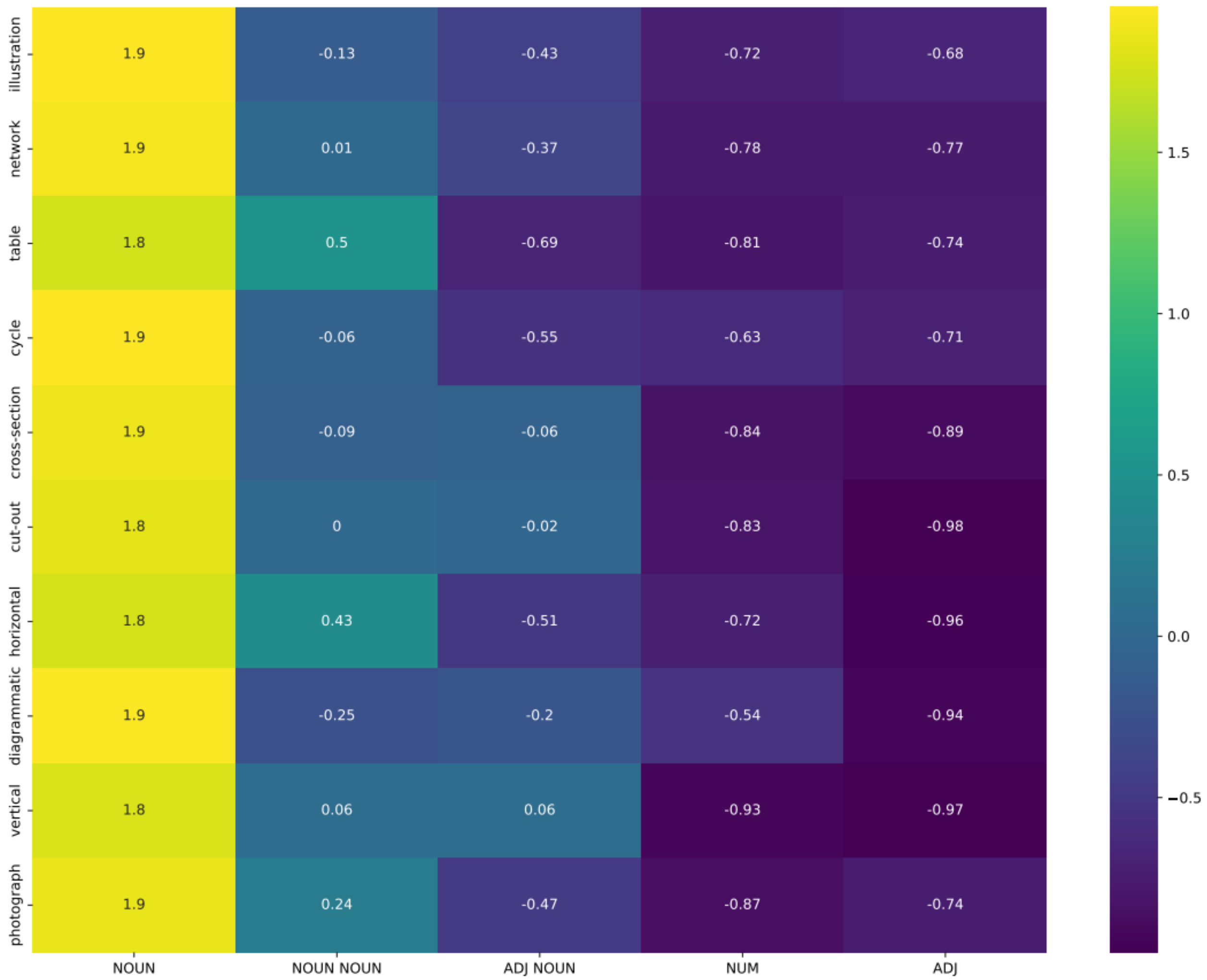

Figure 10. The spread of the five most common POS patterns per macro-group. The heatmap has been normalized per row.

As Figures 9 and 10 demonstrate, "NOUN" by itself is by far the most common pattern in each category except for the PREPARATION, RESTATEMENT and NON-VOLITIONAL CAUSE relations. For the most common relations, this is hardly surprising: both ELABORATION and IDENTIFICATION label objects or their parts, so a single noun is often sufficient for the task. The "NUM" pattern is the fourth most common pattern due to many diagrams simply referring to visual elements by a single number: for example, a row of illustrated bird breeds might have a single digit under each specimen to identify it instead of using its name. The numbers then act as pointers to their respective breeds of bird in some other body of text in the original context of appearance (see Hiippala, 2012, pp. 118–119). The three most common patterns seem to dominate the dataset, while the rest have quite few instances. A single verb, while the tenth most common pattern, is very rare, and mostly appears in the context of PROPERTY-ASCRIPTION most likely due to gerunds. Similarly, the three most common patterns appear consistently by macro-group.

As “NOUN” is the most common pattern and appears in various relations and macro-groups, it can be inferred that a single noun can be used in a variety of ways. The “NOUN NOUN” pattern is the second most prominent pattern across the dataset, consisting of nominal groups (see e.g. Halliday and Martin, 1993, pp. 59–75). As the final of the three most common patterns, “ADJ NOUN” is also a somewhat common pattern, although not as much as the prior two. This pattern can be deployed to assist in classifications (see e.g. Halliday and Martin, 1993, pp. 152–287). See Section 6 for examples of these patterns’ uses.

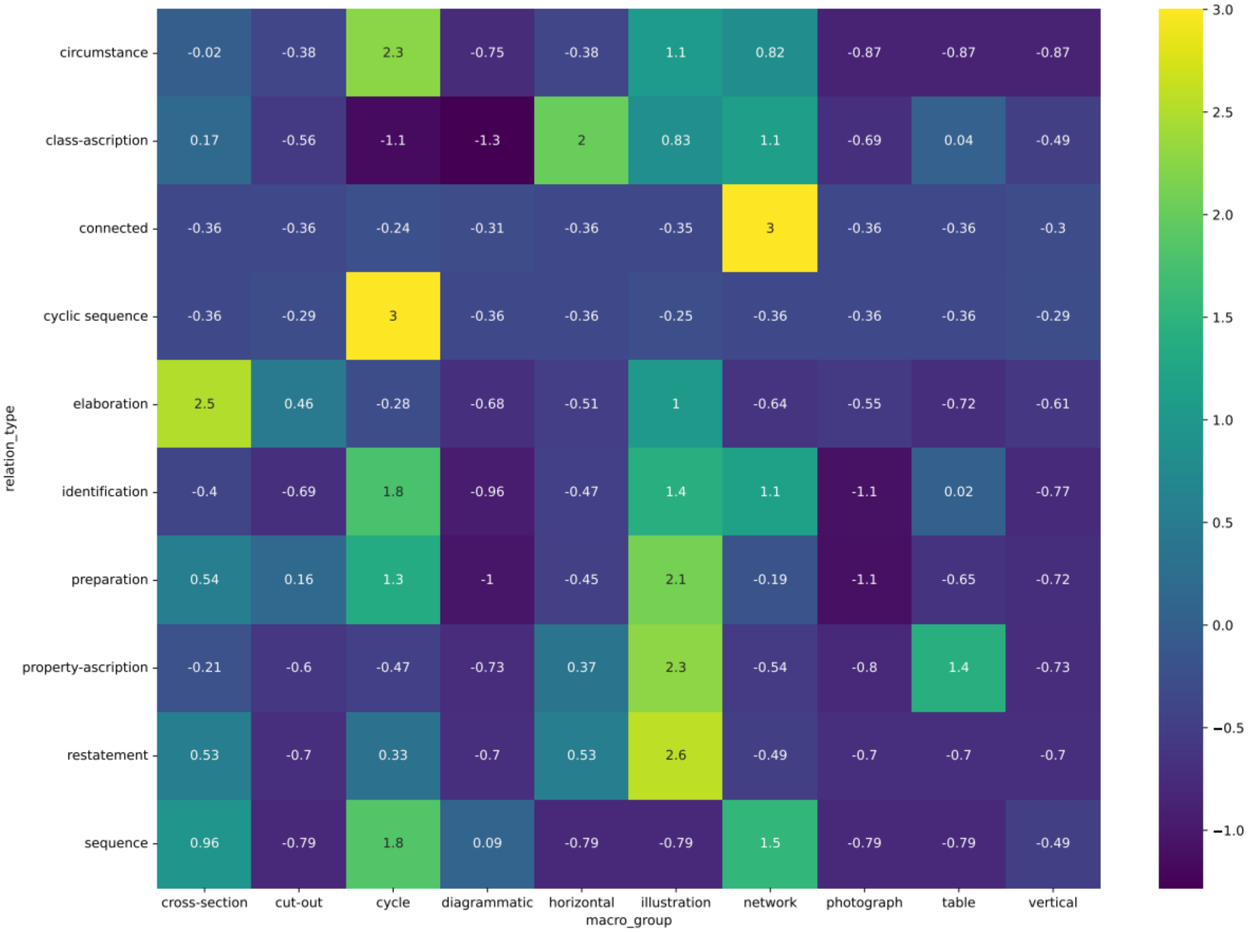

Figure 11. The cross-tabulated heatmap of the relationships between different rhetorical relations and macrostructures.

To contextualize the findings in this chapter, Figure 11 displays the cross-tabulated overlap between different macro-groups and rhetorical relations. As can be seen, certain combinations are prevalent in the data; quite predictably, the network macro-group and CONNECTED relation have the highest possible overlap, as networks in AI2D-RST are annotated with said relation (see Figure 4). Similarly, the cycle macrostructure co-occurs with the CYCLIC SEQUENCE relation. The relationship between ELABORATION and the cross-section macro-group is exemplified by Figure 8, in which a cross-section provides a useful look at the parts pointed out by labels.

5.2 Unique POS patterns across diagram types

Table 3 shows the number of unique POS patterns found in each rhetorical relation and macro-group. This was achieved by appending each pattern (after removing punctuation and changing proper nouns to general nouns) to a list by category and ultimately converting it into a set, only maintaining unique values. The tables are organized by the number of unique patterns.

Relation	Patterns
elaboration	328
identification	167
preparation	126
property-ascription	62
connected	40
circumstance	32
class-ascription	23
cyclic sequence	18
background	15
sequence	13
restatement	12
means	11
nonvolitional-result	6
nonvolitional-cause	5
contrast	3
enablement	2
disjunction	3
list	2
condition	2
volitional-result	1
conjunction	1

Macro-group	Patterns
illustration	201
cycle	191
cross-section	127
cut-out	91
network	97
horizontal	69
table	54
vertical	46
photograph	17
diagrammatic	12
exploded	3

Table 3. Tables of the number of unique patterns found in each RST relation and macro-group.

As Table 3 shows, the ELABORATION relation possesses the largest number of unique POS patterns, almost doubling the number of the second-highest number of IDENTIFICATION. Given that ELABORATION and IDENTIFICATION are somewhat close in terms of total label count, this indicates that ELABORATION uses a wider range of linguistic structures compared to IDENTIFICATION in the AI2D-RST corpus. PREPARATION, which can be used to explain the contents of a diagram to prepare the reader to interpret it, also has a high number of unique POS patterns, as these explanations can take on various linguistic structures.

While the CYCLIC SEQUENCE relation only displays 18 different POS patterns, the cycle macro-group contains 191 unique patterns and is the second most varied of the macro-structures. For a label to appear in a CYCLIC SEQUENCE relation, the sequence must contain the label by itself and not as a member of a lower-level relation, such as IDENTIFICATION which the cycle then consists of. As such, a label may well appear in a cycle macro-structure without appearing directly in a CYCLIC SEQUENCE. Figure 12 illustrates how a CYCLIC SEQUENCE relation may be composed of rhetorical relations, as labels serve to identify the illustrations, and this IDENTIFICATION relation is then annotated as part of the cycle.

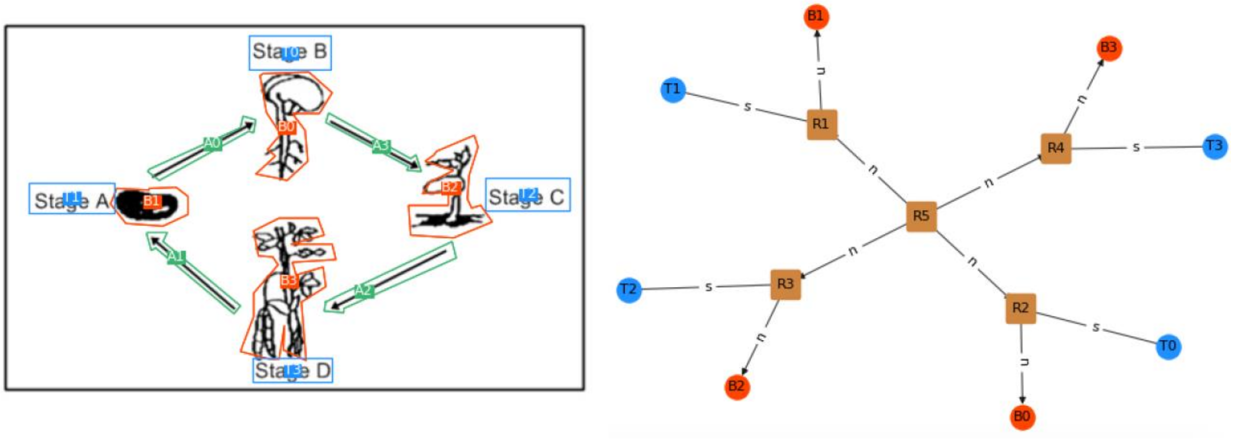

Figure 12. An example of a CYCLIC SEQUENCE (R5) composed of smaller-scale rhetorical relations – in this case, instances of IDENTIFICATION (R1-4) function as the nuclei. Diagram #1287 from AI2D-RST.

In order for a label to appear directly as a part of a CYCLIC SEQUENCE, said label must be part of the cycle by itself, as seen in Figure 13. The number of unique labels in cycles indicates that diverse language may be used in explaining cycles, whether to prepare viewers or further explain individual stages of the cycle. This may be an example of multiple signals being used in unison to enhance the diagram’s coherence and guide its interpretation.

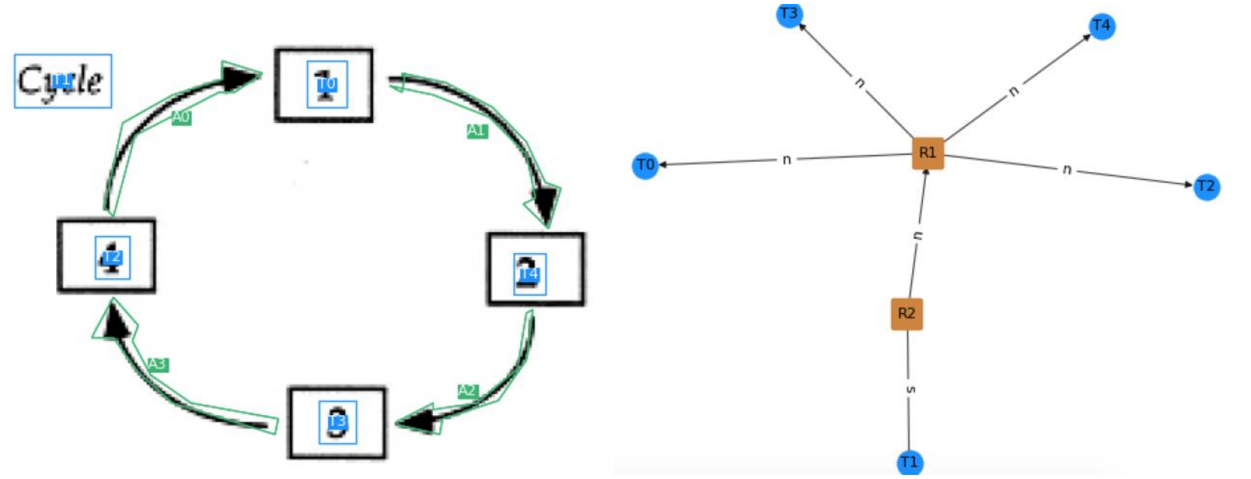

Figure 13. Labels forming a CYCLIC SEQUENCE relation (R1). Relation R2 represents PREPARATION, as the title (“Cycle”) prepares viewers to interpret the diagram. Diagram #502 from AI2D-RST.

Given that illustrations are the second-most occurring macro-structure in the dataset, it is hardly a surprise to see that said structure has the most variation when it comes to linguistic patterns. Illustrations can be used in school textbooks in numerous ways. A table containing the occurrences of labels belonging to each possible combination of relation and macro-structure (see Appendix C) shows the different capacities in which illustrations use written language.

5.3 Average word counts

To aid in the assessment of linguistic complexity across rhetorical relations and macro-groups, I calculate the average word counts in each category. The mean word count by rhetorical relation and macro-group can be seen in Table 4. The rows have been organized according to word count.

Relation	Word Count
enablement	18
background	5
circumstance	4.28
preparation	4.04
means	3.48
restatement	2.62
nonvolitional-cause	2.62
sequence	2.33
elaboration	2.18
nonvolitional-result	2.17
cyclic sequence	1.84
property-ascription	1.77
identification	1.7
class-ascription	1.66
connected	1.59
list	1.5
condition	1.5
contrast	1.5
disjunction	1.25
volitional-result	1
conjunction	1

Macro-group	Word count
cycle	2.56
vertical	2.53
cut-out	2.15
horizontal	2.03
illustration	1.99
table	1.93
cross-section	1.84
network	1.77
photograph	1.66
diagrammatic	1.6
exploded	1.36

Table 4. Tables of the average word count in each RST relation and macro-group.

The most exceptional value in Table 4 is the average word count of 18 in the ENABLEMENT relation. There is, however, a simple explanation to this: the relation is only represented by two labels in the entire AI2D-RST corpus, one of which has a length of 1 word, while the other is 35 words long. This makes the average word count for the relation entirely disproportionate. The other relations meanwhile show much more even numbers. BACKGROUND, CIRCUMSTANCE, PREPARATION and MEANS are on average much longer likely because these attributes are more difficult to visualize and explain, and therefore require longer text to fully relay to receivers.

The ELABORATION relation has an average word count of 2.18, while IDENTIFICATION's is 1.7. The difference is somewhat smaller than expected: prior to processing the JOINT "relation", ELABORATION had almost double the word count of IDENTIFICATION. This shows that examples such as Figure 8, in which shorter labels function to point out parts of an object, have a large impact on the overall word count when counter as parts of an ELABORATION relation rather than a JOINT, even if the non-JOINT labels may be noticeably longer on average.

It is worth mentioning that out of all macro-groups, labels within cycles have the largest average word count. As stated in Section 5.2, a label can appear in a cyclic macro-structure without appearing directly in a CYCLIC SEQUENCE relation; the figure then contains all labels, regardless of relation, that appear in such diagrams. This shows that written language may be more detailed or descriptive on average, serving as ELABORATION or PREPARATION, for example. This spread of relations in cycles can also be seen in Appendix C.

5.4 Phrase classes

Variety in linguistic structures may also be found by examining and comparing phrase classes: notable co-occurrence of these label categories and certain linguistic features could give an answer to my research question of whether different relations and macro-groups consistently deploy different linguistic structures to signal coherence. Table 5 demonstrates the occurrences of verb phrases (VP) and noun phrases (NP) by RST relation and macro-group. As mentioned in Section 4, if a label contains multiple complete phrases, each is accounted for separately.

relation	VP	NP	total
elaboration	155	3364	3665
connected	8	569	591
identification	104	2649	3084
preparation	26	313	351
restatement	3	28	31
class-ascription	2	195	208
background	6	10	18
property-ascription	54	299	440
nonvolitional-result	1	14	18
circumstance	16	52	72
cyclic sequence	1	84	93
volitional-result	0	1	1
means	1	23	24
sequence	2	26	30
enablement	2	1	3
disjunction	0	7	8
list	1	1	2
condition	0	2	2
contrast	0	4	4
nonvolitional-cause	0	8	8
conjunction	0	2	2

macro-group	VP	NP	total
illustration	88	1663	1919
network	36	1250	1307
table	33	371	469
cycle	124	1041	1280
cross-section	33	1787	1920
cut-out	22	688	767
horizontal	29	352	433
diagrammatic	1	85	96
vertical	12	173	203
photograph	1	110	117
exploded	0	11	11

Table 5. Tables of the occurrences of verb phrases (VP) and noun phrases (NP) for each RST relation and macro-group, as well as their total label counts.

As anticipated due to ELABORATION serving both purposes of pointing out parts of a whole and elaborating on the nucleus, the relation contains more verb phrases in relation to its volume than the IDENTIFICATION relation. This difference is slightly lower than expected, as the average word count and number of unique POS patterns are both higher in instances of ELABORATION. 90 percent of all phrases in said relation are noun phrases, while spaCy classifies only 85 percent of IDENTIFICATION phrases as NPs. This is partially explained by the POS pattern of a single numeric appearing in many IDENTIFICATION labels to simply assign a number to a visual element. The lowest relations in the table have so few labels that it is impossible to make conclusions on their uses without a larger corpus.

The number of verb phrases is notable, however, when looking at the cycle macro-structure, in which over 9 percent of all phrases – 123 out of a total of 1280 – are classified as VPs. Since not all labels in cycles are necessarily part of the CYCLIC SEQUENCE relation itself, this table shows the frequency of VPs that appear in them regardless of their rhetorical function such as ELABORATION or IDENTIFICATION. Based on this data, diagrams depicting cycles use generally the most verb phrases out of all macro-structure types.

As arrows often serve similar functions as verbs in diagrams (Alikhani and Stone, 2018), they may be replaced by written language – specifically, verb phrases – to fill the same purpose while also requiring less world knowledge from viewers, as in diagram #4210 (see Figure 5). Verb phrases may also appear alongside arrows to explicate their meaning, hence strengthening cohesion (see Section 2.2); moreover, a verb phrase accompanying an arrow is an instance of multiple signals of the coherence relation in question (see Section 2.4). Figure 14 shows an example of verb phrases occurring alongside arrows.

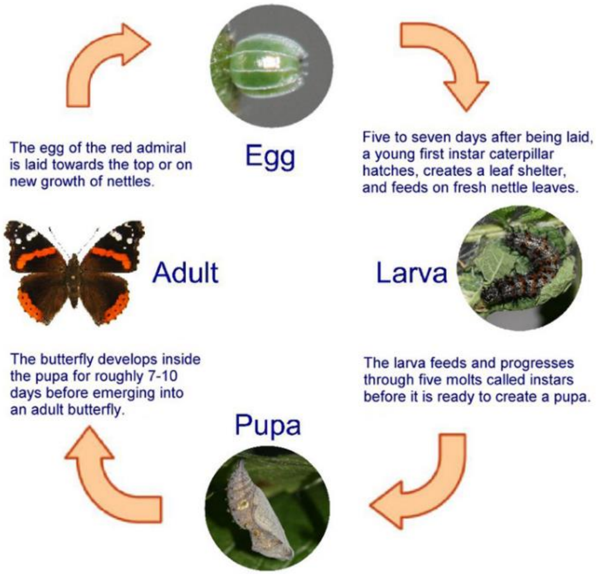
Egg
The egg of the red admiral is laid towards the top or on new growth of nettles.
Five to seven days after being laid, a young first instar caterpillar hatches, creates a leaf shelter, and feeds on fresh nettle leaves.
Adult
Larva
The butterfly develops inside the pupa for roughly 7-10 days before emerging into an adult butterfly.
The larva feeds and progresses through five molts called instars before it is ready to create a pupa.
Pupa

Figure 14. An example of written language accompanying arrows to explicate the process indicated by the arrows. Many similar diagrams of life cycles in the dataset lack written language outside of the labels identifying the illustrations. Diagram #2242.

6 Discussion

In this section, I discuss the results shown in Section 5 and what they might indicate in terms of future computational approaches to diagrams as communicative, multimodal artefacts. Going over each of the subsections, I summarize the findings, explain their relevance for my topic, and give some examples of the analyzed phenomena.

Approaching diagrams as a semiotic mode highlights the importance of their dynamic interpretation, where every element and their combinations contextualize each other on the level of discourse semantics. As coherence relations can be signaled via multiple elements in diagrams, it is appropriate to consider written labels vital for this co-signaling and contextualization where they appear. Consequently, the analysis presented in Section 5 suggests that linguistic structures present in written labels can indeed inform the decomposition of visual expressive resources in diagrams, as indicated by the sub-questions posited in Section 1.2.

Even though certain patterns are very common across the corpus, different macro-structures can deploy quite distinct linguistic patterns. For example, a large portion of labels appearing in cycles contain verb phrases, which can explicitly inform receivers of a given cycle's phases, possibly filling gaps in their world knowledge where needed and co-signaling processes along with diagrammatic elements. Cycles also contain the highest average word count among macro-structures. These findings indicate that cycles, such as the one in Figure 5, may use more descriptive language to convey the possibly complex processes and phenomena within. Similarly, the rhetorical relation of ELABORATION uses verb phrases and diverse linguistic structures to communicate discursive intent to viewers. The findings support my hypothesis in this respect; different macro-structures and rhetorical relations display some co-occurrence with specific linguistic patterns.

Verbs can be used to describe processes, which is quite common in school material (Halliday and Martin, 1996). This, along with how arrows can function in their place or co-occur with them to enhance coherence signaling, can explain their relatively high occurrence percentage in cycles. Even when a cycle does not consist of cyclical diagrammatic elements, it can be signaled by the combination of arrows or lines and written language.

Nouns are very prevalent throughout the entire dataset, unsurprisingly – nominalization and nominal groups have an extensive history in scientific and educational material (see e.g. Halliday and Martin, 1996; Doran, 2017; Martin et al., 2021). Halliday and Martin (1996, p. 161-162) state that “in order to classify and organize with language, we need first of all to turn phenomena into things or nouns” and describe the nominalization of processes as “turning happenings into things which can be technicalized. … Thus, some technical terms are single nominals or things but realize a nominalization, for example, condensation, transpiration.” Nominalization and nominal groups allow for taxonomies, defined by Halliday and Martin (1996, p. 153) as “ordered, systematic classification[s] of some phenomena based on the fundamental principles of superordination (where something is a type of or kind of something else) or composition (where something is a part of something else).” A fine example of taxonomies being used in this way is shown in Figure 6, in which certain parts of a plant constitute their own further segments such as the stamen. Nominalization can also be seen functioning in Figure 14 via grammar such as “respiration”, presented as a phenomenon affecting said cycle, which itself is clearly indicated by a circular diagrammatic element.

Seeing as “NOUN” is the most prominent POS pattern in the data, it must be of use in various contexts. Figure 15 demonstrates how an individual noun can be used to either label an entire object or single out a part thereof. The distinction between these two purposes is made clear by the layout and diagrammatic elements: nouns labeling parts of an object are connected to it via lines, while the noun labeling the entire object is placed in close proximity to the illustration.

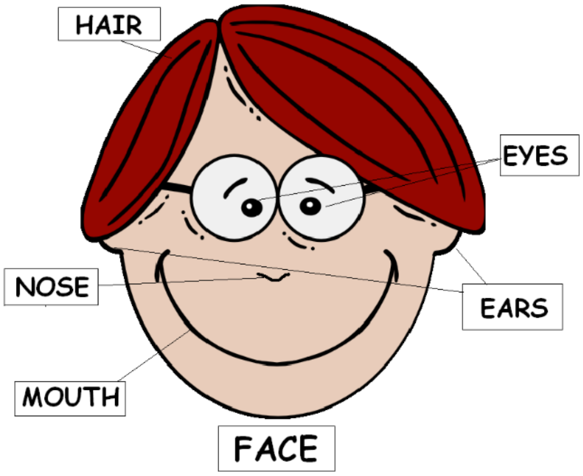

Figure 15. Diagram #0 from AI2D. The diagram illustrates how a single noun can serve both the purpose of ELABORATION through pointing out parts of a whole in unison with diagrammatic elements (lines) and IDENTIFICATION by labeling an entire object (“Face”) via its spatial placement.

The second-most commonly occurring pattern in the corpus is “NOUN NOUN” -- this is due to nominal groups, such as the labels “Plant Respiration” or “Factory Emissions” in Figure 16. Such classifiers in nominal groups are common in scientific and educational material, as discussed by Halliday and Martin (1993, pp. 59–75).

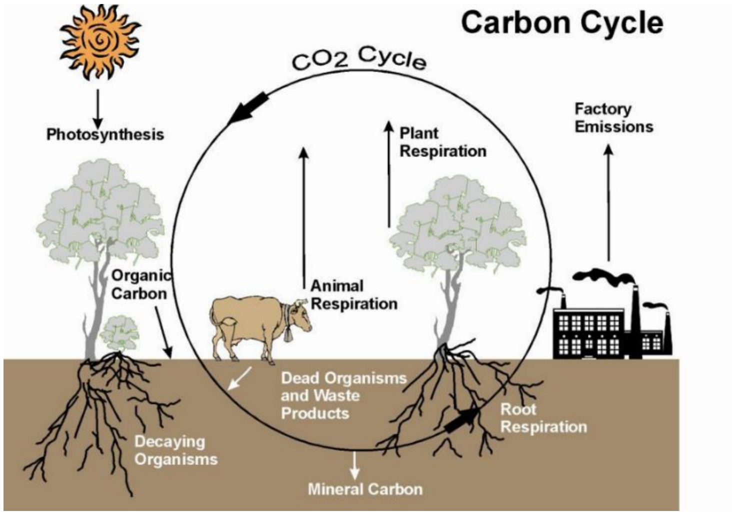

Figure 16. An example of the "NOUN NOUN" pattern in the dataset, with numerous examples such as "Carbon Cycle", "Factory Emissions", "Root Respiration", "Animal Respiration", and "Plant Respiration". Diagram #77.

Finally, "ADJ NOUN" is the third-most recurring pattern. As Figure 17 illustrates, the pattern can be used to categorize and distinguish between different types of phenomena, concepts, or objects, which helps create technical classifications (Halliday and Martin, 1993, pp. 153–250).

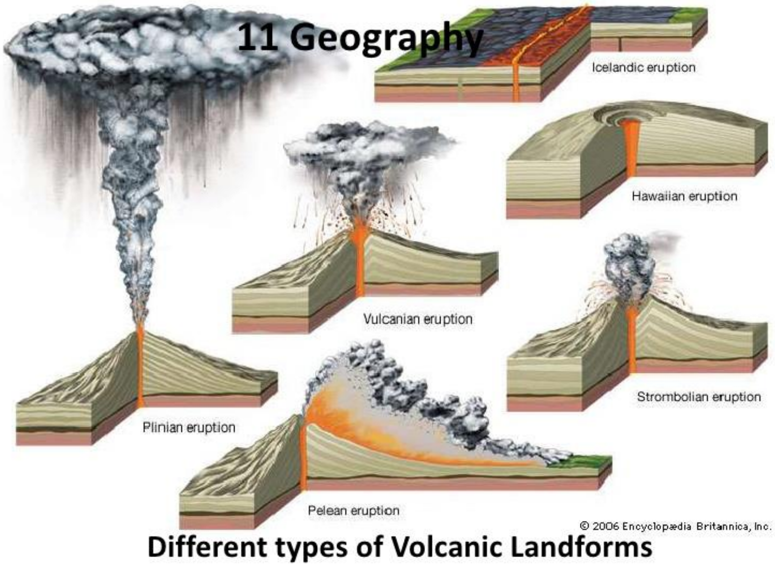

Figure 17. An example of the "ADJ NOUN" pattern in the dataset. Diagram #1456.

These examples demonstrate how each of these common patterns has distinct, yet quite numerous uses in educational material. It bears repeating, however, that these science diagrams have been designed for use in primary schools. It may be of service to consider where nominalization is advantageous; simple cases of IDENTIFICATION (such as in Figure 15), for example, may gain nothing from further guiding viewers' interpretations – but diagrams depicting more complex concepts or processes may benefit from more explicit language such as verb phrases or detailed ELABORATION to help guide younger recipients perhaps not yet familiar with more technical material or with more gaps in their world knowledge.

As Watanabe and Nagao (1998) posit, both written language and layout elements are essential in the interpretation of diagrams. The various ways a single noun can be used can be distinguished via label placement and possible accompanying diagrammatic elements. While other patterns vary by diagram and relation type, the most common patterns seem consistent across categories in AI2D-RST. This relates to my third sub-question in section 1.2: the exact function of some patterns can be co-signaled and made more explicit by other elements present in the diagram. So, the dynamic interpretation of diagrams becomes vital again: it is not only the visual bottom-up makeup of a diagram that constitutes its meaning, but the discourse semantics and co-contextualization provided by labels' linguistic content and structure in addition to their placement near other elements.

As there are co-occurrences between different rhetorical relations, diagram macro-structures, and labels' linguistic structures, it can be inferred that the decomposition of visual expressive resources can be informed to some degree by linguistic structures present in a diagram's written labels. Because of the various forms of written language in AI2D-RST and the diverse ways in which different diagrams use labels, it may be desirable for diagram annotation schemata to emphasize the linguistic content as a possible guide for annotators. If some of the signals used for effectively representing discursive intent are disregarded, the interpretation may not be as cohesive as necessitated by the various fields in which diagrams research is relevant. Thus, simply identifying structural visual units without dynamically considering the entire diagram may not fully take advantage of the discursive flexibility afforded by diagrams.

Using existing NLP libraries to process this data is not without its pitfalls. It must be noted that even though spaCy's current transformer-based model at the time of writing is an improvement over its predecessor, it does make mistakes. Examples from the dataset include the word "rounded" and the phrase "Metamorphic rocks" being incorrectly classified as verb phrases. As a result, there may be certain inaccuracies in the process, especially due to cycles in AI2D-RST including various examples of rock cycles.

Larger corpora would also yield more definitive results on the matter. Even with widely available multimodal corpora such as AI2D-RST, some rhetorical relations are very infrequent. Having access to more comprehensive datasets would help form a clearer picture of linguistic patterns and combinations of signals specific to or more common in certain relations. If further multimodal corpora were to be developed while also implementing some of the suggestions in this thesis, future research of a similar nature would be even easier to conduct.

Further research could successfully use the format set by Watanabe and Nagao (1998), as the two corpora contain the necessary information for accurately identifying labels' location in relation to corresponding visual expressive resources, as well as determining connectivity between said elements. Using layout information accessible via AI2D's polygons and AI2D-RST's connectivity layer, the visual relations of labels and images can be scrutinized in more detail.

It may be beneficial to also seek patterns that occur exclusively in certain rhetorical relations or macro-structures to see if these occur in diagrams consistently for their respective purposes in future datasets. Additionally, spaCy's phrase matcher class can be used to find more specific linguistic patterns, for instance by combining POS information with predetermined words, such as "NOUN + has + NOUN". Combined with the available layout information, a similar ruleset to what Watanabe and Nagao (1998) observe may theoretically be assembled for AI2D-RST – larger corpora, finely targeted linguistic analysis, and detailed layout information may together provide further information as to how specific linguistic patterns function along with placement to create a complete discursive whole.

The interest in diagrams' computational processing has remained somewhat scarce since the work done by Watanabe and Nagao (1998), and so there has been no further pursuit of the ideas they propose. Recent studies in computational analysis of diagrams have yielded mixed results (see e.g. Haehn, Tompkin, and Pfister, 2019). Sachan et al. (2020) demonstrate that NLP approaches can be successfully combined with diagram parsing to extract data from schoolbooks, although the study is only concerned with mathematics. Kim et al. (2019) discuss the recent increase in computational visual question-answering, again in the context of textbooks. Seo et al. (2015), inspired by the growing trend of combining text and vision in NLP, computationally solve SAT geometry problems with modest but promising results. These studies indicate that further research into the linguistic structure of written language in diagrams may prove useful, but do not consider the potential value of a deeper linguistic understanding of written language in diagrams. Some of the methods in this thesis could then be beneficially applied to computational approaches to diagrams; further focus on linguistic structures at the levels of both syntax and discourse may lead to improvements in task-driven fields such as NLP and yield higher accuracy in predictive models.

The implications of this study for empirical multimodality research are equally useful. It functions as an example of the usefulness of new, computationally accessible multimodal corpora as well as the synthesis of various theories from multimodality research, diagrams research, and discourse studies. The data can be approached via the diagrammatic mode to analyze the discourse-semantic aspects of included diagrams effectively to more precisely understand how they convey the complex concepts behind them.

7 Conclusion

This study synthesized theories from multimodality, diagrams research, and discourse studies to examine how written language functions with visual expressive resources to communicate discursive intent. I used natural language processing to computationally analyze two diagram corpora, AI2D and AI2D-RST, for their linguistic features.

The study found that different discourse relations and types of diagrams (macro-groups) use different linguistics structures in their labels to aid receivers in interpreting them without needing to possess extensive world knowledge on the matter. Based on the findings, written language is a vital part of guiding receivers’ interpretation of the diagrams present in AI2D-RST. To maximize coherence and strive for successful receiver interpretation, the diagrams examined use multiple signals in unison, complementing visual expressive resources and diagrammatic elements with written language.

Going forward, the linguistic structure of labels could be used along with layout information to enhance annotation schemata for multimodal corpora and the processing thereof, especially in the context of diagrams research. Empirical multimodality research can benefit notably from employing discourse semantics: because diagrams are gaining relevance in various fields, it may prove invaluable for future studies to emphasize the discursive, multimodal aspects found therein and design annotation schemata and processes accordingly.

Bibliography

Alikhani, M., Hiippala, T. and Stone, M., 2019. 'A coherence approach to data-driven inference in visual communication'. In: Language and Vision Workshop at 2019 Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, United States, 16/06/2019.

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. 3552–3563.

André, E. and Rist, T., 1995. Generating coherent presentations employing textual and visual material. *Artificial Intelligence Review* 9. pp. 147–165. DOI: https://doi.org/10.1007/BF00849177

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., Smith, B.A. (eds) *Multimodal Studies: Exploring Issues and Domains*. London/New York: Routledge. pp. 17–38.

Bateman J.A., 2014a. *Text and Image. A Critical Introduction to the Visual-Verbal Divide*. London: Routledge.

Bateman, J.A., 2014b. 'Multimodal coherence research and its applications'. In: Gruber, H. and Redeker, G., (eds.) *The pragmatics of discourse coherence: Theory and applications.* Amsterdam/Philadelphia: John Benjamins. pp. 145–177.

Bateman, J.A., 2021. 'Dimensions of Materiality'. In: Pflaeging, J., Wildfeuer, J. and Bateman, J.A., (eds.) *Empirical Multimodality Research: Methods, Evaluations, Implications.* Berlin/Boston: De Gruyter. pp. 35–64. DOI: 10.1515/9783110725001-002.

Bateman, J.A. and Delin, J., 2006. Rhetorical Structure Theory. In: Brown, K. (ed.) *Encyclopedia of Language & Linguistics*. Elsevier. pp. 589–597.

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

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

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuglu K. and Kuksa, P., 2011. Natural Language Processing (Almost) from Scratch. *Journal of Machine Learning Research* 12. pp. 2493–2537.

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

Das, D. and Taboada, M., 2019. Multiple Signals of Coherence Relations. *Discours* 24. DOI: https://doi.org/10.4000/discours.10032

De Marneffe, M.-C., Manning, C.D., Nivre, J. and Zeman, D., 2021. Universal Dependencies. *Computational Linguistics* 47(2). pp. 255–308. DOI: https://doi.org/10.1162/coli_a_00402

Doran, Y.J., 2017. *The Discourse of Physics: Building Knowledge Through Language, Mathematics and Image.* 1st ed. New York: Routledge. DOI: https://doi.org/10.4324/9781315181134

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

Engelhardt, Y., 2007. Syntactic structures in graphics. *Computational Visualistics and Picture Morphology* 5. pp. 23–35.

Greenberg, G., 2021. Semantics of Pictorial Space*. Review of Philosophy and Psychology* 12*.* pp. 847–887. DOI: https://doi.org/10.1007/s13164-020-00513-6

Gruber, H. and Redeker, G., (eds.) 2014. *The pragmatics of discourse coherence: Theory and applications.* Amsterdam/Philadelphia: John Benjamins.

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

Hagberg, A.A., Schult, D.A., and Swart, P.J., 2008. ‘Exploring network structure, dynamics, and function using NetworkX’. In: Varoquaux, G., Vaught, T. and Millman, J. (eds). Proceedings of the 7th Python in Science Conference (SciPy2008). Pasadena. pp. 11–15.

Halliday, M.A.K. and Hasan, R., 1976. *Cohesion in English*. London: Longman.

Halliday, M.A.K. and Martin, J.R., 1993. *Writing Science: Literacy and Discursive Power*. London: Taylor & Francis Group.

Hiippala, T., 2012. The Localisation of Advertising Print Media as a Multimodal Process. In: Bowcher, W.L. (ed.) *Multimodal texts from around the world: Cultural and linguistic insights.* New York: Palgrave Macmillan. pp. 97–122.

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

Hiippala, T., Alikhani, M., Haverinen, J., Kalliokoski, T., Logacheva, E., Orekhova, S., Tuomainen, A., Stone, M. and Bateman, J.A., 2021. ’AI2D-RST: A multimodal corpus of 1000 primary school science diagrams’. *Language Resources and Evaluation* 55. pp. 661–688. DOI: https://doi.org.10.1007/s10579-020-09517-1

Hiippala, T. and Bateman, J.A., 2020. *Introducing the diagrammatic mode*. [online] Available at: https://arxiv.org/abs/2001.11224

Hiippala, T. and Bateman, J.A., 2021. Semiotically-grounded distant viewing of diagrams: insights from two multimodal corpora. *Digital Scholarship in the Humanities*. DOI: 10.1093/llc/fqab063

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). Paris: European Language Resources Association (ELRA). pp. 1925–1931.

Holsanova, J., Holmberg, N. and Holmqvist, K., 2009. Reading information graphics: The role of spatial contiguity and dual attentional guidance. *Applied Cognitive Psychology* 23. pp. 1215–1226. DOI: https://doi.org/10.1002/acp.1525

Honnibal, M., Montani, I., Van Landeghem, S. and Boyd, A., 2021. *spaCy: Industrial-strength Natural Language Processing in Python*. Zenodo. DOI: https://doi.org/10.5281/zenodo.4715444

Huang, L., 2021. Toward multimodal corpus pragmatics: Rationale, case, and agenda. *Digital Scholarship in the Humanities* 36(1). pp. 101–114. DOI: https://doi.org/10.1093/llc/fqz080

Hunter, J.D., 2007. "Matplotlib: A 2D Graphics Environment". *Computing in Science & Engineering* 9(3). pp. 90–95. DOI: https://doi.org/10.1109/MCSE.2007.55

Jewitt, C., Bezemer, J., and O'Halloran, K., 2016. *Introducing Multimodality*. London: Routledge.

Kembhavi, A., Salvato, M., Kolve, E., Seo, M., Hajishirzi, H. and Farhadi, A., 2016. A diagram is worth a dozen images. In: *European Conference on Computer Vision (ECCV)*. DOI: https://doi.org/10.1007/978-3-319-46493-0_15

Kim, D., Kim, S. and 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). Florence, Italy: Association for Computational Linguistics. pp. 3568–3584.

Kim, D., Yoo, Y., Kim, J., Lee, S. and Kwak, N., 2017. 'Dynamic Graph Generation Network: Generating Relational Knowledge from Diagrams'. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4167–4175.

Kress, G., Jewitt, C., Ogborn, J., Tsatsarelis, C., 2001. *Multimodal Teaching and Learning: The Rhetorics of the Science Classroom*. New York: Bloomsbury.

Liu, Y. and O'Halloran, K.L., 2009. Intersemiotic Texture: analyzing cohesive devices between language and images. *Social Semiotics*, 19(4). pp. 367–388. DOI: 10.1080/10350330903361059

Mann, W., Matthiessen, C. and Thompson, S., 1989. 'Rhetorical Structure Theory and Text Analysis'. In: Mann, W. and Thompson, S., (eds.) *Discourse Description: Diverse Linguistic Analyses of a Fund-Raising Text*. Amsterdam/Philadelphia: John Benjamins. pp. 39–78. DOI: 10.1075/pbns.16.04man.

Mann, W. and Thompson, S., 1988. Rhetorical Structure Theory: Toward a functional theory of text organization. *Text* 8. pp. 243–281.

Martin, J., Unsworth, L. and Rose, D., 2021. Condensing meaning: Imagic aggregations in secondary school science.

Menendez, D., Rosengren, K.S. and Alibali, M.W., 2020. Do details bug you? Effects of perceptual richness in learning about biological change. *Applied Cognitive Psychology* 34(5). pp. 1101–1117. DOI: https://doi.org/10.1002/acp.3698

O'Halloran, K., & Smith, B. (eds.), 2011. *Multimodal Studies: Exploring Issues and Domains.* 1st ed. New York: Routledge. DOI: https://doi.org/10.4324/9780203828847

Pedersen, J., Kocsis, D., Tripathi, A., Tarrell, A., Weerakoon, R.M.A., Tahmasbi, N., Xiong, J., Deng, W., Oh, O., de Vreede, G-J., 2013. 'Conceptual Foundations of Crowdsourcing: A Review of IS Research'. In: Proceedings of the Annual Hawaii International Conference on System Sciences. pp. 579–588. DOI: 10.1109/HICSS.2013.143.

Purchase, H., 2014. Twelve years of diagrams research*. Journal of Visual Languages and Computing* 25(2). pp. 57–75. DOI: https://doi.org/10.1016/j.jvlc.2013.11.004

Ruslan, M., 2005*. The Oxford Handbook of Computational Linguistics* (Oxford Handbooks). USA: Oxford University Press, Inc.

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

Seo, M., Hajishirzi, H., Farhadi, A., Etzioni, O. and 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). Lisbon, Portugal: Association for Computational Linguistics. pp. 1466–1476.

Steen, F., Hougaard, A., Joo, J., Olza, I., Cánovas, C., Pleshakova, A., Ray, S., Uhrig, P., Valenzuela, J., Woźny, J. and Turner, M., 2018. Toward an infrastructure for data-driven multimodal communication research. *Linguistics Vanguard*, 4 (1). pp. 20170041. DOI: https://doi.org/10.1515/lingvan-2017-0041

Stöckl, H., 2020. 'Linguistic Multimodality – Multimodal Linguistics: A State-of-the-Art Sketch'. In: Wildfeuer, J., Pflaeging, J., Bateman, J., Seizov, O. and Tseng, C. (eds.) *Multimodality: Disciplinary Thoughts and the Challenge of Diversity*. Berlin/Boston: De Gruyter. pp. 41-68. DOI: https://doi.org/10.1515/9783110608694-002

Taboada, M. and Habel, C., 2013. Rhetorical relations in multimodal documents. *Discourse Studies*, 15. pp. 65–89. DOI: 10.1177/1461445612466468.

The pandas development team, 2021. pandas-dev/pandas: Pandas. Zenodo. DOI: http://doi.org/10.5281/zenodo.4681666

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

Vallat, R., 2018. Pingouin: statistics in Python. *Journal of Open Source Software*, 3(31). DOI: https://doi.org/10.21105/joss.01026

Waller, R., 2012. Graphic Literacies for a Digital Age: The Survival of Layout. *The Information Society - TIS* (28). pp. 236–252.

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

Waskom, M. L., 2021. seaborn: statistical data visualization. *Journal of Open Source Software*, 6(60). DOI: https://doi.org/10.21105/joss.03021

Watanabe, Y. and Nagao, M., 1998. Diagram Understanding Using Integration of Layout Information and Textual Information. *COLING-ACL*. pp. 1374-1380.

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

Appendices

Appendix A

						NOUN			**ADJ**	
		NOUN	**ADJ**			**NOUN**	**VERB**	**ADJ ADJ**	**NOUN**	
Relation	**NOUN**	**NOUN**	**NOUN**	**NUM**	**ADJ**	**NOUN**	**NOUN**	**NOUN**	**NOUN**	**VERB**
elaboration	1627	540	553	37	43	33	25	67	43	6
connected	347	88	68	1	4	12	2	8	3	1
identification	1563	391	205	191	66	49	74	23	31	25
preparation	30	57	23	0	2	17	1	5	7	1
restatement	5	6	9	0	0	1	1	0	0	0
class-ascription	95	30	48	0	4	2	3	1	3	1
background	4	0	0	0	1	0	0	0	0	0
property-ascription	161	32	21	9	61	1	8	1	0	38
nonvolitional-result	9	2	3	0	2	0	0	0	0	0
circumstance	22	1	1	0	0	0	0	0	1	2
cyclic sequence	40	8	9	4	6	0	0	0	0	0
volitional-result	1	0	0	0	0	0	0	0	0	0
means	7	4	0	0	0	0	1	0	0	0
sequence	13	2	2	0	1	0	0	1	1	0
enablement	1	0	0	0	0	0	0	0	0	0
disjunction	7	0	0	0	0	0	0	0	0	0
list	1	0	0	0	0	0	0	0	0	0
condition	1	1	0	0	0	0	0	0	0	0
contrast	2	1	1	0	0	0	0	0	0	0
nonvolitional-cause	0	3	1	0	0	1	1	0	1	0
conjunction	2	0	0	0	0	0	0	0	0	0

Appendix B

Macro-group	NOUN	NOUN NOUN	ADJ NOUN	NUM	ADJ	NOUN NOUN NOUN	VERB NOUN	ADJ ADJ NOUN	ADJ NOUN NOUN	VERB
illustration	985	220	152	38	49	20	6	12	8	25
network	700	187	133	6	4	35	6	14	9	9
table	174	98	25	23	25	11	27	3	7	15
cycle	470	133	75	52	55	13	34	6	21	12
cross-section	903	272	319	51	24	20	15	35	25	0
cut-out	335	119	129	31	10	5	13	31	10	2
horizontal	140	71	37	28	18	7	12	4	3	10
diagrammatic	49	12	13	5	0	0	0	0	0	0
vertical	71	24	29	1	3	1	1	1	6	1
photograph	50	18	6	0	1	4	2	0	1	0
exploded	7	2	2	0	0	0	0	0	0	0

Appendix C

	illustration	network	table	cycle	cross-section	cut-out	horizontal	diagrammatic	vertical	photograph	exploded
elaboration	859	45	5	219	1583	614	90	22	53	70	5
connected	2	545	0	20	5	0	0	9	10	0	0
identification	674	602	304	773	202	120	177	48	100	16	6
preparation	88	29	16	69	47	39	22	6	11	4	0
restatement	16	1	0	5	6	0	6	0	0	0	0
class-ascription	32	37	20	2	22	15	50	0	11	5	0
background	2	2	0	8	3	2	1	0	0	0	0
property-ascription	166	15	120	19	33	12	64	5	5	1	0
nonvolitional-result	11	0	0	2	5	0	0	0	0	0	0
circumstance	16	9	0	26	12	4	4	1	0	0	0
cyclic sequence	3	0	0	85	0	3	0	0	2	0	0
volitional-result	0	0	0	1	0	0	0	0	0	0	0
means	2	0	0	11	7	0	0	1	0	0	0
sequence	0	8	0	9	6	0	0	3	1	0	0
enablement	1	0	0	1	0	0	0	0	0	0	0
disjunction	0	0	0	8	0	0	0	0	0	0	0
list	2	0	0	0	0	0	0	0	0	0	0
condition	2	0	0	0	0	0	0	0	0	0	0
contrast	2	0	0	0	0	0	2	0	0	0	0
nonvolitional-cause	2	0	0	6	0	0	0	0	0	0	0
conjunction	0	0	0	2	0	0	0	0	0	0	0