



# Assessing the quality of scientific explanations with networks

S. Wagner  and B. Priemer 

Department of Physics, Humboldt-Universität zu Berlin, Berlin, Germany

## ABSTRACT

This article introduces a network approach to describe the quality of written scientific explanations. Existing approaches evaluate explanations mainly on the level of sentences or as a whole but not on the elementary level of single terms. Moreover, evaluation of explanations is often based on highly inferential scoring techniques. We addressed both issues by converting the elementary structure of terms in explanations into networks (so-called element maps) and analysing these with mathematical measures, thus extracting the size and complexity of an explanation, adequacy, coherence, and use of key terms. A total of 65 explanations of experts and students were analysed quantitatively and qualitatively. Differences between expert and student maps' measures can be interpreted meaningfully against the background of existing research findings. Thus, we argue that our approach using network analysis provides a precise, fine-grained, and low-inferential tool that complements and refines existing approaches. Element maps have the potential to improve teaching and research by precisely revealing the strengths and weaknesses of explanations.

## ARTICLE HISTORY

Received 3 April 2022  
Accepted 20 January 2023

## KEYWORDS

Explanations; assessment; networks

## Introduction

Explaining natural phenomena is a central practice of doing science and for learning science (Braaten & Windschitl, 2011; Forman, 2018; Manz et al., 2020). Therefore, curricula and policy documents for science education place great emphasis on explanations (e.g. NGSS Lead States, 2013; Quinn et al., 2012). According to these, 'to construct logically coherent explanations' (NGSS Lead States, 2013, p. 52) is considered an essential practice to learn science.

As scientific explanations are made up mainly of language, language is crucial for both scientists and students. When explaining a phenomenon scientifically, one must combine terms in order to (e.g. causally and conditionally) describe and connect events of the phenomenon it with theory meaningfully to create a linguistic product and enrich it with appropriate diagrams, pictures, and formulae. Basically, language 'is a system of resources for making meaning' (Lemke, 1990, p. ix) of the world. Thus, 'learning science means learning to *talk* science' (Lemke, 1990, p. 1, emphasis in original). However, often, 'students are not taught *how to talk* science: how to put together

**CONTACT** S. Wagner  [steffen.wagner@physik.hu-berlin.de](mailto:steffen.wagner@physik.hu-berlin.de)  Humboldt-Universität zu Berlin, Berlin, Germany

© 2023 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group  
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives License (<http://creativecommons.org/licenses/by-nc-nd/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited, and is not altered, transformed, or built upon in any way.

workable science sentences and paragraphs, how to combine terms and meanings. [. . .] Is it any wonder that very few succeed?' (Lemke, 1990, p. 22, emphasis in original). Thus, an understanding of how students use language to learn helps to understand their construction of explanations. Since Lemke's concerns, science education researchers have made many efforts to improve learners' explanations and analyse their abilities and difficulties in providing high-quality written explanations.

However, although the practice for learners is to compose explanations by interrelating single terms, no approach in science education research has analysed explanations' quality on this elementary and fine-grained meaning-making level. Instead, characteristics of quality are scored based on coarser levels of analysis. Some scholars evaluate the explanation as a whole (de Andrade et al., 2019; Taber & Watts, 2000; Tang, 2016), whereas others use clauses (Peker & Wallace, 2011) or components of varying size (McNeill et al., 2006; Peel et al., 2019; Ruiz-Primo et al., 2010; Sandoval, 2003; Zarkadis & Papageorgiou, 2020). For example, research has examined whether an explanation as a whole is relevant (de Andrade et al., 2019) or scientifically correct (Taber & Watts, 2000) or if larger parts of an explanation are logically coherent (McNeill et al., 2006). However, this broader level of analysis cannot identify where exactly problems in students' explanations lie. It is evident that any object, property, or process represented as a term in an explanation may or may not be relevant, and any proposition may or may not be correct or coherent. In our view, reducing the grain size of the analysis to the level of individual terms thus offers an opportunity to capture structural features - which point to weaknesses and strength - with particular precision. This not only provides more detailed information on how the explanation was build. It further provides precise hints how explanations can be improved, e.g. by identifying an incorrect proposition in a part of a sentence. So, we see a benefit in a fine-grained analysis because it helps to understand how students' explanations are structured and - building on that - to support students to improve their explanations.

Moreover, the evaluation of explanation features has so far often been based on properties that are assigned to the explanation as a whole or to larger parts of it by interpretative and often more or less explicit coding rules. Yet explanations themselves, as linguistic constructs, have measurable structural properties that can be used for the assessment of features. Thus, we advocate for additional techniques to accurately assess explanations with higher precision (at the fine-grained level of terms) and higher objectivity. One approach is to map the structure of these smallest units of meaning (individual terms) rather than to capture larger components in explanations. Lemke (1983) showed a possibility for structure mapping in the form of thematic patterns, which essentially are networks. In science education, networks have become increasingly prominent tools that help make precise statements about conceptual understanding or knowledge structures (Siew, 2020), for example, with concept or knowledge maps and semantic networks (Koponen & Nousiainen, 2018; Kubsch et al., 2020; Novak, 1990; Thurn et al., 2020; Yun & Park, 2018). One limitation so far is that no mapping tool has been explicitly developed for mapping the fine-grained structure of written explanations with their linguistic specificities. Thus, it is unclear whether and how network measures are suitable for describing the quality of explanations at an elementary level.

Therefore, we aimed to investigate whether and how such network representations and measures are suitable to describe the quality of written scientific explanations in a fine-grained manner based on measurement of their structural properties. To this end,

we developed a (network) tool that maps the fine-grained structure of explanations in an appropriate way from a science education perspective. To introduce it, we elaborate on the fine-grained, elementary structure of explanations and give a short overview of networks, relevant network measures, and facets of explanation quality examined in previous science education research. Then, we briefly describe how to transform explanations into networks. Using the results of a study in which we collected 65 explanations of an optical phenomenon given by physics students and experts, we checked whether and how it is possible to interpret network measures as facets of explanation quality. Finally, we discuss this approach's possible benefits and limitations for science education research and the classroom.

## Theoretical background

Before we describe what we see as the quality of explanation; the fine-grained, elementary structure; and a corresponding network approach, we first narrow down the field of research in which we situate our approach.

Rocksén (2016) distinguished – besides everyday explanations – between explaining as *instructional practice* in the science classroom and *scientific explanations*. In the first concept, researchers analyse explaining as an interaction of the teacher with the learners (e.g. Geelan, 2013; Kulgemeyer, 2018). Concerning scientific explanations, researchers analyse, for example, structural aspects of learners' written explanations, their quality, and obstacles. We take a structural perspective following the latter conceptualisation. Our study asked students to explain a natural phenomenon of apparent depth (see Figure 2, described in the Method section), and we thus categorise our study under *scientific explanation*, according to Braaten and Windschitl (2011) who differentiated between (a) scientific explanation, where learners explain natural phenomena; (b) explanation as explication, where someone explains relevant scientific concepts; (c) explanation as causation, where students must focus on the underlying causes of perceptible events; and finally, (d) explanation as justification, where students must formulate arguments rather than explain phenomena. All four activities are relevant to science education but should be clearly distinguished. For example, while scientific explanations aim to answer questions about a phenomenon that is not in doubt, argumentation in contrast aims to justify a claim; thus, 'there is always a substantial degree of tentativeness associated with any argument' (Osborne & Patterson, 2011, p. 629). However, both can be represented in a written manner, thus have an inner structure and quality, and are relevant to science education. However, because all four notions (a–d) can be presented as discourse products, we did not aim to limit our approach to one of them.

There are three (partly overlapping) lines of research on written explanations. First, some studies introduce new approaches or frameworks to describe (the quality of) explanations (Alameh & Abd-El-Khalick, 2018; Braaten & Windschitl, 2011; de Andrade et al., 2019; Papadouris et al., 2018). Second, there are studies that explore the features and difficulties of learners with explanations from different perspectives (Herman et al., 2019; Peker & Wallace, 2011; Redfors & Ryder, 2001; Ruiz-Primo et al., 2010; Taber & Watts, 2000; Tang, 2016; Wiley et al., 2017; Zarkadis & Papageorgiou, 2020). Third, some studies examine explanation quality in relation to specially designed interventions and support (e.g. Delen & Krajcik, 2015; McNeill et al., 2006; Tang, 2016; Zacharia, 2005). As we introduce a new instrument, explore

its characteristics, and identify difficulties in explanations, we see our work as mainly contributing to the first and second research lines. The application of our instrument in intervention studies, as in the third research line, is a future goal.

### **Quality of explanations**

What constitutes the quality of a written scientific explanation and how it is operationalised varies widely among different approaches and frameworks. For example, some researchers consider causal coherence a relevant characteristic of scientific explanations (Alameh & Abd-El-Khalick, 2018; Braaten & Windschitl, 2011; de Andrade et al., 2019; Sandoval, 2003; Taber & Watts, 2000). Some authors regard the employment of theories or specific concepts as a feature of quality in scientific explanations (Braaten & Windschitl, 2011; de Andrade et al., 2019; Peel et al., 2019; Sandoval, 2003). Further quality characteristics are the relevance of content present in the written explanation (de Andrade et al., 2019; Ruiz-Primo et al., 2010), complexity (Herman et al., 2019; Papadouris et al., 2018; Zarkadis & Papageorgiou, 2020), correctness (Taber & Watts, 2000), sophistication or accuracy (Herman et al., 2019), appropriate use of models (Redfors & Ryder, 2001), completeness of a scientific account present in the explanation (Wiley et al., 2017), or the presence of selected components (McNeill et al., 2006), as well as surface features such as the number of words (Wiley et al., 2017). Some even use teachers' marks for evaluating the quality of an explanation without explicating specific criteria (Tang, 2016). Moreover, even those mentioned individual characteristics are operationalised differently among the approaches.

Thus, no universally valid set of characteristics seems to constitute a high-quality explanation. Instead, the conceptualisation of quality depends on the framework used, the focus, and the respective design in the studies. This makes it difficult to compare individual studies with each other.

Comparing learners and experts in regard to their explanation quality and conceptual understanding, findings describe experts as consistently using key-terms (Lachner et al., 2012; Thurn et al., 2020; Yun & Park, 2018) and stronger interconnecting a phenomenon with theory (Lintern et al., 2018).

### **Elementary structure**

We aim to obtain insights relevant to science learning about the quality of explanations. In order to do so, we investigate the quality of explanations at the level of the smallest units that are meaningful and at the same time can be analysed in a meaningful way (single terms) from science education perspective. Thus, we now outline the elementary structure of explanations.

The basis of the theoretical background of our approach can be traced back to the idea that explanations can be represented as a structure of statements made up of entities and relations (Kuhn, 2000). In written explanations, these basic elements contain words as the smallest meaningful units (Bechtel & Abrahamsen, 2005). Thus, some words stand for *entities* (Halliday & Matthiessen, 2013) or concepts, whereas other words stand for *relations* (Lemke, 1983) between those entities. What these entities and concepts are and what is considered a relation between them again varies depending on the

perspective. For our approach, entities can reasonably be associated with real (observable) or imaginary scientific objects, systems, and properties. Nouns, or pronouns as stand-ins for nouns, commonly represent such entities (Halliday & Matthiessen, 2013). Relationships are identified by other words, especially verbs, adverbs, conjunctions, and so on, and give meaning to entities (Lemke, 1983). We designate both entities and relations as *elements*. Therefore, an element is either an entity or a relation. Because this fine-grained structure of explanations is made up of elements and is thus the fundamental, elementary level that builds up explanations, we call it the *elementary structure*.

## Networks

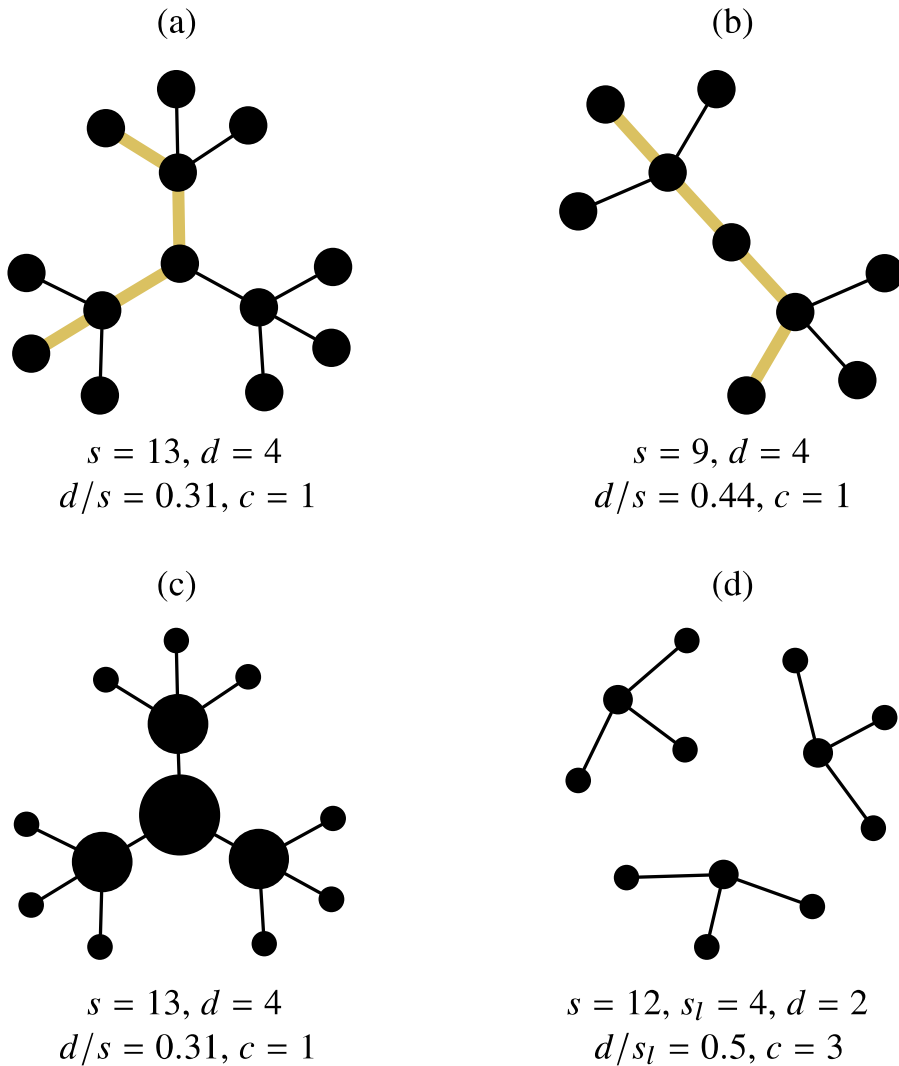
Networks are by definition structural representations (Barabási, 2016) that can be customised for various purposes. They basically consist of vertices (or *nodes*) and connections between them that are called *edges* (lines or arrows). Both nodes and edges can be labelled and thus have the potential to carry semantic content. Therefore, they can represent the elementary structure of explanations, that is, the entities and the relations between them. We call networks representing the elementary structure *element maps*. Element maps can be considered as modified concept maps (Novak, 1990). In concept maps, nodes represent concepts and edges show how these concepts relate to each other. However, usually concept maps are created by learners (and experts) themselves to directly visualise their conceptual structure. They are therefore not designed to map the specific features of written language and thus do not meet the requirements for analysing written scientific explanations.

There are numerous measures to characterise network characteristics, which we illustrate in general in Figure 1. A simple characteristic of networks is the number of all nodes, called the *size* ( $s$ ). Next, the number of steps to take the shortest path between the most distant points in a network is the *diameter* ( $d$ ). A further characteristic of a network is how intertwined or compact it is. A simple measure is to build the *ratio* ( $d/s$ ) of the diameter and size. A network with a high number of nodes and a small diameter is presumably compact and has a low ratio of  $\frac{d}{s} < 1$ . In contrast, a long linear chain of nodes will have a ratio of  $\frac{d}{s} \approx 1$ .

Some networks consist of components<sup>s</sup> that have no connections to each other. The number of such *unconnected components* ( $c$ ) thus describes the fragmentation of a network. To describe compactness of networks consisting of several unconnected components, one can take the  $d/s$  ratio of the component with the largest size.

Size, diameter, ratio, and the number of unconnected components are global measures for a network. However, there are also measures to characterise individual nodes, such as how central a node is, for example, the *betweenness centrality* ( $bc$ ). For the purpose of this paper, it is sufficient to understand how  $bc$  is interpreted<sup>1</sup>: A node with high  $bc$  lies in between many other nodes of the entire network (Freeman, 1977) and has a function of holding different parts of the network together. To consider different network sizes, we use normalised  $bc$  (values lie between 0 and 1).

In previous studies using network techniques, experts are characterised by having more interconnected knowledge elements (Jonassen et al., 2013). Moreover, in research



**Figure 1.** Four example networks with different characteristics. The highlighted path in (a) and (b) represents a possible path of the diameter. In (c) and (d), the visual node size represents the betweenness centrality ( $bc$ ) of the nodes. Numerical values for  $bc$  are omitted from the diagram because otherwise, there would have to be up to 13 values for each map. In diagram (d),  $s_l$  and  $d_l$  represent the size and diameter of the largest component, respectively.

on concept maps, experts create larger (Schaal, 2008; Williams, 1998) and more complex maps (Williams, 1998). Moreover, experts' networks show greater coherence (Koponen & Pehkonen, 2010). Additionally, the size of concept maps of a topic grows with learning (Thurn et al., 2020).

Knowing that there are many more measures to describe network characteristics, we focus on the previously mentioned and investigate if and how we can interpret them as facets of explanation quality.

## Research question

Before we describe our approach in the following Method section, we formulate our research question.

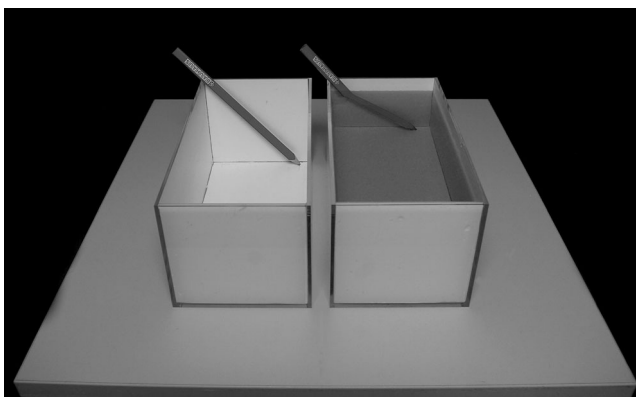
RQ: How can differences in the quality of written scientific explanations in science education given by science learners and experts be described using element maps?

To answer our research question, we apply the introduced measures on the networks that represent the elementary structure of explanations given by science learners and experts in a study. We use two criteria to legitimise these network measures as descriptions of the quality of scientific explanations. First, the measures must show differences between learners and experts. Second, the measured characteristics and observed differences must have a meaningful relationship to relevant existing research in science education.

## Method

### Study design

In our study, we used phenomena relating to the context of apparent depth (see [Figure 2](#)). Participants were presented with the experiments as shown in [Figure 2](#) and asked to write a scientific explanation for their observation in their own words. They were provided with computers and word processing software and given as much time as they needed. The requested level of sophistication should correspond to the level at which this phenomenon is ‘typically’ described in their introductory physics lectures. We collected nine written explanations from experts and 56 explanations from physics or biophysics bachelor students. The student participants attended a lecture and completed exercises about the refraction of light before the survey. All explanations were written in German. Because we do not make statements regarding personal context but only analyse discourse products, we did not collect person-specific data. Based on our experience with this cohort, the participating students can be described as predominantly of



**Figure 2.** Photo of an apparent depth phenomenon. A pencil is placed in each of two open boxes, one filled with water (right) and the other empty (left). The pencil in the right box appears kinked and its tip apparently lifted.

Central European descent, without disparities regarding the distribution of gender, and are slightly older than 20 years, with some outliers on the upper side. We would like to explicitly point out that this sample cannot be considered representative for physics learners in general. However, we have chosen participants with an advanced learning status in physics because we expected text products that are well evaluable in an explorative design on the one hand, and we want to take into account a wider range in the characteristics to be investigated, especially in the direction of possible expertise. Participants were not supported through, e.g. scaffolding. An appropriate intervention should tie in with the abilities, strengths and weaknesses that the learners have naturally, i.e. without support. The aim of our study is to identify these and make them visible. Therefore, we did not provide any scaffolding. Knowing that the tool works in general, we can then adapt it to the specifics of other target groups with text products that are probably harder to evaluate. We chose to employ scholars from physics education as experts in explaining physics content to a given target group. Furthermore, the experts involved have dealt with explanations of these phenomena both in an educational context and in publications. We realise that the number of experts is quite small, especially compared to the number of students. However, our experts' explanations offer enough material for a rich and detailed analysis. Our main concern, further, is to study learners and not experts. Experts serve us only as a reference point. Calculating differences between the two groups of participants can, nevertheless, adequately account for the different sample sizes with the appropriate measures.

In total, we received 65 explanations. All explanations were written using simple text editor software. Participants were allowed to add pencil and paper sketches. The sketches served as support for analysing the maps' content and in the case of unclear meaning of terms but were not analysed separately. All participants were informed about the purpose of this study and agreed on using their given responses for our research purpose. Since it was - at this stage - not our intention to rate the quality of the participants' explanation on a scale, no performance expectation was constructed. Instead, the focus was set on detecting differences between students' and experts' explanations with respect to certain network measures.

The explanations were transformed into (original) element maps using spreadsheet software to extract the elementary structure and the R environment for network visualisation and analysis. Afterwards, a content evaluation was performed, checking all entities for relevance and all relationships for correctness. We developed and refined a coding manual in advance that guided the whole process. The manual describes the four transformation parts (identifying entities, identifying relations, evaluating relevance, and evaluating correctness). Interrater agreement (two raters) was checked on thirteen explanations using Cohen's kappa for the relevance of entities ( $\kappa = 0.68$ ) and relative mutual agreement on the correctness of the relations ( $r = 0.83$ ). A third (senior) research expert rated the content validity based on the manual's evaluation description as positive by evaluating, if the manual uses disciplinary correct rules and examples that lead to a disciplinary adequate evaluation of correctness and relevance.

As we mentioned in the beginning of the section, we use a phenomenon of apparent depth for our study. Apparent depth makes objects in water appear closer to the water surface (above their real position) when viewed from outside the water (see [Figure 2](#)). Instances of this are looking at coins in a fountain or at a paddle protruding out of a



boat diagonally into the water. This phenomenon is usually explained by the refraction of light (Snell's law) that occurs when light passes from one medium (e.g. water) to another (e.g. air). An observer outside the water constructs an image of the object from the refracted light (rays). This results in an apparent object position at a shallower depth in the water than where the image would be without the water (Nassar, 1994). The interpretation of the visual stimuli by an observer thus plays a special role here. Whether the person knows that the object is actually in a different place from where he or she sees it is an important aspect that is essential for an explanation. Therefore, it is important not only to show a picture of an object that appears to be lifted, but also, for example, to show the object with and without water and to make the tactile location of the objects physically accessible to the observer (and not only through a photo), as we have also done both.

### **Creating element maps from written explanations**

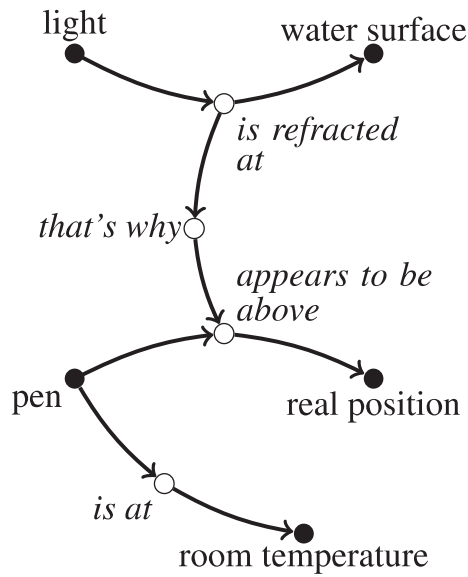
Mapping the elementary structure means assigning the respective components of the elementary structure to those of networks. We have done this by representing elements with nodes and the connections between them with arrows. Because there are two different elements (entities and relations), there will be two different node types: entity nodes and relation nodes.

The mapping is done in three steps. For a brief report of the procedure, here, we use the following short (fictional) example explanation:

'Light is refracted at the water surface. That's why the pen that is at room temperature appears to be above its real position.'

In the first step, we extract *all nouns and pronouns* from the explanation. They are used to build the entity nodes in element maps. In the first sentence of our example, the nouns are 'light' and 'water surface'. Additionally, multiple mentions of the same noun as well as synonymous (pro)nouns are merged into one entity so that an explanation does not grow because the same entity is named several times. In the second step, we check explanations for *all relations between the entities* (e.g. 'is refracted at'). These relations are depicted as relation nodes. The entity and relation nodes become connected by arrows in the direction of reading the explanation. A relation node with its connected entity nodes forms a proposition. The third step captures *all relations between propositions* (e.g. 'that's why'). In this way, the entire elementary structure of an explanation is extracted sentence by sentence and assembled into a map. At this point, we would like to point out why we decided to represent relations as nodes in contrast to concept maps where relations are visualised as labelled arrows. An entity can belong to several propositions, whereas each relation only belongs to precisely one proposition. Therefore, to connect two propositions, we choose relations as the start and endpoints of the connection. A network object that is a start or endpoint becomes a vertex (node) by definition. Thus, we represent it as a node in the element maps.

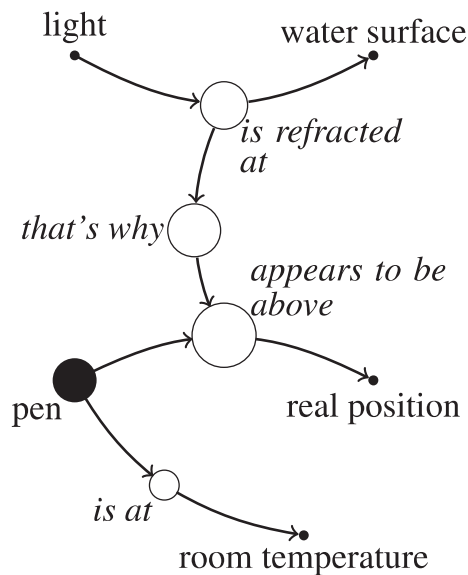
Because element maps so far represent the elementary structure of an explanation as a participant originally wrote it, we call them *original element maps*. For our example explanation, the original element map is shown in [Figure 3](#) (with all nodes having the same size) and [Figure 4](#) (with betweenness centrality shown as the visual node size).



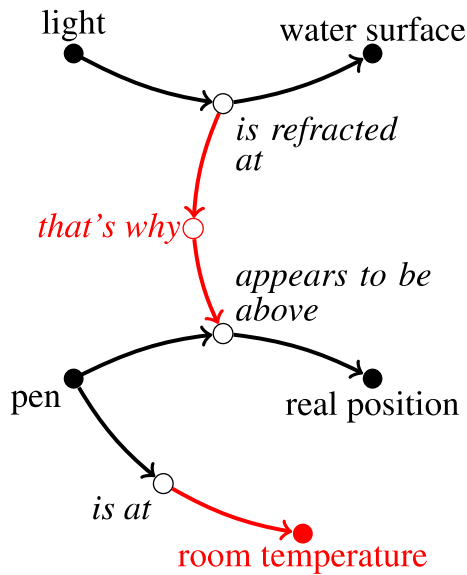
**Figure 3.** Original element map of the example explanation. Filled circle nodes with upright labels represent entities, and empty circle nodes with italic labels represent relations.

### Evaluating content

So far, the map might still contain inappropriate propositions, incorrect relations, and irrelevant entities. If we want to make meaningful statements about the quality of

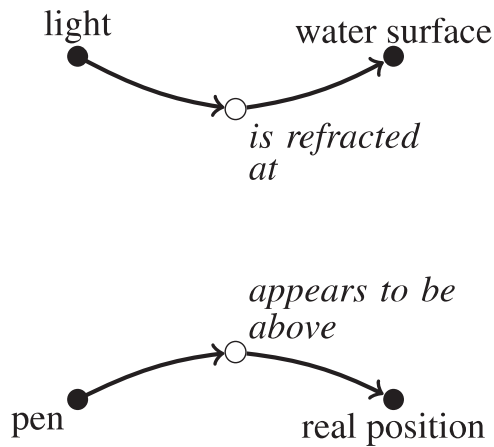


**Figure 4.** Original element map of the example explanation with betweenness centrality shown as the visual node size.



**Figure 5.** Evaluated element map of the example explanation with faulty elements marked in red.

explanations based on network measures, we obviously must consider possible errors. Accordingly, these shortcomings must be identified in a further step and somehow made visible in the element map. Therefore, each element must be evaluated as either relevant or not relevant (entity) and correct or not correct (relation). If an entity is not relevant or a relation is not correct, we call it a *faulty element*. Because propositions consist of these elements, the appropriateness of a proposition can be judged by these two assignments of relevance and correctness. In our example, the room temperature is not relevant to the phenomenon. Furthermore, the relation ‘that’s why’ is an incorrect relation because refraction is a necessary but insufficient condition for the occurrence of an apparent depth phenomenon. Hence, reducing the cause of the appearance phenomenon to refraction is an oversimplification. Rather, there must be an observer in a medium with a refractive index that is different from the one in which the object is placed. We have thus marked these two elements as faulty. The criterion for evaluating the entities as *relevant* was predominantly whether there was any recognisable connection to the phenomenon from a physical point of view. For example, stating that the pen is at room temperature is seen as irrelevant since this has no impact on the phenomenon as observed. The criterion for evaluating the relationships as *correct* was that they did not grossly contradict perception, established principles or common formulations with regard to the phenomenon context. Simply saying that the pen appears to be above its real position because light is refracted at the water surface is rated as incorrect because it oversimplifies an explanation by omitting reasoning with central concepts. We like to clarify that we counted participants’ statements, that a pencil that penetrates the water surface remains ‘unbroken’ even though it appears to be bended, as a correct and relevant part of the explanation. Because this evaluation is an interpretative step, its quality must be assessed using interrater agreement. We call a map with marked



**Figure 6.** Core element map without faulty elements. The map consists of two unconnected components (fragments).

faulty elements an *evaluated element map*. For our example, the evaluated map is shown in Figure 5.

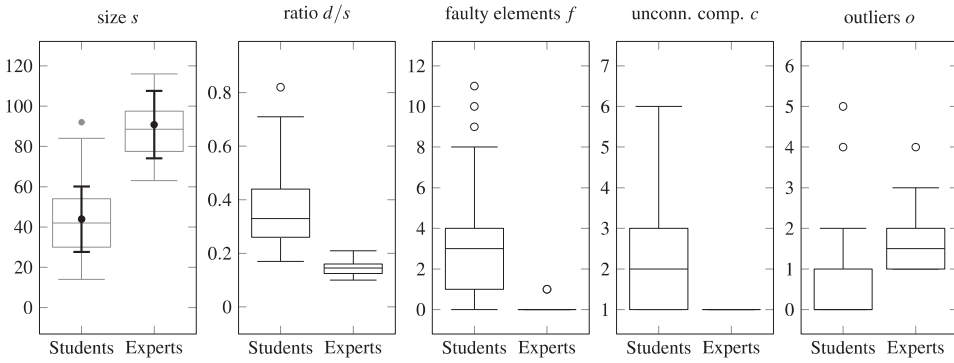
In a third version, we eliminate all propositions affected by faulty elements so that only valid statements with relevant and correct elements are included. This results in a map that represents the disciplinarily relevant and correct core of the explanation. We therefore call it the *core element map* of an explanation. The core map for our example explanation is shown in Figure 6.

### **Applying network analysis**

We analysed size ( $s$ ), ratio ( $d/s$ ), the number of unconnected components ( $c$ ), and betweenness centrality ( $bc$ ) on the core maps and the number of faulty elements ( $f$ ) on the evaluated maps.

## **Results**

In the following, we present both quantitative and qualitative results for each network. Regarding size, the distributions for both groups – learners and experts – do not significantly deviate from a normal distribution. Therefore, we report mean ( $M$ ) and standard deviation ( $SD$ ) and use Welch's  $t$ -test to reveal differences between groups. This test is recommended if one compares the central tendency of two unrelated groups with different sample sizes (Ruxton, 2006). The ratio of diameter to size as well as the numbers of faulty elements, unconnected components, and betweenness centrality outliers deviate significantly from a normal distribution in both groups (Shapiro–Wilk test). Therefore, we report medians ( $Mdn$ ) and analyse group differences using the nonparametric Wilcoxon signed-rank test for these characteristics. For effect sizes, we report Cohen's  $d$  (for size) and  $r$  for the other network measures. All data are visualised in Figure 7.

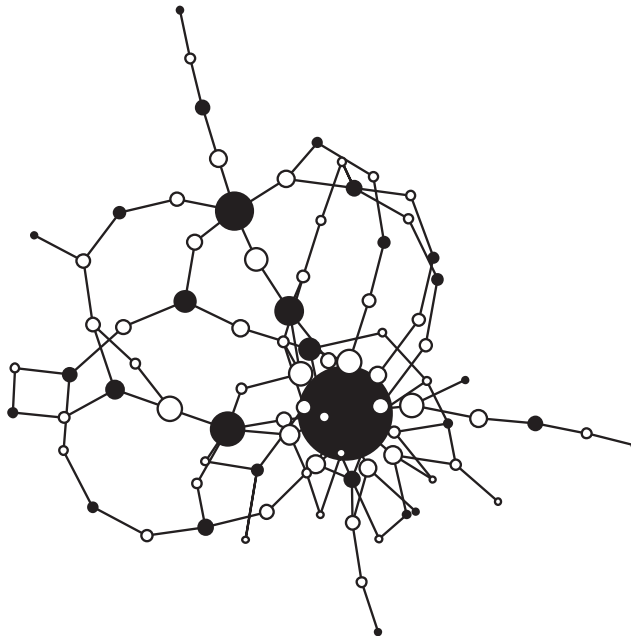


**Figure 7.** Boxplots to compare students’ and experts’ map characteristics. All differences between groups are significant ( $p < 0.001$ ). For the size, mean values and standard deviations are additionally visualised as black dots and lines.

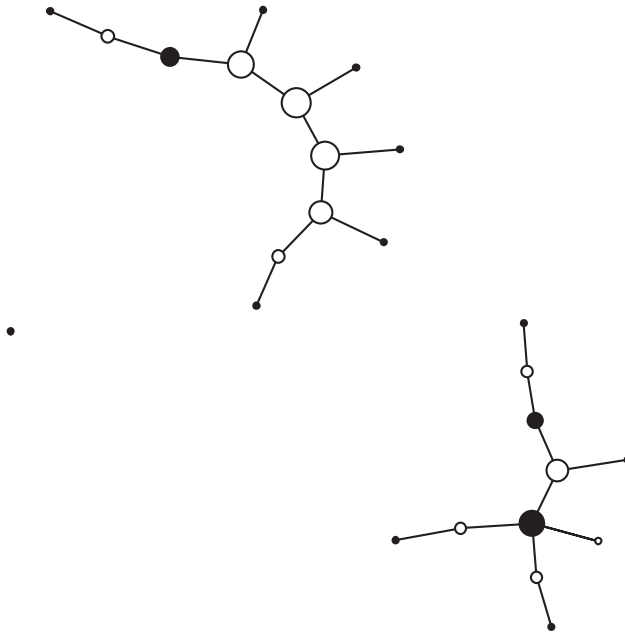
**Size and ratio of diameter to size**

**Quantitative**

Experts’ core maps ( $M = 90.8, SD = 17.26$ ) are of larger size  $d$  than students’ core maps ( $M = 43.9, SD = 16.84$ ), with the difference being significant and of large effect (Welch’s  $t$ -test,  $t = -7.60, p < 0.001$ , Cohen’s  $d = 2.77$ ). The ratio of diameter to size ( $d/s$ ), however, is smaller for experts’ maps ( $Mdn = 0.15$ ) than for those of students ( $Mdn = 0.33$ ), also with a large effect (Wilcoxon signed-rank test,  $W = 499.5$ ,



**Figure 8.** Core element map of an expert’s explanation from our sample.



**Figure 9.** Fragmented core element map of a student's explanation from our sample.

$p < 0.001$ ,  $r = .58$ ). Figure 8 and 9 show two example maps with different sizes and ratios. The expert's core element map (Figure 8) contains more elements that are more intertwined than the student's map does (Figure 9).

### Qualitative

A qualitative analysis reveals what makes experts' maps larger and more intertwined. All experts described the phenomenon in great detail by emphasising spatial and attributive relationships of the observer, the water surface, and so on. For example, one expert described the pencils (Figure 2) in detail: 'The pencils (left and right) are diagonally in box. They are placed before the back walls. Both boxes have the same size. The right pencil has a knee at the water surface, while the left pencil looks straight. The part under water seems to be bent upwards. The observer looks at the box obliquely from above [...]' (translated by the authors). Moreover, they associated a significant part of the entities of the phenomenon with corresponding theoretical properties, for example, water and air with their refractive indices, or the perceived image with the intersection of extended light rays etc. Some of these entities of the phenomenon and the theory are linked via causal connections. In the students' maps, the description of the pencils are much shorter, for example one student only wrote: 'The pencil appears clearly bent to the human eye' (translated by the authors). Further, we repeatedly miss features of the phenomenon (e.g. water surface, observer) and of refraction (directions of light rays, angles, image), as well as spatial, causal, and attributive relations between these features.

## Faulty elements

### Quantitative

Students' explanations contain more faulty elements,  $f$  ( $Mdn = 3$ ), than those of experts ( $Mdn = 0$ ), which is a significant difference between them and has a large effect ( $W = 472$ ,  $p < 0.001$ ,  $r = .52$ ). Of students' faulty elements, two thirds are incorrect relations and one third are irrelevant entities. In two explanations, an expert included an irrelevant entity.

### Qualitative

In students' maps, irrelevant entities often originate from concepts of light that are not relevant to the explanation of apparent depth phenomena. These are the particle concept, the wave concept of light, and the reflection of light at the water surface. For example, one student wrote that at the surface '[...] light photons are refracted differently' (translated by the authors). Furthermore, physical properties that are in our study unnecessary for apparent depth (humidity, temperature) and misinterpretations of perception (optical illusion) occur. Our example from the Method section contained such an unnecessary property (temperature). In two explanations, an expert included an irrelevant entity as well. Here, the expert pointed to the extreme case of looking vertically into the water from above. However, because this case was not part of the phenomenon, we decided to code this as irrelevant. Other than this example, the experts made no errors.

A look at the propositions based on inappropriate relations reveals first that they are predominantly considered to be violations of physics laws, for example, of Snell's law or Fermat's principle. Another general obstacle is oversimplification, where students describe the phenomenon in detail and then explain it only by mentioning the term 'refraction' without explaining this connection more precisely, as in our example explanation. Furthermore, some statements concern inappropriate formulations, for example, refraction takes place *in* water instead of at its surface. The experts did not formulate any sentences that included an inappropriate connection.

## Number of unconnected components

### Quantitative

All experts' core maps appear in one piece. Students' core maps, by contrast, tend to fall into fragments ( $Mdn = 2$ ) after content evaluation and removing faulty elements, with a maximum of six unconnected components. The difference between groups is significant, with a moderate effect ( $W = 432$ ,  $p < 0.001$ ,  $r = .44$ ). [Figure 8](#) and [Figure 9](#) illustrate this difference by showing a coherent, one-piece element map of an expert's explanation ([Figure 8](#)) and a three-fragment map of a student's explanation ([Figure 9](#)).

### Qualitative

Maps of experts' explanations are held together by many causal, spatial, temporal, quantitative, and qualitative relations. A qualitative analysis of the student maps shows that of the fragments of a map, one often contains a description of a phenomenon (e.g. entities such as water, the tank, the pen, etc. and the relations between them), whereas theoretical elements (light, refraction, etc.) are found in another fragment. This peculiarity is also

illustrated in our example core map in Figure 9. Therein, the short phenomenon description and the mention of refraction are completely separated. However, the fragments can vary in size from single-element fragments to large and complex fragments. In students' fragmented maps, different patterns appear. Some students struggle with the conceptual coordination of certain theoretical elements, either because they include irrelevant ones or because they do not know how to appropriately connect those theoretical elements to each other and to the features of the visible phenomenon. Other students' maps have a rather large theory fragment that is not (or weakly) connected to the phenomenon description. Again, other maps lack important elements (e.g. the 'observer'). Some maps show a mixture of these three features. Single-element fragments generally seem to be more often theoretical terms than descriptions of the phenomenon in our sample.

### **Betweenness centrality**

#### **Quantitative**

Regarding *bc*, we focus on nodes that are particularly central for the network because arranging explanations around key elements is expected to be a characteristic of expertise (Lachner et al., 2012). Such nodes can be identified by setting the quantile range of a map's *bc* distribution above which nodes start to appear as outliers. We set the quantile range for betweenness values in the experts' maps as a threshold, so that each expert has at least one outlier. Then, we counted the numbers of outliers *o* in each map and compared those of experts with those of students.

Accordingly, all experts' maps have at least one element with particularly high *bc*. Experts' explanations thus seem to be arranged around a few elements that are central to the explanation (similar to example (a) in Figure 1). By contrast, 36 students' maps are not arranged around elements of exceptional *bc*. Instead, compared with experts' maps, they are often built of many nodes with mid-level centrality. Other students' maps do have many nodes with high *bc* but only a few less important elements. Here, high-*bc* nodes do not appear as outliers. Such maps are more chain-like (similar to example (b) in Figure 1). Accordingly, the maps of experts ( $Mdn = 2$ ) contain more outliers than those of students ( $Mdn = 0$ ), whereby the difference is again significant with a moderate effect ( $W = 74.5$ ,  $p < 0.001$ ,  $r = .46$ ). Figure 8 shows a map of an expert's explanation from our sample, which is arranged around one element with high *bc* (here, 'observer'), whereas the student's map in Figure 9 has no such element and is arranged in rather chain-like fragments.<sup>2</sup>

#### **Qualitative**

When examining experts' central elements qualitatively, we found only a few different ones. In expert maps, the observer (or the eye) generally takes a central role. In only a few students' core maps does the observer also play a central role, and some do not even contain an entity for the observer. In some expert maps, particular light rays and their properties (e.g. directions) play a central role. Here, students tend to use the more unspecific term 'light'. Finally, the object of the apparent depth phenomenon (in our examples, the pencil) is central to expert maps. Interestingly, this object has high centrality in only one of the student maps. Again, some student maps that explain the phenomenon of apparent depth do not even contain that entity. An example is the



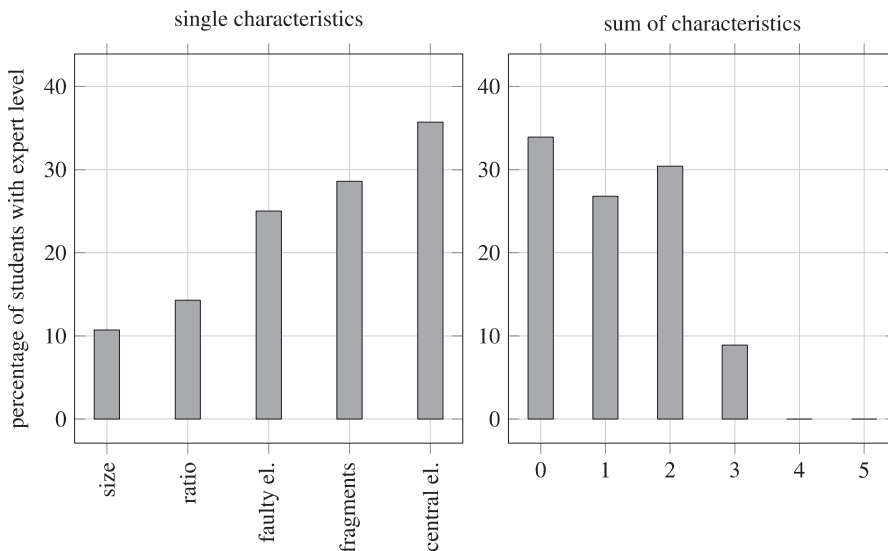
following entire explanation of one student for the observed phenomenon: ‘Due to the different refractive indices, a change in the direction of the light beam occurs at the water surface. The light beam is bent downwards when it leaves the water, and the image appears shifted’ (translated by the authors). In other words, those participants do not mention the object that they are looking. Instead, they speak of refraction, or the phenomenon in a more general sense.

### Quantitative overview

In the left bar chart in Figure 10, we show how many students’ maps achieve an expert-like level in each single network characteristic. We define that an expert-like level is reached when the value in a characteristic is within the range of values of all experts. Only 11% of the students’ maps have an expert-like size. However, about one third of all explanations are arranged around key terms the way experts construct their explanations. For all other characteristics, the percentage lies between these two limits. The right bar chart in Figure 10 shows the percentage of students’ maps achieving an expert-like level in zero to all five network characteristics. About 30% of all students’ explanations reach an expert-like level in none, one, or two of the characteristics and none in more than three characteristics.

### Discussion

In the following, we describe what the network measures of the element maps tell us about the explanations. We analyse how well the measures describe specific quality characteristics of explanations, interpret them, and discuss their benefits. The criterion



**Figure 10.** Bar charts showing the percentage of student core element maps that achieve the expert level in each characteristic (left) and the percentage of student core element maps that achieve an expert level in a sum of characteristics (right).

here is whether the differences found in these measures between the explanations of learners and experts can be meaningfully explained, for example, with the help of results from neighbouring research areas.

### ***Size and ratio of diameter to size***

The network characteristics size and ratio of diameter to size represent the size and complexity of an explanation. The amount of information and its complexity is understood as the sophistication of an explanation (cf. de Andrade et al., 2019; Zangori et al., 2015).

In both characteristics, we see significant differences between the maps of experts and learners. Our results show that experts integrate more extensive, detailed, and strongly interconnected knowledge in their explanations. In this respect, learners have the most significant difficulty reaching an expert level compared with other characteristics.

Science education research on expertise using network approaches supports these findings and the interpretation because experts create more complex and larger maps and (concept) maps created by learners become larger while concept learning. Since conceptual understanding and scientific knowledge of a phenomenon are strongly connected with the ability to explain it (Berthold & Renkl, 2009; Zacharia, 2005), it is reasonable that the elementary structure of written explanations given by experts show similar differences to those of learners.

Compared with other methods that, for example, only count the number of words, element maps offer the advantage of actually capturing only the relevant, disciplinarily correct core of the explanation. In element maps, mere repetitions of words or statements, for example with synonymous terms that increase the word count, do not lead to improvement of an explanation. In other words, the only way to increase size is to add information to the explanation that is relevant for the phenomenon at hand. The only way to decrease ratio of diameter to size is to link existing information in the explanation by correct relations. This supports the internal validity of interpreting size and ratio of diameter to size of element maps as size and complexity of an explanation. In addition, the qualitative investigation of the elementary structure also provides a good account of *how* expert explanations become larger and more complex, namely through a detailed description of the phenomenon, an elaborate rendering of the theory, and a stronger linking of phenomenon and theory, which can plausibly be described as a property of expert knowledge (Lintern et al., 2018) but which has not yet been recorded clearly in the structure of explanations. With these quantitative and qualitative results, element maps enable more objective and precise statements about the size and complexity than, for example, dichotomous and more interpretative scoring.

### ***Faulty elements***

The number of faulty elements can be used to describe the adequacy – or rather inadequacy – of an explanation. A good explanation should only contain appropriate propositions (Alameh & Abd-El-Khalick, 2018) with relevant information (Alameh & Abd-El-Khalick, 2018; Salmon, 2006) and correct relations (i.e. no faulty elements). Here, one must consider that a single error-free proposition is by no means an adequate explanation. Therefore, we use this measure only in connection with other measures.

Our study shows a significant difference between experts and learners concerning this measure. This difference can also be plausibly explained. Previous research has shown that learners find it challenging to use an appropriate concept of light for explanations (Redfors & Ryder, 2001). In particular, various cohorts have shown difficulties with apparent depth phenomena (Andersson & Kärrqvist, 1983; Galili & Hazan, 2000). Among other things, learners find it difficult to link the observer, the object, and the image that the observer sees apparently lifted (Kaewkhong et al., 2010). Our analysis shows one possible reason for this: the observer's absence in the explanation. Therefore, with the help of element maps, detailed content-related error analyses are possible. Thus, we consider the interpretation of faulty elements as the adequacy of scientific explanations as valid, since no other feature of the written explanation increases or decreases the number of faulty elements. However, one must be careful to interpret faulty elements as a learner's conceptual understanding or ability to explain a phenomenon at all, because, for example, problems in the appropriate use of language might lead to use incorrect relations or irrelevant entities. Element maps depict the structure of a written explanation, not of what an individual has in mind. Although there is a strong relation, it is not a one-on-one-mapping between both instances (also see Novak, 1990 for concept maps).

Nevertheless, besides the mention of problems, it should not go unmentioned that in the comparison of the faulty elements with the number of all elements (size of the maps), it is noticeable that the learners' explanations contain much more relevant than irrelevant terms and much more correct than incorrect relationships. Element maps thus clearly reveal both the learners' deficits and potentials in explanations.

### ***Number of unconnected components***

Since previous research lacks a clear and well-established conceptualisation of coherence, we suggest interpreting the number of fragments as the coherence of an explanation because coherence describes how well an explanation holds together (Sandoval, 2003). The more a network falls into components, the less coherent it is.

Our study shows up to six fragments among the students and a clear difference from the experts, all of whom produce coherent explanations. This result confirms other research on concept networks for explanations, according to which experts' networks have shown greater coherence (Koponen & Pehkonen, 2010). So far, we cannot use this measure to make a statement about causal coherence, which has been the focus of many studies in science education (cf. de Andrade et al., 2019; Kang et al., 2014; Taber & Watts, 2000; Zacharia, 2005). In principle, causality could also be captured as an element map feature. However, our analysis shows that coherence in explanations cannot be reduced to the causal aspect alone. What makes expert explanations coherent is causal and, for example, spatial, temporal, attributive, qualitative, and quantitative relationships. A focus only on causal coherence would narrow the view too much to assess the quality of an explanation adequately (Braaten & Windschitl, 2011).

### ***Betweenness centrality***

The betweenness centrality is well qualified to describe the use of key terms in explanations. Identifying which terms are actually central to an explanation is particularly

helpful. Our results confirm previous findings on conceptual understanding, according to which use of particular key terms is a characteristic of expertise (Lachner et al., 2012; Thurn et al., 2020; Yun & Park, 2018). The rather chain-like shape of some student explanations without central terms is also known from studies of conceptual understanding with the help of concept networks (Koponen & Nousiainen, 2018) and, with element maps, becomes evident in the structure of science learners' explanations. In addition, our results show that the experts all use a set of key terms, but which term is particularly central to an explanation differs between experts. This set (i.e. observer, light rays, object) is well interpretable from a physical point of view (see, e.g. Nassar, 1994) and thus supports the validity of betweenness centrality as a measure for the key function of terms. The 'observer' is the end point of the light path and perceives the image of the object. Thus, the observer can be seen as the mediator of theory and phenomenon in the explanation. 'Light rays' are the essential tools with which the constituents (object, image, interface, observer) are connected and therefore have a central importance. The object ('pen') that the observer sees is of particular importance because it is precisely its apparent and actual position that diverge. When applied to the elementary structure, key words mediate between different parts of an explanation and in this way enable experts to create large explanations with a comparatively small diameter, that is, greater complexity or sophistication. This shows that the individual characteristics of the quality of explanations are interrelated and can be advantageously analysed coherently with the help of networks. Moreover, although it is not the focus of our study, the results also show significant differences within the group of learners in two respects. First, in every single facet of explanation quality our results cover a wide range, qualitatively and quantitatively. Second, students create different *kinds* of explanations. While some build a large explanation with many errors, other contain less errors but are more fragmented, again others are much smaller but more complex at the same time etc. Thus, our results show that using element maps also enables detecting potential differences both between explanations of different learners and between explanations of learners before and after interventions.

## Conclusions, limitations, and outlook

We visualised and measured the fine-grained structure of written explanations by learners and experts using network analysis. The analysed characteristics can be well interpreted as facets of the quality of explanations: the size of the maps as the size of the explanation, the ratio of size and diameter as a measure of complexity, the number of unconnected components as coherence, and the elements with high betweenness centrality as key elements of the explanation.

With a combination of qualitative and quantitative analysis of all measures, element maps provide a good description of the quality of explanations based on their elementary structure. The results are precise and objective based on the structural properties of written scientific explanations. Where interpretations are necessary, we have made the procedure transparent by assessing interrater reliability and judged it to be very good. Each measure on its own is conceivable as a support for other procedures to shed light on a facet of the quality of explanations. Thus, we argue that element maps add a new, promising tool to known methods of analysing explanations. More precisely,

element maps provide measures for characteristics of written texts that describe the quality of an explanation precisely and objectively.

However, the current state of our tool is not without limits. The procedure has so far only been used in the context of explanations of the optical phenomenon of apparent depth with a limited number of participants. Explanations of other phenomena could have other characteristics, which is why the procedure might have to be adapted according to the specific context. Moreover, the application of our tool to other phenomena and a discussion of its compatibility with corresponding research results is essential for external validity. What is also still pending is the comparison of our approach with other methods applied to the same sample. A convergence of both methods would also give a better picture of validity. Our samples of the learners and the experts clearly also entail certain limitations. The development of explanations of other phenomena should therefore also be accompanied by a variation of relevant sample parameters, such as age, as well as the task and support given to them. Varying the task can as well unveil if participants may have confused instructional explanations with scientific explanations. In order to address these limitations, we are currently investigating written scientific explanations of an acoustic phenomenon by students aged 13–15 and experts using our tool in parallel with the system of analysis by de Andrade et al. (2019). Beside this, we aim to elaborate more on differences within the group of learners and not only between experts and learners as well as proper interventions. Furthermore, element maps are merely products of a text analysis and, thus, do not include other representation (e.g. figures) used in explanations. Thus, integrating other representations into the network analysis of text would be a fruitful extension of our method. Finally, as it is with other assessment tools as well, our analysis of the written explanations - even though it uses a relatively objective procedure - may be prone to biases towards certain learners. Non-native speakers, students with dyslexia, or students with challenges to express themselves in the „language of physics“ may be rated low despite an adequate understanding. However, element maps - when used as a support for learning and not for simply scoring students' performance - can help students to overcome their challenges.

Nevertheless, according to our results, it is challenging for learners to reach a level of expertise in one of the characteristics of explanation quality. Moreover, it is particularly difficult to produce a comprehensive, high-quality explanation. A promising way to make our process more economical for teachers and researchers alike lies in machine-learning approaches (Zhai et al., 2020), which are becoming more and more popular in dealing with explanations, not least using networks (Wulff et al., 2022). For example, identifying entities and their synonyms based on word types (nouns, pronouns) can be automated using example explanations of experts and learners. Further, identifying incorrect and irrelevant elements can also be automatised. Once the elementary structure is extracted in this way, it can be directly analysed, measured and visualised. This would also make fine-grained analysis instantaneously accessible for direct feedback to the explainer. For example, this would enable pointing out missing entities, missing or incorrect relations between the entities or unrelated parts in the explanation while it is being written. Then, the effectiveness of such direct feedback on the quality of explanations could and should be investigated in science education research to test whether automation is helpful here. In addition, automation can support teachers in the formative assessment of students' explaining skills. For research purposes, automation has the

advantage of being more economical, so larger samples of written explanations can be studied. However, when using machine learning, for example, this also requires phenomenon-specific training to account for phenomenon-specific word meanings and features of explanations. Manually extracting the elementary structure of large written explanations is time-consuming, whereas the process of network visualisation and analysis is comparatively fast. In contrast, short explanations comprising only a few sentences can be visualised quickly, even by hand, and thus offer the possibility of visualising the elementary structure. Up to now, our tool is therefore intended instead for research purposes.

With element maps, it is possible to analyse student performance in each of the core map characteristics, highlight the strengths and weaknesses of an explanation, and design structured supports and interventions that build on the learners' situation at baseline. These skills are, for example, to know, identify, and name the relevant entities of both the phenomenon at hand and the concerned theory. Furthermore, it is necessary to connect these entities to build scientifically appropriate propositions and to arrange these propositions around central aspects. Expert maps can potentially guide this process.

The tool offers the possibility to improve teaching by making appropriate recommendations based on the individual characteristics of the explanation features. Furthermore, if the procedure works well in other contexts, researchers could use element maps to investigate, for example, whether specially designed instructions actually improve explanation quality or which interventions influence which explanation characteristics. We argue that using element maps to capture the elementary structure of explanations is an important step towards improving the quality of explanations, which are an important scientific practice and a central part of science learning across all levels, from primary school to university.

## Notes

1. For a mathematical description, see Freeman (1977).
2. In both maps, we have removed labels and arrowheads to focus on structural differences rather than on content. All content information is still available for every map version.

## Disclosure statement

No potential conflict of interest was reported by the author(s).

## Ethics statement

We assure that all subjects involved in the study could participate voluntarily, that participation or non-participation had no consequences, that all participants were informed about the objectives and that no identifiable personal data were collected according to data protection standards.

## ORCID

S. Wagner  <http://orcid.org/0000-0003-1774-9694>

B. Priemer  <http://orcid.org/0000-0001-5399-7631>

## References

- Alameh, S., & Abd-El-Khalick, F. (2018). Towards a philosophically guided schema for studying scientific explanation in science education. *Science & Education*, 27(9–10), 831–861. <https://doi.org/10.1007/s11191-018-0021-9>
- Andersson, B., & Kärrqvist, C. (1983). How Swedish pupils, aged 12–15 years, understand light and its properties. *European Journal of Science Education*, 5(4), 387–402. <https://doi.org/10.1080/0140528830050403>
- Barabási, A.-L. (2016). *Network science*. Cambridge University Press.
- Bechtel, W., & Abrahamsen, A. (2005). Explanation: A mechanist alternative. *Studies in History and Philosophy of Biological and Biomedical Sciences*, 36(2), 421–441. <https://doi.org/10.1016/j.shpsc.2005.03.010>
- Berthold, K., & Renkl, A. (2009). Instructional Aids to support a conceptual understanding of multiple representations. *Journal of Educational Psychology*, 101(1), 70–87. <https://doi.org/10.1037/a0013247>
- Braaten, M., & Windschitl, M. (2011). Working toward a stronger conceptualization of scientific explanation for science education. *Science Education*, 95(4), 639–669. <https://doi.org/10.1002/sce.20449>
- de Andrade, V., Freire, S., & Baptista, M. (2019). Constructing scientific explanations: A system of analysis for students' explanations. *Research in Science Education*, 49(3), 787–807. <https://doi.org/10.1007/s11165-017-9648-9>
- Delen, I., & Krajcik, J. (2015). What do students' explanations look like when they use second-hand data? *International Journal of Science Education*, 37(12), 1953–1973. <https://doi.org/10.1080/09500693.2015.1058989>
- Forman, E. A. (2018). The practice turn in learning theory and science education. In D. W. Kritt (Ed.), *Constructivist education in an age of accountability* (pp. 97–111). Springer International Publishing. [https://doi.org/10.1007/978-3-319-66050-9\\_5](https://doi.org/10.1007/978-3-319-66050-9_5)
- Freeman, L. C. (1977). A set of measures of centrality based on betweenness. *Sociometry*, 40(1), 35–41. <https://doi.org/10.2307/3033543>
- Galili, I., & Hazan, A. (2000). Learners' knowledge in optics: Interpretation, structure and analysis. *International Journal of Science Education*, 22(1), 57–88. <https://doi.org/10.1080/095006900290000>
- Geelan, D. (2013). Teacher explanation of physics concepts: A video study. *Research in Science Education*, 43(5), 1751–1762. <https://doi.org/10.1007/s11165-012-9336-8>
- Halliday, M. A. K., & Mattheissen, C. M. I. M. (2013). *Halliday's introduction to functional grammar*. Routledge.
- Herman, B. C., Owens, D. C., Oertli, R. T., Zangori, L. A., & Newton, M. H. (2019). Exploring the complexity of students' scientific explanations and associated nature of science views within a place-based socioscientific issue context. *Science & Education*, 28(3–5), 329–366. <https://doi.org/10.1007/s11191-019-00034-4>
- Jonassen, D. H., Beissner, K., & Yacci, M. (2013). *Structural knowledge: Techniques for representing, conveying, and acquiring structural knowledge*. L. Erlbaum.
- Kaewkhong, K., Mazzolini, A., Emarat, N., & Arayathanitkul, K. (2010). Thai high-school students' misconceptions about and models of light refraction through aplanar surface. *Physics Education*, 45(1), 97. <https://doi.org/10.1088/0031-9120/45/1/012>
- Kang, H., Thompson, J., & Windschitl, M. (2014). Creating opportunities for students to show what they know: The role of scaffolding in assessment tasks. *Science Education*, 98(4), 674–704. <https://doi.org/10.1002/sce.21123>
- Koponen, I. T., & Nousiainen, M. (2018). Concept networks of students' knowledge of relationships between physics concepts: Finding key concepts and their epistemic support. *Applied Network Science*, 3(1), 1–21. <https://doi.org/10.1007/s41109-018-0072-5>
- Koponen, I. T., & Pehkonen, M. (2010). Coherent knowledge structures of physics represented as concept networks in teacher education. *Science and Education*, 19(3), 259–282. <https://doi.org/10.1007/s11191-009-9200-z>

- Kubsch, M., Touitou, I., Nordine, J., Fortus, D., Neumann, K., & Krajcik, J. (2020). Transferring knowledge in a knowledge-in-use task—investigating the role of knowledge organization. *Education Sciences*, 10(1), 20. <https://doi.org/10.3390/educsci10010020>
- Kuhn, T. S. (2000). *The road since structure: Philosophical essays, 1970—1993, with an autobiographical interview* (J. Conant & J. Haugeland, Eds.). University of Chicago Press.
- Kulgemeyer, C. (2018). Towards a framework for effective instructional explanations in science teaching. *Studies in Science Education*, 54(2), 109–139. <https://doi.org/10.1080/03057267.2018.1598054>
- Lachner, A., Gurlitt, J., & Nückles, M. (2012). A graph-oriented approach to measuring expertise—detecting structural differences between experts and intermediates. Proceedings of the 34th annual conference of the cognitive science society, 653–658.
- Lemke, J. L. (1983). Thematic analysis: Systems, structures, and strategies. *Semiotic Inquiry*, 2(3), 159–187. [http://static1.1.sqspcdn.com/static/f/694454/12424800/1306519358857/Lemke\\_1983\\_ThematicAnalysis.pdf?token=Pa1RSIxGmRe%2BWbBnTIyNBDqrmgk%3D](http://static1.1.sqspcdn.com/static/f/694454/12424800/1306519358857/Lemke_1983_ThematicAnalysis.pdf?token=Pa1RSIxGmRe%2BWbBnTIyNBDqrmgk%3D)
- Lemke, J. L. (1990). *Talking science: Language, learning, and values*. Ablex Pub. Corp.
- Lintern, G., Moon, B., Klein, G., & Hoffman, R. R. (2018). Eliciting and representing the knowledge of experts. In K. A. Ericsson, R. R. Hoffman, A. Kozbelt, & A. M. Williams (Eds.), *The Cambridge handbook of expertise and expert performance* (pp. 458–522). Cambridge University Press.
- Manz, E., Lehrer, R., & Schauble, L. (2020). Rethinking the classroom science investigation. *Journal of Research in Science Teaching*, 57(7), 1–27. <https://doi.org/10.1002/tea.21625>
- McNeill, K. L., Lizotte, D. J., Krajcik, J., & Marx, R. W. (2006). Supporting students' construction of scientific explanations by fading scaffolds in instructional materials. *Journal of the Learning Sciences*, 15(2), 153–191. [https://doi.org/10.1207/s15327809jls1502\\_1](https://doi.org/10.1207/s15327809jls1502_1)
- Nassar, A. B. (1994). Apparent depth. *The Physics Teacher*, 32(9), 526–529. <https://doi.org/10.1119/1.2344102>
- NGSS Lead States. (2013). *Next generation science standards: For states, by states*. The National Academic Press.
- Novak, J. D. (1990). Concept mapping: A useful tool for science education. *Journal of Research in Science Teaching*, 27(10), 937–949. <https://doi.org/10.1002/tea.3660271003>
- Osborne, J. F., & Patterson, A. (2011). Scientific argument and explanation: A necessary distinction? *Science Education*, 95(4), 627–638. <https://doi.org/10.1002/sci.20438>
- Papadouris, N., Vokos, S., & Constantinou, C. P. (2018). The pursuit of a “better” explanation as an organizing framework for science teaching and learning. *Science Education*, 102(2), 219–237. <https://doi.org/10.1002/sci.21326>
- Peel, A., Zangori, L., Friedrichsen, P., Hayes, E., & Sadler, T. (2019). Students' model-based explanations about natural selection and antibiotic resistance through socio-scientific issues-based learning. *International Journal of Science Education*, 41(4), 510–532. <https://doi.org/10.1080/09500693.2018.1564084>
- Peker, D., & Wallace, C. S. (2011). Characterizing high school students' written explanations in biology laboratories. *Research in Science Education*, 41(2), 169–191. <https://doi.org/10.1007/s11165-009-9151-z>
- Quinn, H., Schweingruber, H., & Keller, T. (2012). *A framework for K-12 science education: Practices, crosscutting concepts, and core ideas*. National Academies Press.
- Redfors, A., & Ryder, J. (2001). University physics students' use of models in explanations of phenomena involving interaction between metals and electromagnetic radiation. *International Journal of Science Education*, 23(12), 1283–1301. <https://doi.org/10.1080/09500690110038620>
- Rocksén, M. (2016). The many roles of “explanation” in science education: A case study. *Cultural Studies of Science Education*, 11(4), 837–868. <https://doi.org/10.1007/s11422-014-9629-5>
- Ruiz-Primo, M. A., Li, M., Tsai, S.-P., & Schneider, J. (2010). Testing one premise of scientific inquiry in science classrooms: Examining students' scientific explanations and student learning. *Journal of Research in Science Teaching*, 47(5), 583–608. <https://doi.org/10.1002/tea.20356>



- Ruxton, G. D. (2006). The unequal variance t-test is an underused alternative to student's t-test and the Mann–Whitney U test. *Behavioral Ecology*, 17(4), 688–690. <https://doi.org/10.1093/beheco/ark016>
- Salmon, W. C. (2006). *Four decades of scientific explanation*. University of Pittsburgh Press.
- Sandoval, W. A. (2003). Conceptual and epistemic aspects of students' scientific explanations. *The Journal of the Learning Sciences*, 12(1), 5–51. [https://doi.org/10.1207/S15327809JLS1201\\_2](https://doi.org/10.1207/S15327809JLS1201_2)
- Schaal, S. (2008). Concept mapping in science education assessment: An approach to computer-supported achievement tests in an interdisciplinary hypermedia learning environment. In A. J. Cañas, P. Reiska, M. Ahlberg, & J. D. Novak (Eds.), *Proceedings of the 3rd international conference on concept mapping* (pp. 228–235).
- Siew, C. S. Q. (2020). Applications of network science to education research: Quantifying knowledge and the development of expertise through network analysis. *Education Sciences*, 10(4), 101. <https://doi.org/10.3390/educsci10040101>
- Taber, K. S., & Watts, M. (2000). Learners' explanations for chemical phenomena. *Chemistry Education Research and Practice*, 1(3), 329–353. <https://doi.org/10.1039/B0RP90015J>
- Tang, K.-S. (2016). Constructing scientific explanations through premise–reasoning–outcome (PRO): An exploratory study to scaffold students in structuring written explanations. *International Journal of Science Education*, 38(9), 1415–1440. <https://doi.org/10.1080/09500693.2016.1192309>
- Thurn, C. M., Hänger, B., & Kokkonen, T. (2020). Concept mapping in magnetism and electrostatics: Core concepts and development over time. *Education Sciences*, 10(5), 129. <https://doi.org/10.3390/educsci10050129>
- Wiley, J., Hastings, P., Blaum, D., Jaeger, A. J., Hughes, S., Wallace, P., Griffin, T. D., & Britt, M. A. (2017). Different approaches to assessing the quality of explanations following a multiple-document inquiry activity in science. *International Journal of Artificial Intelligence in Education*, 27(4), 758–790. <https://doi.org/10.1007/s40593-017-0138-z>
- Williams, C. G. (1998). Using concept maps to assess conceptual knowledge of function. *Journal for Research in Mathematics Education*, 29(4), 414–421. <https://doi.org/10.2307/749858>
- Wulff, P., Buschhüter, D., Westphal, A., Mientus, L., Nowak, A., & Borowski, A. (2022). Bridging the gap between qualitative and quantitative assessment in science education research with machine learning—A case for pretrained language models-based clustering. *Journal of Science Education and Technology*, 31(4), 490–513. <https://doi.org/10.1007/s10956-022-09969-w>
- Yun, E., & Park, Y. (2018). Extraction of scientific semantic networks from science textbooks and comparison with science teachers' spoken language by text network analysis. *International Journal of Science Education*, 40(17), 2118–2136. <https://doi.org/10.1080/09500693.2018.1521536>
- Zacharia, Z. C. (2005). The impact of interactive computer simulations on the nature and quality of postgraduate science teachers' explanations in physics. *International Journal of Science Education*, 27(14), 1741–1767. <https://doi.org/10.1080/09500690500239664>
- Zangori, L., Forbes, C. T., & Schwarz, C. V. (2015). Exploring the effect of embedded scaffolding within curricular tasks on third-grade students' model-based explanations about hydrologic cycling. *Science & Education*, 24(7–8), 957–981. <https://doi.org/10.1007/s11191-015-9771-9>
- Zarkadis, N., & Papageorgiou, G. (2020). A fine-grained analysis of students' explanations based on their knowledge of the atomic structure. *International Journal of Science Education*, 42(7), 1162–1182. <https://doi.org/10.1080/09500693.2020.1751340>
- Zhai, X., Yin, Y., Pellegrino, J. W., Haudek, K. C., & Shi, L. (2020). Applying machine learning in science assessment: A systematic review. *Studies in Science Education*, 56(1), 111–151. <https://doi.org/10.1080/03057267.2020.1735757>